Using Presence-Absence Data on Areal Units to Model the Ranges and Range Shifts of Select South African Bird Species

Kristin M. Broms

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2013

Committee:

Loveday L. Conquest, Chair

Samuel K. Wasser, GSR

Devin S. Johnson

John R. Skalski

Paul D. Sampson

Christian E. Grue

Les Underhill

Program Authorized to Offer Degree:
Quantitative Ecology & Resource Management

University of Washington

Abstract

Using Presence-Absence Data on
Areal Units to Model the Ranges and Range Shifts
of Select South African Bird Species

Kristin M. Broms

Chair of the Supervisory Committee:
Chair Loveday L. Conquest
Quantitative Ecology & Resource Management

The study of where species occur is an important concern in ecology. Over the last decade, the occupancy model has been the primary tool used in attempts to answer where, when and why species occur where they do. In this dissertation, the occupancy model is improved upon by adding a spatial component to account for similarities between adjacent sites. The model was applied to the Southern ground hornbill (*Bucorvus leadbeateri*), an elusive species that occurs in low densities through much of sub-Saharan Africa; South Africa is the southernmost end of its range and the distribution of the species there is unknown. The model uncovered new areas of potentially high occupancies, quantified the strong associations between hornbill occurrences and the availability of protected areas, and provided the appropriate structure to ask additional biological questions.

The spatial occupancy model was then further adapted to include a temporal component in order to quantify range expansions or contractions when data are collected over time. This model was used on the common myna (*Acridotheres tristis*), an invasive species in South Africa whose range has been expanding in recent decades. The results suggest that the range of the myna continues to expand at a rate of 3% a year and that its occurrences are associated with high human population densities.

TABLE OF CONTENTS

Page

List of Figures iii

List of Tables iv

Glossary vi

Chapter 1: Introduction 1
1.1 Objectives and Dissertation Outline 1
1.2 Motivation: Presence-Absence Data and Atlas Maps in Ecology . . . 3
1.3 The Southern African Bird Atlas Project (SABAP) 6
1.4 Occupancy Models to Analyze Presence-Absence Data 8
1.5 Incorporating Neighboring Spatial Relationships 11
1.6 Motivation: Multi-Season Models for Presence-Absence Data 14

Chapter 2: Spatial Occupancy Models for Atlas Data: With Application to the Southern Ground Hornbill 19
2.1 Introduction 19
2.2 Methods 22
2.3 Results 32
2.4 Discussion 33
2.5 Conclusion 37

Chapter 3: A Simulation Study of the RSR and ICAR Occupancy Models . 49
3.1 Introduction 49
3.2 Methods 54
3.3 Results 58
3.4 Summary of Results 80
3.5 Conclusions 82

Chapter 4: A Spatio-Temporal Model, Introduced and Applied to the Common Myna . . . 89
4.1 Introduction . . . 89
4.2 Methods . . . 95
4.3 Model Selection . . . 103
4.4 Simulation Study . . . 104
4.5 Results . . . 106
4.6 Discussion . . . 111

Chapter 5: Conclusions & Implications . . . 118
5.1 Summary . . . 120
5.2 Future Model Applications . . . 122
5.3 Conservation and Management Implications . . . 123
5.4 Recommendations for Sampling Design Improvements . . . 124
5.5 Final Remarks . . . 125

Bibliography . . . 126

Appendix A: Hornbill Supplementary Material . . . 137

Appendix B: Myna Supplementary Material . . . 142
B.1 Myna SABAP 2 data. . . . 143
B.2 Model equations. . . . 144
B.3 List of symbols and their definitions. . . . 146
B.4 Derivation of the neighborhood dispersal vector. . . . 147
B.5 Simulation study results. . . . 148
B.6 Parameter estimates from model variants of the myna analysis. . . . 152

LIST OF FIGURES

Figure Number		Page
1.1	Map of SABAP 1 data.	16
1.2	The spatial coverage of the SABAP 2 data, as of 19 August 2013.	17
1.3	Neighborhood structure of the data.	18
2.1	Ground hornbill analysis area, plus the number of detections per site.	46
2.2	Predicted occupancy probabilities of the ground hornbill, from the ICAR model.	47
2.3	Predicted occupancy probabilities of the ground hornbill, from the RSR-160 model.	48
3.1	Examples of eigenvector mapping.	84
3.2	Simulation study estimates when missing large-scale covariates.	86
3.3	Simulation study estimates when missing clustered covariates.	88
3.4	Bias of the spatial occupancy models.	90
3.5	Simulation study estimated standard errors.	92
4.1	Myna detections and surveys.	114
4.2	Neighborhood structure.	115
4.3	Predicted dispersal surface of the myna.	115
4.4	Estimated occupancy probabilities of the myna.	116
4.5	Estimated occurrences of the myna.	117
A.1	Ground hornbill analysis area, plus the number of surveys per site.	138
A.2	Realized hornbill occupancies from the Full ICAR model.	139
A.3	Realized hornbill occupancies from the Full RSR-160 model.	140
B.1	SABAP 2 surveys and detections of the myna.	143
B.2	Myna occupancy probability standard errors.	158
B.3	Myna occurrence standard errors.	159

LIST OF TABLES

Table Number		Page
2.1	A summary of the models that were fit to the ground hornbill data.	39
2.2	Hornbill parameter estimates, nonspatial occupancy models.	42
2.3	Hornbill parameter estimates, the ICAR occupancy models.	43
2.4	Hornbill parameters estimates, the RSR-400 occupancy models.	44
2.5	Hornbill parameter estimates, the RSR-160 occupancy models.	45
3.1	The posterior predictive loss criterion (PPLC) for model selection.	61
3.2	Nonspatial versus spatial occupancy models.	63
3.3	A comparison of model fit when the spatial random effect is used in place of all covariates.	64
3.4	Accuracy of the reported standard errors.	67
3.5	Trade-off between run times and model error.	70
3.6	Model errors under alternative spatial regimes.	70
3.7	Model results for low occupancy probabilities.	72
3.8	Predicting model coefficients with uncorrelated variables.	73
3.9	Predicting model coefficients with correlated variables.	73
3.10	Biases from collinearity between detection and occupancy probabilities.	77
3.11	Biases from clustered collinearity.	77
3.12	Comparing posterior estimates for different gamma priors for the spatial parameter, τ.	78
3.13	Biases resulting from a sparsely sampled area.	80
3.14	Biases resulting from a sparsely sampled area on a large grid.	80
4.1	Nonhomogeneous model of the myna, parameters and estimates.	109
4.2	Homogeneous model of the myna, parameters and estimates.	110
4.3	Neighborhood dispersal probabilities of the myna, from the homogeneous model.	111
A.1	Ground hornbill predicted occupancy probabilities.	141

B.1 Simulation study results from the nonhomogeneous, spatio-temporal model. 148
B.2 Simulation study results from the nonhomogeneous, spatio-temporal model, with varying long-distance dispersal 149
B.3 Simulation study results from the homogeneous, spatio-temporal model. 150
B.4 Simulation study results from the homogeneous, spatio-temporal model, with varying long-distance dispersal. 151
B.5 Myna results from the nonhomogeneous, spatio-temporal model using latitude as the habitat gradient. 153
B.6 Myna results from the nonhomogeneous, spatio-temporal model using longitude as the habitat gradient. 154
B.7 Myna results from the nonhomogeneous, spatio-temporal model, with long-distance dispersal a function of human populations. 155
B.8 Myna results from the homogeneous, spatio-temporal model, with long-distance dispersal a function of human populations. 156
B.9 Dispersal probabilities associated with Table B.8. 157

GLOSSARY

CAR MODEL: A regression model that includes a *conditional autoregressive* variable to account for residual spatial autorcorrelation.

ICAR MODEL: *Intrinsic conditional autoregressive* model. A specific type of CAR model that allows one to write the joint distribution of the spatial autocorrelation as a set of conditional distributions. One can then use MCMC techniques to estimate the parameters associated with the spatial variable.

MORAN OPERATOR MATRIX (Ω): An equation whose eigenvalues determine the distribution for the Moran's I statistic for that data set. (The values that the Moran's I stastistic may take depend on the values of the sites for which one wants inference and on the number and arrangement of these sites.)

MORAN'S I STATISTIC: A measure of spatial autocorrelation in a data set. Negative values indicate negative spatial autocorrelation (i.e. the data look like a checkerboard); positive values indicate positive spatial autocorrelation (i.e. the data are clumped). A zero value indicates a random spatial pattern.

OCCUPANCY: Is generally reported as a probability. It is the probability that a species will occur at an unsurveyed site.

OCCUPANCY MODEL: A regression-type model that uses data with multiple visits/surveys per site to account for the fact that a species may be present at a site

but go unseen. It uses this data coupled with environmental and observation-process variables to estimate the probabilities of where the species is likely to occur across a given landscape.

OCCURRENCE: The acutal outcomes of species occupancy at each site. Similar to a coin flip, the occurrences would be the heads or tails that result, while the occupancy probability is the probability of these said occurrences arising.

PENTAD: A 5-minute latitude by 5-minute longitude rectangle, which is approximately 8 km x 7.6 km. The data for SABAP 2 is collected at this spatial resolution.

QDGC: *Quarter Degree Grid Cell.* A 15-minute latitude by 15-minute longitude rectangle. The area of one of these sites is approximately 720 km^2. The data for SABAP 1 was collected at this spatial resolution.

RSR: *Residual spatial regression* is a model derived from the ICAR model. Similar to the ICAR model, it uses a random effect variable to model residual spatial autocorrelation. It additionally uses eigenvectors from the Moran Operator Matrix to reduce the dimension of the precision matrix (the inverse of the variance matrix) associated with the spatial random effect.

SABAP: *The Southern African Bird Atlas Project.* A database of checklists submitted by volunteer bird watchers. Each checklist contains a list of bird species that were seen along with the location of the sightings. The project has occurred in two phases. SABAP 1 is data collected primarily between 1987 and 1991. SABAP 2 was started in July 2007 and is ongoing.

ACKNOWLEDGMENTS

Without the help of so many people, this dissertation would not have been completed. First I would like to thank my adviser, Loveday Conquest, for believing in me and being a strong role model, showing me how one can successfully care about education and community at a top research university, and how that attention is appreciated by all of the students who are able to gain from the kindness and guidance.

The rest of my committee was also vital to my dissertation. Devin Johnson informed me about the RSR model that I test in Chapters 2 and 3 and the spatio-temporal model of Hooten and Wikle (2010) that I adapted in Chapter 4. Besides exposing me to these recent, computing advances and novel statistical methods, he wrote an R package, "stocc", which helped me more thoroughly understand these models and easily implement them.

Sam Wasser gave me invaluable advice and suggestions on how to make my work most interesting for non-statisticians and keep my results relevant biologically. Chris Grue taught me the relevancy of occupancy models to the U.S. Fish and Wildlife departments, and the GAP program in the U.S. in particular. Paul Sampson kept me on my statistical toes. And last by not least, I want to thank John Skalski. Without him, I may not have persevered.

Professor Les Underhill was also on my committee. I want to thank him separately for it was his passion and energy that got me started and excited about my dissertation research. He is thoroughly trained in both ecological and statistical studies, and is an asset to the worldwide scientific community for keeping much of the ecological research in Southern Africa statistically rigorous. He is especially an asset to the

birding community of South Africa. His encouragement of the volunteers, and the large numbers of volunteers themselves, have made the Southern African Bird Atlas Project a huge success and a great resource for scientists to learn more about bird communities than they ever have before.

Along a similar vein, I want to thank Dr. Res Altwegg for his contribution and enthusiasm for complex mathematical models to better understand the biological phenomenon of South Africa. Dr. Altwegg was a huge help in directing my research on the value of occupancy models and the role of presence-absence data in the larger ecological community.

Together, Les and Res got me started with my research by inviting me to South Africa to help with an analysis for a booklet for the Intergovernmental Panel on Climate Change conference in Denmark in 2008. Having such a potentially large impact was an eye opening experience and was great motivation for beginning my work. Thank you for that first trip to Cape Town, and thank you to the National Science Foundation of South Africa for helping to fund it.

I want to recognize everyone at the South African National Biodiversity Institute and the Animal Demography Unit at the University of Cape Town for their generosity during my visits. In particular, I want to thank Sue Kuyper. I am not sure if the ADU could exist without her help. And thank you to Sally Hofmeyr for our conversations and her kindness in letting me stay with her for my second visit to Cape Town. It was wonderful to have made such a good friend halfway around the world.

My second trip to South Africa was made possible through the people and institutions of South Africa, but also from the generosity of foundations and agencies. I am grateful for the NASA Space Grant, which allowed me the time and finances to go Cape Town in January–March of 2011. I am grateful for the Wordwide Universities Network (WUN) for additional funds so that my travels were not financially stressful.

Most of my funding came through Teaching Assistantships within the Quantitative Science (Q SCI) undergraduate program at the University of Washington. It was pleasure to get to know more students by helping them with their statistical analyses, and it kept my basic statistical knowledge in top shape.

The Center for Studies in Demography and Ecology (CSDE) provided essential computing resources that allowed me to run large models, simulations of these models, and use ArcGIS from afar; my dissertation would not have been possible without their computer clusters.

Other funding and emotional support came from the Delta Kappa Gamma International Society, an organization of women educators from around the globe. The scholarships that I received helped me through the second two years of my degree, and I leaned heavily on the faith and trust that they had in my abilities and strengths. In particular, I want to thank Kate Grieshaber and the other ladies of the Rho Chapter. I will miss catching up at our monthly meetings and the warmth and good feeling that filled me up at the end of each of these.

Dr. Ellen Pikitch and her lab at Stony Brook University provided me with support during the last two years of my doctoral candidacy. I am grateful to Ellen for exposing me to the mechanics of another lab, teaching me about the global importance of forage fish, and the large, positive influences that science can have on policy. I also want to thank the rest of the lab- Christine Santora, Konstantine Rountos, Tasha Gownaris, Sara Cernadas-Martin, and Tess Geers- for their conversations and curiosities. I will miss going out on the ShiRP trawls with you all!

While I was still in Seattle, Dr. Dee Boersma kindly invited me to sit in on her lab meetings. It was great to get to know other biologists-in-training, to learn more about penguins than I thought was possible, and to have the experience of where my statistical knowledge could be of use.

At the University of Washington, it was also a pleasure to meet and work with Dr. Sievert Rohwer. I look forward to our future collaborations. And I want to express gratitude to everyone within the Quantitative Ecology & Resource Management (QERM) program. It has been a long relationship and I am happy to be a member of such a smart, talented group. A special thanks to Joanne Besch for her support to all of us at QERM and for keeping the program running smoothly.

Thank you to Marc Kéry, Andy Royle, and Mevin Hooten for reading over my work and providing valuable feedback. It is very encouraging to know that such great people work within my chosen field of study!

Finally, I am grateful for my family and friends outside of the academic community for lending an ear all these years. I guess now it is time to join the workforce like the rest of you ...

Chapter 1

INTRODUCTION

1.1 Objectives and Dissertation Outline

Ecology can be described as the study of how organisms interact with their environment, how this relationship evolves over time, and what mechanisms drive these dynamics (Kéry and Schaub, 2012). The answers to these questions require knowledge on species' abundances or occurrences. Sometimes, occurrence data are of interest through its relationship to abundance, as it is generally cheaper and/or more efficient to collect than data on abundance (Royle and Dorazio, 2008). Other times, occupancy (the proportion of a landscape that a species occupies) will be the key state variable (i.e., Scott et al. (2002); Gaston (2003)).

The goal of my PhD research is to use presence-absence data from the Southern Africa Bird Atlas Project (SABAP) to quantify South African bird species' ranges and range shifts via occupancy probabilities, and to correlate them with available environmental variables, with the ultimate goal of improving upon established techniques by utilizing the extra information that is contained with spatially and temporally connected sites and surveys of a large database of presence-absence records. The virtues of this knowledge have been extolled time and again (e.g., see Scott et al. (2002); MacKenzie et al. (2006); Royle and Dorazio (2008); Kéry and Schaub (2012)). In this vein, I have developed a model that incorporates the imperfect detection of a species, the spatial and temporal correlations of the data, and the environmental covariates that affect occupancy and detection. I have also applied the models to two species: the Southern ground hornbill (*Bucorvus leadbeateri*) and the common myna (*Acridotheres tristis*).

Occupancy is a fundamental concept in macroecology, landscape ecology, and meta-population ecology (Brown and Maurer, 1989; Levins, 1969; Hanski, 1998); it is a natural summary of habitat suitability (Boyce and McDonald, 1999); and it is used as an important criterion in determining conservation status for the International Union for Conservation of Nature (IUCN) red list, a tool widely used by governments, NGOs and scientific institutions to set conservation priorities and goals (Kéry and Schaub, 2012; Hilton-Taylor, 2001). In particular, larger sets of occurrence data compiled into atlas projects can and have been used for biodiversity monitoring and to determine the effects of climate change (Kéry and Schaub, 2012; Altwegg et al., 2012).

In this chapter, I describe the data; the base models used to analyze similar data sets; and introduce the spatial components that will be added to those models. Chapter 2 presents an application of a spatial occupancy model to the Southern ground hornbill. The model revealed a larger distribution for the species than would have been identified if one did not account for imperfect detection and spatial autocorrelation. Because the Southern ground hornbill has a Threatened status in South Africa, the results have important management implications for conservation efforts. Chapter 3 is a simulation study of the spatial occupancy models that tests their accuracy and sets boundaries for future model usage. Chapter 4 introduces a new occupancy model with explicit spatial and temporal components for a multi-season analysis of an open population. This model is particularly important for rapid colonizers. Included in this chapter is a simulation study demonstrating the model's accuracy in its predictions and an application of the model to the common myna, an invasive species within South Africa. The model concludes that the range of the myna continues to expand at a rate of approximately 3% a year, with the thrust of this expansion occurring along the coast and northward into Zimbabwe. Chapter 5 summarizes the main results from each chapter and the resource management implications of the work.

Each chapter is connected with a flow that begins here with qualitative descriptions

of the research questions and culminates with the spatio-temporal hierarchical model. However, the dissertation has been written so that each chapter may be enjoyed independently of the others, depending on the reader's interests.

1.2 Motivation: Presence-Absence Data and Atlas Maps in Ecology

The collection of presence-absence data has a long history in ecology. A review of species occurrence modeling from 2001 refers to hundreds of such data sets (Scott et al., 2002). Most often, these data are used to predict where a species occurs, the extent of its distribution across a landscape, and/or the species-habitat relationships. Use of presence-absence data is common because it is noninvasive and cheap to collect; it does not involve capturing and tagging any animals (MacKenzie et al., 2006; Royle and Dorazio, 2008). For example, the MacKenzie et al. (2002 paper, which introduced the occupancy model to analyze such data, has been cited over 1,226 times (Google Scholar search, May 31, 2013).

Recent technological advances have only continued the proliferation of this type of data. For example, scat samples may be collected and analyzed as presence-absence data (i.e., Wasser et al. (2012)); camera traps may be employed and similarly analyzed (i.e., Hines et al. (2010)).

In addition to studies on single species or single ecosystems, many scientists build databases of "atlas data". Atlas data are one of the most popular forms of survey in ornithology (Donald and Fuller, 1998), but it may be collected for any species. While the protocol of the data collection will vary from project to project, an atlas is usually presence-absence data collected on a grid by multiple observers, with the purpose of mapping the geographical patterns of occurrence over a large area (Donald and Fuller, 1998).

The Southern African Bird Atlas Project (SABAP) data used in this dissertation is a prime example. The atlas survey was originally conducted in the late 1980s, and its success inspired many similar atlas programs in Southern Africa: the Bird in Reserves

Project in South Africa, the Mozambique Bird Atlas Project, the Southern African Frog Atlas Project, the Southern African Reptile Conservation Assessment, and the Southern African Butterfly Conservation Assessment (Harrison et al., 2008). More globally, there have been atlas surveys of amphibians (Heikkinen and Högmander, 1994) and breeding birds (Högmander and Møller, 1995) in Finland; breeding bird atlases in Switzerland (Tobalske and Tobalske, 1999); bird surveys statewide in Nebraska and region-wide in the Great Lakes Basin (Scott et al. (2002) (ch 32), Sargeant et al. (2005)), and large atlases of plant species occurrences such as FLORKART in Germany (Bierman et al., 2010). In addition, the counts from line-transect surveys such as the North American Breeding Bird Survey are often simplified to presence-absence and analyzed as atlas data (MacKenzie et al., 2006).

As the collection of presence-absence data grows in popularity, it becomes increasingly important to use the correct methods in its examinations. Prior to the seminal paper by MacKenzie et al. (2002), this type of data was analyzed primarily by logistic regressions (Scott et al., 2002). These models were not ideal because they ignored the possibility of non-detection, also known as false-negatives, and thereby consistently underestimated occurrence rates (Royle and Dorazio, 2008). The model framework introduced by MacKenzie et al. (2002) and Tyre et al. (2003), referred to hereafter as an "occupancy model", accounts for the possibility of non-detection and it is the modeling framework that I adapted in this dissertation.

The occupancy model is a great model for many data sets, but will be inadequate when data are collected from nonrandom sampling. For atlas data, the sampled sites and desired area of inference involve adjacent sites. Allowing neighboring sites to be more similar than sites that are far from one another can improve model predictions. However, ignoring the fact that the sites of the study were not independent may lead to biased results, inflate the significances of the predictors, and in turn, lead to incorrect model selection (Augustin et al., 1996; Moore and Swihart, 2005).

Spatial dependencies have mostly been ignored in wildlife investigations (Scott

et al., 2002) but there have been a few spatially explicit occupancy models and the quality and quantity of the spatial models continues to grow. Augustin et al. (1996); Heikkinen and Högmander (1994); Högmander and Møller (1995); Wu and Huffer (1997); Hoeting et al. (2000) were the first papers to incorporate spatial dependence in models that predict occurrence. Spatial dependence is accounted for with the addition of an autocovariate term to a model. An autocovariate is a weighted sum of the response values in neighboring sites and it is added as an additional explanatory variable to a regression equation (Wu et al., 2009). Augustin et al. (1996) found that adding an autocovariate to their logistic regression reduced the unexplained deviance, made some previously significant parameters become nonsignificant, and produced less uniform occupancy predictions, better reflecting the clustering of the true distribution of the red deer population in Scotland that they were modeling. Similar results could be expected for other regression models that add a spatial component.

Since the 1990s, there have been a few models that both account for non-detection and include a spatial autocorrelation term: Hooten et al. (2003); Moore and Swihart (2005); Sargeant et al. (2005); Magoun et al. (2007); Wintle and Bardos (2006); Royle et al. (2007); Gardner et al. (2009); Hines et al. (2010); Bierman et al. (2010); Aing et al. (2011); Chelgren et al. (2011); Johnson et al. (2013). Some of these studies were designed for data collected at points or along line transects and therefore are not applicable to the SABAP data. Other examples use an autologistic term to account for the spatial autocorrelation. In this dissertation, I follow the example set by Johnson et al. (2013) and use a *conditional autoregressive* (CAR) variable and then an adaptation to the CAR variable, known as a *restricted spatial regression* (RSR), to account for spatial autocorrelation. These models and the reasoning behind their use are described below, in the "Incorporating Neighboring Spatial Relationships" section on page 11.

1.3 The Southern African Bird Atlas Project (SABAP)

The Southern African Bird Atlas Project is a database of bird lists from amateur and professional bird watchers. It has occurred in two phases: SABAP1 and SABAP 2. Avid bird watchers spend time looking for different bird species either on a specific bird-watching trip or on a walk through their neighborhood, then submit the list of species that were seen along with their location noted as the grid cell that they were in. The aggregation of this presence-absence data leads to a coarse map of the distribution of every species that occurs in South Africa for at least part of the year. The first project, SABAP 1, is the database of bird occurrences recorded primarily from 1987–1991; it also incorporates some data from as early as 1970 and continues in some regions into 1992 and 1993 (Harrison et al., 1997a). The survey area was Southern Africa, covering Botswana, Namibia, Zimbabwe, Lesotho, Swaziland, and South Africa. Figure 1.1 shows the extent and coverage of SABAP 1. The sampling unit was a quarter degree grid cell (QDGC), corresponding to a rectangle of 15 minutes of latitude by 15 minutes of longitude; the area of each cell is approximately 719 km^2 (Larsen et al., 2009). These cells differ in size due to the curvature of the Earth, with a minimum area of 641 km^2 in the southern Cape Province and a maximum area of 740 km^2 in northern Zimbabwe (Harrison et al., 1997a). In model analyses, each cell is assumed to be a similar size. SABAP 1 consisted of 4,000 grid cells, 88 of which have no data (Harrison et al., 1997a); 1,196 of the grid cells occur in South Africa.

For SABAP 1, seven thousand people volunteered to be observers. Five thousand of them submitted at least one checklist, and two thousand were considered regular contributors. About 750 observers contributed 80% of the checklists. Altogether 147,605 checklists were submitted for the region; producing 7 million records of bird occurrences (Harrison et al., 1997a). The effort-per-list for SABAP 1 was not recorded explicitly. A checklist could be submitted for any period of time from five minutes up to a maximum of one calendar month, but on each list, species are recorded in

the order in which they were observed. For further details on SABAP 1, the reader is referred to the Introduction and Methods sections of *The Atlas of Southern African Birds, Volume 1* (Harrison et al., 1997a).

The accumulation of the SABAP 1 lists was used to set the range boundaries where birds could be expected to occur in South Africa. It was the first time that general ranges and migration patterns were calculated for most species. The general success of SABAP 1, coupled with interest in whether and how ranges and migration patterns had changed in the decades since, led to a repeat of the atlas project, and SABAP 2 was born.

The second atlas project, SABAP 2, was begun in 2007 and continues through the present day, with the objective to measure the impact of environmental change on the distributions and abundance of birds since SABAP 1 (Harebottle et al., 2007). Figure 1.2 shows the extent and coverage of SABAP 2 as of February 8, 2011 (Animal Demography Unit, 2011). SABAP 2 has the same basic form as SABAP 1, but with a few changes. One modification is a smaller resolution; the sampling unit is a pentad grid cell, which is a 5-minute latitude by 5-minute longitude rectangular cell. Each pentad is approximately 8 x 7.6 km. However, size again varies from cell to cell due to the longitude lines getting closer to each other as they approach the poles. This change within southern Africa is not very large and all units will again be assumed to be of the same size. SABAP 2 consists of 17,444 pentads covering a survey area of South Africa, Lesotho, and Swaziland (Harebottle et al., 2007). As of May 31, 2013, 1,225 volunteers had submitted 87,797 cards for a total of over 4.6 million records of bird occurrences; 68% of the pentads had been surveyed (Animal Demography Unit, 2011).

For each list that is submitted to SABAP 2, the bird watcher is expected to cover all habitat types that occur within the pentad. The time scale is also more rigid. For each list, the volunteer must be actively bird watching for a minimum of two hours and can spend a maximum of five days recoding birds for each list. From this data,

the number of hours spent intensively birding and the total number of hours for the lists are recorded for each survey. The order in which the birds were seen is noted and the observer who submitted the survey is also recorded.

Each bird list that is submitted is seen as a survey that records the occurrences and non-detections for every species that occurs in South Africa, leading to a lattice structure of presence-absence data. Every time that a species is absent from a list, it is equal to a non-detection. I exploited the spatial and temporal information contained within each survey record to quantify the ranges and range shifts of South African bird species, to empirically correlate these ranges with environmental factors, and to determine associations between ranges and environmental factors. Such relationships are more commonly noted anecdotally but not verified statistically.

1.4 Occupancy Models to Analyze Presence-Absence Data

The occupancy model stems from mark-recapture models and was first published by MacKenzie et al. (2002) and Tyre et al. (2003) and further explained in MacKenzie et al. (2006). Occupancy models account for imperfect detection, the effects of environmental variables on occupancy and detection, and the effects of observational variables on detection. I present the models within a Bayesian framework similar to Royle and Dorazio (2008) because the Bayesian version more easily incorporates spatial autocorrelation. Occupancy models assume that N sites from the desired area of inference are randomly and independently sampled. There are multiple, independent surveys of each of the sites. On each survey j of each site i, the species of concern is either detected or not detected. If the species is detected on at least one survey of the site, then that site is occupied. It is assumed that there are no false positives under these circumstances, i.e., a species is never detected at an unoccupied site. If the species is not detected, then either that site is not occupied or it is occupied but the species was not detected. Let $\mathbf{z} = \{z_1, \ldots, z_N\}$ be the true occupancies of the N sampled sites. For site i,

$$z_i \sim \text{Bernoulli}\,(\psi_i) \tag{1.1}$$

where ψ_i is the probability of occupancy. Thus, $z_i = 1$ if the site is occupied and $z_i = 0$ if the site is not occupied. Site-specific variables that may affect occupancy, such as percentage of farmed land, vegetation composition or human density, are incorporated into the probability of occupancy with a logit or probit link (Royle and Dorazio, 2008; Johnson et al., 2013):

$$\text{probit}\,(\psi_i) = \beta_0 + \beta_1 x_{1,i} + \cdots + \beta_U x_{U,i} \tag{1.2}$$

Let $\mathbf{y} = \{y_1, \ldots, y_N\}$ be the observed occupancies of the N sampled sites from our data. For site i,

$$y_i | z_i \sim \text{Binomial}\,(J_i, z_i p_i) \tag{1.3}$$

where J_i is the number of surveys of site i and p_i is the probability of detecting the species. If survey-specific variables that affect detection probability, such as observer skill, effort, or time of day, are included, then we have a matrix $\mathbf{Y} = \{y_{ij}\}$ of observed occupancies and for each survey j of site i, $y_i | z_i \sim \text{Binomial}\,(J_i, z_i p_{i,j})$ where $p_{i,j}$ is defined through:

$$\text{probit}\,(p_{i,j}) = \alpha_0 + \alpha_1 x_{1,i,j} + \cdots + \alpha_V x_{V,i,j} \tag{1.4}$$

Note that both site-specific and survey-specific variables may affect detection prob-

ability and the same site-specific variables may be in both the occupancy and detection probability functions.

Estimates of the parameters are obtained through the posterior distributions. To get these posterior distributions, uninformative priors, such as normal distributions with large variances, are placed on the and parameters. The posterior distribution is created from thousands of iterations of the model, following a burn-in period. The median value is taken as the point estimate from the posterior distribution, standard deviations are calculated directly from the distribution, and 95% credible intervals are created from the 2.5% and 97.5% quantiles of the distribution. More details on the implementation of Bayesian models are described in Royle and Dorazio (2008), with a good example found in their Chapter 3.

Multi-Season Occupancy Model. The multi-season occupancy model version models changes in occupancy over time as a function of the underlying processes of local colonization and persistence rates (MacKenzie et al., 2003). For the multi-season model, the probabilities of occupancy for season or time period 1 are modeled in the single season as in Equations 3.7 and 1.2 above. The occupancy probabilities for the subsequent seasons or time periods are functions of the persistence rates, colonization rates, and the previous state of occupancy. The persistence rate is the probability that a previously occupied cell, site i, remains occupied between seasons t and $t+1$. (The local extinction rate is 1 minus this probability.) The colonization rate is the probability that a previously unoccupied cell becomes occupied between seasons t and $t+1$. Now $\mathbf{Z} = \{z_{i,t}\}$ is the matrix of the true occupancies of the N sampled sites during time periods $t = 1, \ldots, T$. The occupancies of each site are again assumed to be independent Bernoulli random variables. The probability that a site i is occupied is dependent on the occupancy at time period $t-1$. If the site was previously occupied, z_i, $t-1 = 1$, and the probability that it is currently occupied is equal to the persistence rate. If the site was previously unoccupied, z_i, $t-1 = 0$, then the probability that it

is currently occupied is equal to the local colonization rate.

The persistence and colonization rates can be generalized using logit or probit links, as in Equation 1.2, to incorporate habitat or temporal effects on the probabilities. Imperfect detection is included in the model with the same hierarchical structure as before. Let $\mathbf{Y} = \{y_{i,j,t}\}$ be our data, the observed occupancies of site i at survey j during time period t. Then,

$$y_{i,j,t}|z_{i,j,t} \sim \text{Bernouli}\left(z_{i,t}p_{i,j,t}\right) \tag{1.5}$$

1.5 Incorporating Neighboring Spatial Relationships

An assumption of the base occupancy model is that the sites are random, independent samples from across the landscape, and that there are an infinite number of points to theoretically sample (although one can calculate a finite occupancy under the Bayesian model framework, Royle and Dorazio (2008)). The SABAP data are a connected grid, a finite, lattice-structured data set of areal units, with samples of adjacent cells and with cell boundaries that are arbitrary to bird territories. Also, detailed covariate information for each site may not exist. Therefore, there may be a spatial pattern in the residual, known as spatial autocorrelation that should be modeled when estimating occupancy probabilities.

Presence-absence data may be collected at specific points, along line transects, or on areal units. Each type of collection would impose a different correlation and model structure on the data and desired areas of inference. For areal units, like the SABAP data, an autoregressive variable is used to account for the fact that neighboring sites are likely to be more similar than two sites chosen at random and its addition to a model may improve the model's predictive capabilities (Hoeting et al., 2000).

An autoregressive variable is a Markov Random Field, dependent only on its neighbors' values. For the SABAP data, a site's neighbors are those sites that are

adjacent or diagonal to it (Figure 1.3), Most sites have 8 neighbors and edges have fewer. Other data sets may have different definitions of neighbors.

In its most general form, an autoregressive variable has the following conditional distribution:

$$\eta_i|\boldsymbol{\eta}_{-i} \sim \text{Normal}\left(\mu + \sum_j a_{ij}\left(\eta_i - \mu\right), \sigma_i^2\right) \tag{1.6}$$

where η_{-i} is all sites not equal to i, and μ is the overall mean of the η variables. Because of the Hammersley-Clifford Theorem and Brook's Lemma, the above conditional distributions are guaranteed to have a joint distribution (Banerjee et al., 2004). Let $\mathbf{A} = \{a_{i,j}\}$ and $\mathbf{M} = diag(\sigma_i^2)$, then:

$$\eta \sim \text{Normal}\left(\mathbf{0}, \left(\mathbf{I} - \mathbf{A}\right)^{-\mathbf{1}}\mathbf{M}\right) \tag{1.7}$$

Some reasonable restrictions are added to the above model to make it estimable. Let $\mathbf{W} = \{w_{i,j}\}$ be a weights matrix, with $w_{i,j} = 1$ if sites i and j are neighbors $(i \neq j)$, and 0 otherwise. Let $|n(i)|$ be the number of neighbors of site i. Then:

$$\begin{aligned} \mathbf{A} &= \rho\mathbf{W} \\ \mathbf{M^{-1}} &= \sigma^{-2}\text{diag}\left(|n(i)|\right) \end{aligned} \tag{1.8}$$

Equations 1.7 and 1.8 give us a proper conditional autoregressive (CAR) model. Maximum likelihood methods can be used to solve its parameters. But because the estimation process requires inverting matrices, this advantage becomes computationally prohibitive for large numbers of sites. Another disadvantage of the proper CAR model is that it can only model a narrow range of spatial patterns. In general, a

proper CAR model will lead to a maximum Moran's I value of 0.52 for the variable (Banerjee et al., 2004). The Moran's I statistic is commonly employed to measure a variable's spatial connectivity. It ranges from approximately -1 to 1; a nonzero value means that the variable is more similar (if the statistic is greater than 0) or less similar (if the statistic is less than 0) to its neighbors than would be expected by chance (Wu et al., 2009).

Instead, we set $\rho = 1$ and Equation 1.7 becomes an intrinsic conditional autoregressive (ICAR) model, which can model the full range of spatial autocorrelations, leading to Moran's I values of close to -1 to +1. This setting makes the joint distribution improper, although the conditional distributions are still proper. However, the distribution is used as a prior and still leads to proper posterior distributions, which is in line with the priors usually placed on the other parameters. For the ICAR model, Equation 1.6 simplifies to:

$$\eta_i|\eta_{-i} \sim \text{Normal}\left(\frac{\sum_j \eta_{ji}}{|n(i)|}, \frac{\sigma^2}{|n(i)|}\right) \tag{1.9}$$

The ICAR variable lends itself nicely to Bayesian methods and is flexible because it can be used with different definitions of neighbors. (For example, for irregularly shaped sites, neighbors may be defined as all sites whose distance between their centroids is within a certain cut-off.) When an ICAR variable is included in a generalized linear model, it is added as a random effect and it will be set to have a mean of 0 so as to be identifiable.

Starting in the 1990s, the ICAR has been gaining popularity as computing abilities increase and Bayesian methods have become mainstream. But despite all of its advantages, the ICAR is not a panacea for residual spatial autocorrelation. Recent publications have brought its helpfulness into question, as it appears to bias models' fixed effects coefficients and to inflate their standard errors (Reich et al., 2006; Dor-

mann et al., 2007; Paciorek, 2010; Hodges and Reich, 2010; Hughes and Haran, 2010). In Chapter 3, our simulation studies confirm these biases and inflated standard errors with ICAR variables in an occupancy model framework.

Because of these deficiencies, we use an adaptation to ICAR to build a better spatial occupancy model, which we call a *restricted spatial regression* (RSR) (Johnson et al., 2013). The RSR model is described in Chapter 2.

1.6 Motivation: Multi-Season Models for Presence-Absence Data

Occupancy models that span multiple time periods are alternately called dynamic occupancy models or multi-season occupancy models. They predict occupancy probabilities for each time period and the local persistence (i.e., site survival) and colonization probabilities for each site and between each time period.MacKenzie et al. (2003) introduced the first dynamic occupancy model; this model is popular, having been cited 561 times (Google search, May 31, 2013).

The MacKenzie et al. (2003) dynamic occupancy model is unsatisfactory because it does not explicitly match colonization and extinction probabilities to the true patterns of occurrences that have been seen or predicted for the area of inference. Intuitively, one knows that the dynamic processes of colonization and extinction will mostly happen along the edges of a species' range and this fact should be taken into account in the dynamic models.

There have been some recent examples in the literature to more explicitly link colonization and extinction with spatial relationships (Bled et al., 2011; Yackulic et al., 2012). Most notably, Yackulic et al. (2012) used an autologistic term to model the colonization and extinction probabilities as functions of the occupancy rates of neighboring sites, but the model assumed a constant occupancy probability for the first year for the entire area of inference, which is unlikely to be true for other data sets and is not true for the SABAP data.

Therefore we adapted the Hooten and Wikle (2010) model of diffusion to account

for non-detection and added an explicit spatial process for year 1. The model has a strong theoretical background, based on cellular automata and diffusion processes, demonstrating its effectiveness and utility for other, related data. The new model is also extremely flexible due to its Bayesian framework and can easily be adapted for more complex long distance dispersal and/or persistence probabilities. The combination of these factors will make the multi-season occupancy model presented in Chapter 4, the preferred method of analysis for similar data sets going forward.

Figure 1.1: SABAP 1 data were collected in Botswana, Namibia, Zimbabwe, Lesotho, Swaziland, and South Africa. For my dissertation, I used the data from South Africa only.

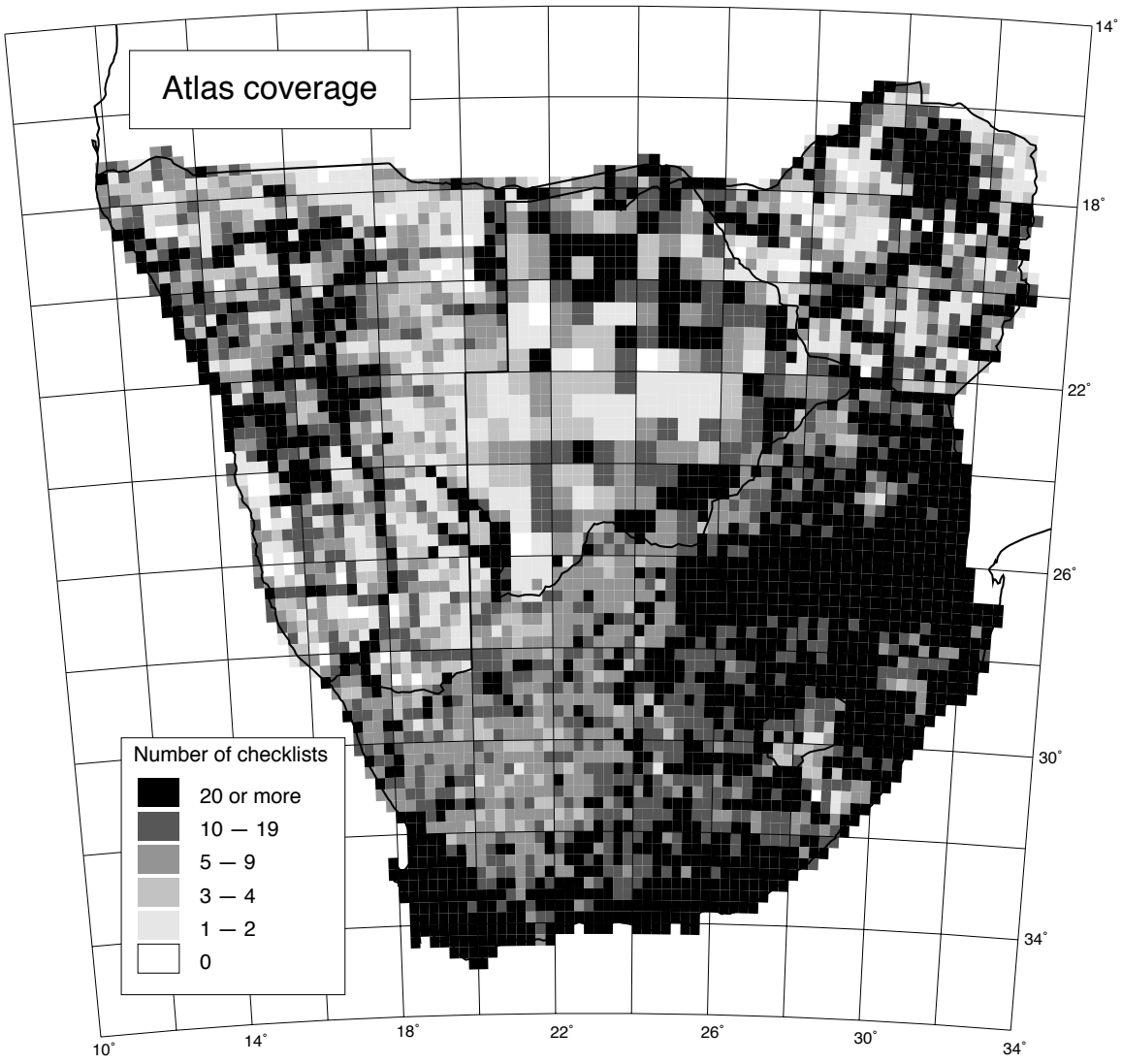


Figure 1.2: The spatial coverage of the SABAP 2 data, as of 19 August 2013.

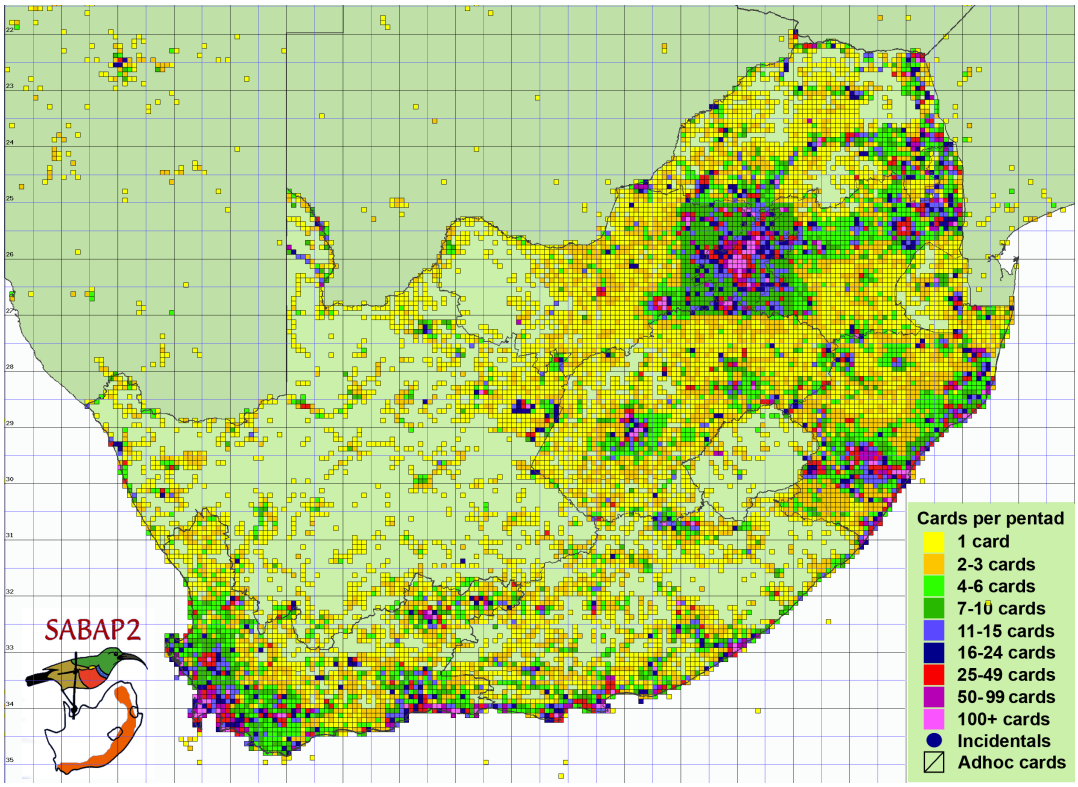


Figure 1.3: Neighborhood structure of the data. A site's neighbors are the sites that are adjacent or diagonal to it. For example, site 5 has 8 neighbors. A site that occurs on an edge of the study area will have fewer neighbors.

1	2	3
4	5	6
7	8	9

Chapter 2

SPATIAL OCCUPANCY MODELS FOR ATLAS DATA: WITH APPLICATION TO THE SOUTHERN GROUND HORNBILL

2.1 Introduction

Predicting species occurrences and investigating species-habitat associations are central questions in ecology, and knowing species' distributions is important to managing species and conserving biodiversity (Block and Brennan, 1993; Jones, 2001; Austin, 2002; Schaefer and Krohn, 2002; Zabel et al., 2002; MacKenzie et al., 2006). For example, management plans for endangered species are based on habitat use (Zabel et al., 2002), and IUCN Red List categories are based on analyses of area of occupancy and extent of occupancy (IUCN, 2001).

Species occurrences and habitat associations are usually modeled with environmental variables, dating back to at least Grinnell (1917). Over time, the qualitative descriptions that correlate species' distributions with their environment have given way to quantitative analysis and a huge variety of models have been developed and used for this aim (Guisan and Zimmermann, 2000; Scott et al., 2002). One type of data that appears often in these analyses is presence-absence data, where the detection/non-detection of the species of interest is recorded along with environmental variables and other relative information about the survey and site. This type of data has often been studied with logistic regression to predict species occurrence (Guisan and Zimmermann, 2000; Scott et al., 2002).

Because the logistic regression cannot account for the possibility of a species being present but not seen, it will usually provide biased estimates of species occurrence

and species-habitat relationships (Tyre et al., 2003; Altwegg et al., 2008; Kéry et al., 2010). In 2002, MacKenzie et al. developed an approach, termed occupancy modeling, which became the new standard for estimating species occurrences while accounting for detection errors. The occupancy model accounts for detection- nondetection in two stages: one component that models true occupancy, and another component that models detection, given that the species occurs at the site.

The occupancy model is often applied to data sets where it is known at the onset that sites are not independent. For example, sites are not chosen randomly, sites are adjacent, and/or the species' territory is larger than the space between the sites. When non-independence is ignored, standard errors may be underestimated, leading to incorrect model-selection and incorrect conclusions on species-habitat associations (Latimer et al., 2006).

For many studies and data sets, adding spatial autocorrelation is a needed improvement to the occupancy model. For point data, one would use geostatistical methods such as kriging to incorporate spatial autocorrelation (Cressie and Wikle, 2011). For areal data, spatial autocorrelation is commonly incorporated in models of occurrence through an intrinsic conditional autoregressive (ICAR) random effect (Banerjee et al., 2004; Cressie and Wikle, 2011). The ICAR variable has intuitive appeal and can readily be incorporated into a hierarchical model and then estimated with general MCMC methods within a Bayesian framework. Among the ICAR model's disadvantages is that it is computationally intensive and it may bias the covariat estimates and standard errors (Hodges and Reich, 2010; Hughes and Haran, 2010). To eliminate the bias arising from the correlations between the fixed effects of interest and the residual spatial autocorrelation, Hodges and Reich (2010) replaced the ICAR random effect with a random effect that is orthogonal to (and therefore uncorrelated with) the fixed effects, a model they called *restricted spatial regression* (RSR). Hughes and Haran (2010) expanded their work by using dimension reduction on this orthogonal spatial effect to reduce the computation time.

In 2013, Johnson et al. introduced the RSR models in the context of occupancy models. Here we further refine the capabilities of these models, guide the reader through their implementation, and provide a more in-depth comparison to the traditional, nonspatial occupancy model and an ICAR spatial occupancy model. Our work broadens the appeal of the RSR by demonstrating its added effectiveness for complex data sets with sparse detections.

We apply the models to the Southern ground hornbill (*Bucorvus leadbeateri*) with data collected through the Southern African Bird Atlas Project (http://sabap2.adu.org.za/). The ground hornbill is the largest of the hornbill family, is the only entirely carnivorous hornbill, and with males measuring 90-130 cm long and 4.2 kg, is one of the largest carnivorous avian species in Africa (Knight, 1990; Hockey et al., 2005). It is an easily identifiable bird that is widespread throughout southern Africa, but its expansive territory sizes exceeding 100 km^2 lead to low densities (Harrison et al., 1997b). Because it is uncommon, moves through thick vegetation, and traverses multiple private farms with inaccessible lands (Theron, 2011), detections tend to be low and the need for occupancy models is particularly acute.

In 2010, the IUCN listed the species as "Vulnerable" throughout its range because its abundance had decreased considerably outside protected areas when compared to historical distributions (Hulley and Craig, 2007; BirdLife International, 2012). Many biologists believe that the ground hornbill should be re-classified as "Endangered" or "Critically Endangered" within South Africa because declines have been especially high at this most southern end of its range (Theron, 2011). But with limited resources to study it, little is known about the ground hornbill, including its habitat preferences and where exactly it occurs (Theron, 2011). This lack of data and lack of understanding of the causes of the hornbill's decline are the key conservation issues associated with the ground hornbill (Morrison et al., 2005).

Our study quantified the habitat requirements of the ground hornbill and assessed the role of protected areas for the persistence of this species. The spatial occupancy

models, and in particular the one with the RSR addition, produced a model with a better fit to the ground hornbill data than the previous models and thus can help guide managers on the importance of protected lands and possible reintroductions of this species. The example can be extended to any other species that occurs in South Africa using the same database, to any other atlas project, or to any presence-absence data that are collected on areal units.

2.2 Methods

2.2.1 Data

The second Southern African Bird Atlas Project (SABAP 2) is a database of bird lists collected by volunteer birders, called "citizen scientists", following a strict protocol and using gridded locations throughout South Africa (found at http://sabap2.adu.org.za/; Greenwood (2007)). It is a record of detection/non-detection data for every bird species occurring in South Africa from June 2007 to present. The gridded sites are 5-minute latitude by 5-minute longitude rectangular cells (Harebottle et al., 2007). Each site is approximately 8 x 7.6 km, and while this size varies slightly due to the curvature of the earth, the change within South Africa is not very large and all units will be assumed to be of the same size. There are 17,444 of these lattice-structured grid cells covering all of South Africa (Harebottle et al., 2007). Each bird list that is collected represents one survey of one site: non-detection records are deduced by a species' absence from the checklist. Bird lists are collected during a minimum observation time period of two hours of intense birding over a maximum of five days, where many hours can be spent only passively birding. For each survey, two measures of effort are recorded: the number of hours spent intensively birding and the total number of hours that the checklist included. As of April 18, 2012, 1,028 volunteers had submitted 68,529 surveys and 60% of all of the gridded sites in South Africa had been surveyed (http://sabap2.adu.org.za/).

In South Africa the ground hornbill's range is restricted to the eastern side of the country, with the predominant detections occurring in Kruger National Park and a second cluster of detections occurring in the southeast part of the country (Figure 2.1). It resides in grassland, woodland and savannah habitats, and lives in co-operative breeding groups whose territory sizes are more than 100 km^2 (Kemp and Kemp, 1980; Vernon, 1986). The territories are determined by nesting sites and food resources: other factors do not seem to influence the distribution and spacing of territories (Theron, 2011). Note that this territory size is much larger than an atlas site; therefore occupancies cannot occur in isolation, and accounting for positive spatial autocorrelation is essential for this species and data set.

The ground hornbill is a resident with fixed territories and little to no seasonal migrations (Kemp and Begg, 1996). Every 2-3 years, one chick is raised within a breeding group; the chick then stays on as a helper to the group until age 3-5 years (Theron, 2011). The dispersal and survival of the sub-adults when they leave the breeding group is poorly understood (Theron, 2011). Because of their long life cycle, we considered the SABAP 2 data as a closed period, i.e., we assume that the species did not go extinct and did not colonize new grid cells between July 2007 and December 2011, the period considered here. This wide temporal scale leads to a broader interpretation of the meaning of occupancy for this species and possibly elevated occupancy rates, but it is in line with the stability and longevity of the family groups.

From the few studies that have been conducted on the ground hornbill, no patterns between its movements and landscape feature have been found and the ground hornbill habitat requirements are poorly understood (Theron, 2011). There is some speculation that grazing degrades the land and may negatively affect the ground hornbill, but no quantitative analyses have been completed to support or refute the claims (Theron, 2011).

Although some detections of the ground hornbill were recorded more inland, the sites considered in this paper are the two core areas of its range (Figure 2.1). The

southern region extends eastward from 29.2° longitude and the northern region extends eastward from 30.7°. There are 2,131 sites of which 205 had at least one ground hornbill detection by 31 December 2011. The sites included in this study were surveyed a median of 3 times; 279 sites were surveyed zero times, 465 sites were surveyed one time, and one site was surveyed a total of 375 times.

The occupancy model and spatial occupancy model both allow unequal sampling of sites by treating each survey as an independent Bernoulli trial, which implicitly gives more weight to the sites that were visited more often, and consequently gives more weight to the values of their site-specific covariates. In order to limit the effect of the unequal samples, we randomly chose 50 surveys from each of the 53 sites that had more than this number of surveys. Choosing 50 surveys was a compromise between using as much data as possible and minimizing the impact of the unequal sampling. Using a subset of the data allowed us to test model robustness and models with fewer and with more surveys were fit, and results were comparable.

Independent variables were incorporated in the models to find relationships between environmental factors, survey-specific details, and the occupancy and detection probabilities. For both components of the model we considered the following site-specific variables: the proportion of the site that was each major vegetation type (grassland, savannah, Indian Ocean Coastal Belt, and forests for this region of South Africa, i.e., biomes as defined by Mucina and Rutherford 2006); the proportion of the site that was cropland (Ramankutty et al., 2010a); proportion of the site that was pastureland (Ramankutty et al., 2010b); proportion of the site that was in a protected area such as a national park, national forest or game reserve (Rouget et al., 2004); the logarithm of the human population per square kilometer; and an indicator variable of whether a major river, defined as a Strahler stream order 5, ran through the site (Silberbauer, 2007). All the covariate data were originally at a higher spatial resolution than the bird surveys; therefore we aggregated it into the sites by taking the area (in km^2) or sum (for the population variable) of the variable of interest divided by the

area of the site, using ArcGIS, Version 10.0. (ESRI, 2011).

The effect of survey-specific variables on the detection probabilities were also considered: year as a continuous variable (with year 2007 = year 1); year as a factor; effort measured as log-intensive hours; effort measured as log-total hours; total number of species that were seen on each survey (standardized); day-of-year (standardized); and (standardized) day-of-year squared, as there may be a peak in detection relating to seasonal changes or breeding season. Year was considered as a continuous variable to see if there was a trend in detection over time and it was separately considered as a factor to be a proxy for changes in annual rainfall. The survey-specific variables were standardized as described above for easier comparisons between coefficients and for numerical reasons.

2.2.2 Models

Occupancy Model. Detailed descriptions of this model can be found in MacKenzie et al. (2006) or Royle and Dorazio (2008). Below, the state-space formulation used with a Bayesian framework, first introduced by Royle and Kéry (2007), is outlined because it can incorporate the spatial autocorrelation component that is added in the following section.

Let $\mathbf{z} = \{z_1, \ldots, z_n\}$ be the true occupancies of the n sampled sites. For site i,

$$z_i \sim \text{Bernoulli}\,(\psi_i) \tag{2.1}$$

where ψ_i is the probability of occupancy, so $z_i = 1$ if the site is occupied and $z_i = 0$ if the site is not occupied. The site-specific covariates ($\mathbf{x_i}$) that may affect occupancy are incorporated into an expression for the probability of occupancy with a logistic equation:

$$\text{probit}\left(\psi_i\right) = \mathbf{x_i^T}\boldsymbol{\gamma} \tag{2.2}$$

where $\mathbf{x_i^T}$ is a vector of the independent variables and $\boldsymbol{\gamma}$ is the vector of regression coefficients. Let $\mathbf{Y} = \{y_{i,j}\}$ be the data, the observed detections of the J surveys for each of the n sampled sites. For survey j of site i,

$$y_{i,j}|z_i \sim \text{Bernoulli}\left(z_i p_{i,j}\right) \tag{2.3}$$

where $p_{i,j}$ is the probability of detecting the species. If the site is not occupied, $z_i = 0$, and the probability of detecting the species is 0. Site-specific *and* survey-specific variables that affect detection probability are incorporated through a link function, as with the occupancy probability:

$$\text{probit}\left(p_{i,j}\right) = \mathbf{x_{i,j}^T}\boldsymbol{\beta} \tag{2.4}$$

Note that the same site-specific variables may be in both the occupancy and detection probability functions. For both ψ_i and $p_{i,j}$, other link functions are possible, but the probit link was used because of its computational efficiency (Johnson et al., 2013).

We first fit the model under a maximum likelihood framework with the "unmarked" package in R (Fiske and Chandler, 2011). The best model was chosen using forward model selection based on p-values from z-tests. We then fit Bayesian versions of the selected models with the "stocc" package in R (Johnson, 2013) for a direct comparison with the spatial occupancy models, which were also analyzed using a Bayesian framework and fit with the "stocc" package. They were run for 40,000

iterations with a burn-in of 10,000 iterations. Every 5th sample was retained for a total posterior sample of 6,000 iterations.

First spatial occupancy model (intrinsic conditional autoregressive, ICAR). Several authors have used spatial occupancy models by adding a CAR-process random effect, $\boldsymbol{\eta}$, in the occupancy probability function (e.g., Högmander and Møller (1995); Sargeant et al. (2005); Magoun et al. (2007)). One way to incorporate the CAR-process random effect would be to augment Equation 2.2 as follows:

$$\text{probit}\left(\psi_i\right) = \mathbf{x_i^T}\boldsymbol{\gamma} + \boldsymbol{\eta}_i \tag{2.5}$$

$$\eta_i|\boldsymbol{\eta}_{-i} \sim \text{Normal}\left(\frac{\sum w_{ik}\eta_{ik}}{|n(i)|}, \frac{\sigma^2}{|n(i)|}\right) \tag{2.6}$$

The distribution of the CAR-process variable, η, is conditional on its value in the other cells, a vector labeled above as $\boldsymbol{\eta}_{-1}$. The expected value of the CAR-variable at a given site is an average of the values of the CAR variables of its neighbors. For our data with its lattice-structure, we used the weights $w_{i,k} = 1$ if cell k is adjacent or diagonal to site i, and 0 otherwise. The number of neighbors that site i has is $|n(i)|$, and it is used to weight the variance for each site. Most sites have 8 neighbors but edges have fewer.

The intrinsic CAR model (Equation 2.6) is improper and may not be written as a full, joint probability model but the set of conditional distributions may be used as a prior model to obtain a proper posterior. For more details on the CAR model, the reader is referred to Cressie and Wikle (2011). This model can be fit using WinBUGs (Lunn et al., 2000); see e.g., (Yackulic et al., 2011).

We instead ran the model using the "stocc" package in R, package version 1.0-5 (Johnson, 2013) to directly compare the ICAR with the RSR model results. By writing Equation 2.6 in matrix form, our spatial occupancy model may be written as

follows:

$$
\begin{aligned}
&z_i \sim \text{Bernoulli}\,(\psi_i) \\
&\text{probit}\,(\psi_i) = \mathbf{x_i^T}\boldsymbol{\gamma} + {}_i \\
&\boldsymbol{\eta} \sim \text{Normal}\,(\mathbf{0}, \tau^{-1}\mathbf{Q^{-1}}) \\
&y_{ij}|z_i \sim \text{Bernoulli}\,(z_i p_{ij}) \\
&\text{probit}\,(p_{ij}) = \mathbf{x_{ij}^T}\boldsymbol{\beta}
\end{aligned}
\tag{2.7}
$$

where τ is the inverse of the variance, i.e., $\tau = \frac{1}{\sigma^2}$. $\mathbf{Q}$ is a precision matrix with elements:

$$
Q_{ik} = \begin{cases} -1 & \text{if } i \neq k \text{ and } i \text{ and } k \text{ are neighbors} \\ 0 & \text{if } i \neq k \text{ and } i \text{ and } k \text{ are not neighbors} \\ |n(i)| & \text{if } i = k \end{cases}
\tag{2.8}
$$

In Equation 2.7, the CAR-process variable is written as a full, improper prior distribution, which is equivalent to the set of conditional priors of Equation 2.6. The "stocc" package samples from the joint distribution (Equation 2.7), while WinBUGS would sample from the conditional distributions (Equation 2.6) to obtain the posterior distributions. Although "stocc" and WinBUGS use different sampling algorithms, the results would be similar using either software option. The advantage of sampling from the joint distribution is that the algorithm is faster than when one uses the conditional distributions.

The parameters were assigned the following priors:

$$\begin{aligned}
\boldsymbol{\gamma} &\sim \text{Normal}\left(\mathbf{0}, \mathbf{Q}_{\boldsymbol{\gamma}}^{-\mathbf{1}}\right) \\
\boldsymbol{\beta} &\sim \text{Normal}\left(0, \infty\right) \\
\tau &\sim \text{Gamma}\left(0.5, 0.0005\right)
\end{aligned} \tag{2.9}$$

where $\mathbf{Q}_{\gamma} = \mathbf{X}^{\mathbf{T}}\mathbf{X}/N$ and N is the number of sites. These priors were chosen for the the $\boldsymbol{\gamma}$ coefficients because they made the model run faster than flat priors but did not seem to affect the posterior distributions. The prior for the spatial parameter, τ, followed the recommendation by Kelsall and Wakefield (1999).

Because of the slower run-time (see Results section), full, forward model selection was not completed for the spatial occupancy models. Instead, we fit a model with the same parameters as the best nonspatial occupancy model. Two more parsimonious models were also fit because the addition of the spatial effects made some previously significant covariates become nonsignificant with 95% credible intervals that overlapped 0. 6,000 iterations were retained for the posterior distribution (a total of 40,000 iterations were run with a thinning rate of 5), and the first 2,000 of the retained iterations were discarded as the burn-in.

Second spatial occupancy model (restricted spatial regression, RSR). We also analyzed the data using a restricted spatial regression (RSR) adaptation to the ICAR model. This method is described in Hughes and Haran (2010) and Johnson et al. (2013). The model is similar to Equation 2.7 but now $\boldsymbol{\eta}$ is replaced by $\mathbf{K}\boldsymbol{\alpha}$. $\mathbf{K}$ is a subset of the eigenvectors for what is known as the Moran operator matrix:

$$\Omega = \frac{N\mathbf{P}^{\perp}\mathbf{A}\mathbf{P}^{\perp}}{\text{sum}(\mathbf{A})} \tag{2.10}$$

where $\mathbf{A}$ is an adjacency matrix with elements:

$$a_{ik} = \begin{cases} 1 & \text{if } i \neq k \text{ and } i \text{ and } k \text{ are neighbors} \\ 0 & \text{otherwise} \end{cases} \tag{2.11}$$

and

$$\mathbf{P}^{\perp} = \mathbf{I} - \mathbf{X}\left(\mathbf{X}^{\mathbf{T}}\mathbf{X}\right)^{-\mathbf{1}}\mathbf{X}^{\mathbf{T}} \tag{2.12}$$

is the projection matrix (Hughes and Haran, 2010). The Moran operator matrix may be used to form a general version of the Moran's I statistic. The eigenvalues of the operator matrix determine the distribution for the Moran's I statistic (Boots and Tiefelsdorf, 2000). Let:

$$\alpha \sim \text{Normal}\left(\mathbf{0}, \tau^{-\mathbf{1}}\left(\mathbf{K}^{\mathbf{T}}\mathbf{Q}\mathbf{K}\right)^{-\mathbf{1}}\right) \tag{2.13}$$

where $\mathbf{Q}$ is the ICAR precision matrix as above (Equation 2.8). Note that if $\mathbf{K}$ is a matrix of all possible eigenvectors, the RSR model reverts back to the ICAR model.

The dimension of $\mathbf{K}$, i.e., the number of eigenvectors to include in the model, is set by the user. In our models, we included a restriction to 400 eigenvectors, which is 1/4 of the number of surveyed sites, and a more restricted model with 160 eigenvectors, which is 1/10 of the number of surveyed sites and is the default restriction suggested by Hughes and Haran (2010). For the rest of the paper, we refer to the restrictions as the number of columns of $\mathbf{K}$ that are included. Table 2.1 summarizes the models that were fit.

Table 2.1 summarizes the models that were fit. As a reminder, there were four model structures that we compared: the nonspatial model, the ICAR model, the RSR

model restricted to 400 eigenvectors, and the RSR model restricted to 160 eigenvectors. For each model structure, the results from three sets of detection covariates were tested and compared. A Full model that included all covariates that were significant at the $p = 0.05$ level when added using forward model selection and likelihood-based inference; a Middle model that removed the detection covariates that were marginally significant; and a Limited model that only included the two most significant detection covariates.

Model Selection Criteria. The models were compared using three different statistics, one of which was the posterior predictive loss criterion (PPLC) (Gelfand and Ghosh, 1998). Similar to other model selection criteria, it is a combination of a goodness-of-fit measure and a penalty for over-fitted models. Because it is unknown how well the PPLC would perform for hierarchical models, two other model selection criteria were implemented to determine the best fitting model. Similar to Moore and Swihart (2005), an area under the curve (AUC) statistic was calculated using the median occurrences, the $\hat{z}_i$'s from Equation 2.7, outputted from the model and then analyzed with the "ROCR" package in R (Sing et al., 2007). If the species has been detected at a site, then its occurrence is known and $\hat{z}_i = 1$ for that site. At the sites where the species was not detected, the true occurrences are not known. Because the AUC statistics is dependent on prediction of these unknown values, this statistic should be interpreted carefully. We created a third statistic that we called the "absolute psi conditional" (APC) to see how model estimates compared when true detection histories were taken into account:

$$APC = \frac{1}{n} \sum_{i=1}^{n} |\hat{\psi}_i - \hat{\psi}_{conditional,i}| \tag{2.14}$$

As before, n is the number of sites, the $\hat{\psi}_i$'s are the chosen model's predicted

occupancy probabilities, and the $\hat{\psi}_{conditional,i}$'s are the realized occupancy probabilities that take the detection histories into account. If the species was detected at a site, $\hat{\psi}_{conditional,i} = 1$ for that site. If the species was not detected, then the probability represents the fraction of the MCMC iterations where the model estimates that the site was occupied and $\hat{z}_i = 1$.

2.3 Results

For the nonspatial occupancy model, the following variables significantly affected detection: number of species on checklist, proportion of cropland, proportion of protected area, number of hours spent intensively birding, proportion of savannah, and the day-of-year of the survey (Table 2.2). Only the proportion of protected area significantly affected occupancy. The models that include all of these covariates are the "Full" models. Occupancy probabilities ranged from 0.178 to 0.787.

Savannah and day-of-year were only marginally significant, so a more parsimonious model that excluded these predictors, a "Middle" model, was also run; the resulting parameter estimates were similar to the Full model (Table 2.2). In order to test the effect of the potential confounding relating to the proportion of protected area affecting both detection and occupancy, a third model, the "Limited" model, was fit with only detection probability covariates that were more significant than protected area: the number of species on the checklist and the proportion of cropland. The parameter estimates were different from the models with more covariates (Table 2.2) but they resulted in a similar pattern of predicted occupancy probabilities (Figure 2.2).

The ICAR models resulted in similar patterns and estimates as the three Bayesian nonspatial models described above (Table 2.3). Each ICAR model took over 19 hours to run and the posterior plots had high correlations between iterations.

The RSR models produced two types of results– some were similar to the nonspatial and ICAR models described above and others picked up more residual spatial

autocorrelation and produced different results (Figure 2.3). The Full RSR-400, Middle RSR-400, Middle RSR-160, and Limited RSR-160 models were the ones that produced results similar to the nonspatial and ICAR models (Table 2.4). The Limited RSR-400 and Full RSR-160 models picked up the residual spatial autocorrelation and produced substantially different results and a substantially different range map (Figure 2.3). When restricted to 400 columns, each RSR model took 1–2 hours to run; when restricted to 160 columns, each model took slightly less than 1 hour to run.

All three model selection criteria chose the RSR models as the best-fitting. In particular, the PPLC model selection criterion picked the Full RSR-160 model, with the Limited RSR-400 model, as a close second. The AUC and the APC model selection criteria both switched the best and second-best order, and picked the Limited RSR-400 model as the best fitting with the Full RSR-160 being a close second (Table 2.1). Because the best model choice between the Limited RSR-400 model and the Full RSR-160 was ambiguous, we will focus our discussions on the Full RSR-160 model.

2.4 Discussion

Neither the nonspatial nor the ICAR occupancy model produced range maps that truly reflected the ground hornbill detections. Because no covariates other than protected area affected occupancy, these models had dichotomous occupancy predictions, with high occupancy probabilities within a protected area ($\hat{\psi} = 0.79$) and a lower occupancy probability of $\hat{\psi} = 0.18$ everywhere else. These models were unable to predict the ground hornbill occurrences in the southern region, as can be seen by comparing Figures 2.1 and 2.2.

In addition to not being able to differentiate between sites where the ground hornbill did and did not occur, these models did not reproduce the spatial patterns that the ground hornbill distribution likely exhibits. Since the territory of a ground hornbill group is larger than a site's area, a detection at a site should cause the neighboring sites to have elevated occupancy probabilities. This was not the case

produced from these models. Instead, maps of the realized occupancies, which are the probabilities of the ground hornbill occupying a site given the detection history of the previous surveys of the site, show many isolated detections of ground hornbill in areas where the expectation of finding them is very low (Appendix, Figure A.2). In addition, the ICAR's long run time (over 20 hours) minimized the model testing that was feasible.

The RSR model used the eigenvectors of the Moran operator matrix to reduce the dimensionality of the covariance matrix and make the spatial process orthogonal to the covariate matrix, which decreased its computation time and removed the collinearity between parameters. These two differences from the ICAR model enabled it to uncover the residual spatial autocorrelation in the southern area that better captured the ground hornbill's presence in this part of its range (Figure 2.3). The restriction to 160 columns, which was 1/10 of the possible eigenvectors as recommended by Hughes and Haran (2010), did not appear to have drawbacks, but rather was beneficial as it uncovered the spatial correlation that was seen in the raw data and expected by the ground hornbill's biology, but not captured by the other models.

As with all smoothing techniques, the desired level of smoothing– equivalent here to the most appropriate level of restriction– is up to the researcher. The models are not particularly sensitive to the exact cut-off for the number of eigenvectors, but the maps gradually become smoother as the number of eigenvectors is decreased. Conversely, if all eigenvectors were included, then the RSR model would be equivalent to the ICAR model. From simulation runs not shown here, we found a minimum of 150 eigenvectors and a maximum of 500 eigenvectors to be appropriate, even for data sets with as many as 10,000 sites.

All three model selection criteria picked the RSR models as the best-fitting, which matched the visual inspections of the predicted occupancy maps. The PPLC statistic was used by Johnson et al. (2013), but has otherwise not been applied in an occupancy model framework. For the ground hornbill data, the statistic did closely match the

models selected by the other two criteria and the visual inspection of the maps. Within a specific model structure, the PPLC tended to have the lowest value for the model with the most parameters and the highest value for the most parsimonious models, suggesting potential bias towards over-fitting the models. It should also be noted that because the statistic tries to take into account model complexity, its value is dependent on the model structure and therefore it cannot be concluded that the PPLC can be used to choose between an ICAR and RSR model. Also, the PPLC is supposed to be used similarly to AIC, to pick between models, but its value does not quantify whether the model actually fits the data.

The AUC statistic was very close to 1 for all models. In an occupancy model framework, the responses are not known for all sites– some occupancies (i.e., the responses) are predicted by the model. Therefore, the AUC statistic is essentially testing whether we correctly predicted our predictions, and hence its tendency to be close to 1. Nevertheless, its model selection matched that of the APC statistic.

Similar to the AUC statistic, the APC model selection criterion tried to test how well our predictions matched the data albeit for the APC statistic, the detection histories of each site were taken into account in its calculations. Since predictive occupancy maps are generally drawn for the same area from which the surveying was conducted, it makes sense to include the detection histories when assessing model fit. Similar to the PPLC statistic, most of the models had very similar APC statistics but the Limited RSR-400 and the Full RSR-160 had much lower values, matching the model selection of the other two criteria. We believe that APC has future value in assessing model fit for occupancy models, whether they be spatial in nature or not. An advantage that it has over PPLC is that its value helps with model selection and assesses model fit– the researcher would want the APC statistic to be as close to 0 as possible. A disadvantage of the APC is that it does not test whether a less complex model would fit the data equally well.

Beyond comparing the ICAR and RSR models, our analysis provided quantitative

evidence of the importance of protected areas to the ground hornbill. The coefficient associated with protected areas was much larger for the RSR model (2.46, SE = 0.522, Table 2.5) than for the nonspatial and ICAR occupancy models (1.75, SE = 0.192, Tables 2.2 and 2.3), suggesting that a protected area has a greater effect on occupancy than the other models realized.

Whether trying to distinguish between savannah and grassland biome preferences, or determining the influence of land development by looking at whether a site was crop or pasture land, none of these habitat variables came close to significantly affecting the occupancy probabilities of the ground hornbill. This result suggests that there is an inherent resource benefit on the protected areas, or it could signify that the ground hornbill's decline on unprotected lands is not related to food resources, but to other causes that have been proposed: poisoning, persecution for breaking windows, loss of nesting sites, or collisions with power lines (Morrison et al., 2005; Theron, 2011).

Several covariates affected the detection probabilities. The number of species on a checklist can be interpreted as the observer's skill– more species on a list increases detection probabilities because it indicates that the observer is able to identify more species. The number of hours spent birding also positively influenced the detection probabilities: the more effort that is put into the survey, the more likely one is to see an elusive bird such as the ground hornbill. Being on a protected area increased the detection probabilities, which may be a reflection of observer bias. Because birders know that ground hornbill populations are healthy on protected areas, they may be more ready to spot and identify a hornbill than when they are on private lands and have no such expectations. Cropland decreased detection probabilities, which may be a reflection of limited access to private farms or low detectability between crops. Savannah vegetation may similarly lower detection due to thick vegetation.

Our results highlighted future survey needs: note that the most southern tip of the data set had very few surveys conducted, with most sites visited 0 or 1 time (Appendix, Figure A.1). The relatively high, predicted occupancy probabilities seen

in the Full RSR model (Figure 2.3) show potential for previously undetected ground hornbill territories in this area. These predicted occurrences correspond well with area where ground hornbills were found during the first Southern African Bird Atlas Project, conducted between 1987 and 1992 (Harrison et al., 1997b).

2.5 Conclusion

We find the RSR models to be a good alternative to ICAR, are easily implemented in R, and produce estimates in a fraction of the time. These models are new and need to be explored further through simulations and applications to obtain guidelines on their use, but our analyses shows them to be very promising when applied to occupancy models. Our data set had 2100 sites; the RSR method and the "stocc" package make a spatial occupancy model analysis feasible for even larger data sets.

The Southern ground hornbill example demonstrates how the RSR model can produce a better fit than the traditional, nonspatial and ICAR occupancy models, neither of which produced results that fit against the known detections. The RSR model can be used on any presence-absence data set consisting of neighboring sites. The SABAP 2 database alone has this type of data on the 800+ species that occur in South Africa and similar atlas data has been collected at a range of spatial scales in countries all over the world. A spatial occupancy model is the appropriate method to analyze this gridded data and the RSR model may be the best version of these methods.

For the Southern ground hornbill, local conservation groups reiterate the need for more scientific data on its biological and ecological needs (Morrison et al., 2005). Our range maps highlight the southernmost portion of the data set as an area with potential for high occupancies. Future SABAP 2 efforts should focus on surveying this area to see whether these high predicted occupancies are true. Alternatively, the results repeatedly conclude that the Southern ground hornbills are found primarily in protected areas; therefore our work provides quantitative support of the value of

protected areas to the ground hornbill. It also reinforces findings of previous studies that the ground hornbill is otherwise a diet generalist (Theron, 2011). Its decline outside protected areas is likely not due to a lack of native vegetation but to the other reasons that have been proposed: loss of nesting sites, poisoning, and prosecution. Future resources should focus on remediating these issues.

We note that our data set only covers a small portion of the Southern ground hornbill's distribution in Africa. Therefore our conclusions on the species-habitat associations are limited to this area and further analyses should be run to determine if the relationships hold for the rest of its range.

Table 2.1: A summary of the models that were fit to the ground hornbill data from the Southern African Bird Atlas Project and compared within the article, with their detection and occupancy covariates, and three model selection criteria: the fitted models' minimum posterior predictive loss (PPLC), an area under the ROC statistic (AUC), and a comparison of predicted and conditional occupancy probability estimates (APC). The PPLC picked the Full RSR-160 model as best fitting and the other two model criteria slightly favored the Limited RSR-400 model. The names "Full", "Middle", and "Limited" are to distinguish between similar models that were fit using the same methods.

Model	**Detection Covariates**	**Occupancy Covariates**	**PPLC**	**AUC**	**APC**
Full Nonspatial	NSPP[a] + CROP[b] + PROTECTED[c] + INTENSIVE [d] + SAV[e] + DAY[f] + DAY2[g]	PROTECTED	923.4	0.935	0.077
Middle Nonspatial	NSPP + CROP + PROTECTED + INTENSIVE	PROTECTED	923.5	0.935	0.077

Continued on next page

Table 2.1 (continued)

Model	**Detection Covariates**	**Occupancy Covariates**	**PPLC**	**AUC**	**APC**
Limited Nonspatial	NSPP + CROP	PROTECTED	926.8	0.94	0.075
Full ICAR	NSPP + CROP + PROTECTED + INTENSIVE + SAV + DAY + DAY2	PROTECTED	922.9	0.956	0.077
Middle ICAR	NSPP + CROP + PPROTECTED + INTENSIVE	PROTECTED	923.6	0.938	0.077
Limited ICAR	NSPP + CROP	PROTECTED	926.2	0.951	0.075
Full RSR-400	NSPP + CROP + PROTECTED + INTENSIVE + SAV + DAY + DAY2	PROTECTED	923.5	0.947	0.077
Middle RSR-400	NSPP + CROP + PROTECTED + INTENSIVE	PROTECTED	923.2	0.956	0.077
Limited RSR-400	NSPP + CROP	PROTECTED	906.7	0.998	0.036

Continued on next page

Table 2.1 (continued)

Model	**Detection Covariates**	**Occupancy Covariates**	**PPLC**	**AUC**	**APC**
Full RSR-160	NSPP + CROP + PROTECTED + INTENSIVE + SAV + DAY + DAY2	PROTECTED	905.8	0.987	0.048
Middle RSR-160	NSPP + CROP + PROTECTED + INTENSIVE	PROTECTED	923.4	0.947	0.077
Limited RSR-160	NSPP + CROP	PROTECTED	926.5	0.948	0.075

[a]NSPP = number of species seen on the survey

[b]CROP = proportion of the site that was cropland

[c]PROTECTED = proportion of site that was protected lands such as a National Park

[d]INTENSIVE = number of hours spent intensively birding

[e]SAV = proportion of the site that was savannah vegetation

[f]DAY = standardized day of year

[g]DAY2 = standardized day of year squared

Table 2.2: The parameter estimates, standard errors, and 95% credible intervals from the nonspatial occupancy models on the Southern ground hornbill data. a.) The Limited model with only two covariates affecting detection. b.) The Middle model which has more detection covariates. c.) The Full set of detection covariates. See Table 2.1 for definitions of the symbols.

c.) **Limited nonspatial model**			
	Mode	SE	95% CI
Occupancy (probit-scale):			
(Intercept)	-1.16	0.084	(-1.34, -1.01)
PROTECTED	2.13	0.186	(1.80, 2.53)
Detection (probit-scale):			
(Intercept)	1.01	0.038	(-1.09, -0.94)
NSPP	0.29	0.024	(0.25, 0.34)
CROP	-1.2	0.243	(-1.71, -0.76)

b.) **Middle nonspatial model**			
	Mode	SE	95% CI
Occupancy (probit-scale):			
(Intercept)	-0.95	0.105	(-1.12, -0.70)
PROTECTED	1.69	0.189	(1.38, 2.12)
Detection (probit-scale):			
(Intercept)	-1.46	0.088	(-1.62, -1.28)
NSPP	0.25	0.028	(0.20, 0.30)
CROP	-0.58	0.273	(-1.11, -0.03)
PROTECTED	0.52	0.088	(0.31, 0.66)
INTENSIVE	0.08	0.028	(0.03, 0.14)

a.) **Full nonspatial model**			
	Mode	SE	95% CI
Occupancy (probit-scale):			
(Intercept)	-0.92	0.107	(-1.15, -0.72)
PROTECTED	1.75	0.192	(1.36, 2.11)
Detection (probit-scale):			
(Intercept)	-1.49	0.104	(-1.68, -1.28)
NSPP	0.24	0.029	(0.19, 0.30)
CROP	-0.57	0.277	(-1.13, -0.05)
PROTECTED	0.48	0.12	(0.28, 0.74)
INTENSIVE	0.09	0.027	(0.04, 0.14)
SAV	-0.02	0.118	(-0.27, 0.19)
DAY	-0.02	0.026	(-0.07, 0.03)
DAY2	0.05	0.03	(-0.01, 0.11)

Table 2.3: The parameter estimates, standard errors, and 95% credible intervals from the ICAR occupancy models on the Southern ground hornbill data. a.) The Limited model with only two covariates affecting detection. b.) The Middle model which has more detection covariates. c.) The Full set of detection covariates. See Table 2.1 for definitions of the symbols.

a.) **Limited ICAR model**			
	Mode	SE	95% CI
Occupancy (probit-scale):			
(Intercept)	-1.18	0.083	(-1.33, -1.01)
PROTECTED	2.15	0.184	(1.80, 2.51)
Detection (probit-scale):			
(Intercept)	-1.01	0.038	(-1.08, -0.93)
NSPP	0.3	0.024	(0.25, 0.34)
CROP	-1.24	0.244	(-1.73, -0.77)
b.) **Middle ICAR model**			
	Mode	SE	95% CI
Occupancy (probit-scale):			
(Intercept)	-0.92	0.104	(-1.14, -0.73)
PROTECTED	1.76	0.192	(1.37, 2.12)
Detection (probit-scale):			
(Intercept)	-1.44	0.088	(-1.62, -1.27)
NSPP	0.25	0.028	(0.19, 0.30)
CROP	-0.62	0.273	(-1.10, -0.04)
PROTECTED	0.48	0.089	(0.32, 0.67)
INTENSIVE	0.08	0.027	(0.03, 0.14)
c.) **Full ICAR model**			
	Mode	SE	95% CI
Occupancy (probit-scale):			
(Intercept)	-0.94	0.106	(-1.13, -0.72)
PROTECTED	1.7	0.192	(1.33, 2.08)
Detection (probit-scale):			
(Intercept)	-1.46	0.101	(-1.68, -1.29)
NSPP	0.24	0.029	(0.18, 0.30)
CROP	-0.57	0.27	(-1.12, -0.06)
PROTECTED	0.52	0.121	(0.28, 0.75)
INTENSIVE	0.09	0.027	(0.04, 0.14)
SAV	-0.04	0.12	(-0.28, 0.18)
DAY	-0.01	0.026	(-0.06, 0.03)
DAY2	0.05	0.03	(-0.01, 0.11)

Table 2.4: The parameter estimates, standard errors, and 95% credible intervals from the RSR occupancy models, restricted to 400 eigenvectors, on the Southern ground hornbill data. a.) The Limited model with only two covariates affecting detection. b.) The Middle model which has more detection covariates. c.) The Full set of detection covariates. See Table 2.1 for definitions of the symbols.

a.) **Limited RSR-400 model**			
	Mode	SE	95% CI
Occupancy (probit-scale):			
(Intercept)	-2.06	0.529	(-3.50, -1.54)
PROTECTED	3.94	1.029	(2.72, 6.52)
Detection (probit-scale):			
(Intercept)	-1.01	0.037	(-1.09, -0.95)
NSPP	0.3	0.023	(0.25, 0.34)
CROP	-1.27	0.248	(-1.73, -0.76)
b.) **Middle RSR-400 model**			
	Mode	SE	95% CI
Occupancy (probit-scale):			
(Intercept)	-0.91	0.109	(-1.12, -0.70)
PROTECTED	1.72	0.195	(1.36, 2.13)
Detection (probit-scale):			
(Intercept)	-1.45	0.088	(-1.63, -1.28)
NSPP	0.25	0.028	(0.20, 0.30)
CROP	-0.64	0.271	(-1.11, -0.05)
PROTECTED	0.51	0.088	(0.33, 0.67)
INTENSIVE	0.09	0.028	(0.03, 0.14)
c.) **Full RSR-400 model**			
	Mode	SE	95% CI
Occupancy (probit-scale):			
(Intercept)	-0.97	0.107	(-1.14, -0.72)
PROTECTED	1.71	0.193	(1.37, 2.12)
Detection (probit-scale):			
(Intercept)	-1.49	0.105	(-1.7, -1.29)
NSPP	0.25	0.029	(0.19, 0.30)
CROP	-0.6	0.274	(-1.09, -0.02)
PROTECTED	0.49	0.119	(0.28, 0.74)
INTENSIVE	0.09	0.027	(0.03, 0.14)
SAV	-0.04	0.118	(-0.26, 0.21)
DAY	-0.02	0.026	(-0.07, 0.03)
DAY2	0.05	0.03	(-0.01, 0.11)

Table 2.5: The parameter estimates, standard errors, and 95% credible intervals from the RSR occupancy models, restricted to 160 eigenvectors, on the Southern ground hornbill data. a.) The Limited model with only two covariates affecting detection. b.) The Middle model which has more detection covariates. c.) The Full set of detection covariates. See Table 2.1 for definitions of the symbols.

a.) **Limited RSR-160 model**			
	Mode	SE	95% CI
Occupancy (probit-scale):			
(Intercept)	-1.16	0.081	(-1.33, -1.01)
PROTECTED	2.12	0.186	(1.78, 2.51)
Detection (probit-scale):			
(Intercept)	-1.01	0.038	(-1.08, -0.94)
NSPP	0.29	0.024	(0.25, 0.34)
CROP	-1.24	0.245	(-1.71, -0.74)
b.) **Middle RSR-160 model**			
	Mode	SE	95% CI
Occupancy (probit-scale):			
(Intercept)	-0.93	0.104	(-1.14, -0.73)
PROTECTED	1.76	0.187	(1.40, 2.14)
Detection (probit-scale):			
(Intercept)	-1.44	0.088	(-1.61, -1.27)
NSPP	0.25	0.027	(0.19, 0.30)
CROP	-0.62	0.272	(-1.1, -0.04)
PROTECTED	0.49	0.089	(0.32, 0.66)
INTENSIVE	0.09	0.027	(0.03, 0.14)
c.) **Full RSR-160 model**			
	Mode	SE	95% CI
Occupancy (probit-scale):			
(Intercept)	-1.23	0.213	(-1.76, -0.89)
PROTECTED	2.46	0.522	(1.72, 3.76)
Detection (probit-scale):			
(Intercept)	-1.48	0.106	(-1.69, -1.28)
NSPP	0.25	0.029	(0.19, 0.30)
CROP	-0.6	0.279	(-1.17, -0.08)
PROTECTED	0.55	0.12	(0.30, 0.76)
INTENSIVE	0.09	0.027	(0.04, 0.15)
SAV	-0.09	0.126	(-0.31, 0.18)
DAY	-0.01	0.026	(-0.06, 0.04)
DAY2	0.04	0.03	(-0.01, 0.10)

Figure 2.1: Our study region, which is the eastern side of South Africa. The dotted line outlines the sites included in the analysis. The areas covered in a crosshatch represent protected areas; the large protected area in the northern region is Kruger National Park. The red squares represent sites where Southern ground hornbills were detected on at least one survey of the Southern African Bird Atlas Project.

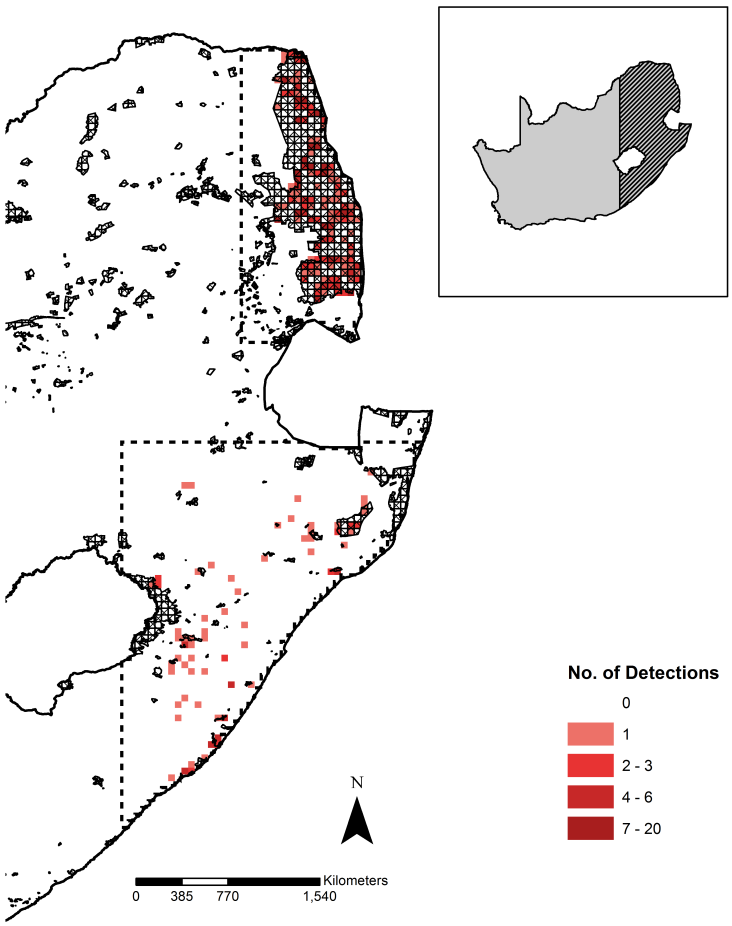


Figure 2.2: Predicted occupancy probabilities of Southern ground hornbills in eastern South Africa estimated by the ICAR model. The range maps for nonspatial occupancy models had almost identical patterns. When compared to Figure 1, the model does not fit the data well as there were several sites with detections in the southern region, but the probabilities of occupancy for these sites are very low (< 0.20).

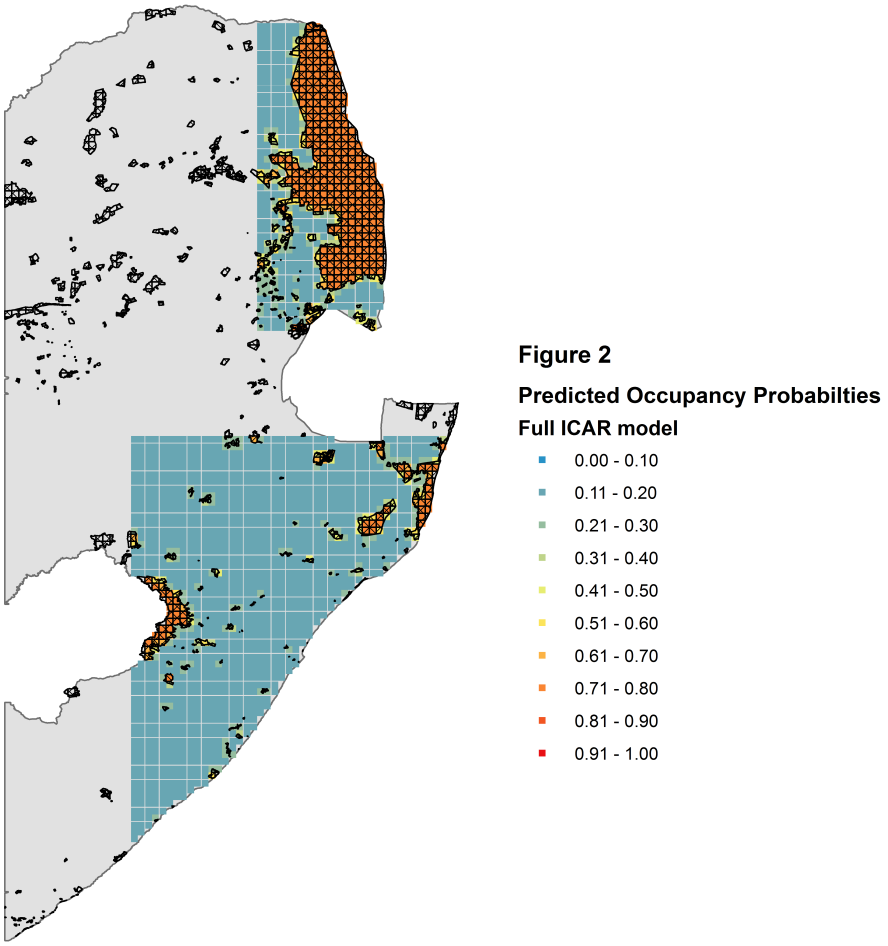
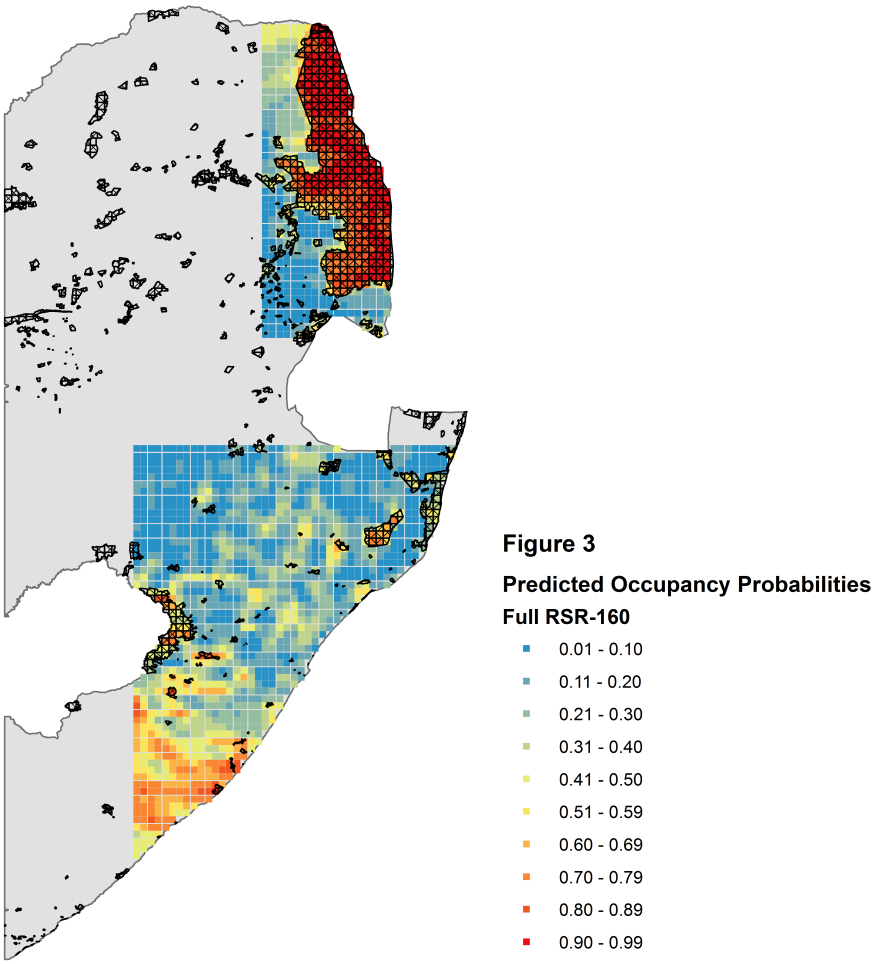


Figure 2.3: Predicted occupancy probabilities of the Southern ground hornbill in eastern South Africa for the Full RSR model, restricted to 160 columns. The southern area has much higher probabilities of occupancy than Figure 2. The most southern section has relatively high predicted occupancy probabilities; this area of the data set had very few surveys.



Chapter 3

A SIMULATION STUDY OF THE RSR AND ICAR OCCUPANCY MODELS

3.1 Introduction

The collection of presence-absence data, or more correctly detection/non-detection data, is popular in ecology because it is noninvasive and is less expensive to collect than other methods (Royle and Dorazio, 2008). Our work focuses on the situation where the detection/non-detection data are collected on a grid. In such studies, surveys are conducted on units with area (as apposed to point counts or line transects) and all species either seen or heard within those units are reported. The study may be focused on a particular species and only report the presences of the one species, or it may be a general study of all species present. Sometimes these studies record counts of the species of interest and sometimes they only record the presence-absence of the species. In this chapter, we assume that presence-absence data were collected, or that the count data was collapsed into presence-absence data.

This grid study design is used often and is popular with atlases; for example, two large such databases are the Southern African Bird Atlas Project (SABAP, http://sabap2.adu.org.za/) and the Swiss Breeding Bird Atlas (http://www.vogelwarte.ch/swiss-breeding-bird-atlas.html). In addition, this type of data occurs in other fields such as disease mapping or image processing, where similar issues of non-detection and spatial autocorrelation must be resolved.

To deal with the issue of non-detection for such binary data, occupancy models are widely used within ecology to estimate probabilities of occurrence, given the fact that the species of interest may be present but go undetected during a survey. As of

20 December 2012, the seminal MacKenzie et al. (2002) paper that introduced the occupancy model has been cited at least 1,117 times; their follow-up book, MacKenzie et al. (2006) has been cited 915 times; and the related Royle and Dorazio (2008) book has been cited 182 times (Google Scholar Search).

Only a few of the papers that address imperfect detection have also addressed the inclusion of spatial autocorrelation (e.g., Bierman et al. (2010); Hines et al. (2010); Johnson et al. (2013); Moore and Swihart (2005); Sargeant et al. (2005)). (Positive) spatial autocorrelation is the notion that sites that are next to or near each other are likely to be more similar than sites that are farther apart. If one does not account for spatial autocorrelation between sites, then parameter standard errors are underestimated, the significance of parameters is inflated, and the wrong model will be chosen as the best-fitting (Dormann et al., 2007). In addition, ignoring the residual spatial autocorrelation removes meaningful information; the fact that neighboring sites will be more alike than two randomly sampled sites can be used to improve model prediction (Latimer et al., 2006).

When data are collected on areal units or on a grid, spatial autocorrelation is usually incorporated into models as a *conditional autoregressive* (CAR) variable, and more specifically, as an *intrinsic conditional autoregressive* (ICAR) variable (Cressie and Wikle, 2011; Banerjee et al., 2004; Hughes and Haran, 2010). Johnson et al. (2013) and in the previous chapter, an advancement on the ICAR model was used, called a *restricted spatial regression* (RSR), to model the spatial autocorrelation. Both models add a spatial random effect to the regression model to help estimate the occupancy probabilities. The added variable follows a Normal distribution with a mean of 0 and a variance that is estimated by the model, same as random effects that are not spatial in nature. Faraway (2005) provides a nice introduction to the estimation process of a random effect in a regression model.

Similar to the way a random effect changes the intercepts in a regression model and therefore the baseline values of the responses, the spatial random effects adjust

the occupancy probabilities, which are the responses in these models. The manner in which this is done is best exemplified by looking at the conditional distributions for the ICAR spatial random effect. As mentioned above, the spatial random effect, η, follows a multivariate Normal distribution:

$$\boldsymbol{\eta} \sim \text{Normal}\left(0, \sigma^2 \mathbf{Q^{-1}}\right) \tag{3.1}$$

Alternatively, we can look at the conditional distributions of the individual variables, η_i, that comprise the vector $\boldsymbol{\eta}$:

$$\eta_i | \boldsymbol{\eta}_{-i} \sim \text{Normal}\left(\frac{\sum w_{ik}\eta_k}{|n(i)|}, \frac{\sigma^2}{|n(i)|}\right) \tag{3.2}$$

$\boldsymbol{\eta}_{-i}$ is all of the elements of the $\boldsymbol{\eta}$ vector except for element i. Therefore, given the neighboring values of site i, η_i will have an expected value that is the average of its neighbors, with a variance associated with this expectation.

If most of site i's neighbors have higher occupancy probabilities than would be expected from the regression coefficients, then their η_i's take on positive values, the expected value of η_i for site i is greater than 0, and site i is expected to have a higher value as well, than what would be expected from the regression coefficients. Somewhat counter intuitively, a higher variance associated with the spatial random effect equates to more residual spatial autocorrelation being accounted for by the variable. A lower variance means that the spatial random effect is not capturing any residual spatial autocorrelation.

The conditional distributions of Equation 3.2 would not necessarily form a well-defined joint distribution, but they do form the joint distribution of Equation 3.1 because of the Hammersley-Clifford Theorem (Banerjee et al., 2004). The spatial

random effect is added to the regression model. If we write the regression model in matrix form, this becomes:

$$\text{logit}\,(\boldsymbol{\psi}) = \mathbf{X}^{\mathbf{T}}\boldsymbol{\beta} + \boldsymbol{\eta} \tag{3.3}$$

(The above generalized linear model uses a logit-link but a probit-link can be used interchangeably at this stage.)

In order to estimate the parameters of a CAR model, one can use maximum likelihood methods but since this involves the inversion of the variance matrix, this method becomes computationally impossible with too many sites. In addition, this proper CAR is only able to model a limited amount of spatial autocorrelation and cannot capture extremely strong spatial autocorrelation (Banerjee et al., 2004). Instead, we use Markov Chain Monte Carlo methods to estimate the spatial random effect.

As may be imagined, recent publications have proven that positive correlations exist between the covariates, $\mathbf{X}$, and the spatial random effect, $\boldsymbol{\eta}$, of an ICAR model. These correlations bias the coefficients, $\boldsymbol{\beta}$, of the fixed effects and *overestimate* their standard errors, implying that covariates are not significant that are indeed significant.

A nonspatial model that ignores the residual spatial autocorrelation *underestimates* the standard errors. In essence, using an ICAR model trades one set of problems for another. However, one may still prefer the ICAR models to a nonspatial model because of its better predictive abilities in determining the occupancies for the unsurveyed sites due to the fact that they incorporate the neighborhood information into their predictions. This extra advantage of spatial models should not be discounted; in ecology, it is most likely that all sites will *not* be surveyed and interpolation between sites is essential.

The RSR model is an attempt to rectify the ICAR's shortcomings while still

taking advantage of the neighborhood information to interpolate to the unsurveyed sites. The derivation of the RSR model is given in Chapter 1. It uses the prominent eigenvectors associated with the precision matrix of $\boldsymbol{\eta}$ instead of the covariance matrix itself in structuring the variance of the spatial random effect. The linear combinaions of the eigenvectors recreate the spatial patterns of the data. For example, one can imagine combining maps like Figure 3.1, which would, in a sense, give the RSR variable. By using only a subset of the eigenvectors, the RSR model is less correlated with the fixed effects, the $\mathbf{X}$, and smooths the spatial pattern across the landscape. When fewer eigenvectors are included in the model, the correlation between $\mathbf{X}$ and $\boldsymbol{\eta}$ decreases and the coefficients of the regression model, the $\boldsymbol{\beta}$'s, are unbiased, but the smoothed response map may or may not be desired.

In some of the simulations of this chapter, we tried to determine what happens when different levels of restriction are applied and when the smoothing of the RSR model is beneficial and when it is a disadvantage. The RSR models were developed by Hughes and Haran (2010). In that publication, they discuss the theory behind the model, how it evolves from the ICAR model, and that it produces unbiased coefficient estimates for binary, count, and normal data, with restrictions as low as 100 eigenvectors for a $30 \times 30 = 900$ sites data set. They observed increased bias in the parameters estimated from the RSR model when the model included fewer eigenvectors; and large credible intervals for the ICAR models, often coupled with biased parameter estimates. In the end, they recommend using the eigenvectors that correspond to eigenvalues > 0.7, which would be approximately 10% of the eigenvectors for a square lattice.

In this chapter, we used simulation studies to test whether these previous conclusions regarding both the ICAR and RSR model remain true when the layer of non-detection of an occupancy model is included in the model structure.

It is important to know how the inclusion of the ICAR or RSR variable affects model predictions because these models have been used in the past and will continue

to be used in the future. The ICAR model is readily implemented using WinBUGS, and both the ICAR and RSR model can be implemented with the "stocc" package in R (Johnson 2012). The modeling of spatial correlation is gaining traction in ecology analyses as more scientists realize they can use the spatial neighborhood information to their advantage to better interpolate the results, as more people understand the intricacies of the spatial models, and as the computation time of the models decreases.

Here we evaluated various aspects of the ICAR and RSR models to see if they produce accurate parameter estimates and correctly predict the occupancy probabilities within the occupancy model framework. We also tested whether the standard errors are correctly estimated to gain insight into the precision of the estimates, and we tested what model selection tool can be used to pick between different spatial model structures.

3.2 Methods

The reader is referred to Chapter 2 for a description of occupancy models, spatial occupancy models with ICAR random effects, and spatial occupancy models with RSR random effects. For the simulations we generated occupancy probabilities on an $n \times n$ grid of sites as:

$$\text{logit}\left(\psi_i\right) = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} \tag{3.4}$$

Even though the "stocc" package uses probit-links to model the occupancy and detection probability functions, we generated the probabilities using the logit-links to explore model robustness to mild departures from the generating functions. Because logit() and probit() functions produce almost identical curves, this difference should not lead to radical changes in bias or precision but does allow us to hint at how such assumptions relate to the model estimates. In reality we will never know the true

distributions, so it makes sense to test model robustness with slightly mismatched data generation to better understand how the models will perform with unknown data generation structures.

In Equation 3.4, $\mathbf{x_1}$ usually represented "broad-scale" variation in the data such as a latitude trend (Figure 3.5a), and $\mathbf{x_2}$ represented a clustering in the data (Figure 3.3a). Sometimes both $\mathbf{x_1}$ and $\mathbf{x_2}$ were broad-scale covariates; sometimes both $\mathbf{x_1}$ and $\mathbf{x_2}$ were clustered covariates; and sometimes only one covariate was used to model the occupancy probability. The exact covariates that were used for each scenario are mentioned within the Results sections.

The broad-scale covariates were the standardized x- (and y-) coordinates of the grid and they are labeled LAT_i and $LONG_i$ for clarity in the rest of the chapter. The clustered covariates were modeled as either a *simultaneous autoregressive*, SAR, variable or a discretized spatial exponential function. A SAR variable follows similar principles to a CAR variable and produces similar spatial patterns, but with a slightly different structure of how the spatial component interacts with the unexplained error and response term (see Banerjee et al. (2004), for more details). It was chosen because data cannot be generated from an improper distribution such as the ICAR model and the proper CAR model only produces limited spatial patterns with a maximum Moran's I statistic around 0.5 (Banerjee et al., 2004). A discretized spatial exponential function based on the Matern clustering process was used in other simulations (Cressie and Wikle, 2011). The variety of spatial explanatory variables was used to address how the structure of the missing spatial autocorrelation affects the model predictions.

The detection probabilities were modeled as:

$$\text{logit}\left(p_i\right) = \alpha_0 \tag{3.5}$$

or:

$$\text{logit}\,(p_i) = \alpha_0 + \alpha_1 x_{i,1} \tag{3.6}$$

The $\mathbf{x_1}$ covariates could be either the same broad-scale (LAT_i) or clustered (u_i) covariate that was used to model the occupancy probabilities, although for most simulations, the detection probability was taken as a constant across all sites as in Equation 3.5. Unless otherwise noted, the detection probability was set equal to 0.40. For all scenarios, we set the number of surveys per site to be three.

For each scenario, or set of simulations, the covariates and hence the occupancy and detection probabilities were generated once, while the maps of actual occurrences and detections varied between simulations as Bernoulli random variables:

$$z_i \sim \text{Bernoulli}\,(\psi_i) \tag{3.7}$$

$$y_{ij} \sim \text{Bernoulli}\,(z_i \cdot p_i) \tag{3.8}$$

Where $\mathbf{z} = \{z_1, \ldots, z_N\}$ are the true occurrences and $\mathbf{Y} = \{y_{ij}\}$ are the true detections at each sites. Each scenario was designed to test a specific question such as whether low occupancy probabilities affected the results. A full factorial design of scenarios was not implemented because of the exponential increase in simulations as the number of possible combinations between scenarios increased.

We ran between 5 and 25 simulations for each scenario. While this is a small number of simulations for each specific run, the overlap between scenarios consistently came to the same conclusions, making us confident that our conclusions represent the true results. These simulations were not intended to be the definitive comparison between the ICAR and RSR models, but do provide a greater exploration of the models than has ever been completed. For example, the only other published simulation

study of the RSR model runs only one simulation per scenario (Hughes and Haran, 2010).

Between scenarios, we changed the broad-scale and clustered covariates being included in the model; the regression coefficients that simulated low occupancy and/or low detection; and the priors placed on the spatial random effect. Some simulations were on a grid of $30 \times 30 = 900$ sites and some were on a grid of $100 \times 100 = 10,000$ sites to examine the performance for large data sets. Unless otherwise noted, the scenario assumed half of the sites were randomly chosen to be surveyed.

After we generated the data, we assumed one or more of the covariates were missing. In that way, each model had a residual spatial random effect that needed to be modeled.

For each generated data set, we ran several models and compared their outputs. All models were run using the "stocc" package in R (Johnson, 2013). We ran a nonspatial occupancy model using the "unmarked" package in R; a nonspatial occupancy model using the "stocc" package in R; a spatial occupancy model with ICAR random effects using the "stocc" package; and spatial occupancy models with RSR random effects using the "stocc" package. The RSR models included several levels of restrictions, i.e., models were tested with different amounts of eigenvectors included.

The "unmarked" and "stocc" nonspatial occupancy model results always matched and therefore only the "stocc" nonspatial occupancy model will be mentioned in the rest of the chapter for easier comparison with the spatial models.

Only representative subsets of all scenarios that were tested are included in the Results section for succinctness.

3.2.1 Model Performance Measures

Model performance was measured by the overall bias:

$$\frac{1}{M \cdot N} \sum_{M} \sum_{N} \hat{\psi}_i - \psi_i \tag{3.9}$$

and by the average, absolute difference between the estimated and true occupancy probability for each site i:

$$\overline{|\hat{\psi}_i - \psi_i|} = \frac{1}{M \cdot N} \sum_{M} \sum_{N} |\hat{\psi}_i - \psi_i| \tag{3.10}$$

where the averages were taken over both the M simulations and N sites. This second statistic was used to reflect the accuracy of the site-specific estimates and we call it the "site-specific error." Lower values are preferred to larger values but note that the statistic will always be positive.

For an example of how the statistic works, assume that two sites both had a true occupancy probability of $\psi = 0.30$, but the model estimated its value as 0.20 for one of the sites and 0.40 for the other site. Overall, the occupancy estimates are unbiased, but the site-specific error is $\frac{|0.20-0.30|+|0.40-0.30|}{2} = 0.10$. This performance measure was used because it had a consistent value between scenarios and between model structures and could therefore be used to compare results between both the scenarios and model structures. Examining occupancy maps was a qualitative confirmation of model selection based on the site-specific errors.

3.3 Results

In each subsection of this section, we propose a question on model performance, explain why the question is important, provide the detailed methods associated with testing the question, and then answer the question.

3.3.1 How did the model selection criterion perform?

The "stocc" package outputs the "posterior predictive loss criterion" (PPLC) from Gelfand and Ghosh (1998) to help with model selection. Like other model selection criteria (e.g., AIC), it is a summation of a goodness-of-fit statistic and a parsimony statistic and lower values are desired. Its performance for hierarchical models such as occupancy models was unknown. We tested whether the PPLC will pick the true, generating model and if it has a tendency to pick one model structure over another (i.e., the RSR versus the ICAR structure).

Data generation. We modeled the occupancy probability as a function of two correlated, clustered covariates:

$$\begin{aligned}
&\text{logit}\left(\psi_i\right) = 1 + u_{i1} + u_{i2} \\
&\mathbf{u_1} \sim \text{Normal}\left(0, \mathbf{\Sigma}\right) \\
&\mathbf{u_2} \sim \text{Normal}\left(0.5\mathbf{u_1}, \mathbf{\Sigma}\right) \\
&\mathbf{\Sigma} = \sigma^2 \left(\left(\mathbf{I} - \alpha\widetilde{\mathbf{W}}\right) \left(\mathbf{I} - \alpha\widetilde{\mathbf{W}}\right)^T \right)^{-1}
\end{aligned} \tag{3.11}$$

where $\widetilde{\mathbf{W}}$ is the row-standardized adjacency matrix, assuming the "queen's" definition of neighbors, $\sigma = 0.25$, and $\alpha = 0.98$ in our simulations. $\mathbf{\Sigma}$ follows a SAR distribution (Banerjee et al., 2004), and $\mathbf{u_2}$ is purposely correlated with $\mathbf{u_1}$ to mimic to the real-life possibility that covariates are likely to be correlated with both the response variable and with each other. In this set of simulations, our $\mathbf{u_1}$ and $\mathbf{u_2}$ had a correlation of 0.73. The detection probability equaled 0.50 and 450 of the $30 \times 30 = 900$ total sites were sampled. 32 simulations were run and the median model fit statistics are given in Table 3.3.1. The models that we applied to the data had the form:

$$\text{probit}\,(\boldsymbol{\psi}) \sim \mathbf{u_1} + \mathbf{u_2} \tag{3.12}$$

for the "true" model. The ICAR model replaced $\mathbf{u_2}$ with a spatial random effect:

$$\text{probit}\,(\boldsymbol{\psi}) \sim \mathbf{u_1} + \boldsymbol{\eta} \tag{3.13}$$

The RSR models used $\mathbf{K}\boldsymbol{\alpha}$ as the random effect:

$$\text{probit}\,(\boldsymbol{\psi}) \sim \mathbf{u_1} + \mathbf{K}\boldsymbol{\alpha} \tag{3.14}$$

with $\mathbf{K}$ restricted to 150 eigenvectors.

Conclusions. The PPLC tended to pick the ICAR model as the best fit and would even choose it over the true model that was used to generate the data (Table 3.3.1). In general, the ICAR models did fit the data well implying that the model selection statistic is usually correct, but the model fit should be ground-truthed with qualitative comparisons of the model predictions against the original data), suggesting a tendency of the PPLC to pick complex models over parsimonious models. The PPLC should not be used to pick between the ICAR and RSR model structures. That said, because of its tendency to over fit, when the PPLC does pick a more parsimonious model, that parsimonious model should be viewed as much better fitting.

3.3.2 Must a spatial random effect be included in the model?

Most models ignore the possibility of residual spatial autocorrelation but it is very possible that not all environmental factors that affect a species' occurrences will be

Table 3.1: The posterior predictive loss criterion (PPLC) for model selection. The selection statistic is the sum of the goodness-of-fit component and the complexity penalty component. Lower values are desired. The design of the true model matched the data generation, so it should fit the data best but the model selection chose the ICAR model.

Model	PPLC statistic, D_m	Goodness-of-fit component, G_m	Complexity penalty component, P_m
True, generating model	466.5	232.7	232.7
ICAR	457.5	224.2	234.5
RSR-150	465.4	229.5	235.4

available to model. Therefore we tested the effects of excluding a spatial variable.

Data Generation. For the "large-scale variation" scenario, occupancy probabilities were a function of the scaled x- and y-coordinates of the data:

$$\text{logit}\left(\psi_i\right) = 1 + LAT_i + LONG_i \tag{3.15}$$

For the "clustered variation" scenario, occupancy probabilities were a function of the x-coordinate and a SAR variable:

$$\begin{aligned} &\text{logit}\left(\psi_i\right) = 1 + LAT_i + u_{i1} \\ &\mathbf{u_1} \sim \text{Normal}\left(0, \mathbf{\Sigma}\right) \\ &\mathbf{\Sigma} = \sigma^2 \left(\left(\mathbf{I} - \alpha\widetilde{\mathbf{W}}\right) \left(\mathbf{I} - \alpha\widetilde{\mathbf{W}}\right)^T \right)^{-1} \end{aligned} \tag{3.16}$$

with $\sigma = 0.75$, and $\alpha = 0.95$ for these simulations. The following nonspatial model:

$$\text{probit}\,(\boldsymbol{\psi}) \sim \mathbf{LAT} \tag{3.17}$$

And ICAR model:

$$\text{probit}\,(\boldsymbol{\psi}) \sim \mathbf{LAT} + \boldsymbol{\eta} \tag{3.18}$$

And RSR model:

$$\text{probit}\,(\boldsymbol{\psi}) \sim \mathbf{LAT} + \mathbf{K}\boldsymbol{\alpha} \tag{3.19}$$

were used to predict the occupancy probabilities. The RSR model was restricted to 90 eigenvectors. The detection probability was set to 0.40 and 25%, 50% or 75% of the $30 \times 30 = 900$ sites were sampled. In Table 3.2, we only report the results for when 50%, or 450, of the sites were sampled. 25 simulations were run for each scenario.

Conclusions. When models omit a covariate that does indeed predict occupancy, as in our nonspatial model, the overall occupancy probability estimates remained unbiased but the individual, site-specific errors were high (Table 3.2).

Adding a spatial random effect dramatically improved the site-specific occupancy probabilities. Including a spatial random effect never resulted in a worse model fit than the nonspatial models, although sometimes the spatial models were unable to pick up the residual spatial autocorrelation and their output was identical to the nonspatial models. Running a spatial occupancy model never resulted in a worse fit than a nonspatial model.

Table 3.2: Nonspatial versus spatial occupancy models. Overall, the nonspatial models were unbiased but when looking at their site-specific estimates, they performed poorly. It is desired for all numbers to be as close to 0 as possible.

Model	Overall Bias	Site-Specific Error
Nonspatial	-0.01	0.15
ICAR	-0.01	0.03
RSR-90	-0.01	0.05

3.3.3 Can the spatial random effect be used in place of the covariates?

For many data sets, meaningful covariates may be missing or may have been collected at a different resolution from the presence-absence data. If collected at a different resolution, there may be change-of-support or modifiable-area-unit-problem issues resulting from their inclusion. Rather than introduce questionable relationships between the species occurrences and the available habitat data, it is of interest to know if predictive occupancy maps can be created without any covariate information.

Data generation. Many scenarios were run that tested model performance without covariates. For illustrative purposes, only one will be described in detail here.

The data were generated through the set-up described in the preceding section, Equations 3.16. The models assumed that one variable was available to help predict the occupancy probabilities, as in Equations 3.18 and 3.19 or that occupancy probabilities were only a function of the spatial random effects. For the ICAR model:

$$\text{probit}\left(\boldsymbol{\psi}\right) \sim \boldsymbol{\eta} \tag{3.20}$$

And for the RSR model:

Table 3.3: A comparison of model fit when the spatial random effect is used in place of all covariates. When a spatial random effect is used to replace either one or two spatially correlated covariates, both the ICAR and RSR models are still able to build accurate, predictive occupancy maps. With heavy restriction, the RSR models being to perform poorly but their fits are comparable when 180 eigenvectors are included.

	Overall Bias		Site-Specific Error	
Model	Fit with one covariate:	Fit with no covariates:	Fit with one covariate:	Fit with no covariates:
ICAR	-0.01	-0.01	0.03	0.04
RSR-180	-0.01	-0.01	0.04	0.06
RSR-90	-0.01	-0.01	0.05	0.06
RSR-30	-0.01	-0.01	0.06	0.07

$$\text{probit}\,(\psi) \sim \mathbf{K}\alpha \tag{3.21}$$

Several levels of restriction with the RSR model were compared: 30, 90, and 180 eigenvectors.

Conclusions. Both the ICAR and RSR models can be used to model occurrence in the absence of any covariates (Table 3.3; Figure 3.5 and 3.3). The models missing both covariates– those that modeled occupancy solely as a function of the spatial random effect– produced occupancy maps as accurate as the other models (Table 3.3). In fact, excluding all covariates sometimes created more accurate occupancy maps.

When a spatial random effect is used in place of a covariate, the extreme occupancy probability values may be shrunk towards the middle, meaning that low occupancy probabilities ($\psi < 0.20$) are overestimated and high occupancy probabilities ($\psi > 0.90$) are underestimated (Figure 3.5).

3.3.4 Are the standard errors estimated correctly?

As noted in the Introduction, if one ignores residual spatial autocorrelation, standard errors are underestimated (Dormann et al., 2007; Latimer et al., 2006). Recent publications have brought attention to the fact that the inclusion of an ICAR variable may overinflate the standard errors, essentially trading one problem for another (Hodges and Reich, 2010; Reich et al., 2006).

The purpose of this set of simulations was to test whether the ICAR and RSR models were overestimating the standard errors for the occupancy model, tested by seeing if the standard errors as calculated from the model output matched the true variation in the estimates of the parameters. We concluded that the ICAR model does overestimate the parameter standard errors and that the RSR model also overestimates the standard errors but by a smaller amount.

Data generation. In order to properly test the variances, the data were generated from probit function instead of the logit function in order to match the model structure of the models fit using the "stocc" package:

$$\begin{aligned} &\text{probit}\left(\psi_i\right) = LAT_i + u_{i1} \\ &\mathbf{u} \sim \text{Normal}\left(\mathbf{0}, \mathbf{H}(\theta)\right) \\ &h_{ij} = \exp\left(\frac{-\|s_i - s_j\|}{\theta}\right) \end{aligned} \tag{3.22}$$

In these simulations, the clustered covariate followed the discretized spatial exponential function, with a range of $\theta = 30$. $\mathbf{H}(\theta)$ is a matrix of elements $\{h_{ij}\}$ and $\|s_i - s_j\|$ is the Euclidean distance between sites i and j. The detection probability was set to 0.40; 450 of 900 sites were sampled; and 100 simulations were run. On this generated data, we ran a "true" model that was nonspatial and included both covariates to give a baseline for the parameter standard deviations:

$$\text{probit}\,(\boldsymbol{\psi}) \sim \mathbf{LAT} + \mathbf{u} \tag{3.23}$$

The spatial models included the LAT variable but modeled the u_1 variable as a spatial random effect instead, as in Equations 3.18 and 3.19.

For each simulation, the standard deviation associated with the posterior distribution's parameter was calculated from the model output and was labeled as $\widehat{SE}(\beta)$. We computed the medians of the $\widehat{SE}(\beta_i)$'s that were outputted from the 100 simulations and denote the medians as $\overline{\widehat{SE}(\beta_i)}$. We compared these values to the standard deviations associated with the parameter estimates themselves. We stored the median parameter values outputted from each simulation, the $\hat{\beta}_i$, and then we take the standard deviation of the (β_i)'s gained from all 100 simulations and labeled it $SD(\hat{\beta}_i)$. If the models are estimating the standard errors correctly then these values should match: $\overline{\widehat{SE}(\beta_i)} = SD(\hat{\beta}_i)$.

Conclusions. The ICAR model consistently overestimated the standard errors while the RSR model had standard errors that were more in line with the true standard deviations of the estimates and were more similar to the standard errors of the true model (Table 3.4). These discrepancies between standard errors for the parameters led to similar error inflation for the occupancy probabilities themselves (Figure 3.5). The standard errors of the detection probability parameters were estimated correctly for all models.

3.3.5 How much restriction can occur in the RSR models?

Once a researcher has decided to use an RSR model, the question remains as to how many eigenvectors should be included in the restriction. While this question is partially tested in all of the simulations, the most thorough set of simulations is

Table 3.4: Accuracy of the reported standard errors. The standard errors associated with the ICAR model are much higher than those reported by the other models (left columns). The true variance in the estimation of the parameters, $SD(\hat{\beta}_i)$, is much lower than the reported values, $\overline{\widehat{SE}(\beta_i)}$ for the ICAR model and is somewhat lower than the reported values for the RSR model. The two types of estimates for the standard errors match for the true model, as expected.

Model	Intercepts		Parameter Coefficients	
	$\overline{\widehat{SE}(\beta_0)}$	$SD(\hat{\beta}_0)$	$\overline{\widehat{SE}(\beta_1)}$	$SD(\hat{\beta}_1)$
True, generating model	0.18	0.18	0.17	0.17
ICAR	0.29	0.19	0.42	0.21
RSR-150	0.23	0.19	0.25	0.19

described here. We tested models with several levels of restriction and several types of missing spatial variables.

Data Generation. In the first set of simulations, the occupancy probabilities were generated as functions of the discretized spatial exponential function:

$$\begin{aligned} &\text{logit}\left(\psi_i\right) = 1 + u_{i1} \\ &\mathbf{u} \sim \text{Normal}\left(\mathbf{0}, \mathbf{H}(\theta)\right) \\ &h_{ij} = \exp\left(\frac{-\|s_i - s_j\|}{\theta}\right) \end{aligned} \tag{3.24}$$

with $\theta = 30$, as in Equation 3.22 in the previous section. We set the detection probability equal to 0.40 and sampled 450 of 900 sites. For each of the ten simulations run, we analyzed the data with a "true" nonspatial model that included the u_1 covariate:

$$\text{probit}\left(\boldsymbol{\psi}\right) \sim \mathbf{u_1} \tag{3.25}$$

The ICAR and RSR models predicted the occupancy probabilities strictly as func-

tions of the spatial random effects as in Equations 3.20 and 3.21, respectively. The RSR models were restricted to 30, 60, 90, 120, 180, 240, 300, and 450 eigenvectors. In the second set of simulations , the occupancy probabilities were generated as functions of a SAR-variable:

$$\begin{aligned} \text{logit}\left(\psi_i\right) &= 1 + LAT_i + u_{i1} \\ \mathbf{u_1} &\sim \text{Normal}(\mathbf{0}, \mathbf{\Sigma}) \\ \mathbf{\Sigma} &= \sigma^2 \left(\left(\mathbf{I} - \alpha \widetilde{\mathbf{W}} \right) \left(\mathbf{I} - \alpha \widetilde{\mathbf{W}} \right)^T \right)^{-1} \end{aligned} \tag{3.26}$$

Or as functions of a large-scale variable:

$$\text{logit}\left(\psi_i\right) = 1 + LAT_i + LONG_i \tag{3.27}$$

The ICAR and RSR models predicted occupancy probabilities strictly as functions of the spatial random effects as in Equations 3.20 and 3.21, respectively. The RSR models were restricted to 30, 90, 180 or 350 eigenvectors.

We also tested how the level of restriction affected model predictions when a larger number of sites was involved. For these tests, we set:

$$\text{logit}\left(\psi_i\right) = 2LAT_i + 2w_{i1} \tag{3.28}$$

for a grid of $100 \times 100 = 10,000$ sites, of which we set half to be sampled. The $\mathbf{w_1}$ were a clustered covariate generated by a Matérn clustering process ($\kappa = 100; r = 0.08; \mu = 400$), which is a generalization of the spatial exponential function and did not require intensive matrix multiplication to calculate. The counts were standardized, leading to occupancy probabilities of 0.5–0.6.

Conclusions. The overall bias remained close to 0 regardless of the amount of restriction but the site-specific error increased as the amount of restriction increased and fewer eigenvectors were included. The models were not sensitive to small changes in the restrictions and there were diminishing returns in terms of the site-specific error as the number of eigenvectors increased. With 300 or more eigenvectors, the RSR models on 900 sites had long run times and began to perform poorer. When many eigenvectors were included, the RSR model showed similar characteristics to the ICAR models– they had the benefit of a better fit for the small-scale variations, but they also overestimated the parameter standard errors and had higher correlations between MCMC iterations.

The RSR models performed equally well to the ICAR models when the missing covariate followed the discrete spatial exponential distribution (Table 3.5) and the level of restriction had little affect on this conclusion.

When the missing covariate followed a SAR distribution or a latitudinal trend, then the ICAR performed better than the RSR models and had lower site-specific errors (Table 3.6; Figure 3.5).

When the larger grid of 100×100 sites was simulated, an RSR model with 500 eigenvectors took almost as long to run as the ICAR model due to the matrix computations necessary to initialize the MCMC iterations. The model with 100 eigenvectors took much less time and produced a map with similar accuracy to the ICAR model albeit much smoother (Table 3.5; Figure 3.3).

Our simulations suggest that the optimal restriction for the RSR models is 100–400 eigenvectors, regardless of the total number of sites or the percentage of those sites that were surveyed. Within that range, fewer eigenvectors should be included if one desires a smoother map for predictions and more eigenvectors should be included if one wants to fit the current data and is less interested in the predictions at the unsurveyed sites.

Table 3.5: Trade-off between run times and model error. For all scenarios, half of the total number of sites were assumed to be sampled.

Model	# Sites	Overall Bias	Site-Specific Error	Median Run Time (min)
Nonspatial	900	0.01	0.01	24
ICAR	900	0.01	0.09	39
RSR-30	900	0.01	0.09	30
RSR-60	900	0.01	0.09	31
RSR-90	900	0.01	0.09	32
RSR-120	900	0.01	0.09	34
RSR-180	900	0.01	0.09	38
RSR-240	900	0.01	0.09	44
RSR-300	900	0.01	0.10	53
RSR-450	900	0.01	0.10	93
ICAR	10,000	0.00	0.04	460
RSR-100	10,000	0.00	0.07	228
RSR-500	10,000	0.00	0.07	374

Table 3.6: Model error under alternative spatial variables. These runs exemplify the discrepancy between the ICAR and RSR models when the spatial structure of the missing covariate does not match the ICAR or RSR spatial random effect structure.

Model	# Sites	Overall Bias	Site-Specific Error
Large-Scale Variation			
ICAR	900	-0.01	0.04
RSR-30	900	-0.01	0.07
RSR-90	900	-0.01	0.06
RSR-180	900	-0.01	0.06
Clustered Variation			
ICAR	900	-0.01	0.14
RSR-30	900	0.00	0.17
RSR-90	900	0.00	0.15
RSR-180	900	0.01	0.15

3.3.6 How do low occupancy and/or low detection probabilities affect model performance?

Data Generation. Scenarios were considered to be low occupancy if the average occupancy probability was 0.31 or less; they had low detection if the detection probability was 0.20. The low occupancy probabilities were generated as:

$$\text{logit}\left(\psi_i\right) = -1 - 2LAT_i \tag{3.29}$$

A normal detection probability of 0.40 was used to test the effects of low ocucupanies and a detection probability of 0.20 was used to test the effects of low detections. As with the previous simulations, we assumed a grid of $30 \times 30 = 900$ sites. 16 simulations were run. The model assumed the covariate was not available, as in Equations 3.20 and 3.21.

Conclusions. Lower occupancy probabilities did not affect the differences between models, but led to less precise maps and possibly, positively biased occupancy probability predictions (Table 3.7).

When detection was low, the models overestimated occupancy, and the spatial models were less able to pick up the residual spatial autocorrelation (Table 3.7). The overestimation of occupancy matched the results in the original MacKenzie et al. (2002) paper that introduced occupancy models.

3.3.7 How well do the models predict the fixed effects of interest?

The residual spatial autocorrelation may be correlated with the covariates of interest (Hodges and Reich, 2010; Reich et al., 2006). We tested the ICAR model in an attempt to confirm these findings and to see if the problem also existed for the RSR models.

Table 3.7: Model results for low occupancy probabilities.

Model	Overall Bias	Site-Specific Error
Normal Detections, $p = 0.40$		
Nonspatial	0.01	0.29
ICAR	0.01	0.08
RSR-300	0.00	0.08
RSR-180	0.01	0.09
Low Detections, $p = 0.20$		
Nonspatial	0.04	0.29
ICAR	0.03	0.09
RSR-300	-0.02	0.10
RSR-180	0.04	0.12

Data Generation. Because the estimation of the fixed effects is related to its correlation with the residual spatial autocorrelation, we generated occupancy probabilities as functions of two correlated variables, as in Equations 3.11, but with a probit-link to more directly compare the parameter estimates from when the models are fit using the "stocc" package with the true parameter values:

$$\begin{aligned} \text{probit}\left(\psi_i\right) &= 1u_{i1} + 1u_{i2} \\ \mathbf{u_1} &\sim \text{Normal}\left(\mathbf{0}, \mathbf{\Sigma}\right) \\ \mathbf{u_2} &\sim \text{Normal}\left(\mathbf{u_1}, \mathbf{\Sigma}\right) \\ \mathbf{\Sigma} &= \sigma^2 \left(\left(\mathbf{I} - \alpha\widetilde{\mathbf{W}}\right) \left(\mathbf{I} - \alpha\widetilde{\mathbf{W}}\right)^T \right)^{-1} \end{aligned} \tag{3.30}$$

with $\sigma^2 = 0.25$ and $\alpha = 0.98$. 32 simulations were run on the $30 \times 30 = 900$ grid, half of which were sampled, with a detection probability of 0.5. The models assumed that ψ_i was a function of the spatial random effect and u_{i1} but not u_{i2}, as in Equations 3.13 and 3.14. A "true" nonspatial model was also run assuming that ψ_i was a function of both u_{i1} and u_{i2}, as in Equation 3.12.

In a second set of simulations, the two covariates were not correlated and oc-

Table 3.8: Predicting the model coefficients with non-correlated variables. The true model correctly estimated the parameters, but the spatial models both overestimated the coefficients.

Model	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	Overall Bias	Site-Specific Error
True Parameter Value	0	1	1		
True, generating model	0.01	1.04	0.95	0.00	0.01
ICAR	0.68	1.30	NA	0.00	0.06
RSR-150	0.63	1.26	NA	0.00	0.06

Table 3.9: Predicting the model coefficients with correlated variables. None of the models correctly predicted the coefficient values, although the underestimated intercepts balanced with the overestimated slopes to give unbiased estimates for the occupancy probabilities overall and kept the site-specific errors reasonable.

Model	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	Overall Bias	Site-Specific Error
True Parameter Value	0	1	1		
True, generating model	-0.24	0.64	0.56	-0.08	0.10
ICAR	-0.49	1.34	NA	-0.08	0.12
RSR-100	-0.48	1.33	NA	-0.08	0.13
RSR-250	-0.51	1.38	NA	-0.08	0.12

cupancy probabilities were generated from Equation 3.22, and the models of Equations 3.18 and 3.19 were applied.

Conclusions. When covariates were not correlated, the true model correctly estimated all coefficients and the spatial models overestimated the coefficients but were still able to maintain accurate site-specific occupancy probability estimates (Table 3.8).

When covariates were correlated, none of the models– including the true, generating model– correctly estimated the coefficients, not even the true model, and the occupancy probabilities were underestimated (Table 3.9). More testing needs to be conducted on the ability of the models to estimate the fixed effects coefficients.

3.3.8 Does model performance change when the area of inference is large?

One may want to run a spatial occupancy model on a large area with many sites, for example the SABAP 2 data has 16,000 sites and the Swiss Breeding Bird Survey has 40,000 sites.

Data Generation. Model performance on large areas of inference was tested in multiple scenarios. One example was data generated from Equations 3.15 and then modeled by Equations 3.18 and 3.19.

Conclusion. When 10,000 sites were included in the gridded data, model performance did not change, but the RSR models did have much faster computation times.

For the larger grid of $100 \times 100 = 10,000$ sites, the ICAR models took an average of 7.7 hours to run, the RSR models with 100 eigenvectors took an average of 3.8 hours, and the RSR models with 500 eigenvectors took an average of 6.2 hours (Table 3.5). The long run times for the RSR models were partially due to the matrix operations that were necessary to compute before the MCMC chains began. For example, calculating the eigenvectors for the RSR model took over one hour for each of these models. Additional efficiencies in the matrix calculations within the function could improve run times further.

3.3.9 How does collinearity affect model results?

Two types of collinearity were considered. The first type of collinearity was when two covariates were correlated with each other and both affected the occupancy probabilities. This scenario and the results were described in the "Predicting fixed effects" section above, page 71.

Another type of collinearity is when the same covariate affects both the occupancy and detection probabilities. For example, one can imagine that vegetation height increases the probability of occupancy of a species while at the same time decreasing

the probability of detection. This covariate may exhibit high spatial autocorrelation itself. We examined what happens to model predictions when the data for such a variable is missing.

Data Generation. Because of the ramifications of this question, we tested it in several ways. For the first set of tests:

$$\text{logit}\left(\psi_i\right) = LAT_i \tag{3.31}$$

$$\text{logit}\left(p_i\right) = -1 + LAT_i \tag{3.32}$$

or

$$\text{logit}\left(p_i\right) = -1 - LAT_i \tag{3.33}$$

The detection probabilities were either positively and negatively correlated with the covariate, and the 95% range for the detection probabilities was (0.06, 0.66) with a median of 0.27 for both cases. Occupancy probabilities were positively related to the covariates and their 95% range was (0.16, 0.84) with a median of 0.50. We used 3 surveys per site, surveyed 450 of 900 sites, and ran the scenario for 15 simulations.

The second scenario that we discuss in this section is described with Equation 3.11 in the PPLC performance section:

$$\begin{aligned}
&\text{logit}\left(\psi_i\right) = 1 + u_{i1} + u_{i2} \\
&\mathbf{u_1} \sim \text{Normal}\left(\mathbf{0}, \mathbf{\Sigma}\right) \\
&\mathbf{u_2} \sim \text{Normal}\left(0.5\mathbf{u_1}, \mathbf{\Sigma}\right) \\
&\mathbf{\Sigma} = \sigma^2 \left(\left(\mathbf{I} - \alpha \widetilde{\mathbf{W}} \right) \left(\mathbf{I} - \alpha \widetilde{\mathbf{W}} \right)^T \right)^{-1}
\end{aligned} \tag{3.34}$$

but with varying detection probabilities:

$$\text{logit}\,(p_i) = 0.5u_{i1} \tag{3.35}$$

or

$$\text{logit}\,(p_i) = -0.5u_{i1} \tag{3.36}$$

Conclusions. If a covariate that affected both occupancy and detection was missing from the model, the site-specific errors were high. The estimates were usually unbiased overall if the detection probabilities remained high for all sites. If the covariate had a positive relationship with both occupancy and detection probabilities, then the models did a better job overcoming the collinearity. If the covariate had a positive relationship with occupancy but a negative relationship with detection, then the models had some difficulty estimating the occupancy probabilities and tended to have a negative bias regarding the occupancy probabilities (Tables 3.10 and 3.11).

For the scenarios that we tested, the negative bias may be exaggerated by the low detection probabilities that resulted from allowing detection to vary. If detection and/or occupancy was low, the results were more inaccurate and more imprecise than if there was not any collinearity.

3.3.10 Do the priors affect the results?

As with any Bayesian analysis, it is important to know if and how the prior distributions affect the posterior results. Most importantly for a spatial model is how the priors associated with the spatial parameter, τ, may be affecting the posteriors. For ICAR and RSR models, several priors have been used in previous publications: gamma(0.5, 0.0005); gamma(0.5, 0.005); gamma(1, 1); gamma(0.01, 0.01); and gamma(0.001, 0.001). In this section, we tested the influence of those priors.

Table 3.10: Biases resulting from collinearity between the detection and occupancy probabilities. This table relates the effects of confounding from a large-scale covariate (Equation 3.31). The true model exhibited similar biases and site-specific errors regardless of whether the detection was positively or negatively correlated with the variable. The biases and errors all increased in magnitude for the spatial models.

	Detection is Positively Correlated		Detection is Negatively Correlated	
Model	Overall Bias	Site-Specific Error	Overall Bias	Site-Specific Error
True, generating model	0.09	0.12	0.04	0.10
ICAR model	-0.05	0.14	-0.09	0.26
RSR-150	-0.05	0.15	-0.09	0.25

Table 3.11: Biases resulting from collinearity between the detection and occupancy probabilities with a spatial exponential covariate (Equations 3.11). All models had increased biases and site-specific errors when the detection probability was negatively correlated with the occupancy probabilities.

	Detection is Positively Correlated		Detection is Negatively Correlated	
Model	Overall Bias	Site-Specific Error	Overall Bias	Site-Specific Error
True, generating model	-0.04	0.04	-0.22	0.28
ICAR model	-0.04	0.06	-0.21	0.26
RSR-150	-0.04	0.07	-0.22	0.27

Table 3.12: Comparing posterior estimates for different gamma priors for the spatial parameter, τ.

Model	Priors	Median Posterior	Overall Bias	Site-specific Error
ICAR	(0.5, 0.0005)	0.46	0.00	0.07
	(0.5, 0.005)	0.47	0.00	0.07
	(1, 1)	0.41	0.00	0.07
	(0.01, 0.01)	0.29	0.00	0.08
	(0.001, 0.001)	0.34	0.00	0.08
RSR-180	(0.5, 0.0005)	5.25	0.00	0.09
	(0.5, 0.005)	0.55	0.00	0.09
	(1, 1)	0.37	0.00	0.08
	(0.01, 0.01)	0.34	0.00	0.09
	(0.001, 0.001)	0.32	0.00	0.09

Data generation. The occupancy probabilities were generated from the following set of equations:

$$\text{logit}\left(\psi_i\right) = 1 + LAT_i + LONG_i \tag{3.37}$$

A detection probability of 0.40 and a sampling of 450 of 900 sites were used for 10 simulations. The models assumed that the LAT_i were known but the $LONG_i$ were unknown, and both ICAR and RSR models were run for all of the priors mentioned above.

Conclusion. There was some evidence that the gamma(0.5, 0.0005) was less able to pick up the residual spatial autocorrelation, but otherwise, there was no consistent patterns in the performance of the priors. In general, the priors did not affect the results and the posterior estimates were independent of the starting values for τ (Table 3.12).

3.3.11 Can the models be used with a sparsely sampled area?

It is unlikely that half of a grid of 10,000 sites will be sampled. More likely, only a fraction of the total grid will be sampled. We performed multiple tests to compare the models when few sites were sampled.

Data Generation. We first compared model outputs when 275, 450, or 675, i.e., one-fourth, one-half, and three-fourths, respectively, of the available 900 sites were sampled. In this scenario, we generated the data from Equation 3.15.

We further tested the results of less sampling on a 10,000 site grid. For this scenario, we generated the occupancy probabilities from:

$$\text{logit}\left(\psi_i\right) = -1 - LAT_i \tag{3.38}$$

Slightly different from the previous scenarios, the detection probability was set to 0.30, but again, 3 surveys per site were used. On this grid, we built models where 100, 500, 1,000 or 5,000 of the 10,000 sites were samples. In terms of percentages, 1–50% of the sites were sampled. In one set of simulations, the sampled sites were chosen completely randomly. In another set of simulations, the sites were sampled in pairs to learn if extra neighborhood information was gained form this alternative sampling design to help with the spatial random effect estimation.

Conclusions. It did not make a difference if the sparsely sampled sites were random across the landscape or if sites were sampled in pairs. Therefore, all conclusions discussed are drawn from the random sampling design.

As the number of samples decreased, the overall bias of the model did not change but the site-specific errors increased (Tables 3.13 and 3.14). Mostly, the ICAR and RSR models perform equally well with fewer surveyed sites, with the ICAR models

Table 3.13: When fewer sites are sampled, the overall occupancy probabilities remain unbiased but the site-specific errors increased. The discrepancies between the ICAR and RSR models do not change.

Model	Overall Bias			Site-Specific Errors		
Number of sites surveyed:	675	450	275	675	450	275
ICAR	0.00	-0.01	0.01	0.02	0.03	0.04
RSR-180	0.00	-0.01	0.00	0.04	0.04	0.06

Table 3.14: Sparse data on a large grid. When only 1% of the sites are surveyed, then the models perform considerably worse but are otherwise robust to fewer sites being sampled, albeit with increasing site-specific errors.

Model	Overall Bias				Site-Specific Errors			
Number of sites surveyed:	5,000	1,000	500	100	5,000	1,000	500	100
ICAR	0.01	0.02	-0.01	-0.05	0.04	0.06	0.08	0.16
RSR-150	0.02	0.02	0.00	-0.03	0.07	0.10	0.13	0.18

slightly outperforming the RSR models, as was the case in the previous scenarios as well.

If 10% or less of the area of inference was surveyed, then model output was imprecise and the site-specific errors were higher than before but the overall occupancy probability was generally unbiased (Table 3.14). Often the models had difficulty detecting the underlying spatial patterns when 10% or fewer sites were sampled. When only 100 sites out of a grid of 10,000 sites, or 1% of the area of inference, were surveyed, the models performed poorly and in general could not detect the underlying, missing spatial patterns (Table 3.14).

3.4 Summary of Results

From the simulations described above, we make the following summary conclusions and recommendations.

- When applicable, researchers should fit spatial occupancy models over nonspa-

tial occupancy models.

- Accurate occupancy probability maps can be produced from an ICAR or RSR spatial occupancy model without any covariates.

- The RSR models were not sensitive to changes in the numbers of eigenvectors; restrictions to 105, 112, 120 eigenvectors all gave very similar results.

- In order to pick which spatial model to use, the researcher should determine their desired level of smoothing *a priori* to fitting the models. Restricting the RSR models to 100–400 eigenvectors is an appropriate range of restriction regardless of the number of sites in the area of inference, with fewer eigenvectors leading to smoother predicted occupancy maps.

- One should use a more restricted RSR model, i.e 100–200 eigenvectors to obtain accurate parameter estimates associated with the fixed efffects of interest and accurate standard errors.

- The PPLC model selection tended to over fit and picked complex models over the true, generating models. It should not be used to choose between fitting an ICAR and RSR model.

- Low occupancy probabilities did not affect model performance but low detection probabilities caused negatively biased occupancy probability estimates, for both the spatial and nonspatial models.

- The spatial models were robust to low sampling coverage, even without any fixed covariates to help explain the variability.

- The spatial models were robust to different priors for the spatial parameter, τ.

- Nonspatial and spatial models had trouble estimating the fixed effects of interest when there was confounding between the independent variables used in the model. However, the models were able to compensate and outputted accurate occupancy probabilities even with the inaccurate covariate coefficients. If there is confounding or it is known that a covariate that affects both the occupancy and detection probabilities is not available, it is important for detection probabilities to be high.

3.5 Conclusions

In this chapter, we attempted to quantify and summarize spatial occupancy model performance. We focused on scenarios where the incoming data were complex, degraded, and/or mismatched to the model assumptions. Data complexity was created with the inclusion of multiple covariates that were sometimes correlated with each other in addition to their relationships with the occupancy probabilities. Data degradation was synthesized through low detection probabilities, low occupancy probabilities, sparse numbers of sampled sites, and through the general concept of nondetection that is the basis of the occupancy model. Sometimes, different equations were used in the models than those that comprised the data generating structures in order to test results from mismatched assumptions. Model performance was determined by the models' abilities to overcome these degradations and complexities.

In general, both the ICAR and RSR models showed great adeptness in the face of these challenges. They used their spatial random effects to pick up missing spatial relationships and accurately predict occupancy probabilities for the surveyed and unsurveyed sites.

Many of the details that resulted from our simulation testing were described in the Summary section above but the questions remain of when, if, and how to use a spatial occupancy model over a nonspatial model. The spatial models always provided more accurate occupancy estimates than the nonspatial occupancy models, although

there were simulations where the spatial models were unable to pick up the missing spatial pattern and the models had the same output. Therefore if data are collected on a grid, one should always apply the spatial occupancy models over the nonspatial occupancy models.

Either the RSR or the ICAR model may be preferred, depending on the situation. If the data are well behaved with high detections, most of the desired area of inference is surveyed, and there are no unmodeled, confounding variables, then the ICAR model would be a better choice. That said, although the ICAR model usually created more accurate, predictive occupancy maps and it may be better at picking up the small-scale variance, it may also bias the parameter estimates of the fixed effects and it exaggerated the standard errors.

More likely than not when looking at ecology data, the data will be degraded (e.g., lower detections, fewer sites surveyed) and there will be confounding between the occupancy and detection probabilities. In these scenarios, the RSR models often picked up the residual spatial autocorrelation when the ICAR model was unable to. In addition, if one is using the model to make predictions on the unsurveyed sites, then the smoothing of the RSR model may be preferred.

For these reasons, we recommend always choosing the RSR models within an occupancy model framework. As with other smoothing techniques, the exact amount of smoothing/restriction to be applied is left up to the researcher. We recommend a restriction of 100–400 eigenvectors, with fewer eigenvectors being used if the researcher wants a smoother map of predictions to extrapolate the results to unsurveyed areas and more eigenvectors being included if one wants to capture the smaller-scale variation related to the specific area that has already been surveyed.

Figure 3.1: Examples of patterns created from the eigenvectors of a map of occupancy probabilities.

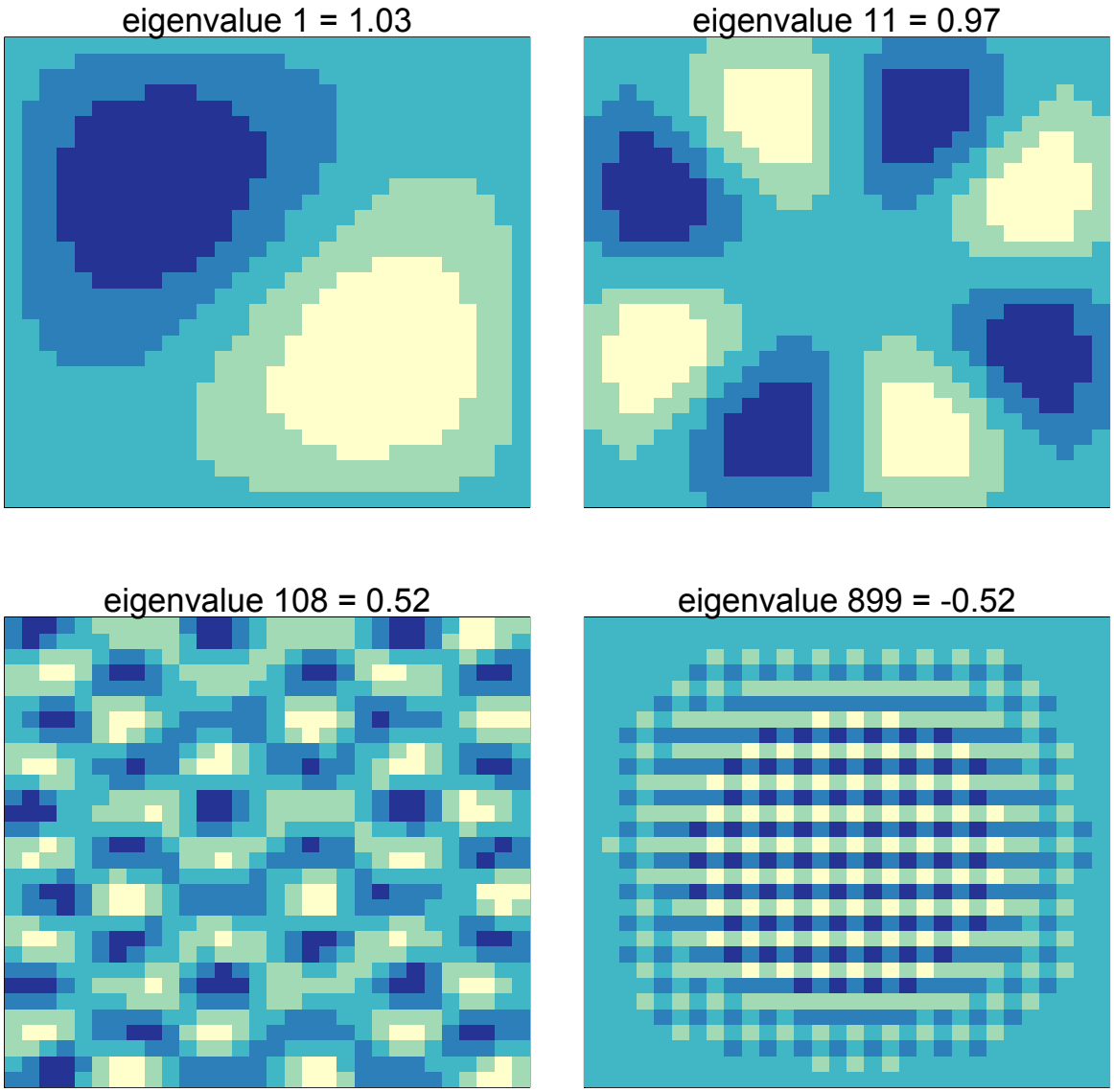


Figure 3.2: Model estimates when missing large-scale covariates a.) The true occupancy probabilities that the models were trying to replicate. This map is from Equation 3.15. b.) Going clockwise from the top-left: The predicted occupancy probabilities from the ICAR model using Equation 3.18; the predicted occupancy probabilities from the ICAR model using Equation 3.20; the predicted occupancy probabilities from the RSR model using Equation 3.21 and restricted to 180 eigenvectors; the predicted occupancy probabilities from the RSR model using Equation 3.19 and restricted to 180 eigenvectors. When the ICAR and RSR models assumed that either one or both covariates of Equation 3.15 were missing, a comparison of the left and right maps show that very similar maps were produced.

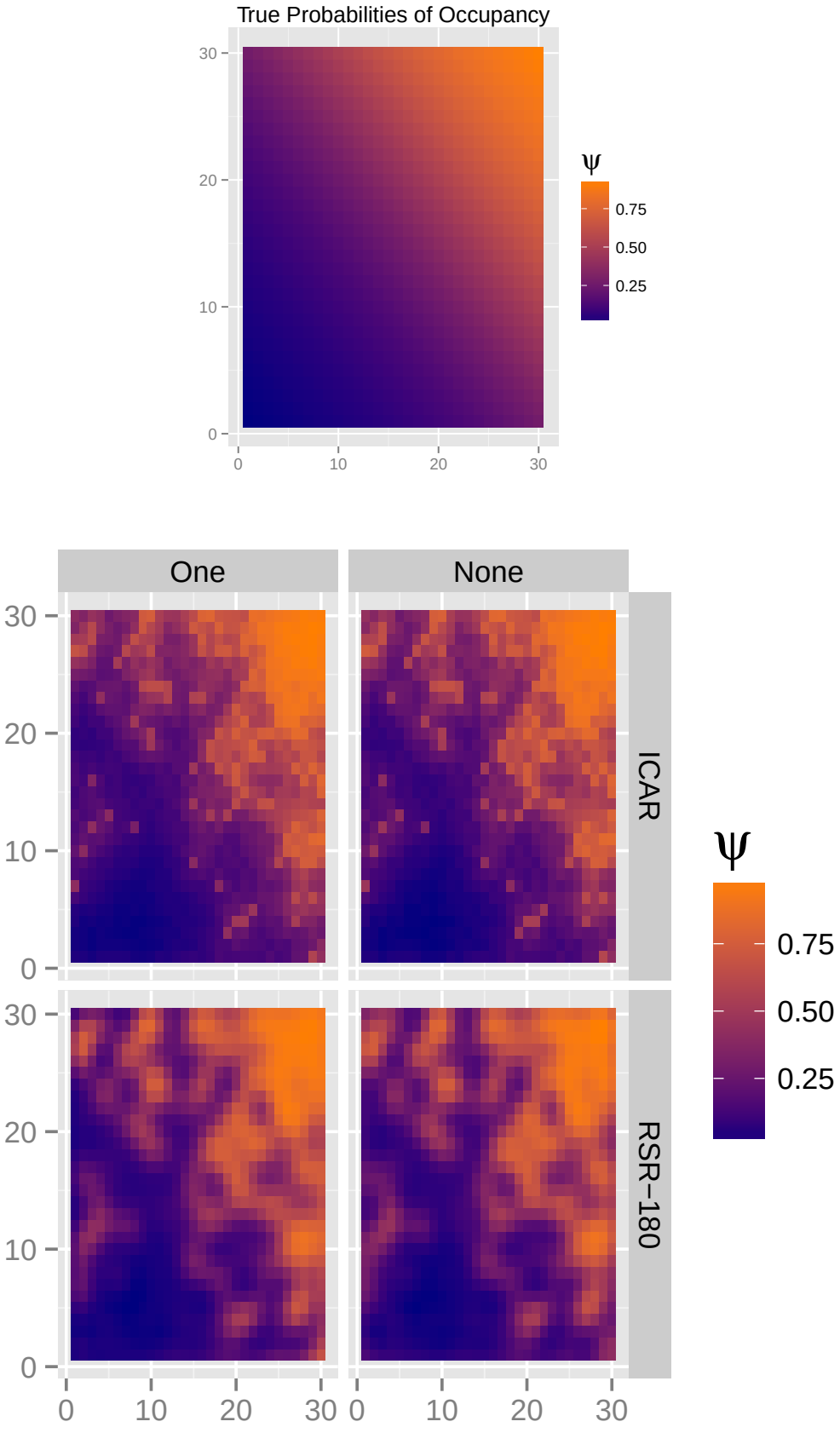
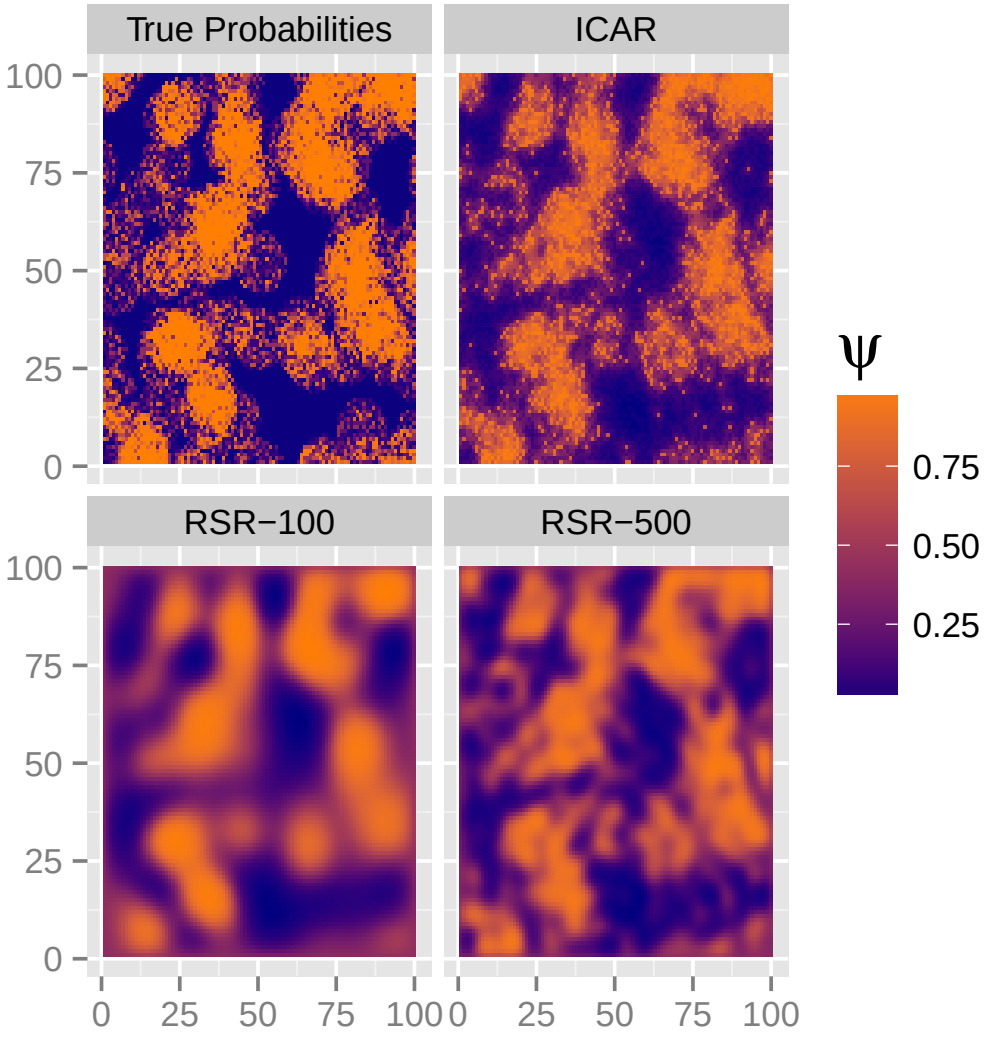

True Probabilities of Occupancy
ψ
0.75
0.50
0.25
One
None
ICAR
RSR−180
0
10
20
30

Figure 3.3: Model estimates when missing clustered covariates for 10,000 sites. Models were fit with occupancy solely being a function of the spatial random effect (Equations 3.20 and 3.21). These plots demonstrate the smoothing that occurs between ICAR and RSR models. Going clockwise from top-left: The true occupancy probabilities that were being estimated; the ICAR model estimates; the RSR model estimates when restricted to 500 eigenvectors; and RSR model estimates when restricted to 100 eigenvectors.



Chapter 4

A SPATIO-TEMPORAL MODEL, INTRODUCED AND APPLIED TO THE COMMON MYNA

4.1 Introduction

Invasive species are a problem worldwide: damaging crops, contributing to the loss of biodiversity, and causing disturbances. One widely referenced publication suggests that invasive species cost the United States alone an estimate $120 billion per year (Pimentel et al., 2005). A better understanding of the causes of an invasive's spread can minimize the damages incurred.

More generally, the knowledge of colonization and extinction patterns has been a quest of ecologists for decades. For many iconic species, scientists have applied radiotelemetry or satellite tags to track the movements of individuals in order to build out population-level inferences. For some species, this information can tell us the causes of population change, but for many species that are rare and elusive, such resources are not available for careful monitoring. In addition, these large-scale questions are difficult to answer when the scale of the data collection is much smaller than the area for which to draw inference. Fortunately, with the advance of crowd-sourcing, modern computing, and remote technologies, we now have the tools to answer these questions for mammals and birds in many regions of the world where the appropriate data were previously unavailable.

Our data comes from the Southern African Bird Atlas Project (SABAP 2). SABAP 2 is a large database of bird detections/non-detections from throughout South Africa from 2007–Present. Volunteer bird watchers following strict protocols on effort and

Figure 3.4: Bias of the spatial occupancy models. The black dots represent the average occupancy probabilities estimates from the true model whose inputs matched the data generating equations, Equation 3.22. The averages were taken over 100 simulations. They show the true model's ability to accurately estimate the occupancy probabilities for the entire range $(0, 1)$ of probabilities. The stars and crosses represent the estimates from the ICAR and RSR-150 models, respectively, and they show similar patterns of inaccurate estimates of the occupancy probabilities. In addition to having high variability between estimates, the models tend to overestimate the extremely low ψ_i values and underestimate the extremely high ψ_i values.

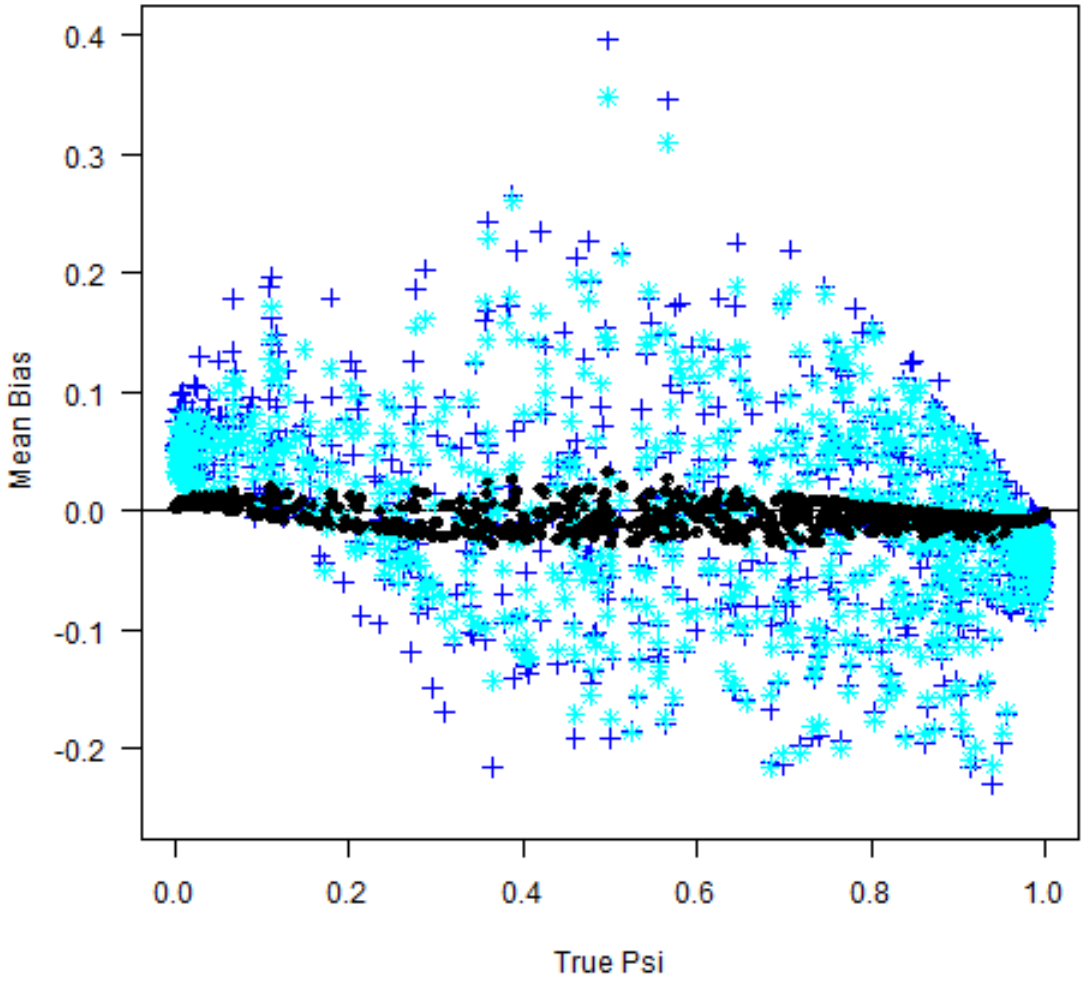
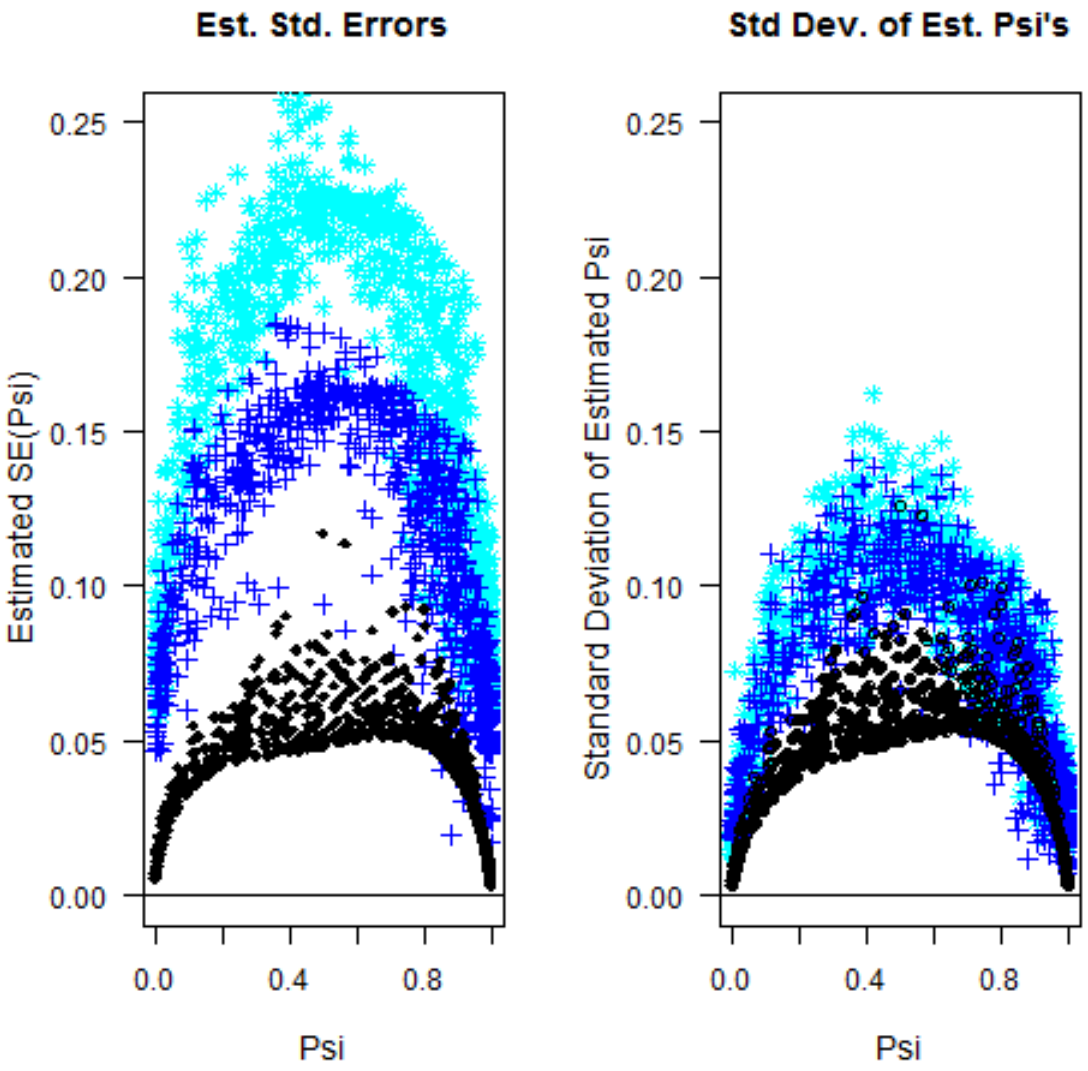

0.4
0.3
0.2
0.1
0.0
-0.1
-0.2
Mean Bias
0.0
0.2
0.4
0.6
0.8
1.0
True Psi

Figure 3.5: Model estimated standard errors. The black dots are the results from a true model whose inputs matched the data generating equations, Equation 3.22. On the left are the results from the median standard errors for the occupancy probabilities as reported by the model output. On the right are the standard deviations related to the median occupancy probabilities from the model output, with the standard deviations taken over the 100 simulations. As desired, the reported standard errors (left graph) match the standard deviations of the parameters (right graph).

The dark blue crosses are the same statistics but for an RSR model, restricted to 150 eigenvectors, and where the fixed u_1 covariate is replaced by a spatial random effect (Equation 3.19). Because the RSR model does not perfectly match the data generating equations, more variance is expected and therefore the graph on the right shows higher standard deviations associated with the parameters.

The RSR model overestimated the standard errors, as seen by the fact that the points on the left graph have larger magnitude than the points on the right graph. In this regard, the ICAR model greatly overestimated its standard errors but the true variance of its parameters matched the variability of the RSR model (right graph).


Est. Std. Errors
Estimated SE(Psi)
0.25
0.20
0.15
0.10
0.05
0.00
0.0
0.4
0.8
Psi
Std Dev. of Est. Psi's
Standard Deviation of Estimated Psi
0.25
0.20
0.15
0.10
0.05
0.00
0.0
0.4
0.8
Psi

habitat coverage collect bird lists. As of May 4, 2013, 86,658 lists have been incorporated into SABAP 2 (http://sabap2.adu.org.za/). Many recent studies similarly use detection/non-detection data amassed by volunteers over large areas and long time periods to get at colonization and extinction patterns (e.g., The Swiss Breeding Bird Survey and The Christmas Bird Counts in the United States).

We applied a recent computing advancement to logistic regressions, the occupancy model, to the SABAP 2 data. Occupancy models use a hierarchical framework to account for the fact that a species may be present at a site but go undetected. Occupancy models are a popular tool for modeling presence-absence data from bird surveys, as evidence by the 1,000+ citations of the MacKenzie et al. (2002) paper that originally introduced this model.

We used a Bayesian version of the occupancy model (see description in Chapter 2) and merged it with an adaptation of the model by Hooten and Wikle (2010) to create a dynamic, multi-season occupancy model that predicts past, current and future species occupancies along with estimates of the parameters that drive these occupancies. Our model is novel in that instead of having the colonization probabilities act independently of the occupancies of a site's neighbors, as in the dynamic occupancy model developed by MacKenzie et al. (2003), we set them to be a function of whether a site's neighbors are occupied. A site is more likely to be colonized if more of its neighbors are currently occupied and if it has better habitat than neighboring sites, but we also allow for the chance of colonization of a site without occupied neighbors through the inclusion of a long-distance dispersal term. Although a few other models do set colonization to be a function of occupied neighbors (Yackulic et al., 2012), our model is more explicit in its connection of these processes while simultaneously allowing for a long-distance dispersal.

Using the SABAP 2 data, we applied the model to the common myna (*Acridotheres tristis*), a starling that is native to Asia but has been introduced to countries around the world. The myna is considered one of the world's top 100 worst invasive species

(Lowe et al., 2000). It was introduced to Durban, a small city in the Southeast corner of South Africa in 1902 (Peacock et al., 2007). After an initial stabilization in that region, its range has undergone spurts of rapid expansions and the myna is now widespread in the eastern half of South Africa (Figure 4.1). The myna's distribution has been noted anecdotally but the drivers of its expansion have not been studied empirically nor statistically.

Our model is the first dynamic occupancy model to determine what may be driving the myna's expansion and what its rates of colonization are. As the myna may outcompete native species, these answers have important conservation implications.

The chapter proceeds as follows. We begin with a description of the data to give context for the subsequent model development. After the data and the model are illustrated, we describe a simulation study that was conducted to test the model. The results, discussion, possible model extensions, and conclusions then follow.

4.2 Methods

4.2.1 Data

The Southern African Bird Atlas Project, SABAP 2, is a large database of bird lists collected by volunteer bird-watchers whom we call citizen scientists (http://sabap2.adu.org.za/). Each bird list is a survey of the bird species detected and not detected within a particular 5-minute latitude by 5-minute longitude grid cell, approximately 8×7.6 km (Harebottle et al., 2007). South Africa is covered by 17,444 of these sites. Each bird list represents one survey of one site with non-detections deduced by a species' absence from the list. Each survey represents a minimum time period of two hours of intensive birding to a maximum time period of five days, and in that time the citizen scientist is expected to cover all habitat types of the grid cell. In this publication, we looked at the checklists as surveys of the common myna, and each checklist is therefore a detection/non-detection record of the myna at a

particular site.

We clumped the data into quarter degree grid cells (QDGC) to compare our model results against an earlier version of the bird atlas project, SABAP 1, which occurred mainly from 1987–1991. Each QDGC is 15-minute latitude by 15-minute longitude and is equal to nine of the smaller grid cells. 1,946 QDGC cover South Africa. SABAP 1 did not have the same, strict protocols on habitat coverage and minimum and maximum time effort per card but the database still provides a good snapshot of the myna distribution in the late 1980s.

The SABAP 2 data were collected from July 2007–present. We divided this multi-year data into five time periods, one for each year of data starting in January 2008 and ending in December 2012. Because 2007 is incomplete, it was excluded from the analysis, leaving five years of complete data.

1,196 of the QDGC sites were included in the analysis and 1,145 of them were surveyed at least once in the five-year period (Figure 4.1; a map of detections and surveys for all South Africa is in Appendix B.1). 209 sites were surveyed more than 30 times in any given year. In order to have more even effort information from each site, the number of surveys per site per year was limited to 30; the surveys were randomly sampled to obtain the 30 to include in the analysis. A benefit of this data truncation is that it allowed us to compare model selection and parameter estimates between slightly different data sets and ensured that our final model was truly representative of the data.

In 2008, the myna was detected at 268 of the 644 sites that were surveyed; in 2009, it was detected at 367 of the 867 surveyed sites; in 2010, it was detected at 395 of 928 surveyed sites; in 2011, it was detected at 370 of 848 sites; and in 2012 it was detected at 401 of 930 sites (Figure 4.1). In each year, the number of detections increased but the number of surveys increased as well, masking any occupancy changes from year to year.

To determine what covariates were likely to affect the myna occurrences and col-

onizations, we first built a single-season spatial occupancy model for the 2008 data. The following site-specific covariates were considered: the logarithm of the human population density (LOG_HUMAN); the latitude and longitude coordinates of the data and their squares (LAT, LONG, LAT2, LONG2); the proportion of the site that was pastureland (PASTURE; Ramankutty et al. (2010b)); the proportion of the site that was cropland (CROP; Ramankutty et al. (2010a)); and the proportion of the site that was in a protected area such as a national park or game reserve (PA; Rouget et al. (2004)). These variables were included because it has been proposed that the distribution of mynas is most associated with human population density and habitat transformations (Peacock et al., 2007). A previous study that analyzed the correlations between the myna distribution and habitat availability clumped all habitat transformations together as one covariate (Peacock et al., 2007); here we separate out agriculturally used land and do not specify an urban habitat variable, as it would be so closely correlated with human density as to be irrelevant as a stand alone variable.

The following survey-specific covariates were considered when building the detection probability function: the number of hours spent intensively birding for the checklist (INTENSIVE); and the total numbers of hours spent on the checklist (TOTAL), which will include the intensive hours birding plus time spent passively birding.

All variables were scaled by their mean and standard deviation for better convergence and ease of comparison between variables. Using the best-fitting single-season occupancy model as the base, we then fit and compared two spatio-temporal models.

4.2.2 Models

Let $\mathbf{z_1} = \{z_{i,1}, \ldots, z_{i,N}\}$ be the true occurrences of sites $i = 1, \ldots, N$ in year 1. If the species was detected at a site, then the species occupies that site and $z_{i,1} = 1$. (We assumed no false-positive detections.) If the species was not detected or the site was not surveyed, then $z_{i,1}$ will be estimated from its occupancy probability, $\psi_{i,1}$. $\boldsymbol{\psi}_1 = \{\psi_{i,1}, \ldots, \psi_{N,1}\}$ is the vector of occupancy probabilities for all sites for year 1.

We modeled the occupancy probabilities as a function of site-specific covariates and a spatial random effect:

$$\begin{aligned} &\mathbf{z_1} \sim \text{Bernoulli}(\boldsymbol{\psi}_1) \\ &\text{probit}(\boldsymbol{\psi}_1) = \mathbf{X}_\psi \boldsymbol{\beta}_\psi + \mathbf{K}\boldsymbol{\alpha} \end{aligned} \tag{4.1}$$

We used the probit-link function to more easily compare the parameter estimates for the spatio-temporal occupancy model with single-season spatial occupancy models fit with the "stocc" package in R (Johnson, 2013), as in Chapters 2 and 3.

Because adjacent sites may be similar to each other even after incorporating the available covariates, residual spatial autocorrelation was accounted for through $\mathbf{K}\boldsymbol{\alpha}$, the spatial random effect. The inclusion of the $\mathbf{K}\boldsymbol{\alpha}$ random effect gives us the *restricted spatial regression*, or RSR, model introduced by Johnson et al. (2013). In this model, the $\mathbf{K}$ are a subset of the eigenvectors for the Moran Operator Matrix, Ω (Hughes and Haran, 2010):

$$\Omega = \frac{N\mathbf{P}^\perp\mathbf{A}\mathbf{P}^\perp}{sum(\mathbf{A})} \tag{4.2}$$

where

$$\mathbf{P}^\perp = \mathbf{I} - \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-\mathbf{1}}\mathbf{X}' \tag{4.3}$$

$$\mathbf{A} = \begin{cases} 1 & \text{if } i \neq k \text{ and sites } i \text{ and } k \text{ are neighbors} \\ 0 & \text{otherwise} \end{cases} \tag{4.4}$$

Within this paper, sites are neighbors if they are adjacent or diagonal from one another (Figure 4.2). It is possible to extend this definition to a second-order neighbor or a non-binary definition of neighbors that is a function of the distance between

centroids of sites. N is the number of sites for which we will draw our inference.

The $\boldsymbol{\alpha}$ follow a Normal distribution:

$$\boldsymbol{\alpha} \sim \text{Normal}\left(\mathbf{0}, \sigma^2 \left(\mathbf{K'QK}\right)^{-1}\right) \tag{4.5}$$

$\mathbf{Q}$ is the precision matrix from the ICAR model and it is similar to the adjacency matrix, $\mathbf{A}$, above:

$$\mathbf{Q} = \begin{cases} -1 & \text{if } i \neq k \text{ and sites } i \text{ and } k \text{ are neighbors} \\ 0 & \text{if } i \neq k \text{ and sites } i \text{ and } k \text{ are not neighbors} \\ |n(i)| & \text{if } i = k \end{cases} \tag{4.6}$$

where $|n(i)|$ is the total number of neighbors (occupied and unoccupied) of site i.

The number of eigenvectors to include in $\mathbf{K}$ is left up to the researcher, but 100–500 has been proposed as the appropriate range of the restriction (Chapter 2). 100 eigenvectors is the minimum to include before the smoothing interferes with the prediction and 500 is the maximum that is necessary due to the diminishing returns with the addition of more eigenvectors (see Chapter 3). We tried several subsets of eigenvectors and found only mild differences between the resulting models; we chose to use 250 eigenvectors as a number that was compromise between minimizing the correlations between covariates and the spatial random effect, and over-smoothing the data.

Occurrences in subsequent years, $t = 2, \ldots, T$, are functions of the occupancies in year 1, the persistence probability, the neighborhood colonization probabilities, and a long-distance dispersal probability:

$$
\begin{aligned}
&z_{i,t} \mid \mathbf{z_{t-1}} \sim \text{Bernoulli}(\theta_{i,t}) \\
&\theta_{i,t} = z_{i,t-1}\phi + (1 - z_{i,t-1})\, I_{\mathcal{N}_{i,t-1}} \bar{d}_{i,t} + (1 - z_{i,t-1}) \left(1 - I_{\mathcal{N}_{i,t-1}}\right) \gamma
\end{aligned}
\tag{4.7}
$$

$\theta_{i,t}$ is the probability of site i being occupied at time t. If site i was occupied at time $t-1$ then $z_{i,t-1} = 1$ and the probability of occupancy is equal to the persistence probability, ϕ (alternatively it can be called the site survival probability). If site i was not occupied at time $t-1$ but at least one of its neighbors was occupied, then $z_{i,t-1} = 0$, $I_{\mathcal{N}_{i,t-1}} = 1$, and the occupancy probability is equal to the probability of being colonized by a neighbor, $\bar{d}_{i,t}$. We interchangeably call this colonization probability the neighborhood dispersal probability. If site i was not occupied and all of its neighbors were also not occupied, then $z_{i,t-1} = 0$, $I_{\mathcal{N}_{i,t-1}} = 0$, and the occupancy probability is equal to the probability of long-distance dispersal, γ.

The persistence probability, ϕ, and the long-distance dispersal, γ, may be estimated as constants or they may be estimated as functions of a time- or space-varying covariate, e.g.

$$
\text{probit}(\gamma_i) = \beta_{\gamma,0} + \beta_{\gamma,1} x_{\gamma,i} \tag{4.8}
$$

The probability of site i being colonized by its neighbors, $\bar{d}_{i,t}$, is a function of the vector $\tilde{\mathbf{z}}_{\mathcal{N}_{\mathbf{i,t-1}}}$, which indicates the neighbors of site i that are occupied at time $t-1$, and of the vector $\mathbf{d_i}$, which is the probability of site i being colonized by one of its occupied neighbors.

$$
\bar{d}_{i,t} = 1 - \exp\left(\tilde{\mathbf{z}}'_{\mathcal{N}_{\mathbf{i,t-1}}} \ln\left(\mathbf{1} - \mathbf{d_i}\right)\right) \tag{4.9}
$$

The formula for Equation (4.9) is derived in Appendix B.4. The dispersal vectors, $\mathbf{d_i}$, are vectors of length 9 and each element represents the probability of site i being colonized by neighbor k, $k = 1, \ldots, 9$ (see Figure 4.2). These probabilities are constant across time but may or may not be constant across space. They may be modeled in one of two ways.

Homogeneous neighborhood dispersal. Here it is assumed that local colonization patterns do not vary across the landscape. The $\mathbf{d_i}$ vector is the same for each site i but each individual d_k that makes up the dispersal vector varies. Under this paradigm, the following priors were placed on the d_k, for every neighbor $k = 1, \ldots, 9$:

$$d_k \sim \text{Uniform}(0, 1) \tag{4.10}$$

Nonhomogeneous neighborhood dispersal. In this scenario, the dispersals are a function of a habitat variable gradient. The colonization probabilities are functions of whether site i has better/worse habitat than its neighbors. The $\mathbf{d_i}$ vectors are different for each site because each site i will have different habitat relationships compared to its neighbors.

$$\begin{aligned} &\text{probit}(\mathbf{d_i}) = \beta_{d,0} + \beta_{d,1}\mathbf{x}_{\mathbf{d},i} \\ &\beta_{d,0} \sim \text{Normal}\,(0, 5) \\ &\beta_{d,1} \sim \text{Normal}\,(0, 5) \end{aligned} \tag{4.11}$$

The $\mathbf{x}_{\mathbf{d},i}$ are the differences in the habitat gradient between site i and its k neighbors. As there are 9 neighbors per site, $\mathbf{x}_{\mathbf{d},i}$ is a vector of length 9. The $\mathbf{x}_{\mathbf{d},i}$ are created through the following relationship: each element of the vector, $x_{\mathbf{d},i,k}$ is equal to site i's k^{th} neighbor's covariate value at that site minus site i's value, and then

divided by the distance between sites to account for the fact that the diagonal neighbors ($k = 1, 3, 7, 9$) are slightly further away than the other neighbors ($k = 2, 4, 6, 8$) (Figure 4.2).

$$x_{\mathbf{d},i,k} = \frac{x_{i,k} - x_i}{\text{dist}(i,k)} \tag{4.12}$$

Observation process. Because the species may be present at a site but go undetected, we included an observation process in the model:

$$\begin{aligned} y_{i,j,t} &\sim \text{Bernoulli}(z_{i,t} \cdot p_{i,j,t}) \\ \text{probit}(p_{i,j,t}) &= \mathbf{X}_p \boldsymbol{\beta}_p \end{aligned} \tag{4.13}$$

The $y_{i,j,t}$ are the observed occurrences of the species and the $p_{i,j,t}$ are the probabilities of detecting the myna on survey j of site i at time t, given it is present at the site. Detection probabilities may be a function of site- and survey-specific covariates. This observation process is the same for all years of data.

Priors. We used the following priors on the rest of the parameters:

$$\begin{aligned} \sigma &\sim \text{Uniform}(0, 100) \\ \boldsymbol{\beta}_\psi &\sim \text{Normal}\left(\mathbf{0}, 5\mathbf{I}\right) \\ \boldsymbol{\beta}_p &\sim \text{Normal}\left(\mathbf{0}, 5\mathbf{I}\right) \\ \gamma &\sim \text{Uniform}(0, 1) \end{aligned} \tag{4.14}$$

Because all variables were standardized by their mean and standard deviation, smaller variances were placed on the β priors so that the prior weights would be more evenly distributed across the spectrum of possible probabilities. More vague priors

and large β's would place a majority of the prior weight on very high and very low occupancy and detection probabilities.

Three chains of the model were each run for 30,000 iterations, a burn-in of 10,000 iterations, and a thinning rate of 5, leaving a total of 12,000 saved iterations. Appendix B.2 gives the model in full and Appendix B.3 can be used as a glossary for the model symbols.

4.3 Model Selection

We fit single-season occupancy models to the data from 2008 to determine the factors that were most correlated with the myna's distribution. To choose the best-fitting single-season model, we first added covariates to the occupancy probability component of the model using forward model selection based on the p-values of the additions. Variable addition stopped when its p-value was greater than 0.05. We then added the site- and survey-specific covariates to the detection probability component of the model, again stopping when an additional variable's p-value was greater than 0.05. The addition of the detection probability covariates rendered some of the occupancy parameters nonsignificant so they were removed from the model one at a time, using high p-values ($p > 0.05$) as the reason for being dropped. After the nonspatial model was fit, the spatial random effect was added to the model and any parameters that became nonsignificant were removed from the model.

The covariates that influence detection were selected with the single-season model for 2008 and then the same covariates were used for the observation process for the subsequent years.

Once the single-season model was chosen, we fit several multi-season models using the single-season model as the base for occurrences for year 1, Equation 4.1. Six multi-season models were fit that varied in the ways that the dispersal probabilities were estimated. Two models used the homogeneous neighborhood colonization as in Equation 4.10; the other models used the nonhomogeneous neighborhood colonization,

as in Equation 4.11, with either human population density, latitude, or longitude as the environmental gradient. The models were fit with the long-distance dispersal being constant and with the long-distance dispersal estimated as a function of the human popultion density. (See Appendix B.2 for the exact equations related to each model.)

We also ran models that estimated a "year 0" of occurrences from the model. In these equations, the year 0 occurrences were estimated solely as a function of the spatial random effect:

$$\begin{aligned} \mathbf{z_0} &\sim \text{Bernoulli}(\boldsymbol{\psi}_0) \\ \text{probit}(\boldsymbol{\psi}_0) &= \beta_0 + K\boldsymbol{\alpha} \end{aligned} \tag{4.15}$$

And Equation 4.7 is used for all the years of actual data, except that $t = 1, \ldots, T$, instead of starting at year 2.

4.4 Simulation Study

A simulation study investigated the convergence and parameter bias for the model renditions described above. We assumed a grid of $30 \times 30 = 900$ sites, of which a random subset of 75% of the sites were selected to be surveyed. Following the setup of Yackulic et al. (2012), we used a constant detection probability of 0.5 and assumed 4 surveys per site.

To generate the data, occupancy for year 1 was a function of the scaled x-coordinate of the data:

$$\text{probit}(\boldsymbol{\psi}_1) = -1 + 1\mathbf{x} \tag{4.16}$$

The true occurrences of year 1 were the outcomes of Bernoulli trials with these probabilities:

$$\mathbf{z_1} \sim \text{Bernoulli}(\boldsymbol{\psi}_1) \tag{4.17}$$

The occurrences in subsequent years were the outcomes of Bernoulli trials from Equation 4.7. In all scenarios, the persistence probability, ϕ, was set as 0.90. Long distance dispersal, γ, was either set constant at 0.05 or as a function of the x-coordinate:

$$\text{probit}(\boldsymbol{\gamma}) = -1.5 + 0.5\mathbf{x} \tag{4.18}$$

These coefficients led to a range of long-distance dispersal probabilities of 0.01 to 0.25, with a median of 0.07. Neighborhood dispersal was homogeneous, with colonization probabilities of (0.20, 0.20, 0.05, 0.20, NA, 0.05, 0.05, 0.05, 0.05) for $k = 1, \ldots, 9$, respectively. Other scenarios had nonhomogeneous neighborhood colonizations with:

$$\text{probit}(\mathbf{d_i}) = -1 + 1\mathbf{x}_{d,i} \tag{4.19}$$

These coefficients let to a range of neighborhood dispersal probabilities of 0.004 to 0.75 with a median of 0.16. The two variations for long-distance dispersal (constant or varying) combined with the two variations for the neighborhood dispersal (homogeneous or nonhomogeneous) led to the testing of four scenarios. Ten simulations were run per scenario.

The above set-up was repeated for another set of four scenarios, the difference being that the year 1 occurrences were generated from Equation 4.16 but there was no detection data included from that year.

For all eight scenarios, the model estimated year 1 occupancy as:

$$\text{probit}(\boldsymbol{\psi}_1) = \beta + \mathbf{K}\boldsymbol{\alpha} \tag{4.20}$$

Therefore, in the models the true covariate was not used and the occupancy probabilities were estimated from the spatial random effect instead. The scenarios where year 1 detection data were unknown/excluded were tested with five simulations each.

Between simulations, the parameters remained constant but the occurrences and detections changed. The given parameter values were chosen to reflect what we believed would be realistic persistence and colonizations from year to year.

4.5 Results

4.5.1 Simulation Study

The detection probability, persistence probabilities, and occurrences for each year were unbiased for all scenarios (Table in Appendix B.5). More variability was seen in the estimation of the dispersal estimates, matching the fact that there were more parameters to be estimated for these components. The long distance dispersal had low bias when the year 1 data were available but exhibited positive bias when the model estimated the year 1 detections and occurrences (right half of Appendix Tables). In particular, the long distance dispersal overestimated the coefficient associated with the x-covariate when long-distance dispersal was varying, as in Equation 4.18 (Tables B.1, B.4). The overestimation of the long distance dispersal was to compensate for the greatly underestimated occurrences in year 1.

In estimating the parameters associated with the neighborhood dispersal, the covariate coefficient in the nonhomogeneous models was slightly underestimated. When the neighborhood dispersal was homogeneous, the probabilities had a slightly positive bias, but none of the neighborhood dispersal biases translated into biased occupancy rates.

The DIC varied widely between the different scenarios. The homogeneous dispersal models always had lower DIC than the nonhomogeneous models; the models with a constant long distance dispersal had lower DIC than the models with varying long-distance dispersal; and models where year 1 data were estimated had higher DIC than models where the year 1 data were available. Therefore, this model selection criteria should not be used to choose between model structures and should be limited to variable selection within a given model structure.

4.5.2 Myna Results

The detection probabilities were positively correlated with the number of hours spent intensive birding and the human population density. The detection probabilities were also functions of the latitude and longitude coordinates of the data (Table 4.1). The myna occupancies of year 2008 were originally correlated with the latitude and longitude coordinates of the data, the human density, and the proportion of agricultural pastureland. Once the detection probability covariates and the RSR spatial random effect were included, only the latitude coordinate and human density affected occupancies. The probability of occupancy increased when either variable increased (Table 4.1).

The spatio-temporal model with homogeneous neighborhood dispersal had a lower DIC (DIC = 32,520.8) than the spatio-temporal model with nonhomogeneous neighborhood dispersals (DIC = 32,579.6), which matches the DIC output from the simulation study. In the homogeneous model, the dispersal probabilities represent the probabilities of the middle site being colonized by its neighbors, these probabilities ranged from 0.006 to 0.561 (Table 4.3). The higher dispersal probability for neighbor 9 means that a site is most likely to be colonized by its southeastern neighbor. Another relatively high rate of dispersal comes from neighbor 4 whose colonization probability is 0.22. Therefore, the range of the myna is mostly expanding northward.

The nonhomogeneous model with neighborhood dispersal being a function of hu-

man density estimated the neighborhood dispersals to range from 0.002 to 0.58, with a median probability of 0.14. The negative coefficient associated with the human population suggests that the myna is dispersing away from the large cities into the less populated surrounding areas, possibly because the myna populations are already saturated in the more heavily populated sites.

A map of the predictive surface of neighborhood dispersals, created from Equation 4.11 but with the habitat differences replaced by the sites' habitat values, shows the routes along which the myna uses for dispersal (Figure 4.3). The dark blue in the northeast corner of the country demonstrates avoidance of Kruger National Park but with potential for colonizations to the north and south of the park.

Two other models with nonhomogeneous neighborhood dispersals were fit where the neighborhood dispersal was a function of latitude or longitude. When dispersal was a function of longitude, its associated coefficient was -1.42, with a 95% credible interval of (-2.012, -0.815); when dispersal was a function of latitude, its associated coefficient was -1.78 with a 95% credible interval of (-3.13, -0.719). Because neither of these models exhibited good convergence, the model that uses the human density as the dispersal parameteter was preferred and chosen as the best-fitting.

In all models, the persistence probability was estimated to be either 0.94 (95% CI: 0.92, 0.95) or 0.93 (95% CI: 0.91, 0.94). The long-distance dispersal was estimated to be 0.01 with a lower bound to its confidence interval equal to 0.

Two additional models were fit with the long distance dispersal as a function of human population densities. The coefficient associated with these covariates was nonsignificant and so models with a constant long distance dispersal were considered to be a better fit (see Appendix B.6).

All models estimated an increase in the number of sites being occupied over time. For the homogeneous model, the estimated number of sites occupied in 2008 was 599 and increased to 724 by 2012 (Table 4.2), suggesting a rate of spread of 4% a year. For the nonhomogeneous model, the estimated number of sites occupied in 2008 was

Table 4.1: The parameter estimates from the model with nonhomogeneous neighborhood dispersals. "Rhat" is a measure of the model convergence; values decrease to 1.00 with more iterations and values below 1.07 are desirable.

NONHOMOGENEOUS DISPERSAL MODEL:					
	Median	**(SE)**	**95 % CI**		**Rhat**
Detection Parameters (probit-scale):					
Intercept	-0.09	(0.02)	-0.12	-0.05	1.00
INTENSIVE	0.11	(0.01)	0.09	0.12	1.00
LAT	0.32	(0.02)	0.29	0.35	1.00
LAT2	-0.48	(0.02)	-0.51	-0.44	1.01
HUMAN_POP	0.57	(0.01)	0.54	0.59	1.00
LONG	0.03	(0.02)	0.00	0.07	1.00
LONG2	-0.32	(0.01)	-0.35	-0.30	1.00
Occupancy Parameters (probit-scale):					
Intercept	0.02	(0.23)	-0.38	0.53	1.04
LAT	2.07	(0.32)	1.50	2.79	1.03
HUMAN_POP	0.84	(0.20)	0.48	1.24	1.01
Spatial Parameter:					
sigma	4.93	(0.81)	3.50	6.76	1.01
Persistance Probability:					
persist.prob	0.94	(0.01)	0.92	0.95	1.00
Long-distance dispersal probability:					
disperse.long	0.01	(0.02)	0.00	0.05	1.00
Neighborhood Dispersal Parameters (probit-scale):					
Intercept	-1.08	(0.12)	-1.31	-0.85	1.00
HUMAN_POP	-0.50	(0.16)	-0.80	-0.20	1.00
Number of Sites Occupied:					
Year 2008	635	(23.1)	592	682	1.03
Year 2009	654	(19.2)	618	693	1.02
Year 2010	675	(18.0)	642	712	1.01
Year 2011	697	(18.7)	662	736	1.01
Year 2012	724	(21.0)	684	766	1.00
Deviance Explained:					
deviance	31656.8	(43.0)	31576.0	31743.3	1.01
DIC	32579.6				

Table 4.2: The estimated coefficient values from the homogeneous model.

HOMOGENEOUS DISPERSAL MODEL:					
	Median	**(SE)**	**95 % CI**		**Rhat**
Detection Parameters (probit-scale):					
Intercept	-0.09	(0.02)	-0.12	-0.05	1.00
INTENSIVE	0.11	(0.01)	0.09	0.12	1.00
LAT	0.31	(0.02)	0.28	0.35	1.00
LAT2	-0.47	(0.02)	-0.50	-0.43	1.00
HUMAN_POP	0.57	(0.01)	0.54	0.59	1.00
LONG	0.03	(0.02)	0.00	0.07	1.00
LONG2	-0.33	(0.01)	-0.35	-0.30	1.00
Occupancy Parameters (probit-scale):					
Intercept	-0.23	(0.18)	-0.56	0.14	1.01
LAT	1.70	(0.32)	1.21	2.43	1.02
HUMAN_POP	0.83	(0.18)	0.51	1.22	1.01
Spatial Parameter:					
sigma	4.63	(1.01)	3.23	7.30	1.04
Persistance Probability:					
persist.prob	0.93	(0.01)	0.91	0.94	1.00
Long-distance dispersal probability:					
disperse.long	0.009	(0.01)	0.000	0.043	1.00
Number of Sites Occupied:					
Year 2008	599	(25.3)	550	649	1.02
Year 2009	643	(18.3)	606	678	1.03
Year 2010	670	(16.6)	637	702	1.03
Year 2011	695	(17.2)	662	730	1.02
Year 2012	724	(19.2)	688	764	1.01
Deviance Explained:					
deviance	31607.23	42.73	31523.54	31691.34	1.00
DIC	32520.77				

Table 4.3: The estimated neighborhood dispersal probabilities (and standard errors) from the homogeneous model. A higher probability means that the site is more likely to be colonized from that direction.

0.08 (0.04)	0.02 (0.02)	0.01 (0.01)
0.22 (0.06)	-	0.16 (0.06)
0.06 (0.05)	0.08 (0.06)	0.56 (0.16)

635 and increased to 724 by 2012 (Table 4.1, Figure 4.5), suggesting a rate of spread of 2.8% a year.

4.6 Discussion

Our models empirically confirmed that human population is a driver of myna occurrences. The models are suggestive that agricultural land transformations to crop or pasture lands do not play as great of a role in the myna dispersals in South Africa as these variables did not significantly affect the occupancy or detection probabilities. Whether or not a site contained protected areas had no affect on the myna occupancy or detection probabilities. Therefore, mynas are not actively avoiding these less disturbed sites but they are also not favoring them.

Previous publications concluded that the myna would avoid highlands and the dry, cold interior of South Africa (Brooke et al., 1986). We did not have data on temperatures at the appropriate spatial and temporal scales to determine if that is indeed a limiting factor to the myna's dispersal. The addition of the spatial random effect was used in place of these potentially significant, missing covariates. In this manner, the models reinforced the need for spatial autocorrelation because these were likely environmental factors that affect myna occupancies but were unavailable for our analysis. The spatial occupancy model used the spatial component to extract information from the residual correlations in the data for better predictive maps.

Figure 4.3 is the surface of potential neighborhood dispersals that shows the flow of colonizations and the spread of the myna northward into Zimbabwe, eastward into Mozambique and Swaziland, and westward along South Africa's coast. It is likely that the myna's range expansion will continue along these routes in the near future.

The myna's capability of long distance dispersals has already been demonstrated, seen in the outpost detections away from its core range in Figure 4.1. In our models, the long distance dispersal probability was estimated to be 1% and its 95% credible interval did overlap 0. Therefore most of the myna's range expansion is through its neighborhood dispersals.

Model Extensions. The simulation study demonstrated the high performance of this dynamic occupancy model. The mild variations of the model can be extended in future applications. For example, one may want some or all of the coefficients to vary for each time period to reflect that colonization patterns may change from year to year. The persistence and long distance dispersal parameters could also vary from year to year. Instead of being constant for all sites, the persistence probability could be a function of a variable and/or the neighborhood dispersals could be a function of more than one environmental gradient variable. With some of these extensions, the models could be used to understand extinction probabilities for species with declining populations.

Management Implications. One of the advantages of building a Bayesian occupancy model as we did here is that the structure easily estimated the actual occurrences for each year of data, along with confidence intervals for these estimates, in addition to the occupancy probabilities. The finite occupancies may be preferred for species management because they relate more closely to the area of study. Occupancy *probabilities* tell the researcher where the species would hypothetically be found if predicting occurrences for an area that has never been surveyed (i.e., Fig-

ure 4.4). If the researcher wants to make further inferences on a larger area from which they have already sampled and surveyed for their analysis, then the estimated *occurrences* explicitly take into account the surveys that have already been conducted (i.e., Figure 4.5).

Other advantages of our model's structures are its two sources of colonizations and the fact that the colonization and extinction probabilities are explicitly linked to a site's occupancy and its neighbors' occupancies from the previous time period. We believe these explicit connections are intuitive, provide more meaning into both processes, and the two sources of colonization more closely mimic the true behavior of the species. The model provides the appropriate structure to ask biological questions.

From the neighborhood colonizations, we created a gradient field for the future spread of the myna. When building a dynamic occupancy model, one of the most interesting results of looking at multiple seasons at once is to concurrently look at the colonization and extinction processes that drive the occupancy map changes. As far as we are aware, previous publications on dynamic occupancy models have not included such gradient maps for the colonization or extinction processes. In the case of the myna, an invasive whose range is expanding, we focused on the colonization process. The map of its dispersal gradient can inform managers on the directions from which the myna will be invading.

Finally, our model concludes that the myna's range continues to expand at a rate of approximately 3% a year. Because the myna population has not yet stabilized, resource managers should continue to be aware of myna expansions and what that may mean for important birding areas and other protected, biodiversity areas.

Figure 4.1: The outline of the map shows the sites included in our analyses. Each square is one quarter degree grid cell (QDGC). The maps on the left give the number of detections at each site in 2008 and 2012; the maps on the right give the number of surveys that occurred at each site in 2008 and 2012.

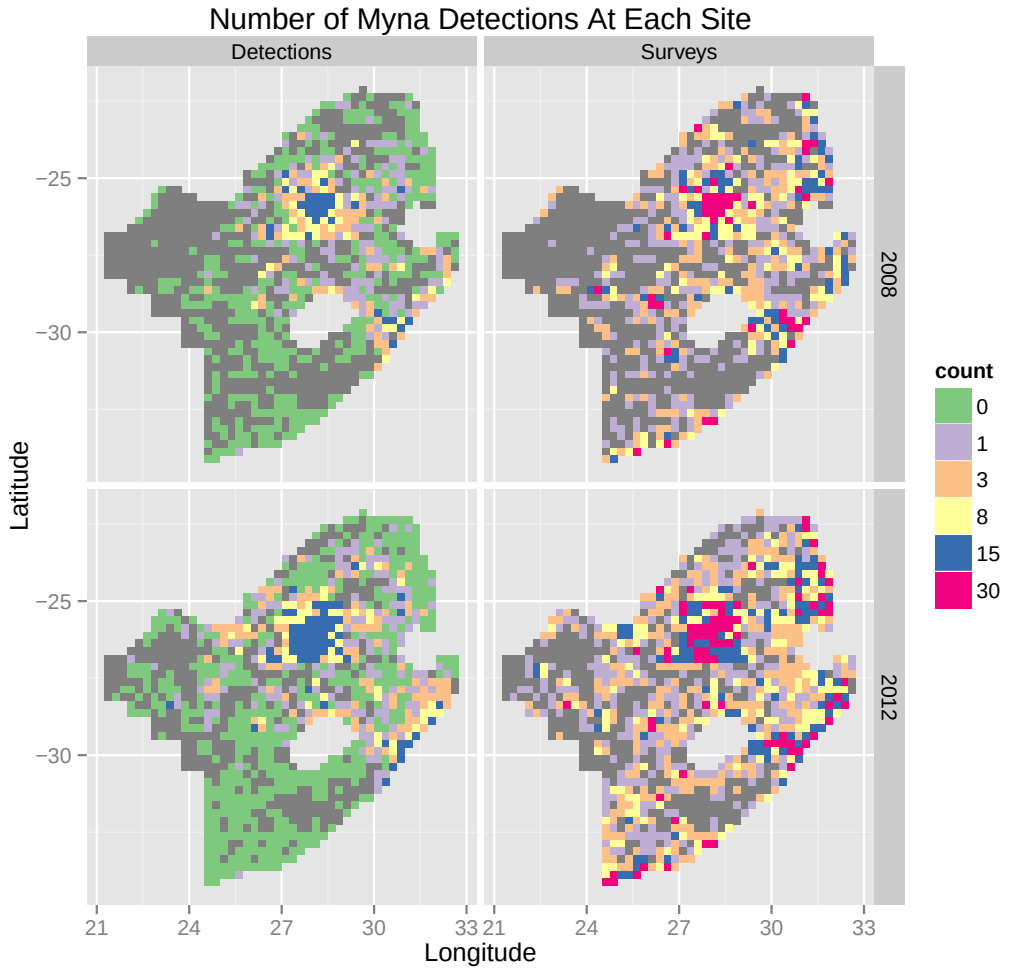


Figure 4.2: Neighborhood structure.

1	2	3
4	5	6
7	8	9

Figure 4.3: Gradient surface of potential neighborhood dispersals. The common myna is likely to disperse to the areas of orange and will avoid the dark blue sites.

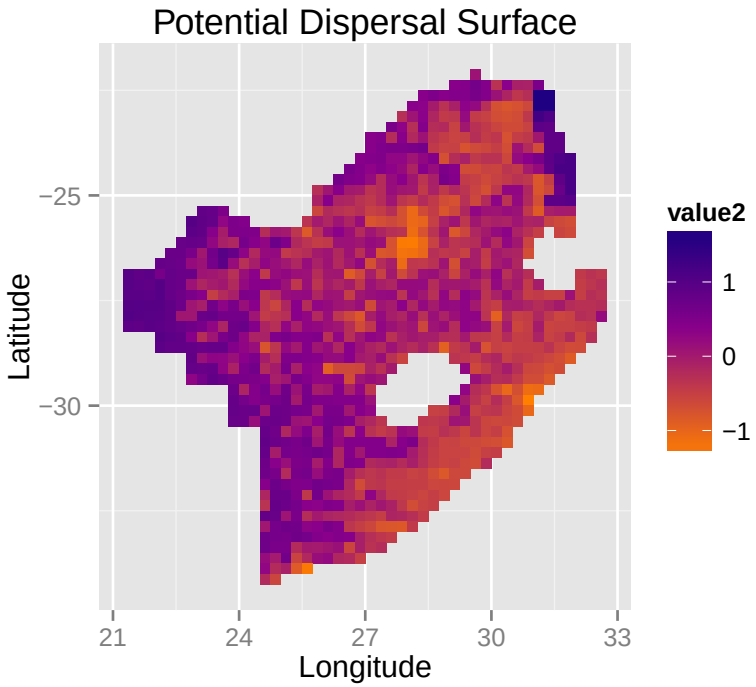


Figure 4.4: Estimated occupancy probabilities. Because 2008 is estimated differently from the other years, the probability map has different shading.

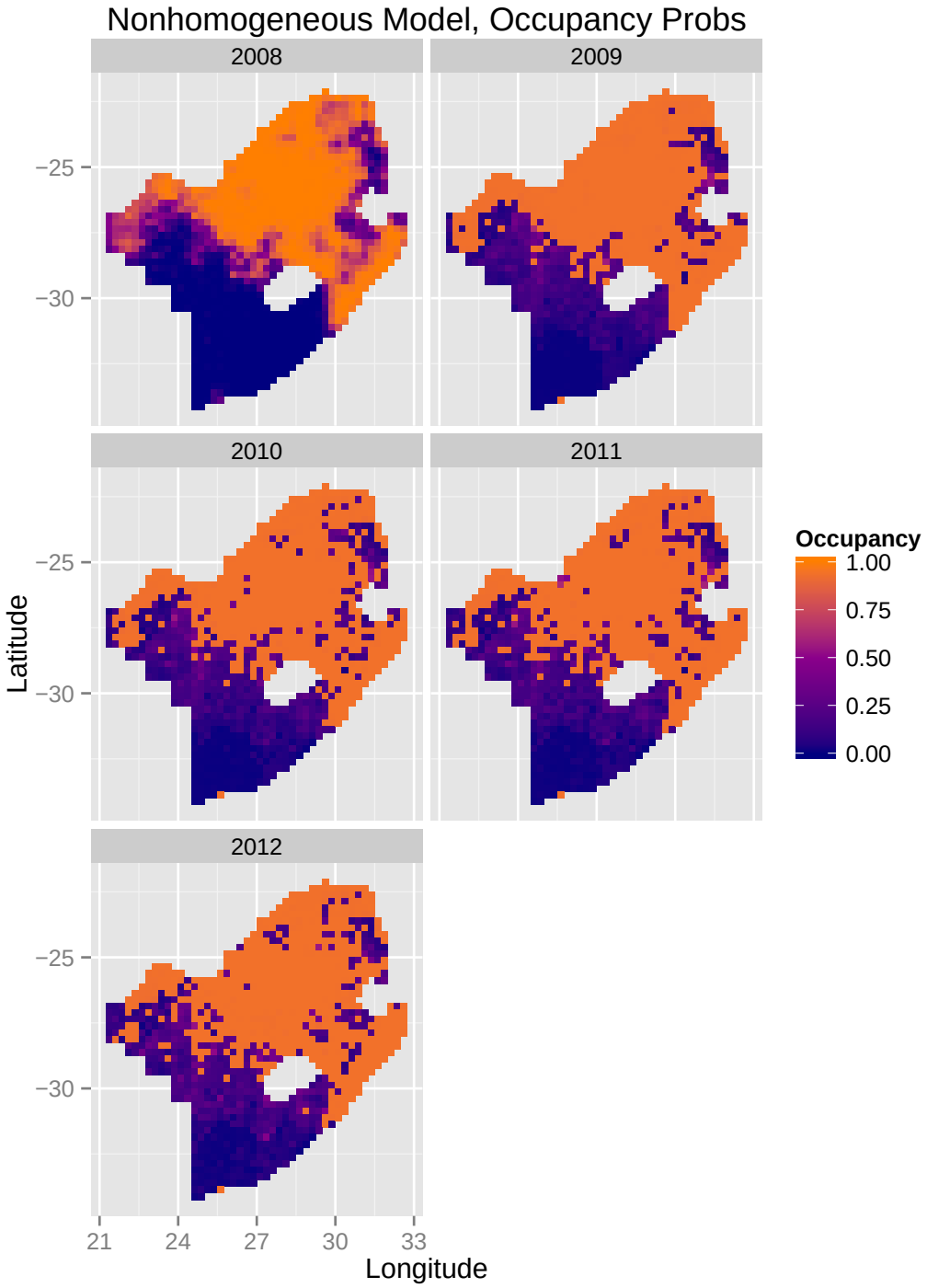
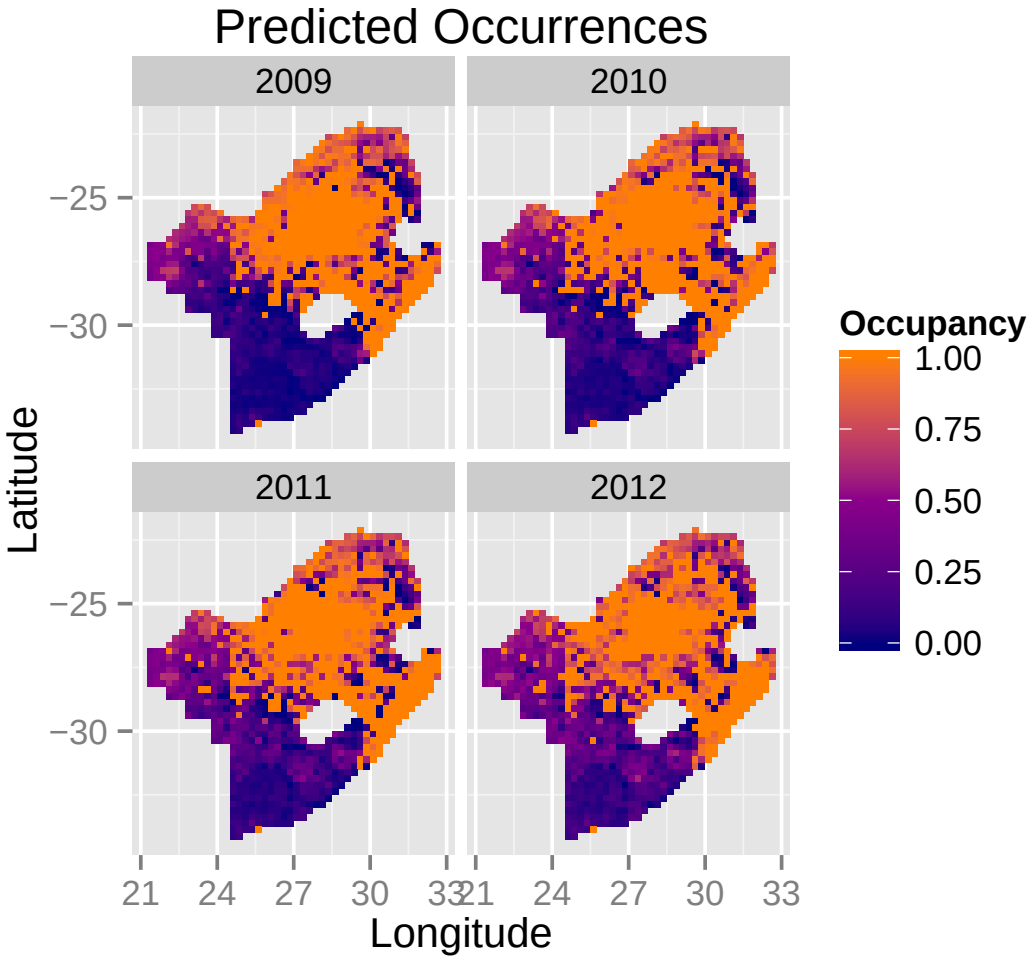


Figure 4.5: The mean occurrence values for each site and all years.



Chapter 5

CONCLUSIONS & IMPLICATIONS

This work has demonstrated the potential of large, volunteer-based databases combined with modern computing power and improved modeling techniques to answer important ecological questions when limited resources are available for studies. Presence-absence data are common in ecology because it is easy to collect, cost efficient, and non-invasive to the species under study. Many techniques have previously been employed to analyze presence-absence data, but most assume the data points are independent. We have updated existing models by adding spatial and temporal components to them. The spatial model introduced in Chapter 2 has provided insight on an elusive species with low detection rates; the spatio-temporal model introduced in Chapter 4 has provided insight on an invasive species whose range is expanding in South Africa. We hope that scientists will decide to take advantage of the spatial and temporal components of their data and expand upon basic occupancy model analyses by using the methods described in the previous chapters.

The methods we constructed were in consideration of the data from the Southern African Bird Atlas Project (SABAP). SABAP is a set of surveys submitted by volunteers and it maps the detections (and non-detections) of all bird species occurring in South Africa. In the 1990s, data from the first phase of the project, SABAP 1, was used to create general range maps for each species in South Africa for the first time. Because the SABAP data involves adjacent sites, we expanded upon a recently developed method, the restricted spatial regression (RSR) model, to combine this spatial information with a hierarchical occupancy model structure. Our spatial occupancy model led to detailed occupancy maps. Adding a spatial component when

modeling occupancy compensated for low occurrences, low sampling coverage, confounding between variables, and allowed us to produce accurate occupancy maps with and without environmental covariates.

The spatial occupancy model can be used on any presence-absence data set consisting of neighboring sites. The SABAP 2 database alone has this type of data on the 800+ species that occur in South Africa; similar atlas data has been collected at a range of spatial scales in countries all over the world. A spatial occupancy model is the appropriate method to analyze this gridded data, and the RSR model may be the best version of these methods to date. Through our simulations, we have shown that researchers should always fit spatial occupancy models rather than nonspatial occupancy models.

The SABAP data also has a temporal component, as it is a database of surveys that have been collected over more than five years. Because the distribution of a population will change over time, we have developed a spatio-temporal occupancy model that includes colonization and extinction processes. This model improves upon previously developed, multi-season occupancy models because colonizations are directly linked to a site's previous occupancy status and its neighbors' occupancy statuses.

The spatio-temporal model estimates occupancy probabilities for year 1 and then derives future occupancy probabilities from the previous year's occupancy probabilities, and the persistence and dispersal probabilities. Rates of population change are derived from the occurrences estimated for each year of data.

The spatio-temporal model was developed specifically for the SABAP data, but it can be applied to any presence-absence data with neighboring sites and with a temporal component. Its application is advantageous to resource managers who wish to learn if, where, and why a species' range is contracting or expanding.

5.1 Summary

In Chapter 2, we demonstrated the utility of adding a spatial random effect to an occupancy model. Traditionally, a CAR variable has been added to a model, but we found an RSR variable to be a better alternative. When applied to the Southern ground hornbill, the RSR occupancy model produced a better fit than either the nonspatial occupancy model or the ICAR occupancy model. In addition, the RSR occupancy models were easily implemented in R, produced estimates in a fraction of the time, and proved to be very promising for large data sets ($> 1,000$ sites).

The analysis concluded that the Southern ground hornbills are found primarily in protected areas; thus providing quantitative support of the value of protected areas to the ground hornbill. The lack of significance of any other vegetation-related variables reinforced findings of previous studies that the ground hornbill is otherwise a diet generalist. Our range maps highlighted the southernmost region of South Africa as an area with potential for high occupancies of ground hornbills.

In Chapter 3, we summarized spatial occupancy model performance. We focused on scenarios where the incoming data were complex, degraded, and/or mismatched to the model assumptions. In general, both the ICAR and RSR occupancy models were robust to these trials. They used spatial random effects to pick up missing spatial relationships and accurately predicted occupancy probabilities for surveyed and unsurveyed sites.

Either the RSR or the ICAR model may be preferred, depending on the desired outcomes. The ICAR model predicted the occupancy probabilities more accurately and was better at picking up small-scale variance, but it biased the parameter estimates of the fixed effects and it inflated the standard errors.

The RSR model is preferred if the data are degraded (e.g. lower detections, fewer sites surveyed) and/or there is confounding between the occupancy and detection probabilities. In these scenarios, the RSR models picked up the residual spatial

autocorrelation when the ICAR model was unable to. In addition, if one is using the model to make predictions on unsurveyed sites, then the smoothing that occurs from the RSR model may be preferred.

For these reasons, we recommend choosing the RSR models within the occupancy model framework. As with other smoothing techniques, the exact amount of smoothing/restriction to be applied when using an RSR model is left up to the researcher. However, we recommend a restriction of 100–400 eigenvectors, regardless of the number of sites in the area of inference. One should use a more restricted RSR model, i.e 100–200 eigenvectors to obtain accurate covariate estimates and standard errors. Fewer eigenvectors will lead to a smoother map of predictions. One should use a less restricted RSR model, i.e. 300–400 eigenvectors, if one wants to more closely mimic the ICAR model results and capture the smaller-scale variation of an area that has already been surveyed.

In Chapter 4, we built a dynamic occupancy model with spatial and temporal components. This model has local and long distance dispersals. The dispersal and persistence probabilities are explicitly linked to a site's occupancy and its neighbors' occupancies from the previous year. We believe these explicit connections are intuitive, provide more meaning for both processes, and that the two types of dispersal closely mimic the true behavior of the species.

A simulation study demonstrated the high performance of this spatio-temporal occupancy model. Mild variations of the model could be extended in future applications. For example, the persistence, neighborhood, and long distance dispersal parameters could vary from year to year. Alternatively, the model could be used to better understand extinction probabilities for species with declining populations.

We applied the model to the common myna, an invasive whose range is expanding in South Africa. The model empirically confirmed that human population is a driver of myna occurrences. The models were suggestive that land transformations from native vegetation to crop or pasture lands do not play a large role in the myna dispersals in

South Africa, and that mynas are not actively avoiding protected areas but they are also not favoring them.

Our model concluded that the myna's range continues to expand at a rate of approximately 3% a year. A map of its neighborhood dispersals shows the potential spread of the myna northward into Zimbabwe, eastward into Mozambique and Swaziland, and westward along South Africa's coast. It is likely that the myna's range expansion will continue along these routes in the near future. Because the myna population has not yet stabilized, resource managers should continue to be aware of myna expansions and what that may mean for important birding areas and other biodiversity hotspots.

5.2 Future Model Applications

No model will be perfect for all data sets. As methods, tools, and data sets advance, we need general models with flexible frameworks that can be specialized for each species, data set, and ecological question. We have developed such models within this dissertation.

In order to demonstrate the models' utility, we applied them to two species: the Southern ground hornbill and the common myna. Both species were chosen because they would not be misidentified, so that there would be no false positive detections, which is a basic assumption of an occupancy model.

The Southern ground hornbill was chosen for the single-season, spatial occupancy model because of its slow reproduction and lack of migrations, allowing us to comfortably combine multiple years of data into one season. Beyond satisfying the model assumptions, the Southern ground hornbill was an interesting example because of the ecological questions that could be answered. Very little is known about the Southern ground hornbill. Therefore, we were able to answer questions on habitat preferences and draw inferences from its low levels of detections outside protected areas. Our model confirmed that the Southern ground hornbill is a diet generalist and that its

populations are doing best in protected areas.

The myna was chosen as the example species for the spatio-temporal occupancy model because it also had interesting ecological questions to be answered. The myna has been described as one of the world's worst invasive species; and its range has been expanding over the last few decades. Our model confirmed that the myna continues to expand, estimated its rate of increase and the direction of expansion, and concluded that the myna resides in sites with high human densities.

Both the spatial occupancy model of Chapter 2 and the spatio-temporal occupancy model of Chapter 4 may be applied to any species that will not be misidentified. The most interesting applications will be to diet generalists, as results may help deduce what other factors drive the species' distribution and dynamics. In line with the recommendations for a nonspatial occupancy model, detection probabilities for the species should average at least 0.30 at the sites where the species does occur, unless more than three surveys have been conducted at the sites.

5.3 Conservation and Management Implications

Occupancy maps provide guidance on land use for resource management. The predictive maps created through this dissertation's models are more insightful and informative than the nonspatial occupancy models and other methods that resource managers currently use.

The inclusion of spatial random effects in our models allow for the prediction of occupancy with missing environmental covariates and even with no covariates. This is advantageous if the survey data are collected on a different scale than the environmental data. It is likely that the scale and resolution of the potential covariates, such as vegetation type or rainfall data, do not match the scale and resolution of the survey data. Other factors such as competition with other species may also affect the distribution of the species of interest, and similarly, the scale and resolution of this data may not match the survey data. In these scenarios, it is very useful to be able

to predict occupancies without having to rely on this "mismatched" data.

In addition, the spatio-temporal model produces a gradient surface of the directions in which dispersal is expected. Causes and patterns of colonization and extinction processes are fundamental concepts of ecology; our spatio-temporal occupancy model is therefore valuable to biologists and theoretical ecologists to empirically test their hypotheses.

5.4 Recommendations for Sampling Design Improvements

The creators of SABAP set strict protocols for their data collection, allowing for complex, detailed analyses that would have otherwise not been an option. They require that each survey be comprised of at least two hours of intensive birding, with a maximum time period of five days spent on the checklist, and that the amount of time spent birding be recorded. This type of observer effort information is important in estimating detection probabilities and it allows one to monitor seasonal trends in the data such as migration patterns.

The other important component of the protocol is to have a tight spatial resolution, setting the sites' sizes to be approximately 7×7 km. This small scale is valuable for resource managers for drawing wildlife-habitat associations; yet, it is a manageable amount of area for a birder to traverse in a survey. Ideally, the size of a site would be smaller to draw even more direct wildlife-habitat associations, but there must be a balance between what is reasonable to ask of volunteers and what is best for the analyses. Both the temporal and spatial requirements of SABAP should be emulated by other atlas projects.

The SABAP administrators regularly put forth "challenges" to their volunteers to encourage them to survey remote areas of South Africa and to increase the number of surveys at some sites, which has led to a decently large sampling coverage. However, it is a challenge for SABAP, indeed for any citizen science project, to get enough surveys, especially for areas with low biodiversity. The occupancy model framework

compensates for some of the uneven sampling of the sites; nonetheless, more surveys need to be completed at the less diverse and/or less accessible areas.

Ideally, the SABAP database would be augmented with surveys from a hired observer to regularly survey the areas of South Africa that are difficult to get to and that are expected to have low biodiversity, i.e. the Karoo desert region. The hired observer would also spend time making sure that there are greater than three surveys per site per year at a selection of sites for each vegetation type and at the full range of latitudes and longitudes of South Africa. If SABAP is unable to hire someone to even out the sampling scheme, the SABAP administrators should continue with their challenges to encourage the volunteers to reach these same goals.

5.5 Final Remarks

Within this dissertation, we have uncovered unique occupancy patterns and processes using new methods for spatial occupancy models. As databases of presence-absence records become larger and computing power increases, new methods can be exploited to provide insight into the occupancy, dispersal, and persistence probabilities. We hope that these models are used in conjunction with the SABAP data and with other presence-absence data sets to answer important ecological questions on the distributions and population changes of wildlife.

BIBLIOGRAPHY

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

Altwegg, R., Broms, K. M., Erni, B., Barnard, P., Midgley, G. F., and Underhill, L. G. (2012). Novel methods reveal shifts in migration phenology of barn swallows in South Africa. *Proceedings of the Royal Society-B*, 279(1733):1485–1490.

Altwegg, R., Wheeler, M., and Erni, B. (2008). Climate and the range dynamics of species with imperfect detection. *Biology Letters*, 4:581–584.

Animal Demography Unit (2011). The Southern African Bird Atlas Project 2. http://sabap2.adu.org.za/. [University of Cape Town. Online; accessed 11-February-2011].

Augustin, N. H., Mugglestone, M. A., and Buckland, S. T. (1996). An autologistic model for the spatial distribution of wildlife. *Journal of Applied Ecology*, 33:339–347.

Austin, M. P. (2002). Case studies of the use of environmental gradients in vegetation and fauna modeling: theory and practice in Australia and New Zealand. In Scott, J. M., Heglund, P. J., Morrison, M. L., Haufler, J. B., Raphael, M. G., Wall, W. A., and Samson, F. B., editors, *Predicting Species Occurrence: Issues of Accuracy and Scale*, pages 73–82. Island Press, Washington, D.C.

Banerjee, S., Carlin, B. P., and Gelfand, A. E. (2004). *Hierarchical Modeling and Analysis for Spatial Data*. Chapman & Hall/CRC Press LLC, Boca Raton.

Bierman, S. M., Butler, A., Marion, G., and Kuhn, I. (2010). Bayesian image restoration models for combining expert knowledge on recording activity with species distribution data. *Ecography*, 33(3):451–460.

BirdLife International (2012). Bucorvus leadbeateri. In IUCN Red List of Threatened Species, V. 2012.1.

Bled, F., Royle, J. A., and Cam, E. (2011). Hierarchical modeling of an invasive spread: the eurasian collared-dove (streptopelia decaocto) in the united states. *Ecological Applications*, 21(1):290–302.

Block, W. and Brennan, L. (1993). The habitat concept in ornithology: theory and applications. *Current Ornithology*, 11:35–91.

Boots, B. and Tiefelsdorf, M. (2000). Global and local spatial autocorrelation in bounded regular tessellations. *Journal of Geographical Systems*, 2:319–348.

Boyce, M. S. and McDonald, L. L. (1999). Relating populations to habitats using resource selection functions. *Tends in Ecology & Evolution*, 14(7):268–272.

Brooke, R., Lloyd, P. H., and De Villiers, A. L. (1986). Alien and translocated terresttrial vertebrates in South Africa. In MacDonald, I., Kruger, F. J., and Ferrar, A. A., editors, *The Ecology and Management of Biological Invasions in Southern Africa*, pages 63–74. Oxford University Press, Cape Town.

Brown, J. H. and Maurer, B. A. (1989). Macroecology: The division of food and space among species on continents. *Science*, 243(4895):1145–1150.

Chelgren, N. D., Adams, M. J., Bailey, L. L., and Bury, R. B. (2011). Using multilevel spatial models to understand salamander site occupancy patterns after wildfire. *Ecology*, 92(2):408–421.

Cressie, N. and Wikle, C. K. (2011). *Statistics for Spatio-Temporal Data*. Wiley.

Donald, P. F. and Fuller, R. J. (1998). Ornithological atlas data: a review of uses and limitations. *Bird Study*, 45:129–145.

Dormann, C. F., McPherson, J. M., Araujo, M. B., Bivand, R., Bolliger, J., Carl, G., Davies, R. G., Hirzel, A., Jetz, W., Kissling, W. D., Kuhn, I., Ohlenmuller, R., Peres-Neto, P. R., Reineking, B., Schroder, B., Schurr, F., and Wilson, R. (2007). Methods to account for spatial autocorrelation in the analysis of species distributional data: A review. *Ecography*, 30:609–628.

ESRI (2011). Arcgis desktop: Release 10. Environmental Systems Research Institute, Redlands, California.

Faraway, J. J. (2005). *Extending the Linear Model with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models*. Chapman & Hall/CRC Texts in Statistical Science.

Fiske, I. and Chandler, R. (2011). unmarked: an R package for fitting hierarchical models of wildlife occurrence and abundance. *Journal of Statistical Software*, 43:1–23.

Gardner, B., Royle, J. A., and Wegan, M. T. (2009). Hierarchical models for estimating density from DNA mark-recapture studies. *Ecology*, 90(4):1106–1115.

Gaston, K. J. (2003). *The Structure and Dynamics of Geographic Ranges*. Oxford University Press, New York.

Gelfand, A. E. and Ghosh, S. K. (1998). Model choice: a minimum posterior predictive loss approach. *Biometrika*, 85:1–11.

Greenwood, J. J. D. (2007). Citizens, science and bird conservation. *Journal of Ornithology*, 148:S77–S124.

Grinnell, J. (1917). The niche-relationships of the California thrasher. *The Auk*, 34:427–433.

Guisan, A. and Zimmermann, N. E. (2000). Predictive habitat distribution models in ecology. *Ecological Modeling*, 135:147–186.

Hanski, I. (1998). Metapopulation dynamics. *Nature*, 396:41–49.

Harebottle, D. M., Smith, N., Underhill, L. G., and Brooks, M. (2007). Southern African bird atlas: Quick-start guide. Accessed through SABAP 2 website: sabap2.adu.org.za/docs/sabap2_instructions_v5.pdf.

Harrison, J. A., Allan, D. G., Underhill, L. G., Herremans, M., Tree, A. J., Parker, V., and Brown, C. J., editors (1997a). *The Atlas of Southern African Birds*, volume 1: Non-passerines. BirdLife South Africa.

Harrison, J. A., Allan, D. G., Underhill, L. G., Herremans, M., Tree, A. J., Parker, V., and Brown, C. J., editors (1997b). *The Atlas of Southern African Birds, Volume 1: Non-passerines*. BirdLife South Africa, Johannesburg, South Africa.

Harrison, J. A., Underhill, L. G., and Barnard, P. (2008). The seminal legacy of the Southern African Bird Atlas Project. *South African Journal of Science*, 104(3-4):82–84.

Heikkinen, J. and Högmander, H. (1994). Fully Bayesian approach to image restoration with an application in biogeography. *Applied Statistics*, 43(4):569–582.

Hilton-Taylor, C., editor (2001). *2000 IUCN Red List of Threatened Species*. International Union for Conservation of Nature and Natural Resources (IUCN), Gland, Switzerland and Cambridge, UK.

Hines, J. E., Nichols, J. D., Royle, J. A., MacKenzie, D. I., Gopalawamy, A. M.,

Kumar, N. S., and Karanth, K. U. (2010). Tigers on trails: Occupancy modeling for cluster sampling. *Ecological Applications*, 20(5):1456–1466.

Hockey, P., Dean, W. R. J., and Ryan, P. G. (2005). *Roberts– Birds of Southern Africa*. The Trustees of the John Voelcker Bird Book Fund, Cape Town, South Africa.

Hodges, J. S. and Reich, B. J. (2010). Adding spatially-correlated errors can mess up the fixed effects you love. *The American Statistician*, 64:325–334.

Hoeting, J. A., Leecaster, M., and Bowden, D. (2000). An improved model for spatially correlated binary responses. *Journal of Agricultural, Biological, and Environmental Statistics*, 5(1):102–114.

Högmander, H. and Møller, J. (1995). Estimating distribution maps from atlas data using methods of statistical image analysis. *Biometrics*, 51:393–404.

Hooten, M. B., Larsen, D. R., and Wikle, C. K. (2003). Predicting the spatial distribution of ground flora on large domains using a hierarchical Bayesian model. *Landscape Ecology*, 18:487–502.

Hooten, M. B. and Wikle, C. K. (2010). Statistical agent-based models for discrete spatio-temporal systems. *Journal of the American Statistical Association*, 105:236–248.

Hughes, J. and Haran, M. (2010). Dimension reduction and alleviation of confounding for spatial generalized linear mixed models. *ArXiv 1101.6649v1 [stat.ME]*.

Hulley, P. and Craig, A. J. (2007). The status of the Southern ground hornbill in the Grahamstown region, Eastern Cape, South Africa. *Ostrich: Journal of African Ornithology*, 78:89–92.

IUCN (2001). *IUCN Red List Categories and Criteria: Version 3.1.* IUCN Species Survival Commission, Cambridge.

Johnson, D. S. (2013). stocc: fit a spatial occupancy model via Gibbs sampling. R package, version 1.0-5.

Johnson, D. S., Conn, P. B., Hooten, M., Ray, J., and Pond, B. (2013). Spatial occupancy models for large data sets. *Ecology*, http://dx.doi.org/10.1890/12-0564.1.

Jones, J. (2001). Habitat selection studies in avian ecology: a critical review. *The Auk*, 118:557–562.

Kemp, A. C. and Begg, K. S. (1996). Nest sites of the Southern ground hornbill, Bucorvus leadbeateri, in the Kruger National Park, South Africa, and conservation implications. *Ostrich: Journal of African Ornithology*, 67:9–14.

Kemp, A. C. and Kemp, M. I. (1980). The biology of the Southern ground hornbill Bucorvus leadbeateri (Vigors)(aves: Bucerotidae). *Annals of the Transvaal Museum*, 32:65–100.

Kéry, M., Gardner, B., and Monnerat, C. (2010). Predicting species distributions from checklist data using site-occupancy models. *Journal of Biogeography*, 37:1851–1862.

Kéry, M. and Schaub, M. (2012). *Bayesian Population Analysis Using WinBUGS. A Hierarchical Perspective.* Academic Press, Waltham.

Knight, G. M. (1990). A preliminary investigation into the status, distribution and some aspects of the foraging ecology of the Southern ground hornbill (Bucorvus cafer) in Natal. Master's thesis, University of KwaZulu-Natal.

Larsen, R., Holmern, T., Prager, S. D., Maliti, H., and Røskaft, E. (2009). Using the extended quarter degree grid cell system to unify mapping and sharing of biodiversity data. *African Journal of Ecology*, 47(3):382–393.

Latimer, A. M., Wu, S., Gelfand, A. E., and Silander, Jr., J. A. (2006). Building statistical models to analyze species distributions. *Ecological Applications*, 16:33–50.

Levins, R. (1969). Some demographic and genetic consequences of environmental heterogeneity for biological control. *Bulletin of the ESA*, 15(3):237–240.

Lowe, S., Browne, M., Boudjelas, S., and De Poorter, M. (2000). *100 of the World's Worst Invasive Alien Species: a Selection from the Global Invasive Species Database*. Species Survival Commission, World Conservation Union, Auckland, New Zealand.

Lunn, D. J., Thomas, A., Best, N., and Spiegelhalter, D. (2000). Winbugs– a Bayesian modeling framework: concepts, structure, and extensibility. *Statistics and Computing*, 10(325–337).

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*, 84:2200–2207.

MacKenzie, D. I., Nichols, J. D., Lachman, G. B., Droege, S., Royle, J. A., 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. L., and Hines, J. E. (2006). *Occupancy Estimation and Modeling: Inferring Patterns and Dynamics of Species Occurrence*. Elsevier Academic Press.

Magoun, A. J., Ray, J. C., Johnson, D. S., Valkenburg, P., Dawson, F. N., and Bowman, J. (2007). Modeling wolverine occurrence using aerial surveys of tracks in snow. *Journal of Wildlife Management*, 71(7):2221–2229.

Moore, J. E. and Swihart, R. K. (2005). Spatial autocorrelation with hierarchically structured data. *Journal of Wildlife Management*, 69(3):933–949.

Morrison, K., Daly, B., Burden, D., Engelbrecht, D., Jordan, M., Kemp, A., Kemp, M., Potgieter, C., Turner, A., and Friedmann, Y. (2005). Southern ground hornbill (Bucorvus leadbeateri) population and habitat viability assessment (phva) workshop report. Technical report, Conservation Breeding Specialist Group (SSC / IUCN) South Africa; Endangered Wildlife Trust.

Mucina, L. and Rutherford, M. C. (2006). *The vegetation of South Africa, Lesotho and Swaziland*. South African National Biodiversity Institute, South Africa.

Paciorek, C. J. (2010). The importance of scale for spatial-confounding bias and precision of spatial regression estimators. *Statistical Science*, 25:107–125.

Peacock, D. S., van Rensburg, B. J., 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(11-12):465–473.

Pimentel, D., Zuniga, R., and Morrison, D. (2005). Update on the environmental and economic costs associated with alien-invasive species in the United States. *Ecological Economics*, 52:273–288.

Ramankutty, N., Evan, A. T., Monfreda, C., and Foley, J. A. (2010a). Global agricultural lands: Croplands, 2000. Data distributed by the NASA Socioeconomic Data and Applications Center (SEDAC).

Ramankutty, N., Evan, A. T., Monfreda, C., and Foley, J. A. (2010b). Global agricultural lands: Pastures, 2000. Data distributed by the NASA Socioeconomic Data and Applications Center (SEDAC).

Reich, B. J., Hodges, J. S., and Zadnik, V. (2006). Effects of residual smoothing on the posterior of the fixed effects in disease-mapping models. *Biometrics*, 62:1197–1206.

Rouget, M., Reyers, B., Jonas, Z., Desmet, P., Driver, A., Maze, K., egoh, B., and Cowling, R. M. (2004). *South African National Biodiversity Assessment 2004: Technical Report, Volume 1: Terrestrial Component*. South African National Biodiversity Institute, Pretoria, South Africa.

Royle, J. A. and Dorazio, R. M. (2008). *Hierarchical Modeling and Inference in Ecology*. Academic Press, 1 edition.

Royle, J. A. and Kéry, M. (2007). A Bayesian state-space formulation of dynamic occupancy models. *Ecology*, 88:1813–1823.

Royle, J. A., Kéry, M., Gautier, R., and Schmid, H. (2007). Hierarchical spatial models of abundance and occurrence from imperfect survey data. *Ecological Monographs*, 77(3):465–481.

Sargeant, G. A., Sovada, M. A., Silvinski, C. C., and Johnson, D. H. (2005). Markov chain Monte Carlo estimation of species distributions: A case study of the swift fox in western Kansas. *Journal of Wildlife Management*, 69(2):483–497.

Schaefer, S. M. and Krohn, W. B. (2002). Predicting vertebrate occurrences from species habitat associations: improving the interpretation of commission error rates. In Scott, J. M., Heglund, P. J., Morrison, M. L., Haufler, J. B., Raphael, M. G., Wall, W. A., and Samson, F. B., editors, *Predicting Species Occurrence: Issues of Accuracy and Scale*, pages 419–428. Island Press, Washington, D.C.

Scott, J. M., Heglund, P. J., Morrison, M. L., Haufler, J. B., Raphael, M. G., Wall, W. A., and Samson, F. B., editors (2002). *Predicting Species Occurrence: Issues of Accuracy and Scale*. Island Press, Washington, D.C.

Silberbauer, M. (2007). Rivers of south africa.

Sing, T., Sander, O., Beerenwinkel, N., and Lengauer, T. (2007). ROCR: Visualizing the performance of scoring classifiers. R package, version 1.0-4.

Theron, N. T. (2011). *Genetic connectivity, population dynamics and habitat selection of the Southern ground hornbill (B ucorvus leadbeateri) in the Limpopo Province*. PhD thesis, University of the Free State.

Tobalske, C. and Tobalske, B. W. (1999). Using atlas data to model the distribution of woodpecker species in the Jura, France. *The Condor*, 101:472–483.

Tyre, A. J., Tenhumberg, B., Field, S. A., Niejalke, D., Parris, K., and Possingham, H. P. (2003). Improving precision and reducing bias in biological surveys: estimating false-negative error rates. *Ecological Applications*, 13(1790–1801).

Vernon, C. J. (1986). The ground hornbill at the southern extremity of its range. *Ostrich: Journal of African Ornithology*, 57:16–24.

Wasser, S. K., Hayward, L. S., Hartman, J., Booth, R. K., Broms, K. M., Berg, J., Seely, E., Lewis, L., and Smith, H. (2012). Using detection dogs to conduct simultaneous surveys of northern spotted (strix occidentalis caurina) and barred owls (strix varia). *PLoS ONE*, 7(8):e42892.

Wintle, B. A. and Bardos, D. C. (2006). Modeling species-habitat relationships with spatially autocorrelated observation data. *Ecological Applications*, 16:1945–1958.

Wu, D., Liu, J., Zhang, G., Ding, W., Wang, W., and Wang, R. (2009). Incorporating spatial autocorrelation into cellula automata model: an application to the dynamics of Chinese tamarisk (Tamarix chinensis Lour.). *Ecological Modeling*, 220(24):3490–3498.

Wu, H. and Huffer, F. W. (1997). Modelling the distribution of plant species using the autologistic regression model. *Environmental and Ecological Statistics*, 4:49–64.

Yackulic, C. B., Reid, J., Davis, R., Hines, J. E., Nichols, J. D., and Forsman, E. (2012). Neighborhood and habitat effects on vital rates: expansion of the Barred Owl in the Oregon Coast Ranges. *Ecology*, 93(8):1953–1966.

Yackulic, C. B., Sanderson, E. W., and Uriarte, M. (2011). Anthropogenic and environmental drivers of modern range loss in large mammals. *Proceeings of the National Academy of Sciences*, 108:4024–4029.

Zabel, C. J., Roberts, L. M., Mulder, B. S., Stauffer, H. B., Dunk, J. R., Wolcott, K., Solis, D., Gertsch, M., Woodridge, B., Wright, A., Goldsmith, G., and Kreckler, C. (2002). A collaborative approach in adaptive management at a large-landscape scale. In Scott, J. M., Heglund, P. J., Morrison, M. L., Haufler, J. B., Raphael, M. G., Wall, W. A., and Samson, F. B., editors, *Predicting Species Occurrence: Issues of Accuracy and Scale*, pages 241–254. Island Press, Washington, D.C.

Appendix A

HORNBILL SUPPLEMENTARY MATERIAL

Figure A.1: The number of surveys per site, from the Southern African Bird Atlas Project. Only data from sites that were included in our analysis are shown. Note that the southernmost portion and an area in the north had no surveys and many other sites were surveyed only once.

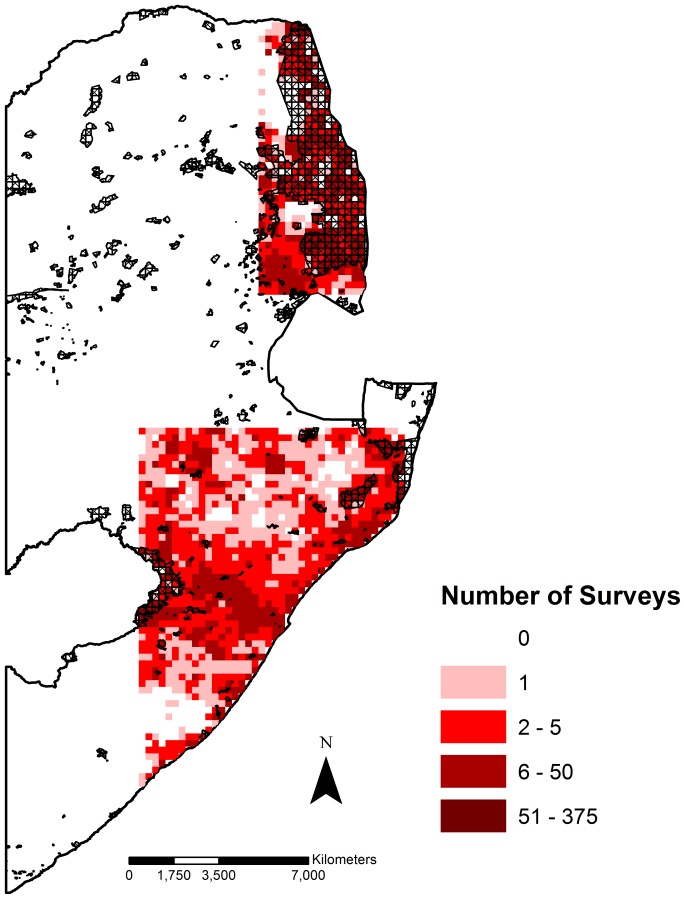


Figure A.2: The realized occupancy probabilities from the full Bayesian occupancy model. The realized occupancy probabilities take into account the detection histories of each site. If the species was detected at a site, then the realized occupancy probability equals one for that site. If the species was not detected, then the probability represents the fraction of the MCMC iterations where the model estimated that the site was occupied. The non-spatial occupancy model and the ICAR occupancy models from Table 1 produced nearly identical, maps. This map highlights the isolated detections in the southern region that the models did not predict.

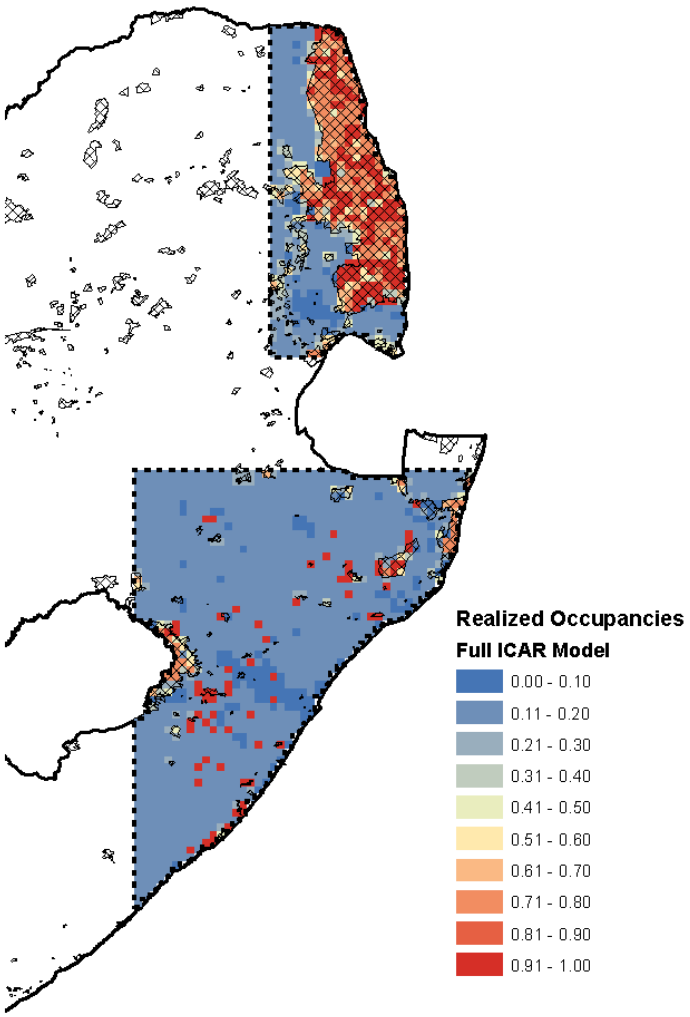


Figure A.3: The realized occupancy probabilities from the full RSR model, restricted to 160 columns. The realized occupancy probabilities take into account the detection histories of each site. Note that compared to Appendix Figure A2, there is much more variation in the predictions in the southern half of the data and that the realized occupancy probabilities exhibit raised values near previous detections.

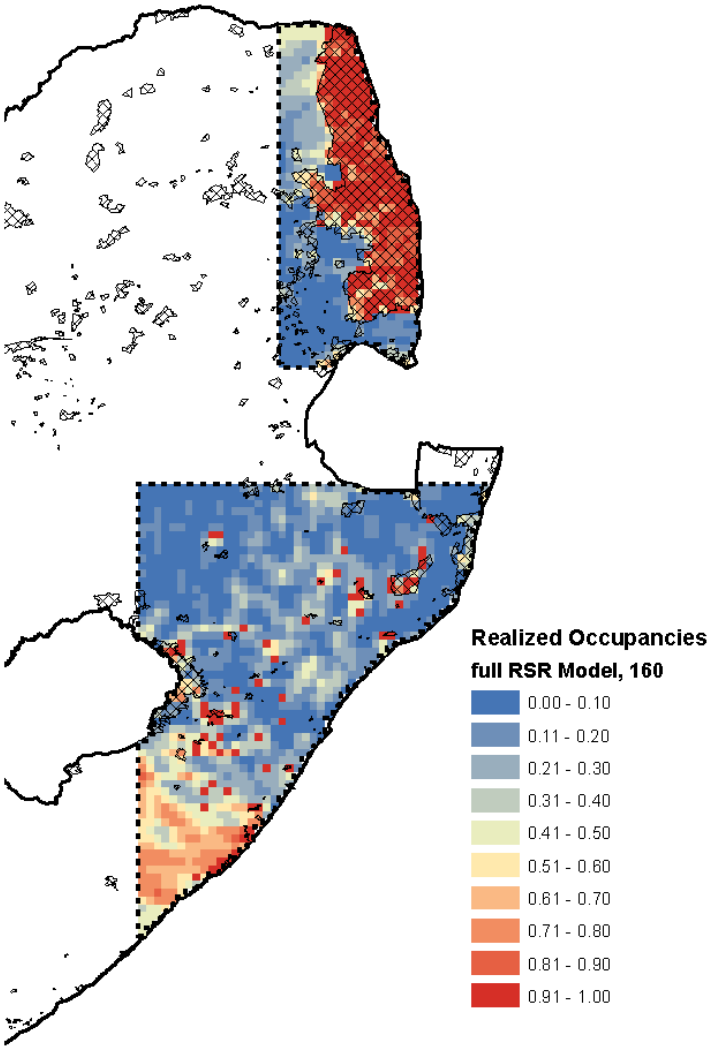


Table A.1: The predicted quantiles and median occupancy probabilities from each of the occupancy models fitted to the Southern ground hornbill data from South Africa. The standard errors associated with the medians are given in parentheses next to the estimate.

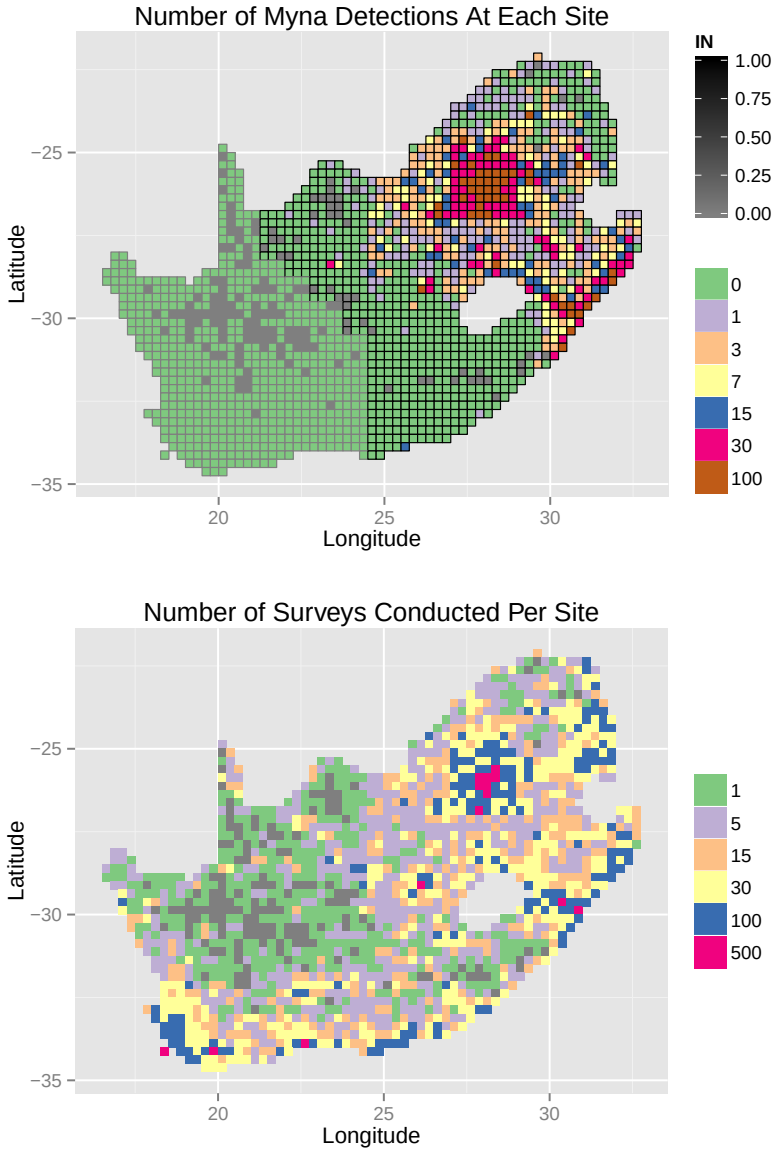
Model	Occupancy probability (SE)		
	25% quartile	50% median	75% quartile
Full nonspatial	0.18	0.18 (0.028)	0.24
Middle nonspatial	0.18	0.18 (0.028)	0.24
Limited nonspatial	0.12	0.12 (0.017)	0.19
Full ICAR	0.18	0.18 (0.029)	0.25
Middle ICAR	0.18	0.18 (0.027)	0.24
Limited ICAR	0.12	0.12 (0.017)	0.19
Full RSR-400	0.18	0.18 (0.028)	0.24
Middle RSR-400	0.18	0.18 (0.029)	0.25
Limited RSR-400	0.06	0.14 (0.19)	0.4
Full RSR-160	0.1	0.23 (0.18)	0.49
Middle RSR-160	0.18	0.18 (0.027)	0.24
Limited RSR-160	0.12	0.12 (0.016)	0.19

Appendix B

MYNA SUPPLEMENTARY MATERIAL

B.1 Myna SABAP 2 data.

Figure B.1: Top: Detections of the common myna. The sites that are included in our analysis are outlined in black. The myna was not detected in South Africa in the SABAP 2 data outside of this area. Bottom: The number of surveys that were conducted throughout South Africa.



B.2 Model equations.

Process model:

Year 1:

$$\begin{aligned} &\mathbf{z_1} \sim \text{Bernoulli}(\boldsymbol{\psi}) \\ &\text{probit}(\boldsymbol{\psi}) = \mathbf{X}_\psi \boldsymbol{\beta}_\psi + \mathbf{K}\boldsymbol{\alpha} \\ &\boldsymbol{\alpha} \sim \text{Normal}\left(\mathbf{0}, \sigma^2 \left(\mathbf{K}'\mathbf{Q}\mathbf{K}\right)^{-1}\right) \end{aligned} \tag{B.1}$$

Years 2 to T:

$$\begin{aligned} &z_{i,t} \mid \mathbf{z_{t-1}} \sim \text{Bernoulli}(\theta_{i,t}) \\ &\theta_{i,t} = z_{i,t-1}\phi + (1 - z_{i,t-1})\, I_{\mathcal{N}_{i,t-1}} \bar{d}_{i,t} + (1 - z_{i,t-1}) \left(1 - I_{\mathcal{N}_{i,t-1}}\right) \gamma \\ &\bar{d}_{i,t} = 1 - \exp\left(\tilde{\mathbf{z}}'_{\mathcal{N}_{\mathbf{i,t-1}}} \ln\left(\mathbf{1} - \mathbf{d_i}\right)\right) \end{aligned} \tag{B.2}$$

Observation model:

$$\begin{aligned} &y_{i,j,t} \sim \text{Bernoulli}(z_{i,t} \cdot p_{i,j,t}) \\ &\text{probit}(p_{i,j,t}) = \mathbf{X}_p \boldsymbol{\beta}_p \end{aligned} \tag{B.3}$$

Priors

$$\begin{aligned} &\boldsymbol{\beta}_\psi \sim \text{Normal}\left(\mathbf{0}, 5\mathbf{I}\right) \\ &\sigma \sim \text{Uniform}(0, 100) \\ &\boldsymbol{\beta}_p \sim \text{Normal}\left(\mathbf{0}, 5\mathbf{I}\right) \\ &\phi \sim \text{Uniform}(0, 1) \qquad (\ \phi = \text{persistence probability}\ ) \\ &\gamma \sim \text{Uniform}(0, 1) \qquad (\ \gamma = \text{out of neighborhood dispersal}\ ) \end{aligned} \tag{B.4}$$

Homogeneous (Neighborhood) Dispersal:

$$\begin{aligned}&\text{for(k in 1:9)\{}\\&\quad d_{i,k} \sim \text{Uniform}(0,1)\\&\}\end{aligned} \tag{B.5}$$

Nonhomogeneous Neighborhood Dispersal:

$$\begin{aligned}&\text{probit}(\mathbf{d_i}) = \beta_{\mathbf{d},0} + \beta_{\mathbf{d},1}\mathbf{x}_{\mathbf{d},i}\\&x_{\mathbf{d},i,k} = \tfrac{x_{i,k}-x_i}{\text{dist}(i,k)}\\&\beta_{\mathbf{d},0} \sim \text{Normal}\,(0,5)\\&\beta_{\mathbf{d},1} \sim \text{Normal}\,(0,5)\end{aligned} \tag{B.6}$$

B.3 List of symbols and their definitions.

$\mathbf{z_1}$	The true occurrences of the myna in year 1 for sites $i = 1, \ldots, n$.
$\boldsymbol{\psi}$	The probabilities of occupancy in year 1.
$\mathbf{X}_\psi$	The site-specific covariates that affect occupancy probabilities.
$\boldsymbol{\beta}_\psi$	The coefficients that determine how site-specific covariates affect occupancy.
$\mathbf{K}\boldsymbol{\alpha}$	The spatial random effects from an RSR model.
$\mathbf{K}$	The eigenvectors from the Moran Operator Matrix.
$\mathbf{Q}$	The precision matrix from an ICAR model.
$z_{i,t}$	The occurence of the myna in year t at site i.
$\theta_{i,t}$	The probability that site i is occupied in year $t = 2, \ldots, T$.
ϕ	The persistence (site-survival) probability.
$\bar{d}_{i,t}$	Probability of site i being colonized by its neighbors at time t. Alternatively, it may be called the neighborhood dispersal probability.
γ	Long-distance/ out-of-neighborhood dispersal probability.
$I_{\mathcal{N}_{i,t-1}}$	An indicator variable that equals 1 if any neighbor of site i is occupied at time $t-1$ and equals 0 otherwise.
$\tilde{\mathbf{z}}_{\mathcal{N}_{\mathbf{i,t-1}}}$	A vector of length 9 that indicates which neighbors of site i are occupied at time $t-1$.
$\mathbf{d_i}$	A dispersal vector of length 9. Each element represents the probability of site i being colonized by neighbor k, $k = 1, \ldots, 9$.
$\mathbf{x}_{\mathbf{d},i}$	Used in the nonhomogeneous model. The habitat differences that influence the dispersal to site i from its neighbors.
$\boldsymbol{\beta}_\mathbf{d}$	Used in the nonhomogeneous model. The coefficients that determine how the habitat differences affect dispersal.
$y_{i,j,t}$	The observed occurrences of the species at site i on survey j at time period t.
$p_{i,j,t}$	Detection probability for survey j of site i at time t.
$\mathbf{X}_p$	The covariates that affect detection probabilities.
$\boldsymbol{\beta}_p$	The coefficients that determine how the covariates affect detection.

B.4 Derivation of the neighborhood dispersal vector.

$$
\begin{aligned}
\bar{d}_{i,t} &= \text{Prob(being colonized)} \\
&= 1 - \text{Prob(not being colonized)} \\
&= 1 - (1-d_1)^{z_1} \ldots (1-d_9)^{z_9} \\
&= 1 - \left(e^{\ln(1-d_1)}\right)_1^z \ldots \left(e^{\ln(1-d_9)}\right)^{z_9} \\
&= 1 - e^{\sum z_i \ln(1-d_i)} \\
&= 1 - exp\left(\tilde{\mathbf{z}}'_{\mathcal{N}_{\mathbf{i,t-1}}} \log\left(\mathbf{1} - \mathbf{d_i}\right)\right)
\end{aligned}
\tag{B.7}
$$

B.5 Simulation study results.

Table B.1: Parameter estimates, true values, and relative biases from the simulation study. Relative bias = (Estimated - True) / True. These simulations are from the nonhomogenous neighborhood dispersal model with constant long distance dispersal. The "Available" columns indicate situations where year 1 was available to be estimated; the "Estimated" columns refer to situations where the true year 1 was excluded.

		Available			Estimated		
Parameter		**Est.**	**True**	**Bias**	**Est.**	**True**	**Bias**
Detection prob.	p	0.50	0.5	0.00	0.50	0.5	0.00
Persistance prob.	ϕ	0.90	0.9	0.00	0.90	0.9	0.00
Long dist. dispersal	γ	0.05	0.05	-0.09	0.20	0.05	2.95
Local dispersal, B1	$\beta_{\gamma,0}$	-0.99	-1	-0.01	-1.05	-1	0.05
Local dispersal, B2	$\beta_{\gamma,1}$	0.75	1	-0.25	1.02	1	0.02
Sites Occupied, Yr 0	$\sum z_{i,0}$	NA	NA	NA	270	273	-0.01
Sites Occupied, Yr 1	$\sum z_{i,1}$	272.5	276.5	-0.01	423	414	0.02
Sites Occupied, Yr 2	$\sum z_{i,2}$	409.5	417	-0.02	550	543	0.01
Sites Occupied, Yr 3	$\sum z_{i,3}$	541.5	547	-0.01	648	636.5	0.02
Sites Occupied, Yr 4	$\sum z_{i,4}$	638.5	642	-0.01	717	708.5	0.01
Sites Occupied, Yr 5	$\sum z_{i,5}$	702.5	708.5	-0.01	752	750	0.00
deviance		6794.3			7616.56		
DIC		8720.9			9710.80		

Table B.2: Parameter estimates, true values, and relative biases from the simulation study. Relative bias = (Estimated - True) / True. These simulations are from the non-homogenous neighborhood dispersal model with varying long distance dispersal. The "Available" columns indicate situations where year 1 was available to be estimated; the "Estimated" columns refer to situations where the true year 1 was excluded.

		Available			Estimated		
Parameter		**Est.**	**True**	**Bias**	**Est.**	**True**	**Bias**
Detection prob.	p	0.50	0.5	0.00	0.50	0.5	-0.01
Persistance prob.	ϕ	0.90	0.9	0.00	0.90	0.9	0.00
Long dist. dispersal	$\beta_{\gamma,0}$	-1.50	-1.5	0.00	-0.07	-1.5	-0.95
	$\beta_{\gamma,1}$	0.92	0.5	0.84	1.53	0.5	2.06
Local dispersal, B1	$\beta_{d,0}$	-1.02	-1	0.02	-0.97	-1	-0.03
Local dispersal, B2	$\beta_{d,1}$	1.11	1	0.11	0.84	1	-0.16
Sites Occupied, Yr 0	$\sum z_{i,0}$	NA	NA	NA	203	284	-0.29
Sites Occupied, Yr 1	$\sum z_{i,1}$	282.5	276	0.02	418	420	0.00
Sites Occupied, Yr 2	$\sum z_{i,2}$	420	419.5	0.00	542	540.5	0.00
Sites Occupied, Yr 3	$\sum z_{i,3}$	537	534.5	0.00	639	628.5	0.02
Sites Occupied, Yr 4	$\sum z_{i,4}$	628	626	0.00	692	693.5	0.00
Sites Occupied, Yr 5	$\sum z_{i,5}$	692	693.5	0.00	747	746	0.00
deviance		6742.4			7545.6		
DIC		8654.4			9804.5		

Table B.3: Parameter estimates, true values, and relative biases from the simulation study. Relative bias = (Estimated - True) / True. These simulations are from the homogenous neighborhood dispersal model with constant long distance dispersal. The "Available" columns indicate situations where year 1 was available to be estimated; the "Estimated" columns refer to situations where the true year 1 was excluded.

		Available			Estimated		
Parameter		**Est.**	**True**	**Bias**	**Est.**	**True**	**Bias**
Detection prob.	p	0.50	0.5	-0.01	0.51	0.5	0.01
Persistance prob.	ϕ	0.90	0.9	0.00	0.90	0.9	0.00
Long distance dispersal	γ	0.06	0.05	0.20	0.09	0.05	0.74
dispersal[1]	d_1	0.19	0.2	-0.07	0.20	0.2	0.00
dispersal[2]	d_2	0.19	0.2	-0.03	0.17	0.2	-0.16
dispersal[3]	d_3	0.04	0.05	-0.25	0.05	0.05	-0.02
dispersal[4]	d_4	0.19	0.2	-0.06	0.19	0.2	-0.03
dispersal[5]	d_5	NA	NA	NA	NA	NA	NA
dispersal[6]	d_6	0.05	0.05	0.08	0.06	0.05	0.27
dispersal[7]	d_7	0.05	0.05	-0.07	0.05	0.05	0.07
dispersal[8]	d_8	0.06	0.05	0.14	0.08	0.05	0.64
dispersal[9]	d_9	0.07	0.05	0.30	0.06	0.05	0.17
Sites Occupied, Yr 0	$\sum z_{i,0}$	NA	NA	NA	249	274	-0.09
Sites Occupied, Yr 1	$\sum z_{i,1}$	279.5	281.5	-0.01	363	359.5	0.01
Sites Occupied, Yr 2	$\sum z_{i,2}$	368	369.5	0.00	457	447	0.02
Sites Occupied, Yr 3	$\sum z_{i,3}$	450	448.5	0.00	536	525	0.02
Sites Occupied, Yr 4	$\sum z_{i,4}$	523	523	0.00	592	584	0.01
Sites Occupied, Yr 5	$\sum z_{i,5}$	587	582	0.01	638	634.5	0.01
deviance		5948.8			6566.9		
DIC		7544.6			8522.9		

Table B.4: Parameter estimates, true values, and relative biases from the simulation study. Relative bias = (Estimated - True) / True. These simulations are from the homogenous neighborhood dispersal model with varying long distance dispersal. The "Available" columns indicate situations where year 1 was available to be estimated; the "Estimated" columns refer to situations where the true year 1 was excluded.

		Available			Estimated		
Parameter		**Est.**	**True**	**Bias**	**Est.**	**True**	**Bias**
Detection prob.	p	0.49	0.5	-0.01	0.50	0.5	0.00
Persistance prob.	ϕ	0.89	0.9	-0.01	0.90	0.9	0.00
Long dist. dispersal	$\beta_{\gamma,0}$	-1.89	-1.5	0.26	-0.37	-1.5	-0.76
	$\beta_{\gamma,1}$	0.48	0.5	-0.04	1.16	0.5	1.33
dispersal[1]	d_1	0.18	0.2	-0.09	0.18	0.2	-0.09
dispersal[2]	d_2	0.21	0.2	0.05	0.22	0.2	0.08
dispersal[3]	d_3	0.06	0.05	0.28	0.05	0.05	0.00
dispersal[4]	d_4	0.22	0.2	0.09	0.20	0.2	-0.02
dispersal[5]	d_5	NA	NA	NA	NA	NA	NA
dispersal[6]	d_6	0.05	0.05	-0.05	0.07	0.05	0.45
dispersal[7]	d_7	0.05	0.05	-0.10	0.06	0.05	0.13
dispersal[8]	d_8	0.05	0.05	-0.01	0.05	0.05	-0.07
dispersal[9]	d_9	0.04	0.05	-0.14	0.05	0.05	-0.03
Sites Occupied, Yr 0	$\sum z_{i,0}$	NA	NA	NA	141	277.5	-0.49
Sites Occupied, Yr 1	$\sum z_{i,1}$	272	272.5	0.00	370	375.5	-0.01
Sites Occupied, Yr 2	$\sum z_{i,2}$	358.5	360	0.00	448	453	-0.01
Sites Occupied, Yr 3	$\sum z_{i,3}$	437.5	437.5	0.00	514	516	0.00
Sites Occupied, Yr 4	$\sum z_{i,4}$	513.5	522	-0.02	579	566.5	0.02
Sites Occupied, Yr 5	$\sum z_{i,5}$	574.5	566.5	0.01	628	609	0.03
deviance		5805.7			6460.0		
DIC		7329.1			8396.7		

B.6 Parameter estimates from model variants of the myna analysis.

Table B.5: The estimated coefficient values from the nonhomogeneous, spatio-temporal model, with latitude acting as the habitat gradient, as aplied to the myna data.

NONHOMOGENEOUS DISPERSAL MODEL, F(LATITUDE)					
	Median	**(SE)**	**95 % CI**		**Rhat**
Detection Parameters (probit-scale):					
Intercept	-0.09	(0.02)	-0.12	-0.06	1.00
INTENSIVE	0.11	(0.01)	0.09	0.12	1.00
LAT	0.31	(0.02)	0.28	0.35	1.00
LAT2	-0.47	(0.02)	-0.51	-0.44	1.00
HUMAN_POP	0.57	(0.01)	0.54	0.59	1.00
LONG	0.03	(0.02)	0.00	0.07	1.00
LONG2	-0.32	(0.01)	-0.35	-0.30	1.00
Occupancy Parameters (probit-scale):					
Intercept	0.04	(0.23)	-0.41	0.54	1.01
LAT	2.20	(0.51)	1.49	3.44	1.03
HUMAN_POP	0.83	(0.23)	0.45	1.34	1.01
Spatial Parameter:					
sigma	5.57	(1.53)	3.69	9.71	1.06
Persistance Probability:					
persist.prob	0.94	(0.01)	0.92	0.95	1.00
Long-distance dispersal probability:					
disperse.long	0.01	(0.02)	0.00	0.06	1.00
Neighborhood Dispersal Parameters (logit-scale):					
Intercept	-1.64	(0.14)	-2.01	-1.44	1.00
LAT	-1.75	(0.66)	-3.29	-0.69	1.00
Number of Sites Occupied:					
Year 2008	638	(21.6)	596	680	1.01
Year 2009	655	(17.9)	620	690	1.00
Year 2010	672	(16.8)	640	705	1.00
Year 2011	689	(17.8)	655	725	1.00
Year 2012	709	(20.3)	671	751	1.00
Deviance Explained:					
deviance	31643.65	(43.5)	31559.216	31730.769	1.00
DIC	32592.2				

Table B.6: The estimated coefficient values from the nonhomogeneous, spatio-temporal model, with longitude acting as the habitat gradient, as applied to the myna data.

NONHOMOGENEOUS DISPERSAL MODEL, F(LONGITUDE)					
	Median	**(SE)**	**95 % CI**		**Rhat**
Detection Parameters (probit-scale):					
Intercept	-0.09	(0.02)	-0.12	-0.05	1.00
INTENSIVE	0.11	(0.01)	0.09	0.12	1.00
LAT	0.32	(0.02)	0.28	0.35	1.00
LAT2	-0.47	(0.02)	-0.51	-0.44	1.00
HUMAN_POP	0.57	(0.01)	0.54	0.59	1.00
LONG	0.03	(0.02)	0.00	0.07	1.00
LONG2	-0.33	(0.01)	-0.35	-0.30	1.00
Occupancy Parameters (probit-scale):					
Intercept	-0.10	(0.21)	-0.53	0.33	1.01
LAT	1.76	(0.39)	1.19	2.67	1.01
HUMAN_POP	0.84	(0.22)	0.49	1.36	1.01
Spatial Parameter:					
sigma	4.78	(1.17)	3.11	7.74	1.01
Persistance Probability:					
persist.prob	0.93	(0.01)	0.92	0.95	1.00
Long-distance dispersal probability:					
disperse.long	0.01	(0.01)	0.00	0.05	1.00
Neighborhood Dispersal Parameters (logit-scale):					
Intercept	-1.33	(0.06)	-1.45	-1.20	1.00
LONG	-1.42	(0.31)	-2.02	-0.80	1.00
Number of Sites Occupied:					
Year 2008	619	(26.2)	569	670	1.01
Year 2009	651	(20.2)	611	689	1.01
Year 2010	677	(18.1)	641	713	1.00
Year 2011	702	(18.1)	667	737	1.00
Year 2012	730	(19.6)	693	770	1.00
Deviance Explained:					
deviance	31650.8	(42.5)	31568.4	31735.8	1.00
DIC	32555.0				

Table B.7: The estimated coefficient values from the nonhomogeneous model, with long distance dispersal varying as a function of human density.

NONHOMOGENEOUS DISPERSAL, LONG-DISTANCE = F(POP)					
	Median	**(SE)**	**95 % CI**		**Rhat**
Detection Parameters (probit-scale):					
Intercept	-0.09	(0.02)	-0.12	-0.05	1.00
INTENSIVE	0.11	(0.01)	0.09	0.12	1.00
LAT	0.32	(0.02)	0.28	0.35	1.00
LAT2	-0.48	(0.02)	-0.51	-0.44	1.00
HUMAN_POP	0.57	(0.01)	0.54	0.59	1.00
LONG	0.03	(0.02)	-0.01	0.07	1.00
LONG2	-0.32	(0.01)	-0.35	-0.30	1.00
Occupancy Parameters (probit-scale):					
Intercept	0.02	(0.21)	-0.38	0.46	1.01
LAT	2.14	(0.41)	1.52	3.13	1.01
HUMAN_POP	0.87	(0.21)	0.51	1.32	1.02
Spatial Parameter:					
sigma	5.02	(1.16)	3.40	7.76	1.01
Persistance Probability:					
persist.prob	0.94	(0.01)	0.92	0.95	1.00
Long-distance Dispersal Parameters (probit-scale:)					
Intercept	-3.69	(1.13)	-6.51	-2.16	1.00
HUMAN_POP	-0.12	(1.07)	-2.32	1.96	1.01
Neighborhood Dispersal Parameters (probit-scale):					
Intercept	-1.06	(0.13)	-1.31	-0.81	1.00
HUMAN_POP	-0.52	(0.16)	-0.84	-0.21	1.00
Number of Sites Occupied:					
Year 2008	636	(22.2)	591	678	1.02
Year 2009	653	(18.6)	616	689	1.02
Year 2010	673	(17.6)	640	708	1.01
Year 2011	695	(18.6)	660	732	1.01
Year 2012	720	(21.1)	681	764	1.01
Deviance Explained:					
deviance	31655.4	(42.3)	31575.4	31740.6	1.00
DIC	32546.3				

Table B.8: The estimated coefficient values from the homogeneous model, with long distance dispersal varying as a function of human density.

HOMOGENEOUS DISPERSAL, LONG-DISTANCE = F(POP)					
	Median	**(SE)**	**95 % CI**		**Rhat**
Detection Parameters (probit-scale):					
Intercept	-0.09	(0.02)	-0.12	-0.05	1.00
INTENSIVE	0.11	(0.01)	0.09	0.12	1.00
LAT	0.31	(0.02)	0.28	0.34	1.00
LAT2	-0.47	(0.02)	-0.50	-0.43	1.00
HUMAN_POP	0.57	(0.01)	0.54	0.59	1.00
LONG	0.03	(0.02)	0.00	0.07	1.00
LONG2	-0.32	(0.01)	-0.35	-0.30	1.00
Occupancy Parameters (probit-scale):					
Intercept	-0.24	(0.18)	-0.60	0.12	1.02
LAT	1.69	(0.34)	1.15	2.48	1.01
HUMAN_POP	0.84	(0.18)	0.53	1.22	1.01
Spatial Parameter:					
sigma	4.57	(1.03)	3.12	7.05	1.01
Persistance Probability:					
persist.prob	0.93	(0.01)	0.91	0.94	1.00
Long-distance Dispersal Parameters (probit-scale:)					
disperse.long	-3.780	(1.15)	-6.552	-2.197	1.00
HUMAN_POP	0.32	(0.77)	-1.23	1.73	1.01
Number of Sites Occupied:					
Year 2008	595	(26.5)	543	645	1.01
Year 2009	638	(19.0)	602	676	1.01
Year 2010	666	(16.9)	633	699	1.01
Year 2011	691	(17.2)	659	726	1.01
Year 2012	721	(19.2)	685	760	1.00
Deviance Explained:					
deviance	31605.94	43.37	31522.96	31693.69	1.00
DIC	32547.25				

Table B.9: The estimated dispersal probabilities (and standard errors) from the homogeneous model with long distance dispersal varying as a function of human density. These dispersal probabilities were estimated along with the parameters of Table B.8.

0.08 (0.04)	0.02 (0.02)	0.01 (0.01)
0.22 (0.06)	-	0.17 (0.06)
0.06 (0.05)	0.08 (0.06)	0.56 (0.16)

Figure B.2: The standard deviations associated with each occupancy probability estimate.

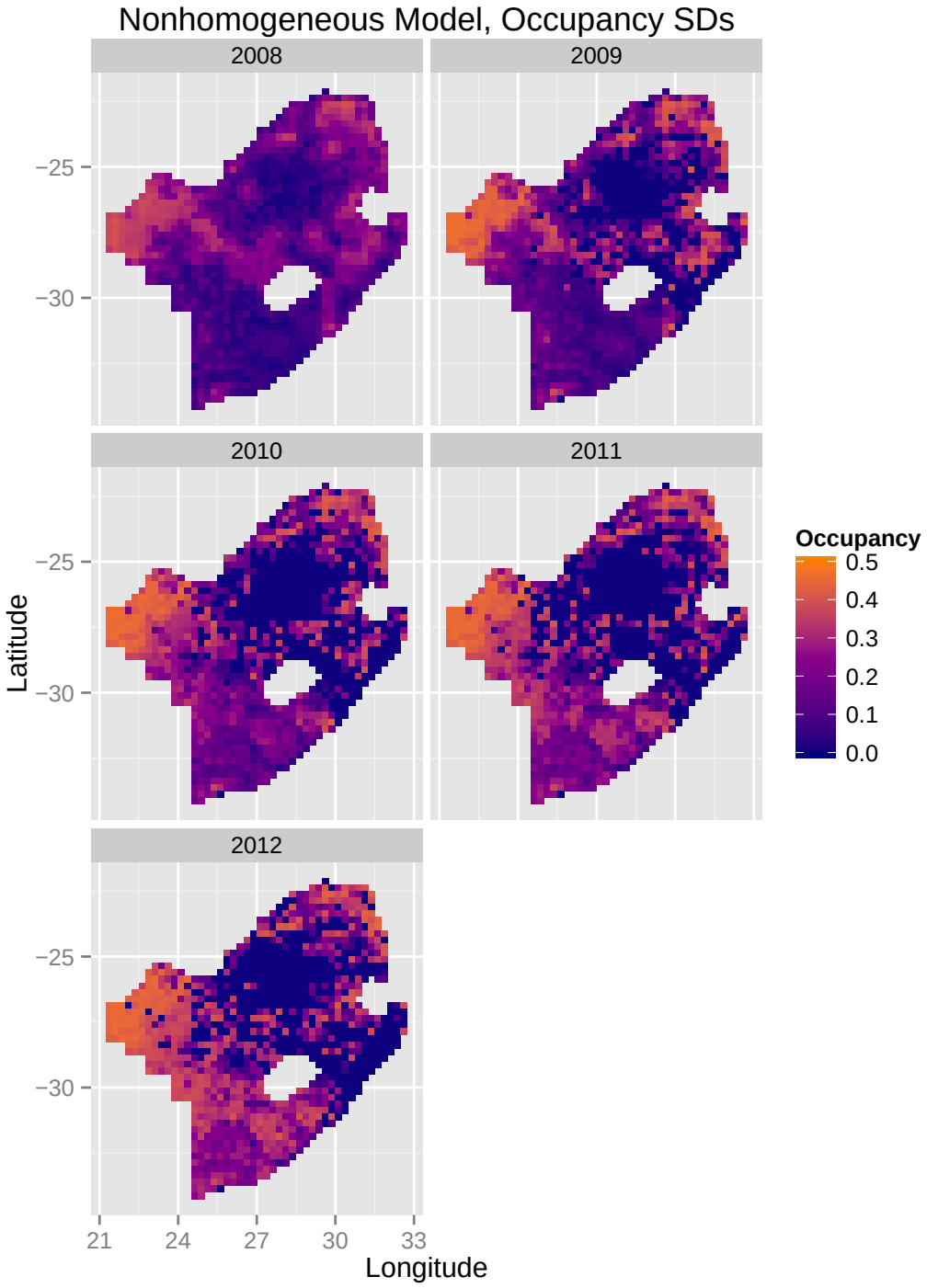


Figure B.3: The standard deviations associated with the occurrence prediction at each site.