Data fusion and spatio-temporal approaches to model species distribution

by

Narmadha Meenabhashini Mohankumar

B.S., University of Peradeniya, Sri Lanka, 2015

AN ABSTRACT OF A DISSERTATION

submitted in partial fulfillment of the
requirements for the degree

DOCTOR OF PHILOSOPHY

Department of Statistics
College of Arts and Sciences

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2022

Abstract

Species distribution models (SDMs) are increasingly used in ecology, biogeography, and wildlife management to learn about the distribution of species across space and time. Determining the species-habitat relationships and the distributional pattern of a species is important to increase scientific knowledge, inform management decisions, and conserve biodiversity. I propose approaches to address some of the most pressing issues encountered in studies of species distributions and contribute towards improving predictions and inferences from SDMs.

First, I present a modeling framework to model occupancy data that accounts for both traditional and nontraditional spatial dependence as well as false absences. Occupancy data are used to estimate and map the true presence of a species, which may depend on biotic and abiotic factors as well as spatial autocorrelation. Traditionally, spatial autocorrelation is accounted for by using a correlated normally distributed site-level random effect, which might be incapable of modeling nontraditional spatial dependence such as discontinuities and abrupt transitions. Machine learning approaches have the potential to model nontraditional spatial dependence, but these approaches do not account for observer errors such as false absences. I combine the flexibility of Bayesian hierarchal modeling and machine learning approaches and present a modeling framework to account for both traditional and nontraditional spatial dependence and false absences. I illustrate the framework using six synthetic data sets containing traditional and nontraditional spatial dependence and then apply the approach to understand the spatial distribution of Thomson's gazelle (*Eudorcas thomsonii*) in Tanzania and sugar gliders (*Petaurus breviceps*) in Tasmania.

Second, I develop a model-based approach for data fusion of distance sampling (DS) and capture-recapture (CR) data. DS and CR are two widely collected data types to learn about species-habitat relationships and abundance; still, they are seldomly used in SDMs due to

the lack of spatial coverage. However, data fusion of the sources of data can increase spatial coverage, which can reduce parameter uncertainty and make predictions more accurate, and therefore, can be used for species distribution modeling. My modeling approach accounts for two common missing data issues: 1) missing individuals that are missing not at random (MNAR) and 2) partially missing location information. Using a simulation experiment, I evaluated the performance of the modeling approach and compared it to existing approaches that use ad-hoc methods to account for missing data issues. I demonstrated my approach using data collected for Grasshopper Sparrows (*Ammodramus savannarum*) in north-eastern Kansas, USA.

Third, I extend my data fusion approach to a spatio-temporal modeling framework to investigate the influence of the temporal support in spatio-temporal point process models to model species distribution. Temporal dynamics of ecological processes are complex, and their influence on species-habitat relationships and abundance operate in multiple spatio-temporal scales. Spatio-temporal point process models are widely used to model species-habitat relationships and estimate abundance across multiple spatio-temporal scales; however, the robustness of the models to changing temporal scales is rarely studied. Understanding the temporal dynamics of ecological processes across the entirety of spatio-temporal scales is key to learning about species' distribution. Therefore, investigating the influence of temporal support on the robustness of spatio-temporal point processes to model species distributions is needed. In my approach, I combine DS and CR data in a spatio-temporal point process modeling framework and investigate the robustness of the model to changing temporal scales. My fused data spatio-temporal model alleviates constraints in individual data sources such as lack of spatio-temporal coverage and enables the study of complex phenomena on multiple-scale species-habitat relationships and abundance. To investigate the impact of temporal support on models' robustness, I conducted a simulation experiment. Then, I illustrate the influence of temporal support to model species-habitat relationships and abundance using data on Grasshopper Sparrows (*Ammodramus savannarum*) in north-eastern Kansas, USA.

Data fusion and spatio-temporal approaches to model species distribution

by

Narmadha Meenabhashini Mohankumar

B.S., University of Peradeniya, Sri Lanka, 2015

A DISSERTATION

submitted in partial fulfillment of the
requirements for the degree

DOCTOR OF PHILOSOPHY

Department of Statistics
College of Arts and Sciences

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2022

Approved by:

Major Professor
Dr. Trevor Hefley

Copyright

Abstract

Species distribution models (SDMs) are increasingly used in ecology, biogeography, and wildlife management to learn about the distribution of species across space and time. Determining the species-habitat relationships and the distributional pattern of a species is important to increase scientific knowledge, inform management decisions, and conserve biodiversity. I propose approaches to address some of the most pressing issues encountered in studies of species distributions and contribute towards improving predictions and inferences from SDMs.

First, I present a modeling framework to model occupancy data that accounts for both traditional and nontraditional spatial dependence as well as false absences. Occupancy data are used to estimate and map the true presence of a species, which may depend on biotic and abiotic factors as well as spatial autocorrelation. Traditionally, spatial autocorrelation is accounted for by using a correlated normally distributed site-level random effect, which might be incapable of modeling nontraditional spatial dependence such as discontinuities and abrupt transitions. Machine learning approaches have the potential to model nontraditional spatial dependence, but these approaches do not account for observer errors such as false absences. I combine the flexibility of Bayesian hierarchal modeling and machine learning approaches and present a modeling framework to account for both traditional and nontraditional spatial dependence and false absences. I illustrate the framework using six synthetic data sets containing traditional and nontraditional spatial dependence and then apply the approach to understand the spatial distribution of Thomson's gazelle (*Eudorcas thomsonii*) in Tanzania and sugar gliders (*Petaurus breviceps*) in Tasmania.

Second, I develop a model-based approach for data fusion of distance sampling (DS) and capture-recapture (CR) data. DS and CR are two widely collected data types to learn about species-habitat relationships and abundance; still, they are seldomly used in SDMs due to

the lack of spatial coverage. However, data fusion of the sources of data can increase spatial coverage, which can reduce parameter uncertainty and make predictions more accurate, and therefore, can be used for species distribution modeling. My modeling approach accounts for two common missing data issues: 1) missing individuals that are missing not at random (MNAR) and 2) partially missing location information. Using a simulation experiment, I evaluated the performance of the modeling approach and compared it to existing approaches that use ad-hoc methods to account for missing data issues. I demonstrated my approach using data collected for Grasshopper Sparrows (*Ammodramus savannarum*) in north-eastern Kansas, USA.

Third, I extend my data fusion approach to a spatio-temporal modeling framework to investigate the influence of the temporal support in spatio-temporal point process models to model species distribution. Temporal dynamics of ecological processes are complex, and their influence on species-habitat relationships and abundance operate in multiple spatio-temporal scales. Spatio-temporal point process models are widely used to model species-habitat relationships and estimate abundance across multiple spatio-temporal scales; however, the robustness of the models to changing temporal scales is rarely studied. Understanding the temporal dynamics of ecological processes across the entirety of spatio-temporal scales is key to learning about species' distribution. Therefore, investigating the influence of temporal support on the robustness of spatio-temporal point processes to model species distributions is needed. In my approach, I combine DS and CR data in a spatio-temporal point process modeling framework and investigate the robustness of the model to changing temporal scales. My fused data spatio-temporal model alleviates constraints in individual data sources such as lack of spatio-temporal coverage and enables the study of complex phenomena on multiple-scale species-habitat relationships and abundance. To investigate the impact of temporal support on models' robustness, I conducted a simulation experiment. Then, I illustrate the influence of temporal support to model species-habitat relationships and abundance using data on Grasshopper Sparrows (*Ammodramus savannarum*) in north-eastern Kansas, USA.

Table of Contents

List of Figures

List of Tables

Acknowledgments

Reaching the finish line of the Ph.D. journey is not possible without the support and encouragement from many people. I want to make use of this opportunity to express my sincere gratitude to those who supported me.

First and foremost, my deepest gratitude goes to my major professor, Dr. Trevor Hefley, for seeing potential in me and continuously encouraging me to reach my highest potential. Your wealth of knowledge and your unassuming approach to research and science are inspiring and from which I greatly benefited. I deeply appreciate every moment you have spent advising, teaching, correcting, and helping me to bring my Ph.D. dissertation to successful completion. I am extremely thankful to my supervisory committee, Dr. Nora Bello, Dr. Perla Reyes, and Dr. Alice Boyle, for their noble guidance, advice, collaboration, and encouragement at every stage of my dissertation research. I would also like to extend my gratitude to Dr. Andres Patrignani, with whom I enjoyed working on my first collaborative research. Thank you for the opportunity and your mentorship.

I extend my sincere thanks to fellow graduate students in Dr. Hefley's research group, Dr. Nelson Walker, Haoyu Zhang, Congxing Zhu, Liying Jin, Aidan Kerns, Camille Rieber, and Andrew Whetten, for being a great source of intellect and creativity that promoted my growth as a professional and a researcher. Many thanks go to department colleagues Hilda Calderon, Linus Addae, and Steephanson Anthonymuthu for providing invaluable support and encouragement to get through each milestone in my graduate education. My Ph.D. dissertation wouldn't be a reality without the collaboration, and valuable subject matter expertise from fellow students in Dr. Boyle's research lab at the Division of Biology, especially Katy Silber, who provided continuous support and contribution to my dissertation research. Thank you for the opportunity to learn and understand the subject matter and further my scientific communication. Further, I thank Dr. Patrignani's research group at the Department of Agronomy for continuously supporting, collaborating, and encouraging me during and after my time working with them.

I'm deeply indebted to Kansas State University and the Department of Statistics for giving me the opportunity to pursue my Ph.D. degree. My sincere gratitude goes to the past and current Department Heads, Dr. Gary Gadbury, Dr. James Neil, and Dr. Christopher Vahl, for providing me with a financial assistantship; without that, pursuing a Ph.D. degree would not have been possible. I would also like to extend my gratitude to the entire faculty and staff at the Department of Statistics for their support, instruction, and encouragement and creating such a warm and welcoming environment that provided me with the best graduate school experience.

I want to acknowledge the research funding from National Science Foundation Division of Environmental Biology under Grant No. 1754491. I'm indebted for the scholarship awards that I received; Howard Siepman Memorial Scholarship, Lynn Ying-Shiang Statistics Graduate Research Scholarship, and Lolafaye Coyne Statistics Graduate Scholarship Award from the Department of Statistics and College of Arts and Sciences; Bluemont Grant, and K-State Proud Scholarship Award from the Office of Student Financial Assistance and K-State Proud; International Student and Scholar Services Scholarship Award from the International Student and Scholar Services; K-State Alumni Association International Student Scholarship Award from the K-State Alumni Association, and travel award scholarships to present my research work at various conferences from K-State Graduate School and K-State College of Arts and Sciences. I'm honored and grateful to have received the leadership awards, the International Leadership Award from the K-State Alumni Association, and the Outstanding Student Member of the Year Award from the Center for Student Involvement & Blue key Honor Society at Kansas State University.

The dissertation alone does not reflect the personal and professional growth during my Ph.D. program; it is also from various on-campus opportunities. I want to thank the Stat Club, Graduate Student Council, Student Governing Association, International Coordinating Council, Sri Lankan Student Association, and Rotaract Club for providing tremendous opportunities for leadership and professional growth and the opportunity to serve the K-State community.

Pursuing a Ph.D. was a journey that required more than the past couple of years of support and motivation. I thank all my teachers in my school, St. Gabriel's Girls College, Hatton, Sri Lanka, for the positivity and encouragement during the early stages of my life that gave me the strength and motivation to pursue my dreams. A special thanks to my teacher and mentor, late Mr. Padmaraj, who was a massive motivation for me to reach my goals. I also thank all the professors and mentors at the Faculty of Science, University of Peradeniya, Sri Lanka, where I completed my undergraduate degree, for all the support and mentorship that led me to pursue a Ph.D. degree.

I cannot stress enough how fortunate I am to have some amazing friends without whom the Ph.D. journey would have been more difficult. Many thanks to Dishan Nahitiya, Prathibha Desman, and Ayyappan Elangovan for being my support system during ups and downs. Words are not enough to express my gratitude to all my friends who stood by me, encouraging and cheering for me. Thank you all for beautiful memories that I will treasure and cherish forever. A special thanks go to the people involved with Bridges International and HIS for creating such a welcoming environment to connect with the Manhattan community, which became my family during the past couple of years. I also thank all my relatives and cousins, especially my aunt Manju Potuhera for the love, support, and cheer throughout this time.

Finally, my deep eternal gratitude goes to my loving parents, Mr. Ramasamypillai Mohankumar and Mrs. Nirmala Mohankumar, and my brother Thiraj Mohankumar for believing in me and helping me to shape my life with positivity and passion. Words are insufficient to thank you for your endless sacrifices, unconditional love, and prayers towards me. Everything I have and everything I am, I owe it all to you.

Dedication

To my parents and brother,

for their endless love, support, and encouragement.

Introduction

The motivations to understand species distribution patterns in nature have a long history dating back to the seventeenth century when early naturalists such as Linnaeus (1781), Humboldt et al. (1805), and Darwin (1859) began the monumental task of investigating the temporal and geographic variation of species diversity on earth (Lomolino et al., 2006). Since then, thousands of studies have contributed to the foundations of modern ecology and biogeography and allowed a more comprehensive view of species diversity, showing that species distributions can be a result of deterministic and stochastic processes (Martiny et al., 2006). While both processes influence species distributions, it still remains to be investigated as to which process is most influential and in which conditions. Scientists that come from sub-disciplines of ecology and biogeography, such as botanists, ornithologists, entomologists, mammologists, mycologists, and anthropologists, are constantly researching and studying the phenomena of species distributions in order to increase scientific knowledge and identify critical threats to the species diversity on earth. Findings from these studies are crucial to informing management decisions and developing conservation strategies and policies.

The spatial and temporal distributions of species and their attributes can emerge from a combination of geographic, ecological, and evolutionary processes. The distribution patterns are not permanent or common for each species and can operate across multiple scales (Fink et al., 2014). For example, at a fine spatial scale, penguins follow more or less a randomly spaced distribution, but they also aggressively defend their territory from their neighbors, so they maximize the distance from neighboring individuals (Le Maho et al., 2014). In areas with patchy resources, such as a patchy distribution of watering holes, species tend to have clumped distributions surrounding the resources (Heithaus et al., 1975; Ostfeld, 1985). A clumped distribution is also common in species that usually serve as prey because grouping acts as a mechanism against predation (Mappes et al., 2005; Reynolds et al., 2009). Moving away from the animal kingdom, the dandelion often exhibits a uniform distribution since the seedlings are dispersed by wind and land in random locations, often uniformly (Derksen et al., 1993). Likewise, species distribution patterns can vary by individual-level, group-level, trait-level, or species-level and are driven by complex biotic and abiotic factors. These factors can change seasonally in response to the availability of resources and also depending upon the scale at which they are studied.

Species distribution models (SDMs) are a fundamental tool that was developed and used by ecologists, biogeographers, and statisticians to learn about spatio-temporal distribution of species (Araujo & Guisan, 2006; Kéry & Royle, 2015). SDMs are used to map geo-referenced time-varying species observations to environmental covariates to predict species distribution and facilitate inference such as species-habitat relationships (Araujo & Guisan, 2006; Kéry & Royle, 2015; Hefley & Hooten, 2016; Koshkina et al., 2017). The development of SDMs originated in the twentieth century to describe a species' niche in environmental and geographic space (Colwell & Rangel, 2009). In the following years, the increasing availability of data and advances in computational tools led to a rapid expansion of modeling developments in SDM literature (Guisan & Thuiller, 2005; Elith & Leathwick, 2009; Hefley & Hooten, 2016). The complexity of these modeling approaches has also increased over the years, from fitting simple regression models that map species-habitat relationships to fitting complex hierarchical models that address data uncertainty, spatio-temporal autocorrelation, and perform data fusion, etc. Moreover, SDMs have been developed for use with many sampling designs and data types that are widely collected in subdisciplines of ecology and biogeography (Araujo & Guisan, 2006; Kéry & Royle, 2015).

Various types of data are being collected through planned or opportunistic surveys to study the distribution of species across space and time (Dorazio, 2014; Fletcher et al., 2019). Data comes in various sizes with various spatio-temporal resolutions, from counts of individuals within a geographic space over a time period to individual observations at precise times and locations. Common data types include presence-absence data (i.e., occupancy data) (Hepler et al., 2018; Joseph, 2020), counts (Elith & Leathwick, 2009; Aarts et al., 2012), presence-only data (Dorazio, 2014; Fletcher et al., 2019), distance sampling data (Burnham et al., 1980; Burnham & Anderson, 1984; Buckland et al., 2001), and capture-recapture data (Otis et al., 1978; Seber, 1982; Pollock et al., 1990). Each data types include distinct types of attributes, strengths, and limitations. For example, presence-absence data arise from planned surveys and are collected by making replicated visits to sites and recording the presence or absence of at least one individual (Hepler et al., 2018; Joseph, 2020). They are high quality compared to data from opportunistic surveys; however, due to the cost, they often cover a small geographic region (MacKenzie, 2005; Koshkina et al., 2017). Furthermore, presence-absence data often suffer from false absence or false presences (Hoeting et al., 2000; MacKenzie et al., 2002; Tyre et al., 2003; Mohankumar & Hefley, 2021a). Meanwhile, distance sampling and capture-recapture data are two classic types of high-quality planned survey data. However, these two sources of data require a large amount of effort and cost to collect data over a large geographic region (Chandler et al., 2018; Mohankumar et al., 2022). Further, distance sampling data often suffer from detection and location uncertainty (Buckland et al., 2004; Farr et al., 2020; Hefley et al., 2020; Mohankumar et al., 2022). Presence-only data, on the other hand, arise from opportunistic surveys and are recorded in museum collections or online public databases (Graham et al., 2004; Dickinson et al., 2010). Therefore, presence-only data usually have broad spatial coverage and provide attractive sources of information to fit SDMs. However, presence-only data are often low in quality as they lack information on species absences (Pearce & Boyce, 2006; Koshkina et al., 2017) and often suffer from imperfect detection, and sampling bias (Dorazio, 2012; Hefley et al., 2013; Dorazio, 2014; Fithian et al., 2015; Koshkina et al., 2017). Presence-only data may also contain location uncertainty (Hefley et al., 2014). It is crucial to develop model-based approaches to account for the limitations in these data because most of these limitations are unfeasible to address at the survey level.

Chapter 1 of this dissertation proposes a hierarchical modeling framework that accounts for the false absences of occupancy data (i.e., site-level presence-absence data) and accounts for non-traditional spatial dependence. In recent years, in SDM literature, “spatial dependence,” also referred to as “spatial autocorrelation,” has received a great deal of attention. The history of analysis of spatial dependence goes back to the work of statisticians such as Moran (1948, 1950), Geary (1954), and Whittle (1954) in the late 1940s and early 1950s. In recent years, there has been a bewildering number of approaches in SDM literature, but not that many explicitly address spatial dependence. Furthermore, the approaches that are used to model spatial dependence in occupancy data while addressing false absences only account for traditional spatial dependence, which is assumed to have been generated from a correlated normally distributed random effect and may be incapable of modeling non-traditional spatial dependence such as discontinuities and abrupt transitions (Hoeting et al., 2000; Johnson et al., 2013b). Machine learning approaches have the potential to model non-traditional spatial dependence, but these approaches do not account for observer errors such as false absences. By combining the flexibility of Bayesian hierarchal modeling and machine learning approaches, a general framework is presented to model occupancy data that accounts for both traditional and non-traditional spatial dependence as well as false absences.

Over the past two decades, many SDM approaches have been developed to account for limitations in data and improve model prediction and inference. These studies range from model developments for individual data types (Hooten & Hobbs, 2015; MacKenzie et al., 2002; Tyre et al., 2003; Phillips et al., 2006; Hefley et al., 2014; Fithian et al., 2015) to combining multiple types of data (Dorazio, 2014; Fithian et al., 2015; Koshkina et al., 2017; Fletcher et al., 2019; Farr et al., 2020; Mohankumar et al., 2022). The literature on combining multiple data sources into a single model has been referred to as "data fusion," "data integration," "data reconciliation," etc. Data fusion utilizes information from multiple types of data to reduce the uncertainty associated with limitations in individual data sources, hence improving the model predictions and inferences (Fletcher et al., 2019; Hooten & Hefley, 2019). Yet reliably fusing data sources can be challenging because data sources can vary considerably in their design, gradients covered, and potential sampling biases (Fletcher et al., 2019). Fletcher et al. (2019) gives an explicit review of recent developments in data fusion for species distribution modeling and emphasizes some potential challenges to combining data. As the availability of ecological and biogeographical data from various sub-disciplines across multiple spatio-temporal scales is increasing, model developments to properly combine multiple data sources are in high demand.

Chapter 2 of this dissertation proposes a data fusion approach that combines distance sampling and capture-recapture data to model species distribution. Distance sampling and capture-recapture are two widely collected data types to learn about species-habitat relationships and abundance; still, they are seldomly used in SDMs due to the lack of spatial coverage. These two data sources alone may suffer from the lack of spatial coverage, but the fusion of the two data sources can increase spatial coverage, which can reduce parameter uncertainty and provide more accurate predictions and inference regarding species distributions (Fletcher et al., 2019; Hooten & Hefley, 2019). This work develops a hierarchical modeling framework for data fusion of distance sampling and capture-recapture data using spatial point processes. The proposed models are built accounting for missing data issues that are unique to each source of data. There is well-developed statistical theory and tools in missing data literature to account for missing data issues (Rubin, 1976; Little, 1992; Little & Rubin, 2019), and they are used to build the proposed models. Chapter 2 emphasizes the advantages of the proposed modeling approach compared with existing approaches in data fusion that use ad-hoc methods to account for missing data issues.

Chapter 3 of this dissertation extends the spatial modeling framework presented in chapter 2 to a spatio-temporal modeling framework. By doing so, the influence of the temporal support on the robustness of the spatio-temporal point process models to model species distribution is investigated. Spatio-temporal point process models are widely used to model species distribution across geographic space and time (Hefley & Hooten, 2016; Renner et al., 2015); however, the notion of support is extremely important to model species distribution using spatio-temporal models. Since ecological processes are inherently complex and operate across a wide range of temporal scales, identifying the appropriate temporal support at which species interact with ecological processes may not always be known a priori. Many studies discuss the inadequacies of the spatial support in point process models, especially in the context of location error and spatial aggregation (Walker et al., 2020; Hefley et al., 2020; Mohankumar et al., 2022). However, the impact of temporal support on point process models to model species distributions is seldom discussed. This is because, unlike temporal support, spatial support is often known, and the individual locations of a species can be mapped into the spatial covariates to adequately model the species distribution. In this study, the advantage of the data fusion approach discussed in chapter 2 is leveraged to enable the modeling of species distribution in multiple temporal scales, from fine scales to coarse scales. Then, the influence of temporal support in spatio-temporal point process models to model species- habitat relationships and estimate abundance is studied.

The research in this dissertation tackles some of the most pressing issues encountered in species distribution studies and contributes toward improving predictions and inferences regarding species distributions across geographic space and time. Chapter 1 in this dissertation is published as Mohankumar & Hefley (2021a), and the two data sets that are used to exemplify the approaches in this work are occupancy data on Thomson's gazelle (*Eudorcas thomsonii*) in Tanzania and sugar gliders (*Petaurus breviceps*) in Tasmania. Chapter 2 in this dissertation is available as Mohankumar et al. (2022). The data set that is used to exemplify the approaches in chapter 2 and chapter 3 is transect data and mist-net data on Grasshopper Sparrows (*Ammodramus savannarum*) in north-eastern Kansas, United States. As understanding species distributions is essential to many areas such as ecology, biogeography, evolution, conservation biology, and wildlife management, the research in this dissertation contributes to many ongoing and future studies to gain critical ecological insights in regard to understanding species distribution patterns.

Chapter 1

Using machine learning to model nontraditional spatial dependence in occupancy data

1.1 Abstract

Spatial models for occupancy data are used to estimate and map the true presence of a species, which may depend on biotic and abiotic factors as well as spatial autocorrelation. Traditionally researchers have accounted for spatial autocorrelation in occupancy data by using a correlated normally distributed site-level random effect, which might be incapable of modeling nontraditional spatial dependence such as discontinuities and abrupt transitions. Machine learning approaches have the potential to model nontraditional spatial dependence, but these approaches do not account for observer errors such as false absences. By combining the flexibility of Bayesian hierarchal modeling and machine learning approaches, we present a general framework to model occupancy data that accounts for both traditional and nontraditional spatial dependence as well as false absences. We demonstrate our framework using six synthetic occupancy data sets and two real data sets. Our results demonstrate how to model both traditional and nontraditional spatial dependence in occupancy data which

enables a broader class of spatial occupancy models that can be used to improve predictive accuracy and model adequacy. This chapter is published in *Ecology* as Mohankumar & Hefley (2021a).

1.2 Introduction

Many ecological studies collect occupancy data to understand the dynamics of species occurrence over space and time (e.g., Hepler et al., 2018; Joseph, 2020). Occupancy data are collected by making replicated visits to sites and recording the presence or absence of at least one individual. During a site visit, individuals may go undetected even when present, resulting in the detection of no individuals (i.e., a false absence). Failure to account for false absences can have a significant impact on parameter estimates and predictions (Hoeting et al., 2000; MacKenzie et al., 2002; Tyre et al., 2003).

To facilitate the analysis of occupancy data that contain false absences, Hoeting et al. (2000), MacKenzie et al. (2002), and Tyre et al. (2003) introduced a zero-inflated Bernoulli model that specifies a distribution of the observed data given the true presence at a site. Using familiar notation for Bayesian hierarchical models, the conditional distribution of the data is

$$y_{ij}|z_i, p_{ij} \sim \begin{cases} \text{Bernoulli}(p_{ij}) & , z_i = 1 \\ 0 & , z_i = 0 \end{cases} , \tag{1.1}$$

where $y_{ij} = 1$ denotes the presence and detection of one or more individuals at the i^{th} site $(i = 1, 2, ..., n)$ during the j^{th} sampling period $(j = 1, 2, ..., J_i)$ and $y_{ij} = 0$ denotes that no individuals were detected. Detection of at least one individual depends on the probability p_{ij}. The z_i is the true presence $(z_i = 1)$ or absence $(z_i = 0)$ at the i^{th} site, which is assumed to be constant during all J_i sampling periods and modeled as

$$z_i|\psi_i \sim \text{Bernoulli}(\psi_i) \; . \tag{1.2}$$

In (1.2), the probability of true presence, ψ_i, is modeled using an intercept term and q site-level covariates with the equation

$$g(\psi_i) = \mathbf{x}_i^{'}\boldsymbol{\beta} \ , \tag{1.3}$$

where $g(\cdot)$ is an appropriate link function (e.g., logit or probit), $\mathbf{x}_i \equiv (1, x_1, x_2, ..., x_q)^{'}$, and $\boldsymbol{\beta} \equiv (\beta_0, \beta_1, \beta_2, ..., \beta_q)^{'}$. Within the vector $\boldsymbol{\beta}$, β_0 is the intercept parameter and $\beta_1, \beta_2, ..., \beta_q$ are regression coefficients.

Since the introduction of the occupancy model in (1.1–1.3), many extensions were developed to address model inadequacies. For example, to account for spatial dependence Johnson et al. (2013a) added a correlated normally distributed site-level effect, η_i (i.e., $(\eta_1, \eta_2, ..., \eta_n)^{'} \sim \text{N}(\mathbf{0}, \boldsymbol{\Sigma})$; see ch. 26 in Hooten & Hefley, 2019) to (1.3) that resulting in

$$g(\psi_i) = \mathbf{x}_i^{'}\boldsymbol{\beta} + \eta_i \ . \tag{1.4}$$

The approach by Johnson et al. (2013a) has been effective in accounting for occupancy model inadequacies caused by traditional spatial dependence (e.g., Wright et al., 2019), which is assumed to have been generated from a correlated normally distributed random effect that imparts varying levels of smoothness on the spatial process. Discontinuities, abrupt transitions, and other "non-normal" spatial processes are common in ecological data, and the traditional spatial random effect may fail to capture such dynamics (e.g., Hefley et al., 2017b). Unfortunately, ecologists lack alternative occupancy model specifications that would allow them to check for and, if needed, model nontraditional spatial dependence.

We demonstrate a framework for occupancy data to model both traditional and nontraditional spatial dependence. Our framework takes a machine learning approach to model the site-level effect in (1.4) and can model both traditional and nontraditional spatial dependence. We illustrate this framework using six synthetic data sets containing traditional and nontraditional spatial dependence and then apply our approach to understand the spatial dynamics of Thomson's gazelle (*Eudorcas thomsonii*) in Tanzania and sugar gliders (*Petaurus* *breviceps*) in Tasmania.

1.3 Materials and methods

1.3.1 Occupancy data requirements

Our proposed modeling framework builds upon the occupancy model of MacKenzie et al. (2002) and Tyre et al. (2003) and therefore is intended for use with occupancy data that was collected with repeated site visits during which the true presence or absence of individuals at a site does not change. In addition, we require that false negative detections are the only observational error. However, our framework is adaptable to accommodate other types of occupancy data (see "Model extensions" in section A.6 of Appendix A for additional detail). For example, our framework can be adapted to account for false presence, which occurs when individuals are not present at a site but are recorded as occurring at a site.

1.3.2 Spatial occupancy model framework

Our proposed framework involves lifting the normal distributional assumption in the spatial component that accounts for the spatial dependence. To accomplish this, we replace the site-level effect in (1.4) with

$$g(\psi_i) = \mathbf{x}_i^{'}\boldsymbol{\beta} + f(\mathbf{s}_i) \ . \tag{1.5}$$

Conceptually, this is an important change; the $f(\mathbf{s}_i)$ is an unknown spatially varying process that is a function, $f(\cdot)$, that depends on the coordinate vector, $\mathbf{s}_i$, of the i^{th} site. The function $f(\cdot)$ is always unknown and is approximated.

This change in perspective is common in the field of machine learning, where the goal is to "learn" or approximate an underlying function using data (see ch. 5 in Hastie et al., 2009). This simple change in (1.5) expands the types of model specifications for the spatially varying process, $f(\mathbf{s}_i)$. For example, regression trees are used to learn about underlying functions that have discontinuities and abrupt transitions, and using regression trees to approximate $f(\mathbf{s}_i)$ could model nontraditional spatial dependence.

Many approaches from machine learning, such as support vector regression, neural networks, boosted regression trees, and Gaussian processes, could approximate $f(\cdot)$. These approaches have been widely used by ecologists to make predictions and inferences about species distributions from abundance and presence-absence data (e.g., De'ath & Fabricius, 2000; Cutler et al., 2007; Elith et al., 2008; Golding & Purse, 2016). However, machine learning approaches are not widely used to model occupancy data because of the issues associated with false absences. Furthermore, approximating the spatial dependence within the occupancy model using machine learning approaches requires custom programming and a level of technical knowledge that hinders widespread use. The existing approaches that blend machine learning approaches with occupancy models are approach specific (e.g., Hutchinson et al., 2011; Joseph, 2020), and therefore switching among the different types of approaches to approximate $f(\cdot)$ is a challenge. For example, switching from a neural network to a regression tree to approximate $f(\cdot)$ in (1.5) would require extensive retooling of computer code, thus hindering model checking, comparisons, and selection.

Fortunately, Shaby & Fink (2012) developed a model-fitting algorithm based on Markov chain Monte Carlo (MCMC) that enables off-the-shelf software for machine learning approaches, such as those available in R (e.g., `rpart(...)`, `svm(...)`, `gam(...)`), to be embedded within hierarchical Bayesian models. Once the initial computer code is written for the occupancy model, switching among machine learning approaches to approximate $f(\cdot)$ requires modifying only a few lines of code. Details associated with model fitting are provided in Appendix A of this dissertation.

1.3.3 Modeling spatial dependence

To identify the spatial dependence and evaluate model adequacy, we use a model selection and model checking approach. First, we use a wide variety of approaches to model spatial dependence and then use a measure of predictive accuracy to determine which approach most accurately models the spatial process. We supplement this predictive approach with a measure of model adequacy (e.g., Wright et al., 2019).

Following Hooten & Hobbs (2015), we measure the predictive accuracy using $-2 \times \text{LPPD}$, where LPPD is the log posterior predictive density. The $-2 \times \text{LPPD}$ is similar to the information criterion used for model selection but uses out-of-sample data rather than in-sample data (Hooten & Hobbs, 2015). As such, $-2 \times \text{LPPD}$ and the difference in $-2 \times \text{LPPD}$ among models can be interpreted similarly to the information criterion that attempts to approximate $-2 \times \text{LPPD}$ using in-sample data (e.g., Watanabe-Akaike information criteria). For example, if model A produced a $-2 \times \text{LPPD}$ score less than the $-2 \times \text{LPPD}$ score produced by model B, then model A has higher predictive accuracy. As a standard of comparison, we fit an occupancy model that does not account for spatial dependence (i.e., (1.3); hereafter nonspatial occupancy model).

In addition, we use Moran's I correlogram to check model adequacy. Moran's I has been used to detect traditional spatial dependence in the residuals of fitted occupancy models (Wright et al., 2019). However, if traditional approaches fail to capture spatial dependence, then Moran's I may identify such inadequacies.

1.4 Synthetic data examples

For our synthetic data examples, we show the probability of occupancy in Fig. 1.1, which includes the three scenarios of nontraditional spatial dependence and the three scenarios of traditional spatial dependence listed below.

1. Spatial dependence that has discontinuities and abrupt transitions generated by a step-wise function (nontraditional; Fig. 1.1a).

2. Spatial dependence forming a circle with the probability of occupancy being low in the center and smoothly increases towards the edges (nontraditional; Fig. 1.1b).

3. Spatial dependence defined by a cosine function (nontraditional; Fig. 1.1c).

4. Normally distributed random effect with a correlation matrix specified by a conditional autoregressive process (traditional; Fig. 1.1d).

5. Normally distributed random effect with a correlation matrix specified by an exponential covariance function (traditional; Fig. 1.1e).

6. Normally distributed random effect with a correlation matrix specified by a squared exponential covariance function (traditional; Fig. 1.1f).

For each scenario, we generate synthetic data using (1.1), (1.2), and (1.5) on a unit square study area (i.e., $\mathcal{S}$ =$[0,1] \times [0,1]$). We divided the study area, $\mathcal{S}$, into 900 grid cells (sites). We set the true values for the parameters to $p_{ij} = 0.5$ and $\beta_0 = 0$. We exclude covariates and regression coefficients in our synthetic data so that the spatial process is unobstructed when ψ_i is mapped onto $\mathcal{S}$, which aids when visual and numerical comparisons are made among the machine learning approaches. From the 900 grid cells, we consider a random sample of $n = 200$ sites as the study area with $J_i = 3$ visits for model fitting.

We apply our spatial occupancy modeling framework to the six synthetic data sets and compare the performance of four embedded machine learning approaches, which include regression trees, support vector regression, a low-rank Gaussian process, and a Gaussian Markov random field. The low-rank Gaussian process and Gaussian Markov random field are approaches that model traditional spatial dependence for data sets with a large number of sites and have been used in models for occupancy data (Johnson et al., 2013a; Heaton et al., 2019). The regression tree and support vector regression are nontraditional approaches and may be capable of modeling nontraditional types of spatial dependence.

We assess the performance of each approach to model spatial dependence using $-2 \times$ LPPD calculated at 200 sites with $J_i = 3$ visits that were not used for model fitting (hereafter out-of-sample sites) and by using Moran's I correlogram. In addition, we visually compare the true probability of occupancy (ψ_i) to the posterior mean of the probability of occupancy ($\text{E}(\psi_i|\mathbf{y})$; Fig. 1.2; see section A.7 of Appendix A).

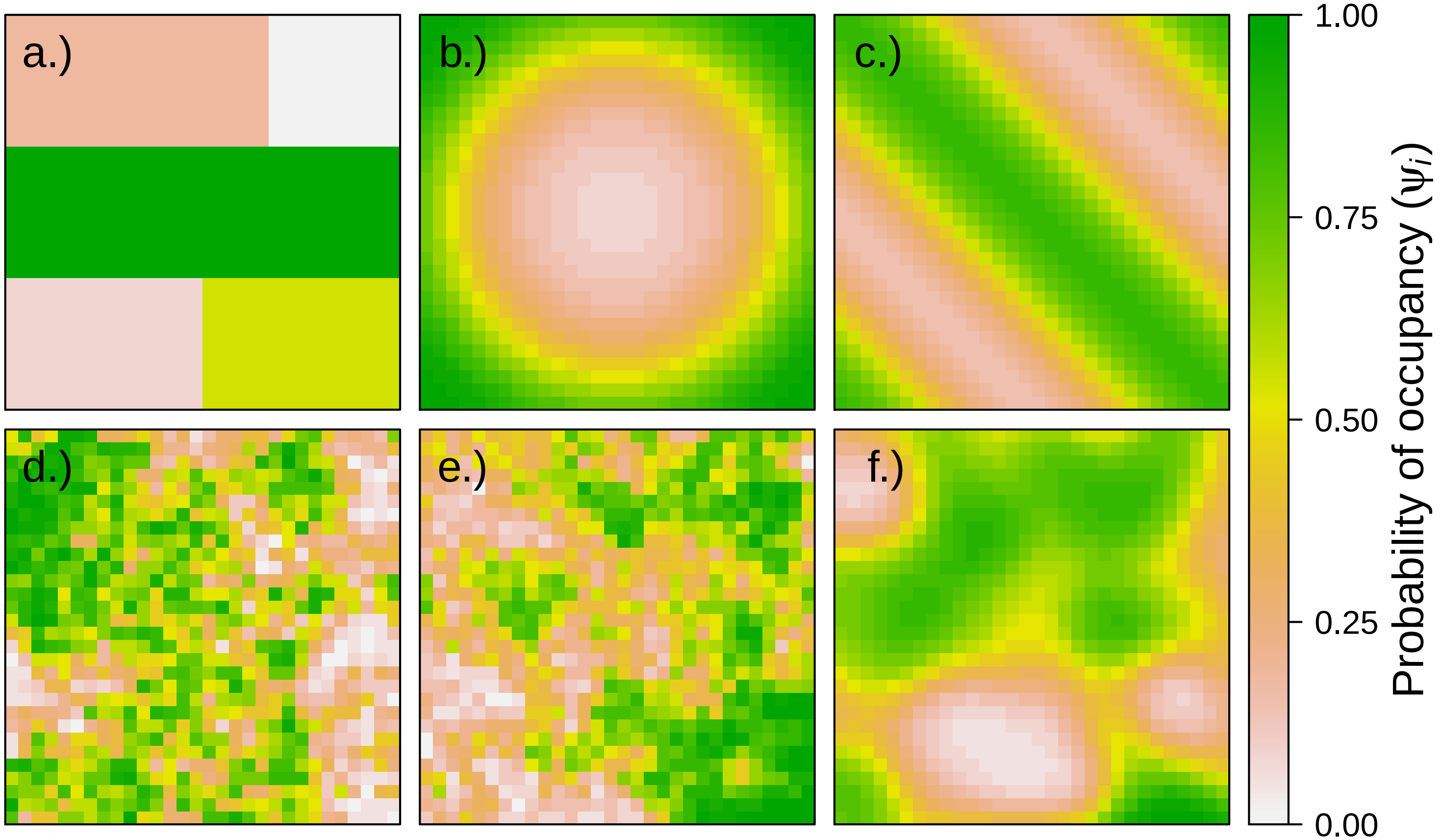

Figure 1.1: Synthetic data examples showing the probability of occupancy (ψ_i in (1.5)) at 900 potential sites (pixels) for six scenarios of traditional and nontraditional spatial dependence. The nontraditional scenarios include spatial dependence having discontinuities and abrupt transitions (panel a), forming a circle (panel b), and defined by a cosine function (panel c). The traditional scenarios include spatial dependence generated from a normally distributed random effect with a correlation matrix specified using a conditional autoregressive process (panel d), an exponential covariance function (panel e), and a squared exponential covariance function (panel f).

1.5 Thomson's gazelle data

We illustrate our spatial occupancy modeling framework using a data set from Hepler et al. (2018), who reported the presence and absence of Thomson's gazelle at 195 sites within Serengeti National Park, Tanzania (Fig. 1.3a). The sites were sampled using a network of motion-sensitive and thermally activated cameras. Images were classified by participants on the citizen science website Snapshot Serengeti. A site visit consisted of an 8-day period during the year 2012 (e.g., January 1–8, 2012). Each site was visited between 1 and 46 times (the mean number of visits was 29). Following Hepler et al. (2018), $y_{ij} = 1$ (from (1.1)) was recorded if an image of at least one Thomson's gazelle was captured at the i^{th} site within the j^{th} 8-d window. A value of $y_{ij} = 0$ was recorded if the site was sampled, but no individuals were observed. Of the 195 sites, 141 had at least one detection. We use 100 randomly selected sites for model fitting and reserve the remaining 95 sites to calculate $-2 \times \text{LPPD}$.

Similar to our synthetic data example, we apply our spatial occupancy modeling framework by embedding four machine learning approaches, which include regression trees, support vector regression, a low-rank Gaussian process, and a Gaussian Markov random field. We exclude site-level covariates in our data example to illustrate our approaches ability to model multiple processes that generate spatial dependence (e.g., missing site-level covariates and spatial autocorrelation) and to illustrate the ability of our method to serve as a "spatial interpolator" for occupancy data (i.e., similar to indicator or binomial kriging, but accounting for false absences). However, as with traditional occupancy models, we can easily include site-level covariates into our spatial occupancy models. The data used in our data examples are available from the Dryad Digital Repository (Mohankumar & Hefley, 2021b).

1.6 Sugar glider data

We illustrate our modeling framework using a second data set from Allen et al. (2018), who reported the presence and absence of sugar gliders. The data were collected during four or five site visits made to 100 sites in the Southern Forest region of Tasmania (Fig. 1.4a). Of the 100 sites, 79 had at least one sugar glider detected. Since this data set has a relatively small number of sites, we used 75 randomly selected sites for model fitting and reserve the remaining 25 sites to calculate $-2 \times$ LPPD. We use the same modeling approaches for this example as we did in the Thomson's gazelle example. The data used in our data examples are available from the Dryad Digital Repository (Mohankumar & Hefley, 2021b).

1.7 Results

1.7.1 Synthetic data examples

In scenario 1, the occupancy model with an embedded regression tree performed best because the other embedded machine learning approaches didn't capture the abrupt transition created by the step-wise spatial process (Fig. 1.2). The $-2 \times$ LPPD was 348.5, 377.2, 377.5, and 384.0 for the embedded regression tree, support vector regression, low-rank Gaussian process, and Gaussian Markov random field, respectively. For comparison, the $-2 \times$ LPPD obtained from the nonspatial occupancy model was 433.1. Similarly, for scenario 1, the comparison of the Moran's I between the occupancy models suggested that spatial dependence must be accounted for using a regression tree; all other approaches resulted in lingering spatial dependence (see section A.7.3 of Appendix A).

Detailed results for scenarios 2–6 are presented in section A.7 of Appendix A. For example, in scenario 2, the spatial dependence forms a circle with the probability of occupancy being low in the center and smoothly increases towards the edge of the circle (Fig. 1.1b). For scenario 2, we expected and found that the embedded support vector regression performed best (see section A.7 of Appendix A). This was expected because this machine learning approach is best suited to learn about smoothly varying deterministic functions. In total, the results from the scenarios clearly demonstrated that if the spatial process is a discontinuous step function, then the approaches used to model traditional spatial dependence are not adequate, and the approaches such as regression trees should be used. If the spatial dependence is traditional, the differences among the approaches are less distinct; nevertheless, in general, support vector regression performs superior for smoothly varying processes (see section A.7 of Appendix A).

1.7.2 Spatial occupancy dynamics of Thomson's gazelle

Across the four embedded machine learning approaches, the probability of occupancy at a site ranged from 0.45 to 0.95 (Fig. 1.3b–1.3e). Generally, the probability of occupancy was high across the entire study area. However, there was a distinct band running from the southwest to the northeast of the study area where the probability of occupancy was much lower (Fig. 1.3b–1.3e).

The measure of predictive accuracy, $-2 \times$ LPPD, was 669.4, 668.8, 671.0, and 668.7 for embedded regression trees, support vector regression, a low-rank Gaussian process, and a Gaussian Markov random field, respectively. For comparison, the $-2 \times$ LPPD obtained from the non-spatial occupancy model was 676.7. Comparison of the Moran's I between the non-spatial and spatial occupancy models suggested that accounting for spatial dependence improves model adequacy; although, the utility of Moran's I is questionable because the differences among approaches are trivial, which may be due to the small number of sites (see section A.8 of Appendix A; Carrijo & da Silva, 2017). In total, the $-2 \times$ LPPD and Moran's I suggest that spatial dependence should be accounted for in the model. However, Moran's I and $-2 \times$ LPPD suggested that the differences among machine learning approaches are less distinct; therefore, it is unclear if the spatial dependence is traditional or nontraditional.

1.7.3 Spatial occupancy dynamics of sugar gliders

For the sugar glider data example, the probability of occupancy at a site ranged from 0.48 to 0.97 (Fig. 1.4b–1.4e) across the four embedded machine learning approaches. The probability of occupancy was generally high across the entire study area; however, there was an area in the eastern and southeastern portion of the study area where the probability of occupancy was relatively low (i.e., $\psi_i < 0.60$), and there were clear visual differences in the probability

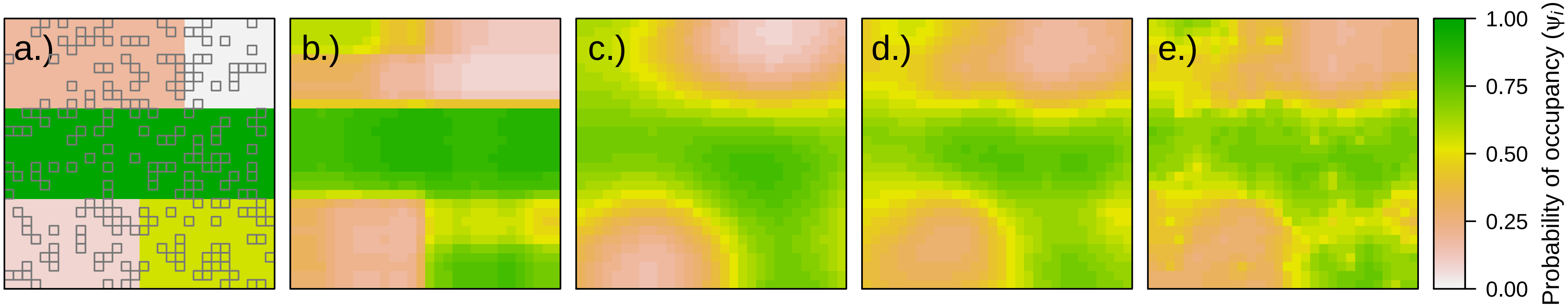

Figure 1.2: The probability of occupancy from scenario 1 of the synthetic data example (panel a) and the posterior mean of the probability of occupancy ($\mathrm{E}(\psi_i|\mathbf{y})$) obtained by fitting spatial occupancy models that included an embedded regression tree (panel b), a support vector regression (panel c), a low-rank Gaussian process (panel d), and a Gaussian Markov random field (panel e). The gray squares in the panel are the locations of the 200 sampled sites used for model fitting.

of occupancy among the four machine learning approaches (Fig. 1.4b–1.4e). The measure of predictive accuracy, $-2\times$LPPD, was 78.2, 80.4, 79.6, and 78.9 for embedded regression trees, support vector regression, a low-rank Gaussian process, and a Gaussian Markov random field, respectively. For comparison, the $-2\times$LPPD obtained from the non-spatial occupancy model was 80.3. Similar to Thomson's gazelle example, the comparison of the Moran's I between the occupancy models suggested that accounting for spatial dependence improves model adequacy (see section A.9 of Appendix A). In total, the $-2 \times$ LPPD and Moran's I suggest that the spatial process (i.e., $f(\cdot)$ in (1.5)) is best modeled using a regression tree. Using Moran's I and $-2 \times$ LPPD as evidence, the results suggested that the spatial dependence is nontraditional.

1.8 Discussion

The use of occupancy models has increased rapidly since the early 2000s. Occupancy data are inherently spatial, but unfortunately, only a limited number of approaches existed to model the spatial process (i.e., Hoeting et al., 2000; Johnson et al., 2013a). This lack of spatial modeling options for occupancy data is in contrast to species distribution models (SDM) that predict the spatial distribution of a species using statistical and machine learning approaches applied to presence-only, count, and presence-absence data. There is a bewildering number of approaches within the SDM literature that are used to model the spatial process. Unfortunately, many of the SDM approaches do not account for contamination in the response variable (e.g., false absences). Understandably ecologists may feel forced to choose between SDM approaches that do not account for contamination in the response variable (e.g., regression trees) and approaches that do, but with a lack of spatial modeling (e.g., occupancy models).

The crux for ecologists planning to use our framework is to determine which machine learning approaches are likely to capture the spatial process, which will require a level of familiarity with the properties of a wide range of machine learning approaches. We recommend James et al. (2013) for a gentle introduction and Hastie et al. (2009) and Murphy

(2012) for more advanced and broad presentations. Within the ecological literature, there are also several excellent guides to machine learning approaches (e.g., De'ath & Fabricius, 2000; Cutler et al., 2007; Elith et al., 2008).

Recently, the hierarchical modeling framework commonly used in ecology has been expanded to include some types of machine learning approaches such as neural networks (Wikle, 2019; Joseph, 2020). Our study builds upon this previous work and expands the types of spatial models ecologists can use for data that fit within the occupancy model framework. Although our work is focused on spatial dependence among the true presence at a site, the approach is easily generalizable. For example, (1.5) implies a linear effect of the site-level covariates (i.e., $\mathbf{x}_i'\boldsymbol{\beta}$). Shaby & Fink (2012) show how machine learning approaches can be used to capture nonlinear and unknown relationships between covariates and the probability of occupancy, thus alleviating the linear assumption in (1.5). Furthermore, many studies that use occupancy models perform covariate selection using model selection techniques (e.g., Hooten & Hobbs, 2015). While model selection techniques work for a small number of covariates, machine learning approaches may be superior when there are a large number of covariates. Another important generalization is that the machine learning approaches can be embedded to model the probability of detection as a function of predictor variables such as Julian date and observer effort (e.g., similar to the use of cubic splines used by Johnston et al., 2018). To facilitate these extensions, we explain in section A.6 of Appendix A how to generalize our framework for other popular ecological models, which is a direct application of the work by Shaby & Fink (2012).

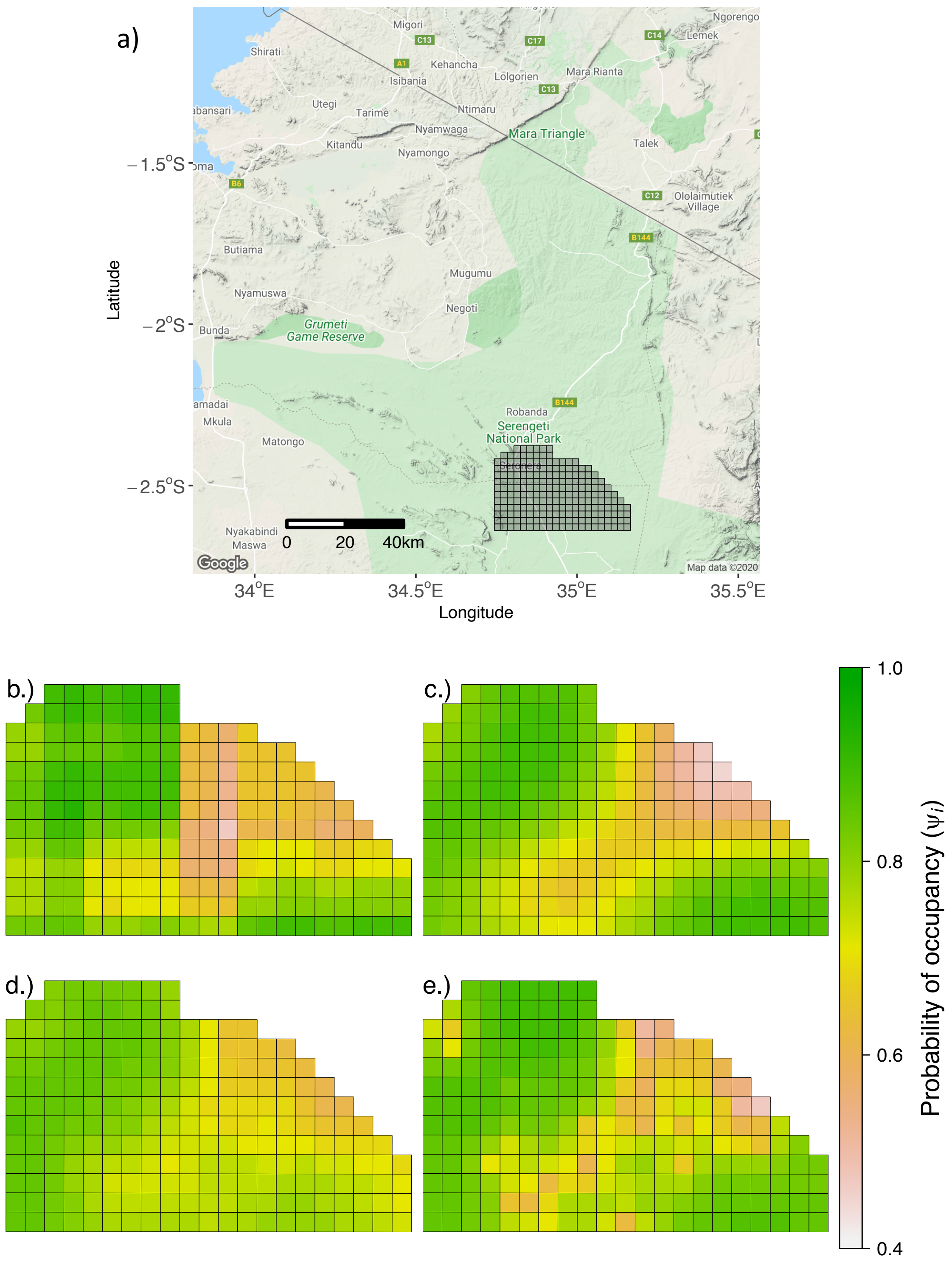

Figure 1.3: Thomson's gazelle data from Hepler et al. (2018) collected at 195 sites within Serengeti National Park, Tanzania (panel a) and the posterior mean of the probability of occupancy ($\mathrm{E}(\psi_i|\mathbf{y})$; panels b–e). Panels b–e show $\mathrm{E}(\psi_i|\mathbf{y})$ obtained by fitting spatial occupancy models that included an embedded regression tree (panel b), a support vector regression (panel c), a low-rank Gaussian process (panel d), and a Gaussian Markov random field (panel e).

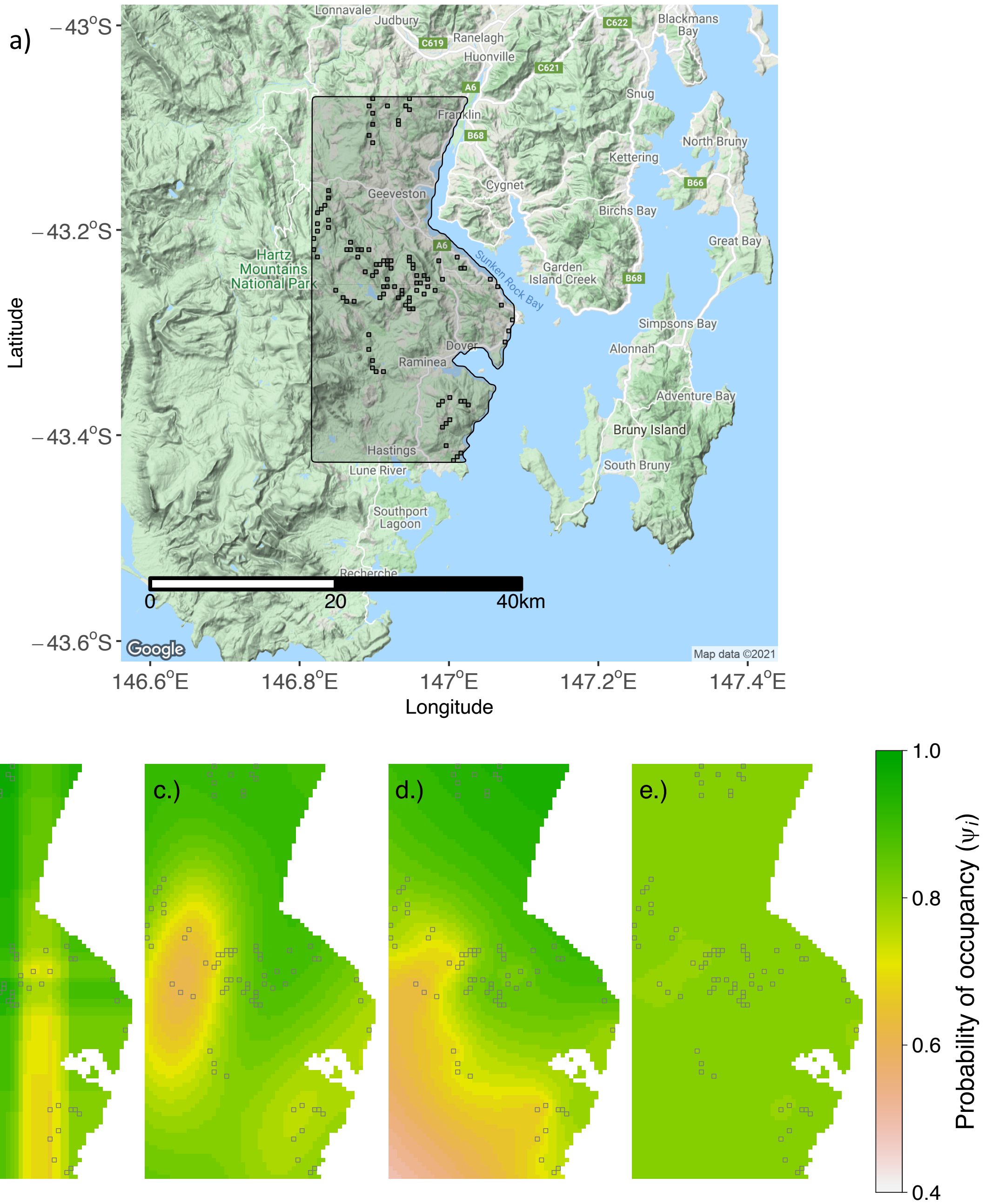

Figure 1.4: Sugar glider data from Allen et al. (2018) collected at 100 sites in the Southern Forest region of Tasmania and the posterior mean of the probability of occupancy ($\text{E}(\psi_i|\mathbf{y})$; panels b–e). Panels b–e show $\text{E}(\psi_i|\mathbf{y})$ obtained by fitting spatial occupancy models that included an embedded regression tree (panel b), a support vector regression (panel c), a low-rank Gaussian process (panel d), and a Gaussian Markov random field (panel e).

Acknowledgements

I thank Dr. Trevor Hefley, who is the co-author of the manuscript published from this chapter. I also thank Drs. Staci Hepler, Michael Anderson, and Robert Erhardt for assistance with the Thomson's gazelle data. I thank Dr. Drew Tyre, Dr. Daniel Fink, an anonymous reviewer, and Dr. Julia Jones for comments that improved our manuscript. Funding for this project was provided by the National Science Foundation via grant DEB 1754491.

Supporting Information

The Thomson's gazelle data and sugar glider data used in our data example are available in the Dryad Digital Repository (Mohankumar & Hefley, 2021b). A tutorial showing how to implement our statistical model and the annotated computer code that can be used to reproduce all results and figures associated with the simulation experiment and data examples are provided within the supporting material associated with Mohankumar & Hefley (2021a).

Chapter 2

Data fusion of distance sampling and capture-recapture data

2.1 Abstract

Species distribution models (SDMs) are increasingly used in ecology, biogeography, and wildlife management to learn about the species-habitat relationships and abundance across space and time. Distance sampling (DS) and capture-recapture (CR) are two widely collected data types to learn about species-habitat relationships and abundance; still, they are seldomly used in SDMs due to the lack of spatial coverage. However, data fusion of the two data sources can increase spatial coverage, which can reduce parameter uncertainty and make predictions more accurate, and therefore, can be used for species distribution modeling. We developed a model-based approach for data fusion of DS and CR data. Our modeling approach accounts for two common missing data issues: 1) missing individuals that are missing not at random (MNAR) and 2) partially missing location information. Using a simulation experiment, we evaluated the performance of our modeling approach and compared it to existing approaches that use ad-hoc methods to account for missing data issues. Our results show that our approach provides unbiased parameter estimates with increased efficiency compared to the existing approaches. We demonstrated our approach using data collected

for Grasshopper Sparrows (*Ammodramus savannarum*) in north-eastern Kansas, USA.

2.2 Introduction

Species distribution models (SDMs) are widely used in ecology, biogeography, and wildlife management to learn about species-habitat relationships and estimate abundance across geographic space and time. Inference and predictions from SDMs are increasingly used to inform conservation and management (Araujo & Guisan, 2006; Kéry & Royle, 2015; Hefley & Hooten, 2016; Koshkina et al., 2017). For example, conflicts between sustaining human activities and preserving biological diversity can be understood by identifying species-habitat relationships across space and time (Hefley et al., 2015). The SDMs are fitted to geo-referenced observations on species such as presence-only, presence-absence, count, distance sampling, and capture-recapture data. Spatially referenced covariates such as elevation, rainfall, soil properties, and vegetation characteristics are used in SDMs to enable statistical inference on species-habitat relationships and obtain spatially heterogeneous abundance estimates (Kéry & Royle, 2015).

Distance sampling (DS) and capture-recapture (CR) are two classic types of planned surveys that collect geo-referenced observations on species. The DS data are collected by recording distances to an individual in the study area from a point or transect (Burnham et al., 1980; Burnham & Anderson, 1984; Buckland et al., 2001). The CR data are collected by capturing an individual in the study area, which involves physically capturing the individual using a trap (e.g., mist nets) or taking a picture (e.g., camera traps; Otis et al., 1978; Seber, 1982; Pollock et al., 1990). The CR data often contain individual identification where DS data do not. There is a long history of collecting these two types of high-quality planned survey data in the field of ecology and wildlife management. However, DS and CR data are seldomly used in SDMs due to the large amount of effort and cost required to collect data that densely covers a large study area (McShea et al., 2016). These two data sources alone may suffer from the lack of spatial coverage, but fusion of the two data sources can increase spatial coverage, which can reduce parameter uncertainty and provide more accurate predictions. Therefore, a fused SDM of DS and CR can provide useful statistical inference regarding the species distribution and abundance more than using any of the data sources alone (see section 25.1 in Hooten & Hefley, 2019).

Construction of an adequate fused data SDM for DS and CR data relies upon accounting for missing data issues that are unique to each source of data. Failure to properly account for these missing data issues may lead to misleading inferences and predictions from the SDMs (Little, 1992; Kéry, 2011; Dorazio, 2012; Hefley et al., 2013). The statistical theory and tools to account for missing data issues are well developed in missing data literature, which can be applied to SDMs (Rubin, 1976; Little, 1992; Mason et al., 2012; Little & Rubin, 2019), but such tools are rarely explicitly employed in SDM literature (Hefley et al., 2013). Therefore, practitioners often use ad-hoc approaches to account for missing data issues and fit SDMs, which will adversely affect inferences and predictions. In some cases, the ad-hoc techniques may produce biased parameter estimates that invert the inferred species-habitat relationship, which is a critical consequence when making inferences (Hefley et al., 2014, 2017a).

Two of the common missing data issues in DS, and CR data are individuals that are missing not at random (MNAR) (Little & Rubin, 2019) and partially missing location information. The MNAR individuals can occur because of two reasons: 1) limited spatial coverage due to the required large amount of effort and cost, limited accessibility, researcher preferences, or previous knowledge regarding the individual locations, or 2) the individuals in a sampled geographic region being unobserved due to the distance to the individual from the point, transect or the trap, observer's experience level, or environmental or geographical features. The partially missing location information occurs when DS and CR only record partial location information of individuals in contrast to complete location information (e.g., the exact geographic coordinates of the location of the individual). Such partially recorded location information makes spatial covariates unrecoverable because the spatial covariate values are usually obtained from a geographic information system that requires the individuals' exact locations. For example, DS surveys only record the distance to an individual from a point or transect and do not record the exact location of the individual. As another example, CR surveys often use tools to attract the individual to the trap, which results in the original, natural location of the individual being unrecoverable because only the location of the trap is recorded (Gerber et al., 2012; Williams & Boyle, 2018). Therefore, the spatial covariate values that may influence the locations of the individuals cannot be obtained.

The missing individuals that are MNAR is implicitly addressed by many DS and CR developments using thinned point process models (e.g., Johnson et al., 2010; Borchers et al., 2015; Fletcher et al., 2019; Farr et al., 2020). Many of these developments use an inhomogeneous Poisson point process (IPPP) which can accommodate spatial inhomogeneity (Diggle et al., 1976; Cressie, 1991; Kéry & Royle, 2015) and enable inferences on the species-habitat relationship and abundance (Warton & Shepherd, 2010; Renner et al., 2015; Hefley & Hooten, 2016). The use of the IPPP enables the estimation of an inhomogeneous intensity function that can produce spatial maps showing the species distribution across the study area. In fact, the field of study of SDMs is almost entirely focused on building, fitting, and using models that are capable of estimating an inhomogeneous intensity function to estimate the species distribution (e.g., Warton & Shepherd, 2010; Renner et al., 2015; Hefley & Hooten, 2016). The spatial maps produced from estimating the species distributions are an essential tool used in conservation reserve planning and administrative regulation implementation (e.g., Hefley et al., 2015). However, the crux in applying existing IPPP based approaches for DS and CR data is that they may not explicitly address the missing data issues in DS and CR data. For example, the approaches may require complete location information regarding the individuals; however, DS and CR data often contain only partial location information. In practice, researchers use ad-hoc methods to circumvent the limitation of partially recorded locations of individuals and fit the models. For example, Fletcher et al. (2019) transformed the DS data to presence-absence data at sites using change of support and fitted a model to the presence-absence data. For another example, Farr et al. (2020) treated DS data as count data at sampling sites and fitted the model to count data. Both of these approaches do not require complete location information, and the partial location information does not pose an issue since the models are fitted to the transformed DS data. As another example, Borchers et al. (2015) proposed an IPPP based unified model for DS and CR data; however, they used a homogeneous point process in all of their applications and did not implement the model for the inhomogeneous case. The homogeneous case contains a constant intensity function where the partial location information is not an issue, but the model is not designed to model the species distribution, which is our primary interest. The inhomogeneous case can model the species distribution; however, the intensity function typically depends on spatially referenced covariates, where the complete location information of the individuals is critical. Therefore, partially recorded location information becomes an issue. In contrast to the ad-hoc approaches, Hefley et al. (2020) proposed a model-based approach to account for the partial location information in DS data. However, their model is merely constructed for DS data, and a subsequent model that accounts for the partial location information in CR data is lacking.

In contrast to properly accounting for missing data issues, constructing a fused data SDM requires adequate model representations for DS and CR data that facilitates data fusion. A fused data SDM utilizes information from both types of data to reduce the uncertainty associated with limitations in individual data sources, hence improving the model predictions and inferences (Dorazio, 2014; Fithian et al., 2015; Koshkina et al., 2017; Fletcher et al., 2019; Hooten & Hefley, 2019; Farr et al., 2020). However, existing IPPP based modeling approaches do not provide model representations for DS and CR data that can be adequately used for data fusion. For example, the unified model proposed by Borchers et al. (2015) represented the model for DS data based on the locations of the individuals and represented the model for CR data based on home range centers which are hypothetical centroids for individuals' activity. The locations of home range centers in CR data are irreconcilable with the locations of the individuals in DS data. For example, the model fitted for CR data would estimate the intensity of home range centers, and the model fitted for DS data would estimate the intensity of the locations of the individuals. Therefore, building a fused data SDM where both data sources share parameters in the underlying IPPP targetting the same inference is not achievable.

A second main issue with existing IPPP based SDMs that involve data fusion is that they often perform spatial aggregation. Spatial aggregation involves partitioning the study area and transforming the locations of the individuals to counts in each of the partitions (e.g., Dorazio, 2014; Koshkina et al., 2017; Farr et al., 2020). However, a significant drawback of spatial aggregation is determining the spatial resolution for the partitions. If the spatial resolution does not adequately represent the sampled region, the model may yield biased estimates of parameters and abundance.

We propose a hierarchical modeling framework that provides adequate model representations for DS and CR data, thereby enabling data fusion and targeting the equivalent inference regarding species-habitat relationship and abundance. We use theory and tools from the missing data literature to build models for the missing data mechanism and account for missing data issues. Our modeling framework can be viewed as a unified framework that can be applied to many other data sources (e.g., presence-only data) and a fusion of them addressing critical issues with missing data. Therefore, our approach advances the types of models developed for species distribution studies. In our work, we propose two fused data SDMs for DS and CR data, one SDM incorporating the recorded distances from DS data and the other SDM without incorporating the recorded distances. We compare the two SDMs and investigate the efficiency gain of the estimated parameters by incorporating additional information regarding the observed individuals, such as the recorded distances. We conduct a simulation experiment to evaluate the performance of our two SDMs compared to existing approaches that use spatial aggregation. We assess the accuracy and the efficiency of the estimated parameters for the species-habitat relationship and obtain an estimate for the expected abundance in the study area. Finally, we demonstrate the approaches using data collected for Grasshopper Sparrows (*Ammodramus savannarum*) in North-Eastern Kansas.

2.3 Materials and methods

2.3.1 Hierarchical modeling framework

Our proposed fused data SDM relies on a hierarchical modeling framework that is based on an IPPP. The models for the observed DS and CR data are conditioned on a common underlying IPPP that represents the underlying point pattern of individuals in the study area.

The underlying IPPP

The underlying IPPP describes the random number and the locations of individuals across the study area based on a continuous inhomogeneous intensity function, a function of spatially referenced covariates (e.g., elevation, temperature, soil attributes, vegetation, etc.). The intensity describes the expected number of individuals per infinitely small unit area and is usually defined as $\lambda(\mathbf{s}) = e^{\mathbf{x}(\mathbf{s})'\boldsymbol{\beta}}$, where, $\mathbf{s}$ represents a vector containing coordinates of a location within the study area $\mathcal{S}$, $\mathbf{x}(\mathbf{s}) \equiv (1, x_1(\mathbf{s}), x_2(\mathbf{s}), ..., x_q(\mathbf{s}))'$, and $\boldsymbol{\beta} \equiv (\beta_0, \beta_1, \beta_2, ..., \beta_q)'$. The $x_1(\mathbf{s}), x_2(\mathbf{s}), ..., x_q(\mathbf{s})$ represent the spatial covariates at the location $\mathbf{s}$, β_0 represents the intercept parameter, and $\beta_1, \beta_2, ..., \beta_q$ represent the regression coefficients associated with the species-habitat relationship. Using the above notation, the probability distribution function (PDF) for the IPPP can be written as (Cressie, 1991)

$$[\mathbf{u}_1, \mathbf{u}_2, ..., \mathbf{u}_N, N|\lambda(\mathbf{s})] = \frac{e^{-\int_{\mathcal{S}} \lambda(\mathbf{s})d\mathbf{s}}(\int_{\mathcal{S}} \lambda(\mathbf{s})d\mathbf{s})^N}{N!} \times N! \prod_{i=1}^{N} \frac{\lambda(\mathbf{u}_i)}{\int_{\mathcal{S}} \lambda(\mathbf{s})d\mathbf{s}}, \tag{2.1}$$

where $\mathbf{u}_1, \mathbf{u}_2, ..., \mathbf{u}_N$ are the locations of all N individuals (missing and observed) in the study area $\mathcal{S}$ (i.e., $\mathbf{u}_i \in \mathcal{S}$). A property of IPPP is that an estimate of the expected abundance in any sub-region $\mathcal{B}$ in the study area can be represented by $\bar{\lambda} = \int_{\mathcal{B}} e^{\mathbf{x}(\mathbf{s})'\boldsymbol{\beta}} d\mathbf{s}$.

Accounting for missing individuals that are MNAR

The missing individuals that are MNAR can be accounted for by identifying and modeling the missing data mechanism. To model the missing data mechanism, we can label the random locations of all individuals in the study area as missing or observed (Gelfand & Schliep, 2018). We can define a vector $\mathbf{m} = (m(\mathbf{u}_1), m(\mathbf{u}_2), ...m(\mathbf{u}_N))$, where $m(\mathbf{u}_i)$ labels the i^{th} individual as missing (i.e., zero) or observed (i.e., one). Employing the missing data mechanism, we can write the distribution of $m(\mathbf{u}_i)$ as a zero-inflated Bernoulli distribution conditioned on $\mathbf{u}_i$.

$$[m(\mathbf{u}_i)|\mathbf{u}_i, q(\mathbf{s}), r(\mathbf{s})] = \begin{cases} q(\mathbf{u}_i)^{m(\mathbf{u}_i)}(1-q(\mathbf{u}_i))^{1-m(\mathbf{u}_i)} & \text{, if } r(\mathbf{u}_i) = 1 \\ 0 & \text{, if } r(\mathbf{u}_i) = 0 \end{cases}, \tag{2.2}$$

where, $q(\mathbf{u_i})$ denote the probability of observing the individual, $r(\mathbf{u}_i) = 1$ denotes that the $\mathbf{u}_i^{\text{th}}$ location is sampled within the study area, and $r(\mathbf{u}_i) = 0$ denotes that the $\mathbf{u}_i^{\text{th}}$ location is not sampled within the study area. The functional form of $q(\mathbf{s})$ and $r(\mathbf{s})$ at a location $\mathbf{s}$ can be defined based on the missing mechanism. The $r(\mathbf{s})$ accounts for the missing individuals that are MNAR due to unsampled geographic regions in the study area, and $q(\mathbf{s})$ accounts for the missing individuals that are MNAR when the corresponding geographic region is sampled, but the individuals are not detected or captured.

By using the distribution of $m(\mathbf{u}_i)$, we can derive the PDF for the location of the i^{th} individual conditioned on the label $m(\mathbf{u}_i)$ as

$$[\mathbf{u}_i|m(\mathbf{u}_i), \lambda(\mathbf{s}), q(\mathbf{s}), r(\mathbf{s})] = \begin{cases} \frac{q(\mathbf{u}_i)^{m(\mathbf{u}_i)}(1-q(\mathbf{u}_i))^{1-m(\mathbf{u}_i)}\lambda(\mathbf{u}_i)}{\int_{\mathcal{S}} q(\mathbf{s})^{m(\mathbf{s})}(1-q(\mathbf{s}))^{1-m(\mathbf{s})}\lambda(\mathbf{s})d\mathbf{s}} & \text{, if } r(\mathbf{u}_i) = 1 \\ 0 & \text{, if } r(\mathbf{u}_i) = 0 \end{cases}. \tag{2.3}$$

An important property of the distributional representation in (2.3) is that it enables the estimation of the locations of unobserved individuals in addition to the locations of the observed individuals. The locations of unobserved individuals can be estimated by augmenting the unobserved individuals and modeling using a Bayesian framework. Many recent model-based approaches based on IPPP use the so-called thinned IPPP (Diggle et al., 1976; Chakraborty et al., 2011; Cressie, 1991; Kéry & Royle, 2015), an implicit representation of the data to account for missing individuals as opposed to the complete distributional representation in (2.3).

Accounting for partially missing location information

The distributional representation in (2.3) accounts for the missing individuals that are MNAR; however, it does not account for the partially missing location information. We propose two models to account for the partially observed location information in data; 1) a model that doesn't incorporate the recorded distances from DS, and 2) a model that incorporates the recorded distances from DS.

The DS and CR surveys each contain a sampled region in the study area, a region surrounding the points, transects, or the traps where the probability of detection or capture is greater than zero. We denote this region as the detection/capture region. In our first proposed model, we assume that the observed location of an individual is uniformly distributed in the detection/capture region that surrounds the point, transect, or the trap it was observed. Under this assumption, we can write the PDF of the observed location of the i^{th} individual conditioned on the actual location of the individual as

$$[\mathbf{y}_i|\mathbf{u}_i] = \begin{cases} |A_{u_i}|^{-1} I(\mathbf{y}_i \in A_{u_i}) & \text{, if } m(\mathbf{u}_i) = 1 \\ 0 & \text{, if } m(\mathbf{u}_i) = 0 \end{cases}, \tag{2.4}$$

where, $\mathbf{y}_i$ denote the observed location of the i^{th} individual, $\mathbf{u}_i$ is the actual location of the i^{th} individual, and A_{u_i} is the detection/capture region surrounding the point, transect or the trap where the individual was observed.

We then propose a second model by incorporating the recorded distances from DS data into the model. We expect that adding additional information regarding the observed locations of the individuals may increase the efficiency of the model parameter estimates. Hefley et al. (2020) account for the partial location information in DS data by incorporating the recorded distances. Based on their approach, and under the assumption that the distances are recorded perfectly, we can assume that the observed location of an individual from a transect is uniformly distributed along the parallel lines to the transect (L_{u_i}) with a perpendicular distance that is equal to the recorded distance d_i. Under this assumption, we can write the PDF of the observed location of the i^{th} individual conditioned on the actual location of the individual as

$$[\mathbf{y}_i|\mathbf{u}_i] = \begin{cases} |L_{u_i}|^{-1} I(\mathbf{y}_i \in L_{u_i}) & \text{, if } m(\mathbf{u}_i) = 1 \\ 0 & \text{, if } m(\mathbf{u}_i) = 0 \end{cases}. \tag{2.5}$$

For a point, L_{u_i} is the perimeter of the circle, where the radius is equal to the recorded distance, d_i. The $|L_{u_i}|$ is the length of the lines or the length of the perimeter of the circle.

2.3.2 Model implementation

The distributions in (2.4) and (2.5) represent the observed location of the i^{th} individual conditioned on the actual location of the observed individual, $\mathbf{u}_i$; however, the actual location of the observed individual is of little interest in our study. Therefore, we can remove $\mathbf{u}_i$ from the model by integrating the joint likelihood of $\mathbf{y}_i$ and $\mathbf{u}_i$. The resulting PDFs representing the observed location of the i^{th} individual are

$$[\mathbf{y}_i|m(\mathbf{u}_i), \lambda(\mathbf{s}), q(\mathbf{s}), r(\mathbf{s})] = \begin{cases} \frac{\int_{A_{u_i}} |A_{u_i}|^{-1}\lambda(\mathbf{u}_i)q(\mathbf{u}_i)d\mathbf{u}_i}{\int_{\mathcal{S}} \lambda(\mathbf{s})q(\mathbf{s})d\mathbf{s}} & \text{, if } r(\mathbf{u}_i) = 1 \text{ \& } m(\mathbf{u}_i) = 1 \\ 0 & , otherwise \end{cases} , \tag{2.6}$$

$$[\mathbf{y}_i|m(\mathbf{u}_i), \lambda(\mathbf{s}), q(\mathbf{s}), r(\mathbf{s})] = \begin{cases} \frac{\int_{L_{u_i}} |L_{u_i}|^{-1}\lambda(\mathbf{u}_i)q(\mathbf{u}_i)d\mathbf{u}_i}{\int_{\mathcal{S}} \lambda(\mathbf{s})q(\mathbf{s})d\mathbf{s}} & \text{, if } r(\mathbf{u}_i) = 1 \text{ \& } m(\mathbf{u}_i) = 1 \\ 0 & , otherwise \end{cases} . \tag{2.7}$$

Moreover, our objectives in the study do not focus on estimating the locations of the unobserved individuals. Therefore, we can retain the PDF for the observed individual locations from (2.6) and (2.7) by setting $m(\mathbf{u}_i)) = 1$. The resulting PDF is a simple marginal distribution that can be fitted using a likelihood-based or Bayesian approach. If practitioners are interested in estimating the locations of unobserved individuals, they can fit the model using a Bayesian hierarchical modeling approach from (2.3–2.5). Details associated with deriving our models are provided in Appendix B.

2.3.3 Fused data SDM

The distributional representations in (2.6) and (2.7) can be used to construct a fused data SDM for DS and CR data. Our proposed distributional represntations represent both DS and CR data based on observed locations of the individuals; therefore, the models share parameters in the underlying IPPP that target the same inference. We assume that the observed locations in the DS and CR data are independent across points, transects and traps within and between the surveys. Representing DS and CR data using our proposed distributional representations and jointly modeling them leads to the following two fused data SDMs. The distribution in (2.8) does not incorporate the recorded distances from DS data, and the distribution in (2.9) incorporates the recorded distances.

$$\begin{aligned}
&[\mathbf{y}_1, ..., \mathbf{y}_{n_{ds}}, \mathbf{y}_{n_{ds}+1}, ..., \mathbf{y}_{n_{ds}+n_{cr}}, n_{ds}, n_{cr} | \lambda(\mathbf{s}), q_{ds}(\mathbf{s}), r_{ds}(\mathbf{s}), q_{cr}(\mathbf{s}), r_{cr}(\mathbf{s})] = \\
&\qquad e^{-\int_{\mathcal{S}} \lambda(\mathbf{s}) q_{ds}(\mathbf{s}) I(r_{ds}(\mathbf{s})=1) d\mathbf{s} - \int_{\mathcal{S}} \lambda(\mathbf{s}) q_{cr}(\mathbf{s}) I(r_{cr}(\mathbf{s})=1) d\mathbf{s}} \times \\
&\qquad \prod_{i=1}^{n_{ds}} \int_{A_{u_i}} |A_{u_i}|^{-1} \lambda(\mathbf{u}_i) q_{ds}(\mathbf{u}_i) I(r_{ds}(\mathbf{u}_i) = 1) d\mathbf{u}_i \times \\
&\qquad \prod_{i=n_{ds}+1}^{n_{ds}+n_{cr}} \int_{A_{u_i}} |A_{u_i}|^{-1} \lambda(\mathbf{u}_i) q_{cr}(\mathbf{u}_i) I(r_{cr}(\mathbf{u}_i) = 1) d\mathbf{u}_i,
\end{aligned} \tag{2.8}$$

$$\begin{aligned}
&[\mathbf{y}_1, ..., \mathbf{y}_{n_{ds}}, \mathbf{y}_{n_{ds}+1}, ..., \mathbf{y}_{n_{ds}+n_{cr}}, n_{ds}, n_{cr} | \lambda(\mathbf{s}), q_{ds}(\mathbf{s}), r_{ds}(\mathbf{s}), q_{cr}(\mathbf{s}), r_{cr}(\mathbf{s})] = \\
&\qquad e^{-\int_{\mathcal{S}} \lambda(\mathbf{s}) q_{ds}(\mathbf{s}) I(r_{ds}(\mathbf{s})=1) d\mathbf{s} - \int_{\mathcal{S}} \lambda(\mathbf{s}) q_{cr}(\mathbf{s}) I(r_{cr}(\mathbf{s})=1) d\mathbf{s}} \times \\
&\qquad \prod_{i=1}^{n_{ds}} \int_{L_{u_i}} |L_{u_i}|^{-1} \lambda(\mathbf{u}_i) q_{ds}(\mathbf{u}_i) I(r_{ds}(\mathbf{u}_i) = 1) d\mathbf{u}_i \times \\
&\qquad \prod_{i=n_{ds}+1}^{n_{ds}+n_{cr}} \int_{A_{u_i}} |A_{u_i}|^{-1} \lambda(\mathbf{u}_i) q_{cr}(\mathbf{u}_i) I(r_{cr}(\mathbf{u}_i) = 1) d\mathbf{u}_i,
\end{aligned} \tag{2.9}$$

where, n_{ds} and n_{cr} are the number of detected and captured individuals from DS and CR respectively, $q_{ds}(\cdot)$ is the probability of detection from a point or transect which depends on the distance from the point or transect to the individual, $q_{cr}(\cdot)$ is the probability of capture from a trap, $r_{ds}(\mathrm{s})$ and $r_{cr}(\mathrm{s})$ are indicator functions defining the detection/capture regions of the DS and CR data respectively, and $n = n_{ds} + n_{cr}$ is the total number of observed individuals from surveys. In our study, we define the probability of detection for DS data by a half-normal function, that is $q_{ds}(\mathbf{u}_i) = e^{-d_i^2/\phi}$, where, d_i is the distance between the point or transect and the the i^{th} detected individual, and ϕ is a scale parameter. The indicator function truncating the detection region from a point or transect is defined as, $r_{ds}(\mathbf{u}_i) = I(\mathbf{u_i} \in A_{ds})$, where A_{ds} is the detection region surrounding a point or transect where probability of detection is greater than zero. We define the probability of capture from a trap as $q_{cr}(\mathbf{u}_i) = \theta$. The indicator function truncating the capture region of a trap is defined as $r_{cr}(\mathbf{u}_i) = I(\mathbf{u}_i \in A_{cr})$, where A_{cr} is the capture region surrounding a trap where probability of capture is greater than zero.

In principle, including additional information regarding the observed individual locations ought to increase the efficiency of parameter estimates from a model. Therefore, we expect the fused data SDM in (2.9) to provide more efficient parameter estimates than the fused data SDM (2.8) since the SDM in (2.9) incorporates the recorded distances from DS data. We investigate this fact in both the simulation experiment and the data example that follows.

2.4 Simulation experiment

We conducted a simulation experiment to evaluate the performance of our two proposed fused data SDMs and compare them to standard approaches that use spatial aggregation. We assessed the performance of the models using the five scenarios listed below.

1. The model from (2.3) fit to DS and CR data containing complete location information of the individuals.

2. The model proposed by Farr et al. (2020) for spatially aggregated data fit to DS and CR data containing partial location information of the individuals.

3. The model from (2.3) tranformed for spatially aggregated data using change of support fit to DS and CR data containing partial location information of the individuals.

4. Our proposed fused data SDM from (2.8) without incorporating recorded distances fit to DS and CR data containing partial location information of the individuals.

5. Our proposed fused data SDM from (2.9) incorporating recorded distances fit to DS and CR data containing partial location information of the individuals.

In our simulation experiment, we simulated a single spatial covariate, $x(\mathbf{s})$ using a reduced rank Gaussian process on an unit square study area (i.e., $\mathcal{S} = [0,1] \times [0,1]$, where $\mathbf{s} \in \mathcal{S}$). We simulated the actual locations of the individuals using the IPPP represented by (2.1) with the intensity $\lambda(\mathbf{s}) = e^{\mathbf{x}(\mathbf{s})'\boldsymbol{\beta}}$. We set the parameter values as $\beta_0 = 9, \beta_1 = 1, \theta = 0.2$, $\phi = 0.025$. We placed 15 points and 65 traps in the study area to obtain DS and CR data, respectively (Fig. 2.1a). We set non-overlapping detection/capture regions to ensure the independence of the observed data across surveys and within surveys (Fig. 2.1c). We constructed the detection region surrounding each point by defining that the individual has to be within a maximum distance of 0.04 from the point to be detected. We constructed the capture region surrounding each trap by defining that the individual has to be within a maximum distance of 0.02 from the trap to be attracted and captured. We obtained spatially aggregated data required to fit the models in scenario 2 and scenario 3 by dividing the study area into 100 non-overlapping partitions and obtaining the number of observed individuals in each partition (Fig. 2.1b). If a partition does not consist of a survey point or a trap, we defined the partition as an unsampled partition.

We simulated 1000 data sets and fitted the models described in scenarios 1–5. We used the complete location information of the individuals in scenario 1, whereas the partial location information of the individuals in scenarios 2–5. Scenario 1 acts as the benchmark scenario where the data with complete location information matches the process described by the fitted model. We evaluated the performance of the models in scenarios 2–5 for data containing partial location information and compared them to benchmark scenario 1. For each simulated data set, we obtained the parameter estimates for the intercept (β_0), the relationship to the spatial covariate (β_1), and the expected abundance ($\bar{\lambda}$). We assessed the reliability of the parameter estimates by calculating the coverage probabilities of the 95%

Wald-type confidence intervals (CIs). We included side-by-side box plots to visually compare the empirical distributions of the parameter estimates. We obtained the relative efficiency of the parameter estimates under scenarios 2–5 with reference to the efficiency of parameter estimates obtained under benchmark scenario. The relative efficiency is calculated by dividing the standard deviation of the respective empirical distribution of the estimates by the standard deviation of the empirical distribution of the estimates under scenario 1.

The integrals in the likelihood functions and the integrated intensity function are approximated using numerical quadrature. We used the Nelder-Mead algorithm in R to numerically maximize the likelihoods and obtain the parameter estimates $\hat{\beta}_0$ and $\hat{\beta}_1$. The estimate for the expected abundance is obtained using $\hat{\bar{\lambda}} = \int_{\mathcal{S}} e^{\mathbf{x}(\mathbf{s})'\hat{\boldsymbol{\beta}}} d\mathbf{s}$. We inverted the Hessian matrix to approximate the standard errors of the parameter estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ and then calculated the 95% Wald-type CIs for $\hat{\beta}_0$ and $\hat{\beta}_1$. We approximated the standard error of the parameter estimate $\hat{\bar{\lambda}}$ using the delta method under first-order Taylor expansion and then calculated 95% Wald-type CI for $\hat{\bar{\lambda}}$.

2.5 Grasshopper Sparrows at Konza Prairie Biological Station, Kansas

We illustrated our proposed models and the existing approaches using data on Grasshopper Sparrows (*Ammodramus savannarum*) from Konza Prairie Biological Station (KPBS). KPBS is a long-term ecological research site in northeastern Kansas, comprised of native tallgrass prairie (Knapp et al., 1998; Williams & Boyle, 2018, 2019). Grasshopper Sparrows are a migratory grassland songbird species that winter in the Southern United States and Northern Mexico and breed throughout grasslands in the United States and Southern Canada. However, the loss of prairie habitat has contributed to a long-term population decline in Grasshopper Sparrows (Herse et al., 2018). Therefore, identifying suitable habitats and investigating the abundance of Grasshopper Sparrow populations is essential for directing conservation efforts.

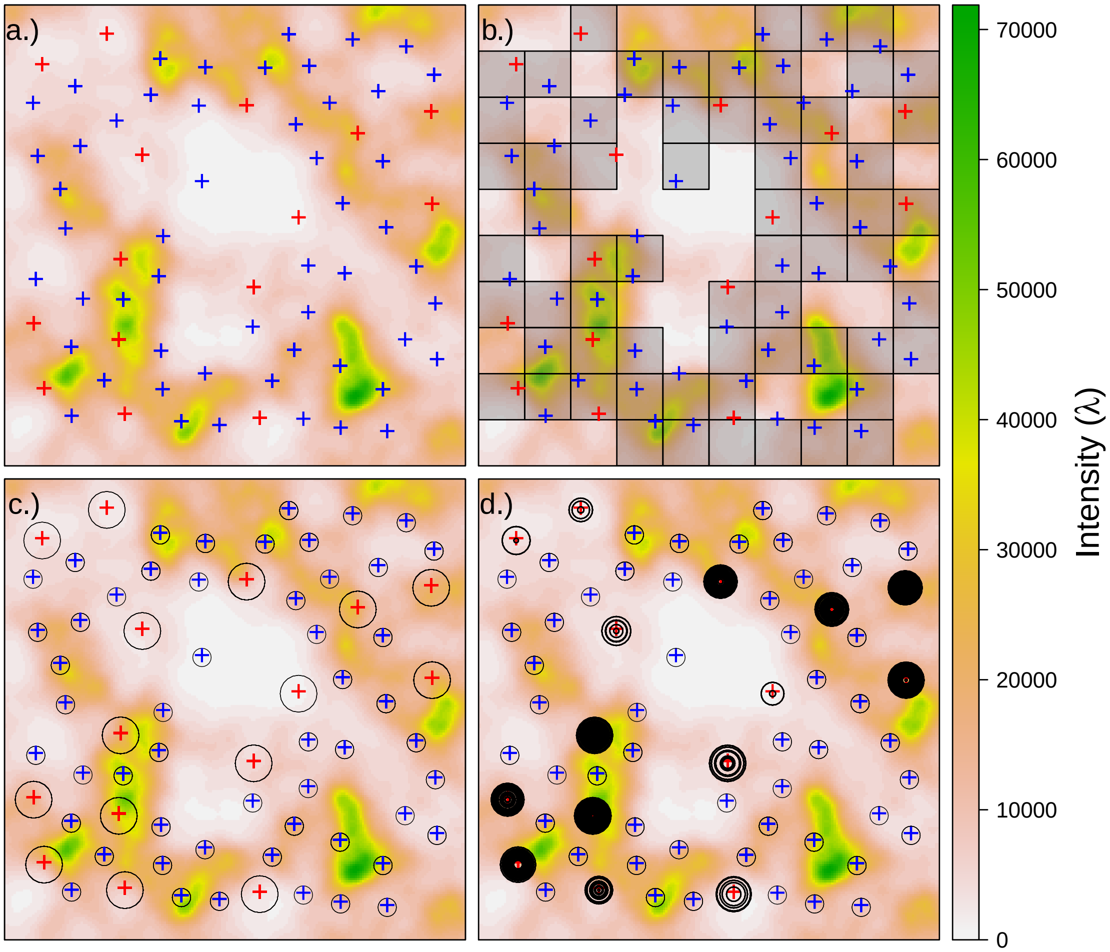

Figure 2.1: Panel (a) displays the points (red +) and traps (blue +) placed in the study area to collect DS and CR data. Panel (b) shows the partitioning of the study area to obtain spatially aggregated DS and CR data (for scenario 2 and scenario 3). Spatially aggregated data are obtained by dividing the study area into 100 non-overlapping partitions and choosing the partitions that include a point or a trap. Panel (c) displays the detection and capture regions of DS and CR data (for scenario 4). Panel (d) displays the circle's perimeter surrounding the points, where the radius is equal to each individual's recorded distance (for scenario 5). Panel (d) also displays the capture regions of the traps.

We used observations from the 2019 breeding season for our analysis. The data consist of 72 observations from 53 transects and 160 observations from 137 mist-net locations (Fig. 2.2a). The transects were surveyed during the month of June as part of the long-term monitoring efforts of birds at the Konza Prairie. Within 24 experimentally-managed pastures, one to four 300m long transects bisect the topographic gradients within the sampling site. A single observer slowly walks the transect, recording the individuals seen or heard on either side of the transect, with the distance to each individual (Boyle, 2019). The mist-nets were used to capture individuals during the entire breeding season from shortly after the adult male birds arrive in April until nests complete in August. The mist net locations were selected to maximize chances of capturing the adult male birds within their territories, and the birds were attracted to nets using a small speaker broadcasting a territorial song (Williams & Boyle, 2018).

Male adult birds sing territorial songs from conspicuous perches in suitable habitats and actively defend 0.5 ha territories from other male birds (Winnicki et al., 2020). Female birds select and build nests within the territories of male birds. Their behavior is very secretive, making them difficult to detect. Thus, both detections and captures consist of male adult birds only. Upon arrival, the male adult birds establish breeding territories at the site. These individual male adult birds may select territories based on many environmental cues such as vegetation, topography, location of conspecifics, and land management (Andrews et al., 2015; Shaffer et al., 2021). To illustrate our approach, we use elevation as the spatial covariate.

We illustrate our approach for DS and CR data using the detections from transects and captures from mist-nets. We assume that the individual has to be within a maximum distance of 150m from the transect to be detected, which is realistic given the topography, song attenuation, and realized distance values (Fig. 2.2c). For captures from mist-nets, we assume that the individual has to be within a maximum distance of 25m to elicit a response and be attracted to the mist-net, a distance reasonable given the speaker volume and observed behavior of the species (Fig. 2.2c). Furthermore, we assume that the observations from the transects and the mist-nets are independent within and between the surveys.

As in scenarios 2–5 in the simulation experiment, we fit the four models to the observed data: 1) the model proposed by Farr et al. (2020) for spatially aggregated data, 2) the model from (2.3) transformed for spatially aggregated data using change of support, 3) our proposed fused data SDM from (2.8) without incorporating recorded distances, and 4) our proposed fused data SDM from (2.9) incorporating recorded distances. We obtain the spatially aggregated data by dividing the study area into non-overlapping partitions and counting observed individuals in each partition. If a partition does not consist of a transect or a mist net, we define the partition as an unsampled partition which led to 66 non-overlapping sampled partitions (Fig. 2.2b). Finally, we fit the models to the data, compare the maximum likelihood estimates and the corresponding 95% Wald-type CIs for β_0, β_1, and $\bar{\lambda}$.

2.6 Results

2.6.1 Simulation experiment

As expected, the benchmark scenario (i.e., scenario 1) yielded an unbiased estimate for β_0, with a high coverage probability of the 95% CIs, 0.942. When the data contained partial location information, scenario 2 and scenario 3 yielded biased estimates for β_0, whereas scenario 4 and scenario 5 yielded unbiased estimates (see Fig. 2.3 for graphical comparison). The coverage probabilities of the 95% CIs for β_0 under scenarios 2–5 were 0.190, 0.180, 0.761, and 0.925, respectively. The relative efficiencies of estimates for β_0 obtained from scenarios 2–5 were 23.204, 15.949, 13.907, and 1.007, respectively. We noticed that the efficiency of the estimate for β_0 under scenario 5, surprisingly reaches the efficiency obtained under the benchmark scenario 1 (see Table 2.1).

Similar to the parameter estimate for β_0, scenario 1 yielded an unbiased estimate for β_1 with a high coverage probability of the 95% CIs, 0.948. However, when the data contained partial location information, scenario 2 and scenario 3 yielded biased estimates for β_1, whereas scenario 4 and scenario 5 yielded unbiased estimates for β_1 (see Fig. 2.3 for graphical comparison). The coverage probabilities of the 95% CIs for β_1 under scenarios 2–5 were 0.749, 0.838, 0.942 and 0.942, respectively. The relative efficiencies of estimates for β_1 obtained from scenarios 2–5 were 1.891, 1.394, 1.089, and 1.041, respectively, where scenario 5 provides the most efficient parameter estimate for β_1 (see Table 2.1).

Scenario 1 yielded an unbiased estimate for $\bar{\lambda}$ with a high coverage probability of the 95% CIs, 0.944. When the data contained partial location information, scenario 4 and scenario 5 yielded unbiased estimates for $\bar{\lambda}$. The coverage probabilities of the 95% CIs for $\bar{\lambda}$ under scenarios 2–5 were 0.343, 0.430, 0.783, and 0.944, respectively. The relative efficiencies of the estimates for $\bar{\lambda}$ obtained from scenarios 2–5 were 265.921, 285.819, 141.896, and 1.038, respectively. We noticed that scenario 5 provides the most efficient parameter estimate for $\bar{\lambda}$, which surprisingly reaches the efficiency obtained under benchmark scenario 1 (see Table 2.1).

Table 2.1: Estimated coverage probability (CP) for the 95% confidence interval (CI) and the relative efficiency (RE) for the parameters β_0, β_1, and expected abundance ($\bar{\lambda}$) obtained under scenario1, scenario 2 , scenario 3, scenario 4, and scenario 5 in the simulation experiment. The parameter estimates are obtained by fitting the models to 1000 simulated data sets.

Scenarios	β_0		β_1		$\bar{\lambda}$	
	CP	RE	CP	RE	CP	RE
Scenario 1	0.942	-	0.948	-	0.944	-
Scenario 2	0.190	23.204	0.749	1.891	0.343	265.921
Scenario 3	0.180	15.949	0.838	1.394	0.430	285.819
Scenario 4	0.761	13.907	0.942	1.089	0.783	141.896
Scenario 5	0.925	1.007	0.942	1.041	0.944	1.038

2.6.2 Grasshopper Sparrows at Konza Prairie Biological Station, Kansas

The estimates obtained for the intercept parameter (β_0) under our two proposed models were similar, with narrow 95% CIs. The models that use spatially aggregated data yielded similar estimates for β_0, but with approximately 12 times wider CIs than our proposed models (see Fig. 2.4a, and 95% CIs in Table 2.2). The estimates obtained for β_1 under all four models yielded similar inference regarding the relationship between species abundance and elevation; however, the estimate for β_1 under the model proposed by Farr et al. (2020) was twice as large as the estimates obtained from the other models (see Fig. 2.4b, and 95% CIs in Table reftable2.2). The crucial outcome from our fitted models is the estimates obtained for $\bar{\lambda}$. The models that use spatially aggregated data yielded inexplicable estimates for $\bar{\lambda}$ with an approximate 163000 times wider 95% CIs than our proposed models (see Fig. 2.4c, and 95% CIs in Table reftable2.2). Altogether, the parameter estimates $\hat{\beta}_0$, $\hat{\beta}_1$, and $\hat{\bar{\lambda}}$ from our proposed two models were similar and yielded narrower 95% CIs. The similarity of the estimates obtained from our two models may be due to the smooth surface of the spatial covariate "elevation."

Table 2.2: Parameter estimates and the width of the 95% CIs for the intercept (β_0), the relationship between the abundance and elevation (β_1), and the log of the expected abundance ($\bar{\lambda}$) for Grasshopper Sparrows at Konza Prairie Biological Station, Kansas. The parameter estimates are obtained from the model proposed by Farr et al. (2020) for spatially aggregated data (Spatially aggregated: FARR), the model from (2.3) transformed for spatially aggregated data using change of support (Spatially aggregated: from (2.3)), our proposed fused data SDM from (2.8) without incorporating recorded distances (Fused SDM: from (2.8)), and our proposed fused data SDM from (2.9) incorporating recorded distances (Fused SDM: from (2.9)).

Models	β_0		β_1		$log(\bar{\lambda})$	
	$\hat{\beta}_0$	Width of 95% CI	$\hat{\beta}_1$	Width of 95% CI	$log(\hat{\bar{\lambda}})$	Width of 95% CI
Spatially aggregated: FARR	-4.767	6.034	0.022	0.015	12.766	6.033
Spatially aggregated: from (2.3)	-4.751	5.616	0.012	0.015	12.669	5.619
Fused SDM: from (2.8)	-11.669	0.486	0.011	0.015	5.742	0.463
Fused SDM: from (2.9)	-11.663	0.484	0.010	0.015	5.743	0.463

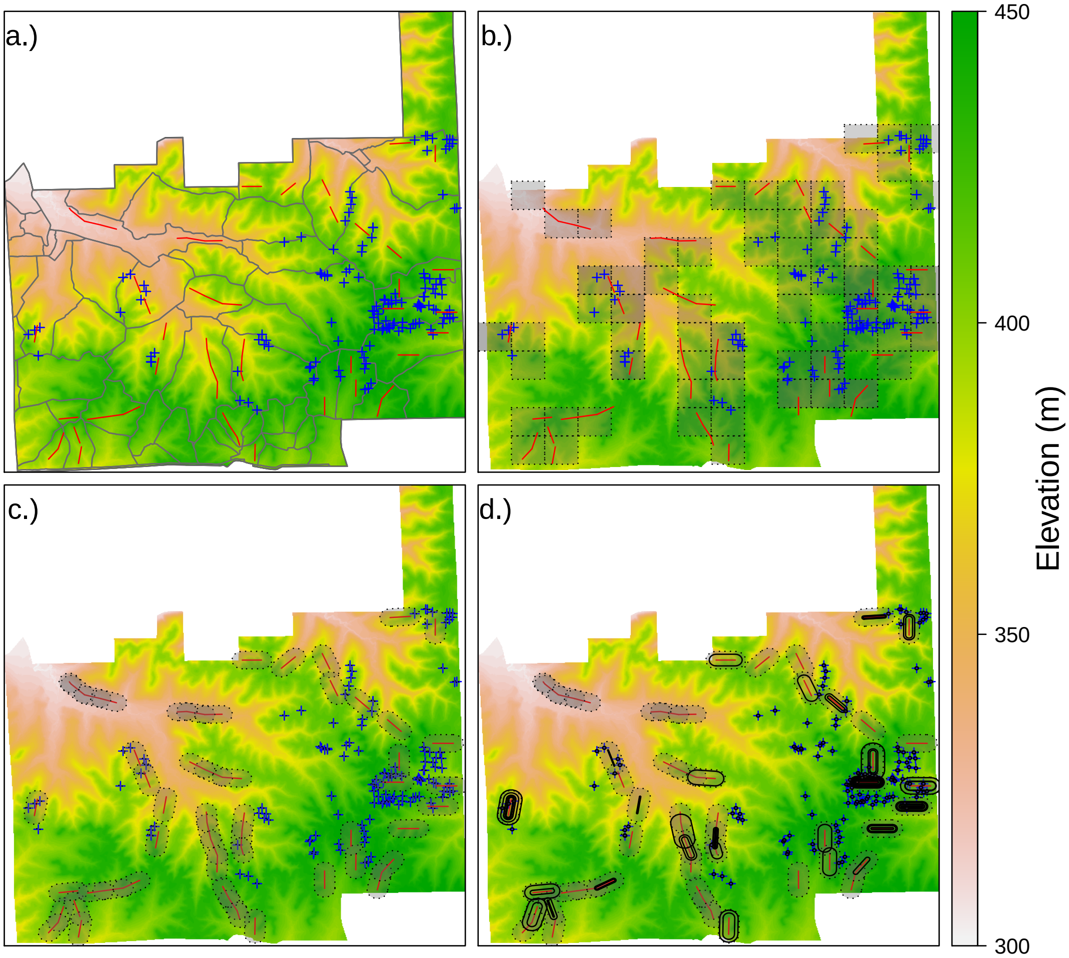

Figure 2.2: Panel (a) displays the transects (red –) and mist nets (blue +) that are used to collect data on Grasshopper Sparrows at Konza Prairie Biological Station (KPBS). The surveys are conducted at watershed-level (grey – in panel (a)). Panel (b) shows the partitioning of the study area (66 partitions) to obtain spatially aggregated data (dashed line) to fit the two models; the model proposed by Farr et al. (2020) for spatially aggregated data, and the model from (2.3) transformed for spatially aggregated data using change of support. Panel (c) displays the detection and capture regions of transects and traps (dashed line) used for our proposed fused data SDM from (2.8) without incorporating recorded distances. Panel (d) displays the parallel lines to the transect with a perpendicular distance equal to each individual's recorded distance, which is used for our proposed fused data SDM from (2.9) incorporating recorded distances. Panel (d) also displays the capture regions of the traps (dashed line).

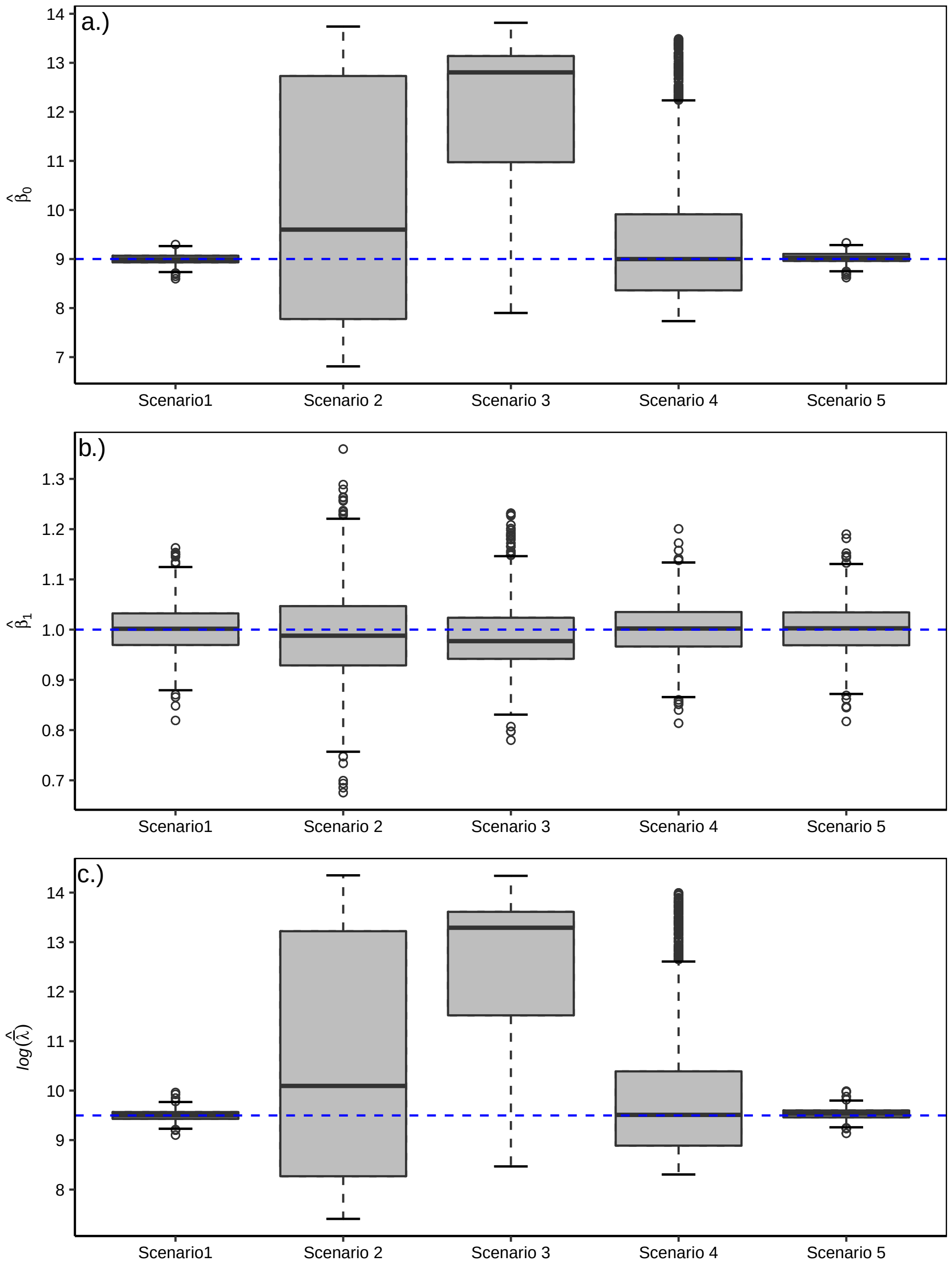

Figure 2.3: The box plots display the estimates of parameters β_0 (panel a), β_1 (panel b), and $log(\bar{\lambda})$ (panel c) obtained under scenarios 1–5 for 1000 simulated data sets. The true values of the parameters ($\beta_0 = 9$, β_1 =1, $log(\bar{\lambda})$= 9.5) are shown by the blue dash line (–).

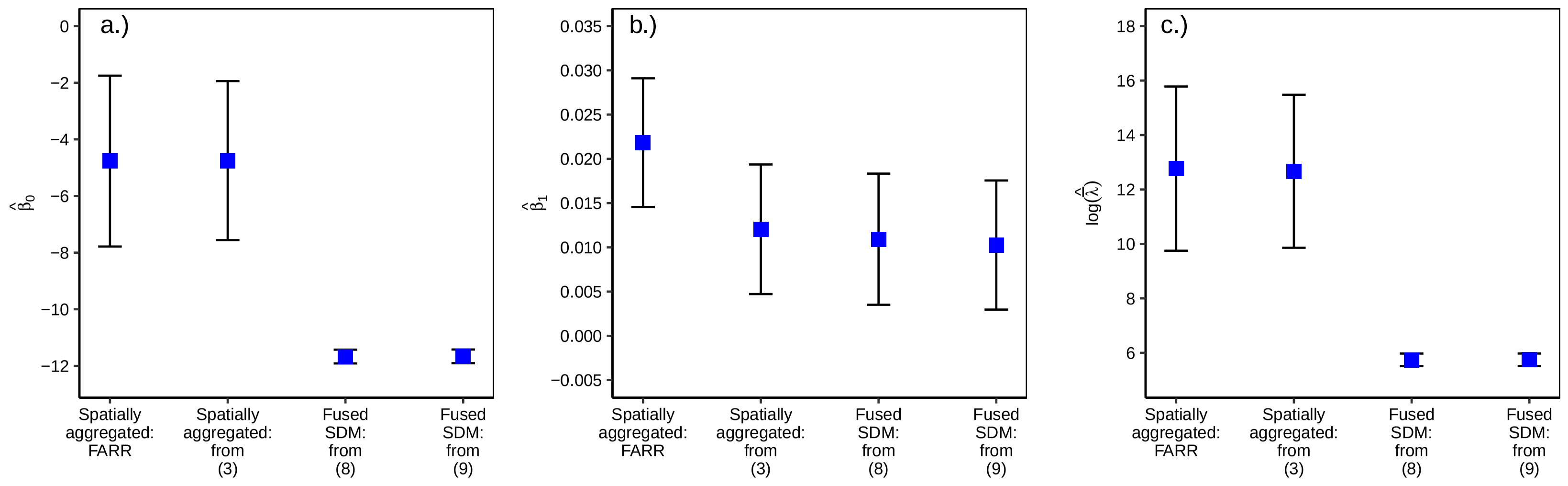

Figure 2.4: Panel (a), panel (b), and panel (c) display the parameter estimates and the 95% CIs for the intercept (β_0), the relationship between the abundance and elevation (β_1), and the expected abundance ($\bar{\lambda}$) for Grasshopper Sparrows at Konza Prairie Biological Station, Kansas. The parameter estimates are obtained from the model proposed by Farr et al. (2020) for spatially aggregated data (Spatially aggregated: FARR), the model from (2.3) transformed for spatially aggregated data using change of support (Spatially aggregated: from (2.3)), our proposed fused data SDM from (2.8) without incorporating recorded distances (Fused SDM: from (2.8)), and our proposed fused data SDM from (2.9) incorporating recorded distances (Fused SDM: from (2.9)). The parameter estimates are shown by the blue square (■), and the 95% CIs are shown by whisker ends.

2.7 Discussion

2.7.1 IPPP generalization for DS and CR data that enables data fusion

A critical aspect of data fusion is providing model representations for multiple data types that target the same inference. The existing point process based models for DS data use individual location information to infer about species-habitat relationship and abundance. In contrast, the existing point process based models for CR explicitly use home range centers. Therefore, the parameters in the underlying point process for the two data sources do not target the same inference. This incompatibility in the underlying process model may explain the lack of approaches for data fusion of DS and CR data. Our proposed approach provides a generalization of Borchers et al. (2015)'s IPPP based model with model representations for DS and CR data that share parameters in the underlying process that target the same inference, hence enabling data fusion. Therefore, our approach enables the use of these two types of high-quality planned survey data to obtain useful statistical inference regarding the species-habitat relationship, accurate estimates for the expected abundance, and more accurate spatial maps for species distributions.

2.7.2 Improvement of inference regarding species-habitat relationship and estimate for the expected abundance by properly accounting for missing data issues

Efficiently acquiring reliable parameter estimates for both β_0 and β_1 is of utmost importance. However, many recent studies only attempt to improve the estimate of β_1, focusing on species-habitat relationships or relative abundance (a measure of expected abundance relative to other species within a community). These approaches do not improve estimates of β_0. In contrast to relative abundance, expected abundance plays a vital role in studying the dynamics of species populations, and estimating the expected abundance depends on both

β_0 and β_1. It is also important to note that a small deviation of $\hat{\beta}_0$ and $\hat{\beta}_1$ from the true parameter value would significantly affect the estimate for the expected abundance due to the exponential function (i.e., $\hat{\lambda}(\mathbf{s}) = e^{\mathbf{x}(\mathbf{s})'\hat{\boldsymbol{\beta}}}$). Our study shows that obtaining reliable, more efficient parameter estimates for β_0 and β_1 crucially relies upon properly accounting for the missing data issues. Our modeling framework explicitly acknowledges and accounts for the missing data issues in DS and CR data using theory and tools from missing data literature.

Our results show that when the data contain partial location information, ad-hoc approaches such as spatial aggregation result in bias parameter estimates with poor efficiency (see Table 2.1). Our proposed models provide reliable, more efficient parameter estimates than existing approaches that use spatial aggregation (see Table 2.1). Furthermore, our simulation experiment led to an important finding: the inclusion of additional information regarding individual locations into the model, such as recorded distances, led to a significant efficiency gain in the parameter estimates. In fact, the efficiency gain surprisingly reaches the efficiency of the parameter estimates obtained under the benchmark scenario with complete location information.

2.7.3 A spatio-temporal fused data SDM

In our simulation experiment, the non-overlapping detection/capture regions ensure the independence of observations across and within surveys. In our data example, we assumed that the observations are independent across and within the surveys. However, we can strengthen the independence assumption by extending our model to a spatio-temporal model. A spatio-temporal model enables the modeling of species abundance patterns across both time and space. By using a continuous-space discrete-time model with short time periods, we can strengthen the independence assumption. However, a spatio-temporal model may have to address the spatio-temporal autocorrelation, which can be addressed by adding a spatio-temporal random effect. A bewildering number of approaches within the SDM literature are developed to model the spatial and spatio-temporal autocorrelation (e.g., Chakraborty et al., 2011; Renner et al., 2015; Mohankumar & Hefley, 2021a), which can be used to incorporate a spatial or a spatio-temporal random effect.

2.7.4 Detection and capture functions

In our study, we defined the probability of detection by a half-normal function of the distance between the point or the transect and the location of the individual. We defined the probability of capture as a constant parameter. However, the probability of detection can be defined by other functions such as uniform, hazard-rate, negative exponential, etc. Similarly, the probability of capture can be defined as a function of covariates such as the observer's experience level or environmental or geographical features. Such extensions of the model enable identifying the factors that influence the probability of detection or capture.

It is possible that the parameters in the detection function or capture function are confounded with the parameters in the intensity function. For example, in a model in which the underlying intensity and the probability of capture are both functions of the same spatial covariate, the underlying point process is confounded with the capture process. For another example, if the underlying intensity function is a function of the distance from the transect, the underlying point process is confounded with the detection process. Accounting for such confounding of the underlying intensity and the detection/capture probability is an area that needs further research. In most situations, we can avoid such confounding during the design of the surveys.

2.7.5 Inclusion of the spatial and non-spatial covariates

The intensity function, probability of detection, and probability of capture can depend on many covariates that are spatial or non-spatial. For instance, in our Grasshopper sparrow data example, the practitioners may want to include ”effort” to define the probability of detection, which is a non-spatial covariate, or they may want to include ”vegetation,” which is a spatial covariate. A non-spatial covariate that is measured during the survey can be easily incorporated into our model. However, for the spatial covariate, our approach requires the spatial covariate values for the entire study region. In most cases, they can be obtained from a geographical information system. However, obtaining the spatial covariate values in the entire study region can be trivial in some situations. In such situations, we can employ an auxiliary model to utilize the available data to predict the spatial covariate values for the entire region and use the predicted values as the input values for the spatial covariate in our models.

Acknowledgements

I thank Dr. Trevor Hefley, Dr. Alice Boyle, and Katy Silber, who are co-authors of the manuscript produced from this chapter. I thank all individuals, including R. Donnelly, J. Gresham, K. Kersten, A. Mayers, E. Smith, and M. Winnerman, who contributed to the Grasshopper Sparrows data used in our data example. This material is based upon work supported by the National Science Foundation under Grant No. 1754491 and the Konza Prairie Long-Term Ecological Research (LTER) Grant No. 1440484. This work was permitted by the US Geological Survey's Bird Banding Lab (23836), the Kansas Department of Parks, Wildlife, and Tourism, and the Kansas State University Institutional Animal Care and Use Committee (IACUC) (protocol 4250).

Chapter 3

Robustness of spatio-temporal point process models to misspecified temporal support

3.1 Abstract

Temporal dynamics of ecological processes are complex, and their influence on species-habitat relationships and abundance operate on multiple spatio-temporal scales. Spatio-temporal point process models are widely used to model species-habitat relationships and estimate abundance across multiple spatio-temporal scales; however, the robustness of models to changing temporal scales is rarely studied. Understanding the temporal dynamics of ecological processes across the entirety of spatio-temporal scales is the key to learning about species-habitat relationships and abundance. Therefore, investigating the influence of temporal support on the robustness of spatio-temporal point process models is critical to understanding species distributions. We use a data fusion approach, as it lifts constraints in individual data sources such as lack of spatio-temporal coverage and expands the spatio-temporal scope of study to enable the study of complex phenomena on species-habitat relationships and abundance across multiple time scales. In our approach, we fuse distance

sampling and capture-recapture data in a spatio-temporal point process modeling framework and investigate the robustness of the model to changing temporal scales. To evaluate the performance of our modeling framework and evaluate the impact of temporal support on models' robustness, we conduct a simulation experiment with four scenarios where species interact with spatio-temporal covariates on continuous, daily, weekly, and monthly temporal scales. We illustrate the influence of temporal support to model species-habitat relationships and abundance using data on Grasshopper Sparrows (*Ammodramus savannarum*) in north-eastern Kansas, USA.

3.2 Introduction

Studying the temporal dynamics of ecological processes and how they drive species-habitat relationships is crucial to understanding the key impacts of environmental and climate changes on species distributions. Conservation and management heavily rely upon these findings to implement conservation reserve planning strategies and administrative regulations (Pressey et al., 2007; Hefley et al., 2015). However, ecological processes are inherently complex, and they operate across a wide range of temporal scales (Fink et al., 2014). Species may interact with these processes at various temporal scales that vary from species to species (Fink et al., 2014). The ecological processes consist of biotic factors (e.g., birth, death, competition, predation, immigration, and extinction) and abiotic factors (e.g., climate change, resource availability, environmental heterogeneity, and human-caused habitat disturbances), and they may directly or indirectly influence species-habitat relationships and cause short to long-term fluctuations in species abundance. For example, soil microbe populations are extremely sensitive to environmental changes, and events such as soil burning or wildfire can cause immediate short-term fluctuations in their populations (Vázquez et al., 1993). Also, long-term degradation of grasslands due to increased land use may influence long-term population decline in grassland birds (Coppedge et al., 2001). Highly variable or heavy rainfall can have direct and indirect physiological consequences on food supply, competition, predation, and pathogens, leading to short-time fluctuations in grassland bird populations

(Williams & Boyle, 2019; Boyle et al., 2020). Studying the temporal dynamics of ecological processes on multiple scales can reveal the ecological context in which species interact with and respond to their environment.

Spatio-temporal point process models are widely used to model species distribution across geographic space and time (Hefley & Hooten, 2016; Renner et al., 2015). The notion of support that is used interchangeably as scale throughout this chapter is extremely important to model species distribution using spatio-temporal models. However, the ultimate obstacle in existing modeling approaches is identifying processes across the entirety of spatio-temporal scales relevant to understanding dynamic changes in species-habitat interactions (Michener & Jones, 2012; Worm & Tittensor, 2018; Rapacciuolo & Blois, 2019). Some studies have emphasized the significance of addressing scale dependence in habitat-selection models (Mayor et al., 2009; McGarigal et al., 2016), but the impact of the scale dependence on point process models to model species distributions is seldom discussed. The spatial scale can be adequately dealt with because the spatial support is often known, and the locations of the observed individuals can be adequately mapped into spatial covariates using appropriate spatial support. Unlike the spatial scale, the influence of the temporal scale is critical and more complex since ecological processes drive biodiversity patterns across multiple temporal scales (Fink et al., 2014). There is rarely a single temporal scale that best identifies how specific ecological processes drive species-habitat relationships and abundance patterns. Therefore, adequately specifying the temporal support in the point process model is unfeasible, and therefore the temporal support in the spatio-temporal point process model may affect the model's accuracy in identifying species-habitat relationships and abundance patterns.

Practitioners often define the temporal support in spatio-temporal point process models based on the resolution at which the phenomenon is being studied (Cressie & Wikle, 2015). Cressie & Wikle (2015) state a motivating example in this context, where the spatio-temporal process under study was the monthly average of maximum temperatures over an area surrounding Iowa for 240 months. Here the spatial support was continuous and the temporal support was discrete on a monthly scale, and the modeling was used to predict the seasonal (monthly) pattern of temperatures. An equivalent species distribution study using spatio-temporal point process models may attempt to learn the seasonal (monthly) pattern of species abundance over an area as a function of monthly averaged covariates such as monthly average temperature. However, questions that remain to be answered include whether species interact with the covariates at monthly scales, whether the use of monthly temporal support in the point process model is adequate, and whether the estimated monthly abundance pattern explains the true temporal dynamics of the species. With enough observations at all appropriate temporal and spatial scales without missing data issues, one can investigate the robustness of the models with varying temporal supports to identify the true underlying species-habitat relationships that are inherent in the process. However, the process cannot be observed at all temporal and spatial scales, and the data may often contain missing data issues (Cressie & Wikle, 2015). For example, data from planned surveys may suffer from lack of spatial and temporal coverage because it requires a large amount of effort and cost to densely cover a large geographic region throughout a long period of time. Opportunistic surveys may contain missing data issues such as sampling bias and location uncertainty that are challenging to account for with unknown sampled regions. Therefore, it is challenging to use individual data sources to investigate the temporal support of the point process models across the entirety of spatio-temporal scales relevant to understanding the dynamic ecological processes.

Data fusion is an increasingly discussed concept in recent years as the availability of ecological data from various ecological sub-disciplines across multiple spatio-temporal scales is increasing. It involves incorporating multiple data sources in a unified modeling framework that can account for the limitations associated with each data source alone and provide more accurate model predictions and inference (Dorazio, 2014; Fletcher et al., 2019; Hooten & Hefley, 2019; Zipkin et al., 2021; Strebel et al., 2022). Therefore, data fusion offers practitioners an opportunity to expand the spatio-temporal scope of research, enabling the investigation of complex phenomena on species-habitat relationships and abundance in multiple temporal scales.

Many of the data fusion approaches developed use spatio-temporal point processes, and most of them are built based on an inhomogeneous Poisson point process (IPPP) (Mohankumar et al., 2022). An IPPP can accommodate spatio-temporal inhomogeneity (Diggle et al., 1976; Cressie, 1991; Kéry & Royle, 2015) and enable inferences on the species-habitat relationship and abundance across space and time (Warton & Shepherd, 2010; Renner et al., 2015; Hefley & Hooten, 2016). Few IPPP based studies discuss models' robustness to the spatial support in which habitat preference of the species may operate; however, the influence of temporal support is seldom discussed (Lin et al., 2011; Ramon et al., 2018). Another crux in most IPPP-based approaches is that they often involve aggregating the individual-level data into counts or presence-absence data without addressing the key scaling dynamics associated with aggregating data across scales (Fletcher et al., 2019; Farr et al., 2020). As we mentioned before, accounting for inadequacies in spatial support is straightforward because spatial support is often known. In fact, many IPPP-based approaches discuss spatial support in the context of location error and spatial aggregation and use the change of support technique to address inadequacies in spatial support (Walker et al., 2020; Hefley et al., 2020; Mohankumar et al., 2022). Unlike the spatial support, identifying the appropriate temporal support at which species interact with ecological processes is more complex and may not always be known a priori. Expanding the scope of IPPP-based data fusion approaches to investigate temporal support would provide a thorough understanding of the effect of misspecified temporal support on the model parameter estimates.

We use an IPPP-based data fusion approach using distance sampling (DS) and capture-recapture (CR) data to investigate the effect of temporal support on the robustness of the spatio-temporal point process models. DS and CR are two types of widely used high-quality data used to study species-habitat relationships and abundance. DS data are collected by recording distances to an individual in the study area from a point or transect (Burnham et al., 1980; Burnham & Anderson, 1984; Buckland et al., 2001). CR data are collected by capturing an individual in the study area, either by physically capturing the individual using a trap (e.g., mist nets) or by taking a picture (e.g., camera traps; Otis et al., 1978; Seber, 1982; Pollock et al., 1990). Due to the high cost and effort required to collect DS and CR data, they may lack the spatio-temporal coverage to model fine-scale species-habitat relationships and abundance patterns. However, fusing the DS and CR data sources will have increased spatio-temporal coverage rather than using the data sources alone. There are many IPPP-based approaches developed in the DS and CR literature individually (Borchers et al., 2015; Fletcher et al., 2019; Farr et al., 2020; Hefley et al., 2020; Mohankumar et al., 2022). Mohankumar et al. (2022) proposed an IPPP-based approach that fuses DS and CR data in a unified modeling framework. They incorporate the missing data mechanism in building their models that account for missing data issues unique to each data source. By doing so, their approach provided more accurate model predictions and inference regarding species-habitat relationships and abundance. Nevertheless, their modeling framework only involved the study of species-habitat relationship and abundance across space, but not time. To estimate species-habitat relationships and abundance across time, an extension of the approach (Mohankumar et al., 2022), including the temporal component, is needed.

Our fused data spatio-temporal modeling framework proposed here is built upon the approach by Mohankumar et al. (2022). We extend their modeling framework by incorporating a temporal component that enables the estimation of species-habitat relationships and abundance patterns across both space and time. We implement models in continuous and discrete time scales and evaluate the models' robustness to the temporal support using a simulation experiment. Finally, we fit the proposed fused data spatio-temporal models to data on Grasshopper Sparrows (*Ammodramus savannarum*) in north-eastern Kansas and illustrate the influence of the temporal support to learn about species-habitat relationships and the temporal abundance pattern.

3.3 Materials and methods

3.3.1 Underlying continuous-time point process

Our modeling framework is based on an underlying IPPP that depends on continuous space and continuous time. The IPPP describes the random number of individuals and the time and the locations of the individuals in the study area at a given time period based on a continuous inhomogeneous intensity function $\lambda(\mathbf{s}, t)$, described by a function of spatio-temporal covariates. The $\lambda(\mathbf{s}, t)$ describes the expected number of individuals per infinitely small unit area at a unit time scale and is usually defined as $\lambda(\mathbf{s}, t) = e^{\mathbf{x}(\mathbf{s},t)'\boldsymbol{\beta}}$, where the $\mathbf{s}$ represents a vector containing coordinates of a location within the study area $\mathcal{S}$, t represents a time point in $\mathcal{T}$, $\mathbf{x}(\mathbf{s}, t) \equiv (1, x_1(\mathbf{s}, t), x_2(\mathbf{s}, t), ..., x_q(\mathbf{s}, t))'$ and $\boldsymbol{\beta} \equiv (\beta_0, \beta_1, \beta_2, ..., \beta_q)'$. The $x_1(\mathbf{s}, t), x_2(\mathbf{s}, t), ..., x_q(\mathbf{s}, t)$ represent q spatio-temporal covariates at the location $\mathbf{s}$ at time t, β_0 represents the intercept parameter, and $\beta_1, \beta_2, ..., \beta_q$ represent the regression coefficients associated with the species-habitat relationship. The probability distribution function (PDF) of the IPPP can be written as

$$\begin{aligned}
[(\mathbf{u}_1, k_1), (\mathbf{u}_2, k_2), ..., (\mathbf{u}_N, k_N), N|\lambda(\mathbf{s}, t)] &= \frac{e^{-\int_{\mathcal{S}}\int_{\mathcal{T}}\lambda(\mathbf{s},t)dtd\mathbf{s}}(\int_{\mathcal{S}}\int_{\mathcal{T}}\lambda(\mathbf{s}, t)dtd\mathbf{s})^N}{N!}\times \\
&\qquad \prod_{i=1}^{N}\frac{\lambda(\mathbf{u}_i, k_i)}{\int_{\mathcal{S}}\int_{\mathcal{T}}\lambda(\mathbf{s}, t)dtd\mathbf{s}} \qquad (3.1) \\
&= e^{-\int_{\mathcal{S}}\int_{\mathcal{T}}\lambda(\mathbf{s},t)dtd\mathbf{s}}\prod_{i=1}^{N}\lambda(\mathbf{u}_i, k_i),
\end{aligned}$$

where, N is the total number of individuals in the study area $\mathcal{S}$ during the study period $\mathcal{T}$. The $(\mathbf{u}_1, k_1), (\mathbf{u}_2, k_2), ..., (\mathbf{u}_N, k_N)$ are the coordinates and the time points of all N individuals such that $\mathbf{u}_i \in \mathcal{S}$ and $k_i \in \mathcal{T}$. The expected abundance in the study area $\mathcal{S}$ at any given time t can be obtained by $\bar{\lambda}_t = \int_{\mathcal{S}}\lambda(\mathbf{s}, t)dtd\mathbf{s} = \int_{\mathcal{S}} e^{\mathbf{x}(\mathbf{s},t)'\boldsymbol{\beta}}d\mathbf{s}$.

3.3.2 Underlying discrete-time point process

The distributional representation in (3.1) assumes that the species interact with the temporal covariate in a continuous time scale and that the temporal support is continuous. However, species may interact with the ecological process at discrete time scales, where now the underlying point process becomes discrete in time. In this context, the spatio-temporal point process can be thought of as a spatial process at a temporal aggregation (Cressie & Wikle, 2015), where the aggregation can be any given time period. Therefore, the underlying discrete-time point process can be defined by partitioning the total time period $\mathcal{T}$ into J time periods $K_1, K_2, ..., K_J$ (e.g., days, weeks, or months) and transforming the temporal support in (3.1) from continuous to discrete. For each time period, an intensity function $\lambda_{K_j}(\mathbf{s})$ can be defined as discrete in time. Now the PDF for the locations of individuals at discrete time period K_j can be written as

$$[(\mathbf{u}_1, K_j), (\mathbf{u}_2, K_j), ..., (\mathbf{u}_{N_{K_j}}, K_j), N_{K_j}|\lambda_{K_j}(\mathbf{s})] = e^{-\int_{\mathcal{S}} \lambda_{K_j}(\mathbf{s})d\mathbf{s}} \times \prod_{i=1}^{N_{K_j}} \lambda_{K_j}(\mathbf{u}_i), \tag{3.2}$$

where, N_{K_j} is the number of individuals in the study area during the time period K_j. In the following subsections we discuss two of the most common representations to define the discrete-time intensity function $\lambda_{K_j}(\mathbf{s})$.

Discrete-time intensity function using a temporally aggregated covariate

A discrete-time intensity $\lambda_{K_j}(\mathbf{s})$ can be defined as a function of a vector of covariates at a location $\mathbf{s}$ during time period K_j (i.e., $\mathbf{x}_{K_j}(\mathbf{s})$), whose values are constant within K_j. The $\mathbf{x}_{K_j}(\mathbf{s})$ can be discrete with levels associated to each time period K_j, or it can be a continuous covariate that is transformed into a discrete covariate by aggregating over each time period K_j such that $\mathbf{x}_{K_j}(\mathbf{s}) = \int_{K_j} \mathbf{x}(\mathbf{s}, t)dt$. In fact, implementing discrete spatio-temporal models by transforming continuous covariates to discrete is extremely common in species distribution studies (Dorazio, 2014; Farr et al., 2020; Farr, 2021). Using the vector of covariates $\mathbf{x}_{K_j}(\mathbf{s})$, the intensity at a given location $\mathbf{s}$ during a time period K_j can be written as

$$\lambda_{K_j}(\mathbf{s}) = e^{\mathbf{x}_{K_j}(\mathbf{s})'\boldsymbol{\beta}}. \tag{3.3}$$

The expected abundance in the study area $\mathcal{S}$ during the time period K_j can be obtained as the integral over the continuous spatial domain $\mathcal{S}$, $\bar{\lambda}_{K_j} = \int_{\mathcal{S}} \lambda_{K_j}(\mathbf{s})d\mathbf{s} = \int_{\mathcal{S}} e^{\mathbf{x}_{K_j}(\mathbf{s})'\boldsymbol{\beta}}d\mathbf{s}$.

Discrete-time intensity function using Change of support (COS)

Change of support (COS) is a technique that can be used to transform the temporal process from continuous support to discrete support. COS is used in many IPPP-based studies that involve spatial aggregation to model species distribution (Fletcher et al., 2019; Farr et al., 2020). Recent studies include using the COS to account for location error in data (Walker et al., 2020; Hefley et al., 2020; Mohankumar et al., 2022). Many of these studies argue that modeling species distribution across space using a surrogate predictor such as a spatially aggregated covariate may lead to location error problems and yield biased parameter estimates about species habitat relationships and abundance (Hefley et al., 2020; Mohankumar et al., 2022). This is primarily because the spatial support is often known, and the locations of the observed individuals can be adequately mapped into spatial covariates using COS using appropriate spatial support. However, this concept is not quite clear in the temporal domain because the temporal support is hardly known, and species may interact with temporal covariates across multiple scales. Therefore, the adequacy of using COS over a temporally aggregated covariate merely depends on the underlying temporal process by which species interact with the environment.

By applying COS, the intensity at a given location $\mathbf{s}$ during a time period K_j can be written as

$$\lambda_{K_j}(\mathbf{s}) = \int_{K_j} e^{\mathbf{x}(\mathbf{s},t)'\boldsymbol{\beta}} dt. \tag{3.4}$$

The expected abundance in the study area $\mathcal{S}$ at the time period K_j can be obtained by $\bar{\lambda}_{K_j} = \int_{\mathcal{S}} \lambda_{K_j}(\mathbf{s})d\mathbf{s} = \int_{\mathcal{S}} \int_{K_j} e^{\mathbf{x}(\mathbf{s},t)'\boldsymbol{\beta}} dt d\mathbf{s}$.

3.3.3 Accounting for missing individuals and location uncertainty

We propose distributional representations for observed DS and CR data extending upon the modeling framework proposed by Mohankumar et al. (2022) that accounts for missing individuals and location uncertainty. For observed data from DS survey, under the assumption that distances are recorded perfectly, the observed location of the i^{th} individual is assumed to be uniformly distributed along the parallel lines to the transect with a perpendicular distance that is equal to the recorded distance (Hefley et al., 2020). Suppose DS data are collected using points, the observed location of the i^{th} individual is assumed uniformly distributed along the perimeter of a circle, with a radius equal to the recorded distance (Hefley et al., 2020). For observed data from CR survey, the observed location of an individual is assumed to be uniformly distributed in the capture region that surrounds the trap where the individual was captured. The capture region is the region where the probability of capture is greater than zero. By adding a temporal component to the modeling framework of Mohankumar et al. (2022) and assuming that the time at which the individual was observed is recorded perfectly, we can write the PDF of the observed location and time of the i^{th} individual conditioned on the actual location and time of the individual as

$$[(\mathbf{y}_i, k_i)|(\mathbf{u}_i, k_i)] = |A_{u_i}|^{-1} I(\mathbf{y}_i \in A_{u_i}), \tag{3.5}$$

where, $(\mathbf{y}_i, k_i)$ is the observed location and time of the i^{th}individual, $(\mathbf{u}_i, k_i)$ is the actual location and time of the i^{th}observed individual represented in the continuous-time underlying point process in (3.1). For DS data, A_{u_i} indicates the parallel lines to the transect with a perpendicular distance that is equal to the recorded distance or the perimeter of the circle surrounding the point with radius equal to the recorded distance, whereas, $|A_{u_i}|$ is the length of the parallel lines or the length of the perimeter of the circle. For CR data, $A_{u_i}$is the capture region surrounding the trap, and $|A_{u_i}|$ is the area of the capture region. Since the distributional representation in (3.5) does not depend on time, the PDF of the observed location of the individual in discrete time period K_j (i.e., $[(\mathbf{y}_i, K_j)|(\mathbf{u}_i, K_j)]$) is equivalent to the distributional representation in (3.5).

Now, we can derive the PDF for the locations and time of observed individuals as,

$$\begin{aligned}[(\mathbf{y}_1, k_1), (\mathbf{y}_2, k_2), ..., (\mathbf{y}_n, k_n), n|\lambda(\mathbf{s}, t), q(\mathbf{s}, t), r(\mathbf{s}, t)] = e^{-\int_{\mathcal{S}}\int_{\mathcal{T}} \lambda(\mathbf{s},t)q(\mathbf{s},t)I(r(\mathbf{s},t)=1)d\mathbf{s},t} \times \\ \prod_{i=1}^{n} \int_{A_{u_i}} |A_{u_i}|^{-1} \lambda(\mathbf{u}_i, k_i) q(\mathbf{u}_i, k_i) I(r(\mathbf{u}_i, k_i) = 1) d\mathbf{u}_i,\end{aligned} \tag{3.6}$$

where, n is the number of observed individuals. For DS data, $q(\cdot)$ represents the probability of detection and for CR data, $q(\cdot)$ is the probability of capture. The representation for $q(\cdot)$ can be written as a function of covariates such as distance to the individual, height of the mist-net, observer's experience level, or environmental or geographical features. In our study, we defined $q(\cdot)$ for DS data using a half-normal function (Hefley et al., 2020), that is $q(\mathbf{u}_i, k_i) = q(\mathbf{u}_i) = e^{-d_i^2/\phi}$, where, d_i is the distance between the actual location of the i^{th}detected individual and the transect, ϕ is a scale parameter (Hefley et al., 2020). However, $q(\cdot)$ can be defined by other functions such as hazard-rate, negative exponential, etc. We define $q(\cdot)$ for CR data as $q(\mathbf{u}_i, k_i) = q(\mathbf{u}_i) = \theta$. The $r(\mathrm{s}, t)$ is an indicator function defining whether the location $\mathbf{s}$ and time t is sampled within the study area and the study period. In our study, we define $r(\mathbf{u}_i, k_i) = r(\mathbf{u}_i) = I(\mathbf{u_i} \leq B)$, where B is the region surrounding a point, transect, or a trap where probability of detection or capture is greater than zero. Now, the continuous-time fused data spatio-temporal model for DS and CR data can be obtained by representing DS and CR data using the distributional representation in (3.6) and obtaining the joint distribution.

Similarly, PDF for the locations of observed individuals in discrete time can be obtained using the PDF represented in (3.2) and the discrete-time intensity functions described in (3.3) and (3.4). The PDF for the locations of observed individuals at discrete time period K_j can be written as

$$\begin{aligned}[(\mathbf{y}_1, K_j), (\mathbf{y}_2, K_j), ..., (\mathbf{y}_{n_{K_j}}, K_j), n_{K_j} | \lambda_{K_j}(\mathbf{s}), q(\mathbf{s}, t), r(\mathbf{s}, t)] = e^{-\int_{\mathcal{S}} \lambda_{K_j}(\mathbf{s}) q(\mathbf{s}) I(r(\mathbf{s})=1) d\mathbf{s}} \times \\ \prod_{i=1}^{n_{K_j}} \int_{A_{u_i}} |A_{u_i}|^{-1} \lambda_{K_j}(\mathbf{u}_i) q(\mathbf{u}_i) I(r(\mathbf{u}_i) = 1) d\mathbf{u}_i,\end{aligned} \tag{3.7}$$

where, n_{K_j} is the number of individuals observed in the study area $\mathcal{S}$ during the time period K_j. The discrete-time fused data spatio-temporal model for DS and CR data can be obtained by representing DS and CR data using the distributional representation in (3.7) and obtaining the joint distribution.

3.4 Simulation experiment

We conducted a simulation experiment under four scenarios in which the temporal covariates drive species abundance in 1) continuous, 2) daily, 3) weekly, and 4) monthly time scales. We simulated 300 datasets for each scenario and fitted seven fused data spatio-temporal models to each simulated dataset: 1) the continuous-time fused data spatio-temporal model using the distributional representation in (3.6), 2) the discrete-time fused data spatio-temporal model using the distributional representation in (3.7) with daily support using a temporally aggregated covariate, 3) the discrete-time fused data spatio-temporal model using the distributional representation in (3.7) with daily support using COS, 4) the discrete-time fused data spatio-temporal model using the distributional representation in (3.7) with weekly support using a temporally aggregated covariate, 5) the discrete-time fused data spatio-temporal model using the distributional representation in (3.7) with weekly support using COS, 6) the discrete-time fused data spatio-temporal model using the distributional representation in (3.7) with monthly support using a temporally aggregated covariate, and 7) the discrete-time fused data spatio-temporal model using the distributional representation in (3.7) with monthly support using COS (Table 3.1).

Since the focus of our study was to investigate the influence of temporal support on models' robustness, we simulated a single temporally varying covariate, $x(\mathbf{s}, t)$ using a reduced rank Gaussian process on a unit square study area (i.e., $\mathcal{S} = [0, 1] \times [0, 1]$, where $\mathbf{s} \in \mathcal{S}$) and a unit time scale (i.e., $\mathcal{T} = [0, 1]$). In our first scenario, we simulated the actual locations and time of the individuals using the model represented in (3.6) with intensity $\lambda(\mathbf{s}, t) = e^{\mathbf{x}(\mathbf{s},t)'\boldsymbol{\beta}}$ assuming that the temporal covariate drives species abundance in a continuous time scale (Fig. 3.1). In our second, third, and fourth scenarios, we assume that the temporal covariate drives species abundance in discrete time scales; daily, weekly, and monthly and we simulated the actual locations and time of the individuals using the model represented in (3.7) with the intensity $\lambda_{K_j}(\mathbf{s}) = e^{\mathbf{x}_{K_j}(\mathbf{s})'\boldsymbol{\beta}}$ (Fig. 3.1). Here the time periods K_j are obtained by partitioning $\mathcal{T}$ into daily, weekly, and monthly partitions, and the discrete-time intensity function is defined by temporally aggregating $\mathbf{x}(\mathbf{s}, t)$ such that, $\mathbf{x}_{K_j}(\mathbf{s}) = \int_{K_j} \mathbf{x}(\mathbf{s}, t)dt$. In all four scenarios, the observed data for DS and CR was obtained by placing 15 transects and 65 traps in the study area. We assume that the individual must be within a maximum distance of 0.04 to be detected from a transect and within a maximum distance of 0.02 from a trap to be attracted and captured. The parameter values used to simulate the data were $\beta_0 = 9, \beta_1 = 1, \theta = 0.2$, and $\phi = 0.025$.

We obtained the maximum likelihood estimates for the intercept (β_0), species-habitat relationship (β_1), and the expected abundance across time ($\bar{\lambda}_t$) for each simulated data set. We assessed the reliability of the parameter estimates for the species-habitat relationship (β_1) by calculating the bias and the coverage probabilities of the 95% Wald-type confidence intervals (CIs). To assess the reliability of the estimates for the expected abundance across time ($\bar{\lambda}_t$), we calculated the mean integrated absolute error (MIAE) and the average coverage probabilities of the 95% Wald-type CIs. The 95% Wald-type CIs for the estimated expected abundance was obtained using the delta method under first-order Taylor expansion. To compare the predictive accuracy of the models, we calculated the proportion of the times Akaike information criterion (AIC) selected each fitted model as the best model among the fitted models.

Table 3.1: Models fitted to data under the scenarios in the simulation experiment.

Model	Description
Model 1	Continuous-time fused data spatio-temporal model using the distributional representation in (3.6)
Model 2	Discrete-time fused data spatio-temporal model using the distributional representation in (3.7) with daily support using the temporally aggregated $x(\mathbf{s}, t)$
Model 3	Discrete-time fused data spatio-temporal model using the distributional representation in (3.7) with daily support using COS
Model 4	Discrete-time fused data spatio-temporal model using the distributional representation in (3.7) with weekly support using the temporally aggregated $x(\mathbf{s}, t)$
Model 5	Discrete-time fused data spatio-temporal model using the distributional representation in (3.7) with weekly support using COS
Model 6	Discrete-time fused data spatio-temporal model using the distributional representation in (3.7) with monthly support using the temporally aggregated $x(\mathbf{s}, t)$
Model 7	Discrete-time fused data spatio-temporal model using the distributional representation in (3.7) with monthly support using COS

3.5 Case study: Grasshopper Sparrows at Konza Prairie Biological Station, Kansas

Grasshopper Sparrows are a migratory grassland bird species that has been increasingly studied in recent years due to their population decline (West et al., 2016; Herse et al., 2018). In fact, they are one of the most threatened groups of birds in the United States (Brennan & Kuvlesky Jr, 2005). The Grasshopper Sparrows' annual migratory cycle includes wintering in the southern United States and northern Mexico and breading throughout grasslands in the United States and southern Canada (Macías-Duarte et al., 2017). Loss of grassland habitat from agricultural intensification and urbanization and unfavorable climatic conditions contributed to a significant decline in Grasshopper Sparrows populations over the years (West et al., 2016; Herse et al., 2018). It is assumed that Grasshopper Sparrows are affected by these environmental processes across multiple spatial and temporal scales; however, how they respond to the temporal variation in the environmental processes is hardly understood. Precipitation is identified as a key driver that affects the Grasshopper Sparrows population both directly and indirectly (Williams & Boyle, 2019) through predation, competition, foraging behavior, vegetation, and growth rate in multiple temporal scales that are often difficult to identify. Furthermore, Grasshopper Sparrows are highly mobile and interact with ecological processes at small time scales; therefore, small-scale fluctuations in precipitation can have large effects on the species population. The timing and magnitude of storms are also varying over the years, and scientists seldom know how they affect Grasshopper Sparrow populations. Therefore, identifying the effect of temporal variation of precipitation on Grasshopper Sparrows' habitat interactions and their abundance across multiple temporal scales may reveal micro and macro scale dynamics influencing Grasshopper Sparrow populations.

Here, we investigate and illustrate how temporal support in models may affect estimating the species-habitat relationships and abundance using data on Grasshopper Sparrows from Konza Prairie Biological Station (KPBS), a long-term ecological research site in northeastern Kansas (Knapp et al., 1998; Williams & Boyle, 2018, 2019). Daily data with 651 observations from 54 transects and 944 observations from 790 mist-net locations were collected during the breeding seasons from 2013 to 2018 (Fig. 3.2a). We assume that the individual had to be within a maximum distance of 150m from the transect to be detected (DS data) and within a maximum distance of 25m to elicit a response and be attracted to the mist-net (CR data).

To implement our models, we used precipitation as a temporal covariate. KPBS annually receives 83.5 cm of precipitation, with the majority of precipitation occurring between April and September (Fig. 3.2b). It is assumed that the Grasshopper Sparrows' interaction with the precipitation is not confined to a single spatio-temporal location. Instead, the species seems to interact with precipitation on a larger time scale (e.g., weekly or monthly precipitation) or as a lagged effect (e.g., running average of precipitations through the prior month or week). Therefore, to fit our continuous-time fused data spatio-temporal models we used a temporally integrated covariate, $\tilde{x}(\mathbf{s},t) = \int_{t-14}^{t} x(\mathbf{s},t^*)dt^*$, where $x(\mathbf{s},t^*)$ is the precipitation at a location $\mathbf{s}$ at time t^* and $\tilde{x}(\mathbf{s},t)$ is the precipitation at location $\mathbf{s}$ integrated over the prior two weeks. Using an integrated covariate over space and time allows the point-level relationship to extend across space and time. An extension of this approach is the inclusion of kernel functions to weight the spatio-temporal covariate surface (Heaton & Gelfand, 2011, 2012). Using the integrated covariate, we define the continuous-intensity function as $\lambda(\mathbf{s},t) = e^{\tilde{x}(\mathbf{s},t)\boldsymbol{\beta}}$ and the discrete-intensity functions from (3.3) and (3.4) as $\lambda_{K_j}(\mathbf{s}) = e^{\int_{K_j}\tilde{x}(\mathbf{s},t)dt}$ and $\lambda_{K_j}(\mathbf{s}) = \int_{K_j} e^{\tilde{x}(\mathbf{s},t)'\boldsymbol{\beta}}dt$, respectively.

We fitted five fused data models to the data 1) the continuous-time fused data spatio-temporal model represented in (3.6), 2) the discrete-time fused data spatio-temporal model represented in (3.7) with weekly support by temporally aggregating $\tilde{x}(\mathbf{s},t)$ over each week, 3) the discrete-time fused data spatio-temporal model represented in (3.7) with weekly support using COS, 4) the discrete-time fused data spatio-temporal model represented in (3.7) with monthly support by temporally aggregating $\tilde{x}(\mathbf{s},t)$ over each month, and 5) the discrete-time fused data spatio-temporal model represented in (3.7) with monthly support using COS. We obtained the maximum likelihood estimates, $\hat{\beta}_0$, and $\hat{\beta}_1$ and calculated the 95% Wald-type CIs. We obtained the estimates for expected abundance across the study region covered by the surveyed watersheds. We used the delta method under first-order Taylor expansion to calculate the 95% Wald-type CIs for estimated expected abundance. We obtained the AIC values to compare the predictive accuracy of the models.

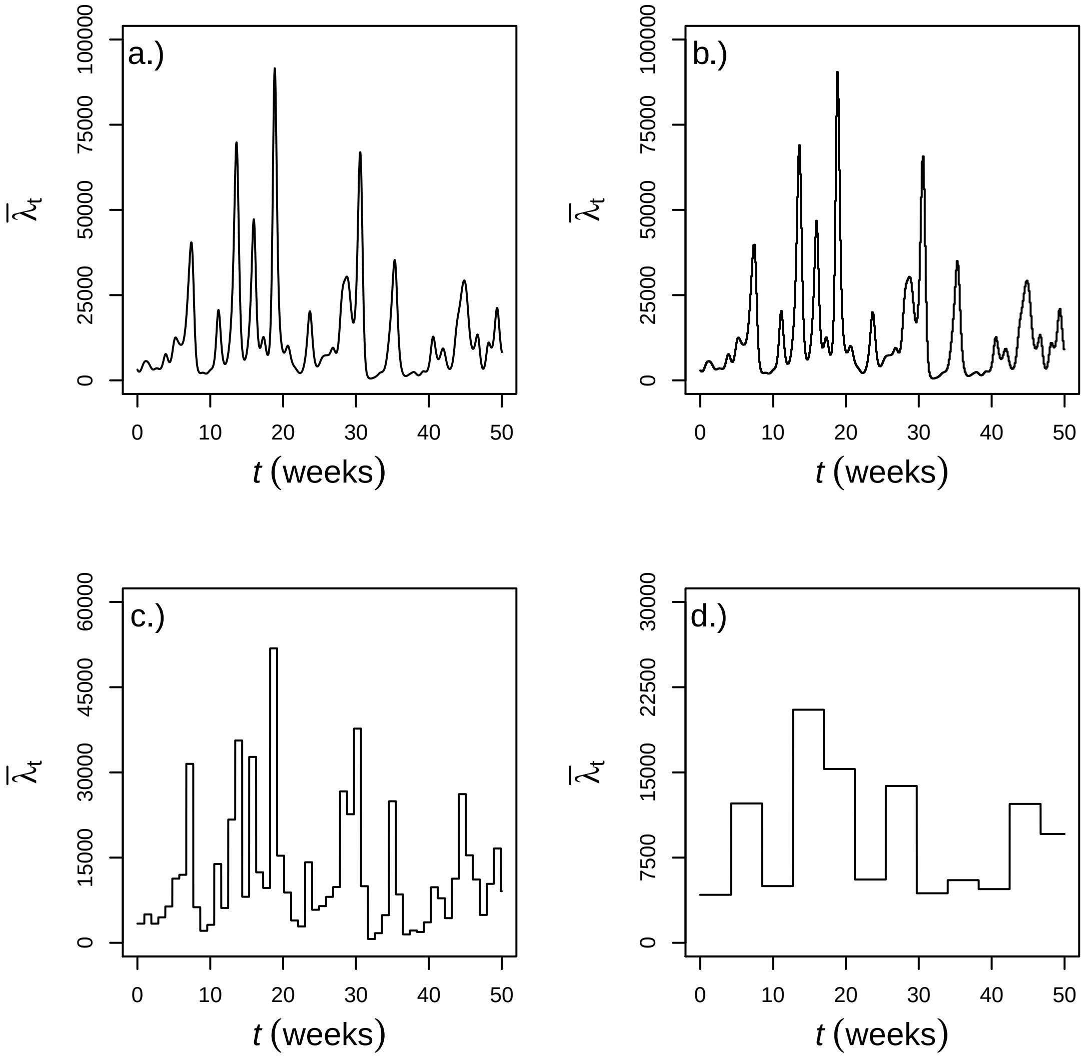

Figure 3.1: A single simulated data set for each of the four scenarios represented in the simulation experiment. Panel (a) represent expected abundance under scenario 1 for a single data set, where the temporal covariate drives species abundance on a continuous time scale; panel (b) represents expected abundance under scenario 2 for a single data set, where the temporal covariate drives species abundance on a daily time scale; panel (c) represent expected abundance under scenario 3 for a single data set, where the temporal covariate drives species abundance in weekly time scale, and panel (d) represent expected abundance under scenario 4 for a single data set, where the temporal covariate drives species abundance in monthly time scale.

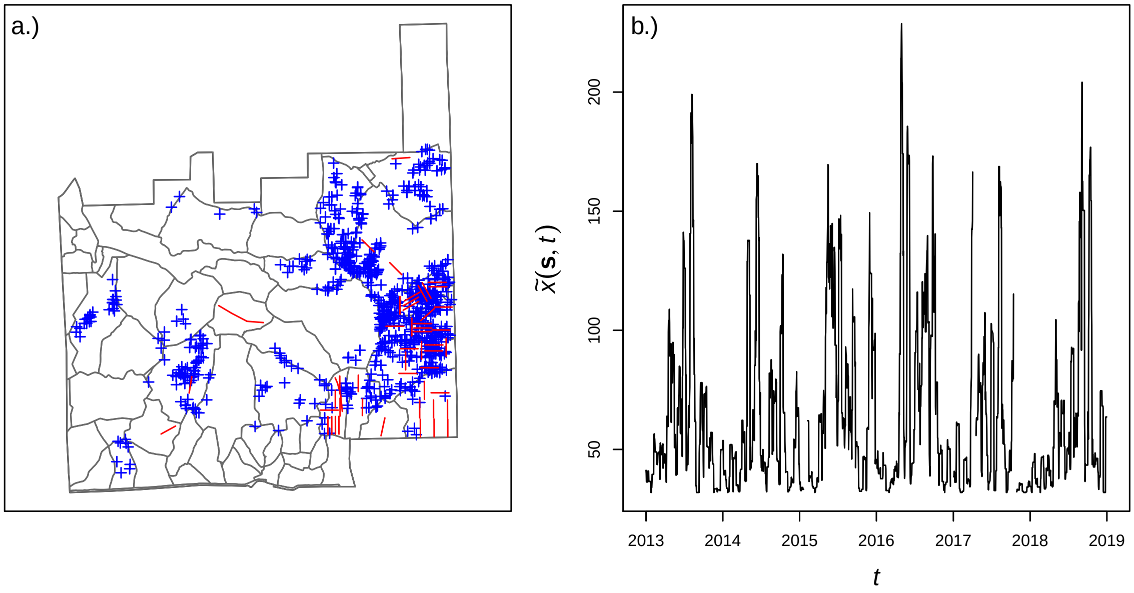

Figure 3.2: Panel (a) displays the line transects (red –) and mist nets (blue +) that are used to collect data on Grasshopper Sparrows at Konza Prairie Biological Station, Kansas. The DS and CR surveys are conducted at watershed-level (grey –). Panel (b) displays the distribution of the time-dependent covariate, precipitation integrated over the prior two weeks (i.e., $\tilde{x}(\mathbf{s}, t)$).

3.6 Results

3.6.1 Simulation experiment

Under the models fitted to data under scenario 1, the continuous-time fused data spatio-temporal model reported the lowest bias, highest coverage probability of the 95% CIs, and the highest proportion of the times that AIC selected as the best model. This was expected since the data generating mechanism in scenario 1 matches the process described by the continuous-time fused data spatio-temporal model. In fact, the discrete-time fused data spatio-temporal model with daily support using the temporally aggregated $\mathbf{x}(\mathbf{s}, t)$, and the discrete-time fused data spatio-temporal model with daily support using COS also reported similar bias and coverage probability of the 95% CIs (see Table 3.2). Furthermore, all three models yielded lower MIAE values for the estimated expected abundance ($\bar{\lambda}_t$) with higher average coverage probability of the 95% CIs compared to other fitted models that used weekly and monthly temporal support (see Table 3.3). The similar performances of the three models with continuous and daily temporal support was because there was no substantial variation in the temporal covariate between continuous and daily temporal scales.

Under the models fitted to data under scenario 2, the discrete-time fused data spatio-temporal model with daily support using the temporally aggregated $\mathbf{x}(\mathbf{s}, t)$ reported the lowest bias, highest coverage probability of the 95% CIs and the highest proportion of the times that AIC selected as the best model (see Table 3.2). This was expected since the data generating mechanism in scenario 2 matches the process described by the discrete-time fused data spatio-temporal model with daily support using the temporally aggregated $\mathbf{x}(\mathbf{s}, t)$. In fact, the discrete-time fused data spatio-temporal model with daily support using COS also reported similar bias and coverage probability of the 95% CIs (see Table 3.2). Furthermore, these two models with daily temporal support yielded the lowest MIAE values for the estimated expected abundance ($\bar{\lambda}_t$) with the highest average coverage probability of the 95% CIs compared to other fitted models with alternate temporal supports (see Table 3.3).

Under the models fitted to data under scenario 3, the discrete-time fused data spatio-temporal model with weekly support using the temporally aggregated $\mathbf{x}(\mathbf{s}, t)$ reported the lowest bias, highest coverage probability of the 95% CIs, and the highest proportion of the times that AIC selected as the best model (see Table 3.2). This is expected since the data generating mechanism in scenario 2 matches the process described by the discrete-time fused data spatio-temporal model with weekly support using the temporally aggregated $\mathbf{x}(\mathbf{s}, t)$. Furthermore, the model yielded the lowest MIAE values for the estimated expected abundance ($\bar{\lambda}_t$) with the highest average coverage probability of the 95% CIs compared to other fitted models (see Table 3.3).

Under the models fitted to data under scenario 4, the discrete-time fused data spatio-temporal model with monthly support using the temporally aggregated $\mathbf{x}(\mathbf{s}, t)$ reported the lowest bias, highest coverage probability of the 95% CIs, and the highest proportion of the times that AIC selected as the best model (see Table 3.2). This is expected since the data generating mechanism in scenario 2 matches the process described by the discrete-time fused data spatio-temporal model with monthly support using the temporally aggregated $\mathbf{x}(\mathbf{s}, t)$. Furthermore, the model yielded the lowest MIAE values for the estimated expected abundance ($\bar{\lambda}_t$) with the highest average coverage probability of the 95% CIs compared to other fitted models (see Table 3.3).

Under all scenarios, the proportion of the times that AIC selected each model as the best model was higher for the models that match or closely match the temporal support in which the data were simulated. However, suppose the models have the same temporal support, one with a temporally aggregated covariate and one without aggregation, AIC may not identify whether the covariate that influences species abundance is temporally aggregated or not.

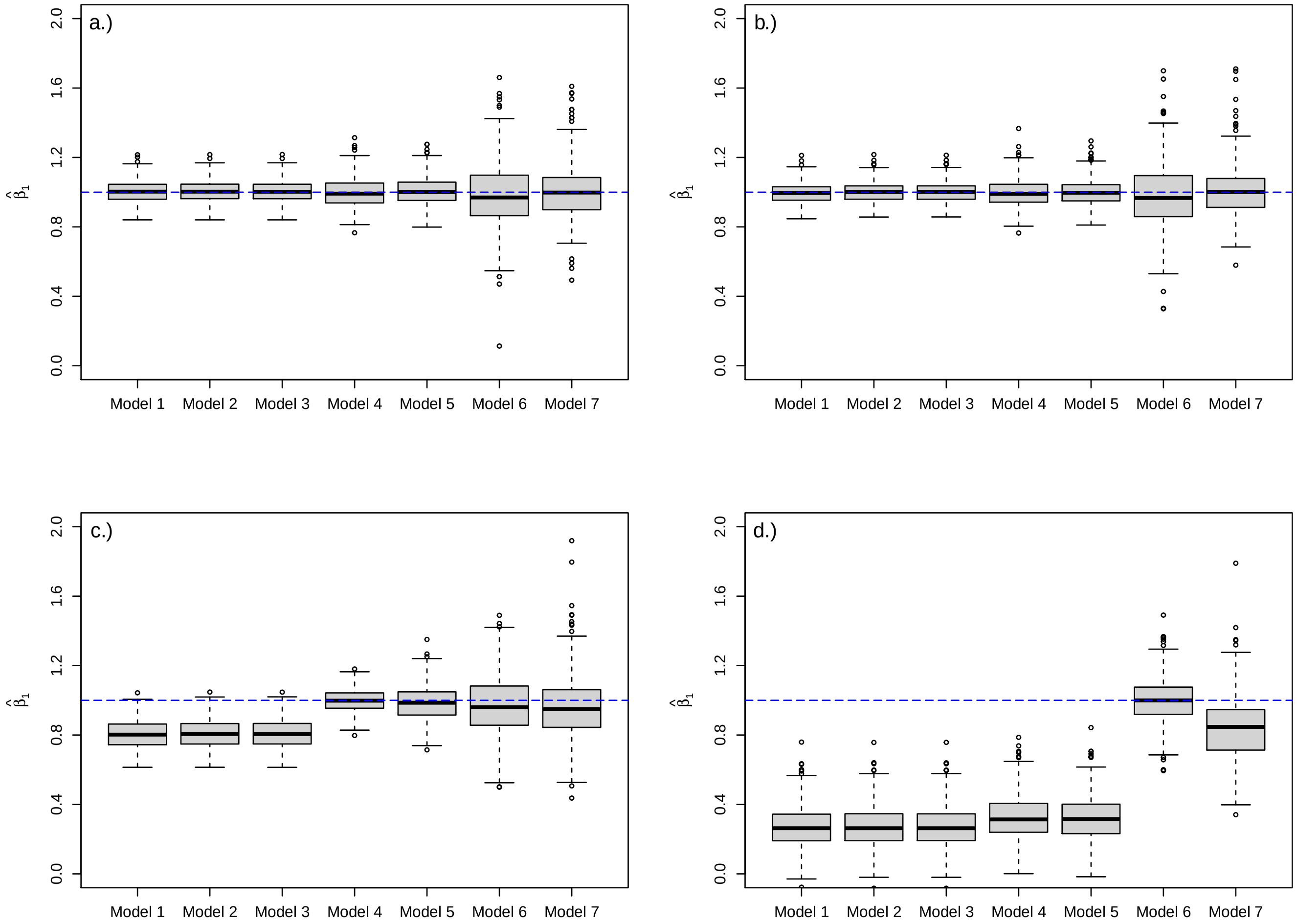

Figure 3.3: The box plots display the estimates for the parameter β_1 obtained from fitting the models to 300 simulated data sets under scenario 1 (panel a), scenario 2 (panel b), scenario 3 (panel c), and scenario 4 (panel d) in the simulation experiment. The fitted models are 1) the continuous-time fused data spatio-temporal model, 2) the discrete-time fused data spatio-temporal model with daily support using a temporally aggregated covariate, 3) the discrete-time fused data spatio-temporal model with daily support using COS, 4) the discrete-time fused data spatio-temporal model with weekly support using a temporally aggregated covariate, 5) the discrete-time fused data spatio-temporal model with weekly support using COS, 6) the discrete-time fused data spatio-temporal model with monthly support using a temporally aggregated covariate, and 7) the discrete-time fused data spatio-temporal model with monthly support using COS. The true value of the parameter β_1 is shown by the blue dash line (–).

Table 3.2: Estimated bias and the coverage probability (CP) for the 95% CI for the parameter β_1 obtained under scenario 1, scenario 2 , scenario 3, and scenario 4, in the simulation experiment. We report the proportion that the Akaike information criterion selected each model as the best model (%AIC). The parameter estimates are obtained by fitting the models to 300 simulated data sets. The highlighted cells represent the case in which the simulated data in the scenario matches the process and the temporal support described by the model.

Scenario	Model 1			Model 2			Model 3			Model 4			Model 5			Model 6			Model 7		
	Bias	CP	%AIC	Bias	CP	%AIC	Bias	CP	%AIC	Bias	CP	%AIC	Bias	CP	%AIC	Bias	CP	%AIC	Bias	CP	%AIC
Scenario 1	0.004	0.937	0.690	0.004	0.937	0.183	0.004	0.937	0.120	0.004	0.847	0	0.005	0.880	0.007	0.017	0.760	0	0.013	0.790	0
Scenario 2	0.004	0.930	0.233	0.002	0.940	0.407	0.002	0.940	0.353	0.003	0.860	0	0.001	0.880	0.007	0.019	0.750	0	0.012	0.787	0
Scenario 3	0.196	0.173	0	0.192	0.180	0.003	0.191	0.180	0	0.001	0.957	0.573	0.014	0.787	0.423	0.026	0.807	0	0.036	0.720	0
Scenario 4	0.729	0	0	0.728	0	0	0.728	0	0	0.678	0	0	0.678	0	0	0.001	0.970	0.887	0.159	0.573	0.113

Table 3.3: Estimated mean integrated absolute error (MIAE), and the average coverage probability (ACP) for the 95% CI for the expected abundance across time ($\bar{\lambda}_t$) obtained under scenario 1, scenario 2, scenario 3, and scenario 4 in the simulation experiment. The estimates for the expected abundance across time ($\bar{\lambda}_t$) are obtained by fitting the models to 300 simulated data sets.The highlighted cells represent the case in which the simulated data in the scenario matches the process and the temporal support described by the model.

Scenario	Model 1		Model 2		Model 3		Model 4		Model 5		Model 6		Model 7	
	MIAE	ACP	MIAE	ACP	MIAE	ACP	MIAE	ACP	MIAE	ACP	MIAE	ACP	MIAE	ACP
Scenario 1	1388.373	0.954	1387.634	0.955	1387.440	0.955	1849.272	0.874	1437.685	0.936	2652.903	0.697	1591.330	0.910
Scenario 2	1608.970	0.915	1428.451	0.946	1428.760	0.946	1860.016	0.874	1478.517	0.941	2643.095	0.701	1646.080	0.906
Scenario 3	3729.425	0.549	3699.968	0.552	3697.117	0.553	1217.996	0.966	1635.285	0.887	2152.429	0.777	1628.700	0.893
Scenario 4	3485.602	0.468	3482.451	0.468	3482.293	0.468	3196.002	0.503	3205.108	0.500	1094.967	0.966	1708.427	0.806

3.6.2 Case study: Grasshopper Sparrows at Konza Prairie Biological Station, Kansas

All five models yielded similar parameter estimates for β_0 with overlapping 95% CIs (see Table 3.4). The estimates for β_1 and the associated 95% CIs obtained from all five models indicated that the species abundance is positively influenced by precipitation (see Table 3.4). Even though the width of the 95% CIs for $\hat{\beta}_1$ across the fitted models were similar, the CIs from the fitted models with monthly temporal support did not overlap with the CIs from the fitted models with continuous or weekly support (see Table 3.4). Therefore the strength of the inferred species' relationship to precipitation was different between the models with monthly temporal support and the models with continuous or weekly support. The discrete-time fused data spatio-temporal model with monthly support using the temporally aggregated $x^*(\mathbf{s}, t)$ reported the lowest AIC value that is $>$200 lower than other fitted models, indicating that using monthly temporal support in the model increases predictive accuracy (see Table 3.4). Even though it is tempting to use AIC to compare the predictive accuracy of the two models with monthly temporal support, our simulation experiment suggested that AIC may not be able to identify the best model between them. Based on the estimates obtained from the discrete-time fused data spatio-temporal model with monthly support using the temporally aggregated $x^*(\mathbf{s}, t)$, the relationship between the estimated expected abundance of Grasshopper Sparrows and the temporally aggregated $x^*(\mathbf{s}, t)$ is visualized in Fig. 3.4f, including the 95% CIs of the estimated expected abundance.

The estimated expected abundance of Grasshopper Sparrows under all five models showed a seasonal variation. Each year, during the breeding season, the estimated expected abundance gradually increased and peaked and gradually decreased towards the latter period of the breeding season (Fig. 3.4). Across the models that use monthly temporal support, the estimated expected abundance in the study area ranged from 205 to 665 with a maximum width of 95% CIs of 200. Across the models that use daily and weekly temporal support, the range was 299 to 862 with a maximum width of 95% CIs of 750.

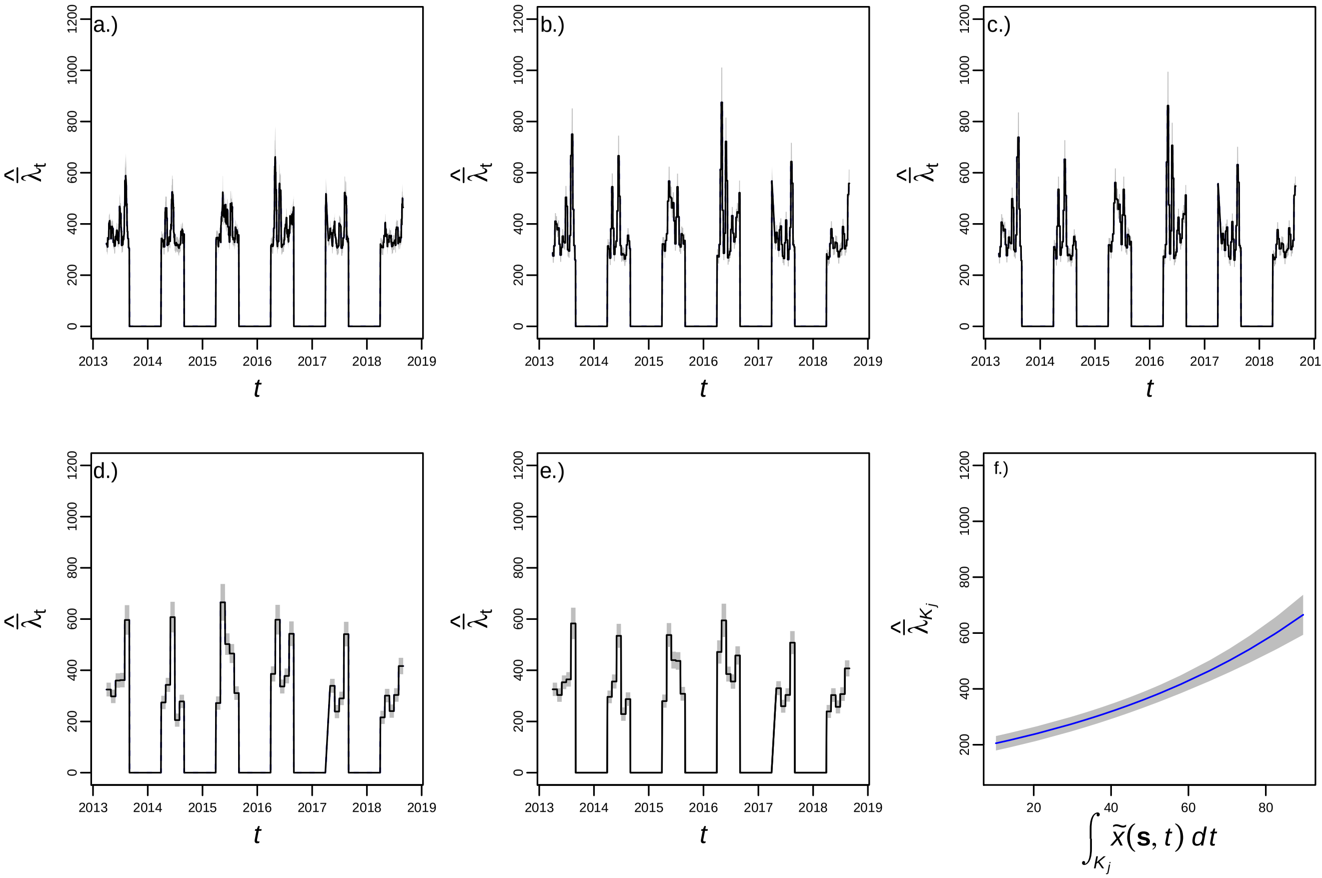

Figure 3.4: Panels (a–e) represents the estimated expected abundance ($\hat{\hat{\lambda}}_t$) in the study region obtained by fitting the models to data on Grasshopper Sparrows at Konza Prairie Biological Station, Kansas. The fitted models are 1) continuous-time fused data spatio-temporal model (panel a), 2) the discrete-time fused data spatio-temporal model with weekly support using the temporally aggregated $x^*(\mathbf{s}, t)$ (panel b), the discrete-time fused data spatio-temporal model with weekly support using COS (panel c), the discrete-time fused data spatio-temporal model with monthly support using the temporally aggregated $x^*(\mathbf{s}, t)$ (panel d), and the discrete-time fused data spatio-temporal model with monthly support using COS (panel e). In panels (a–e), the gray shaded region represents the 95% CIs for $\hat{\hat{\lambda}}_t$. Panel (f) represents the relationship between the estimated expected abundance and the temporally aggregated $x^*(\mathbf{s}, t)$ over monthly scales (i.e., $\int_{K_j} \tilde{x}(\mathbf{s}, t) dt$) and 95% CIs (gray shaded region), obtained from the fitted model that provided the highest predictive accuracy based on the AIC values (i.e., the discrete-time fused data spatio-temporal model with monthly support using the temporally aggregated $x^*(\mathbf{s}, t)$). The K_j represents the j^{th} monthly partition obtained by partitioning the total study period by 30 months (i.e., j=1, 2, ..., 30 months).

Table 3.4: Maximum likelihood estimates and the width of the 95% CIs for β_0, β_1, and the expected abundance ($\bar{\lambda}$) from fitting the models: 1) continuous-time fused data spatio-temporal model, 2) discrete-time fused data spatio-temporal model with weekly support using the temporally aggregated $\tilde{x}(\mathbf{s}, t)$, 3) discrete-time fused data spatio-temporal model with weekly support using COS, 4) discrete-time fused data spatio-temporal model with monthly support using the temporally aggregated $\tilde{x}(\mathbf{s}, t)$, and 5) discrete-time fused data spatio-temporal model with monthly support using COS to data on Grasshopper Sparrows at Konza Prairie Biological Station in north-eastern Kansas, USA. The covariate under consideration is the integrated precipitation over the prior two weeks (i.e., $\tilde{x}(\mathbf{s}, t)$). Also, AIC values are reported for the five models.

Models	β_0		β_1		AIC
	$\hat{\beta}_0$	Width of 95% CI	$\hat{\beta}_1$	Width of 95% CI	
Continuous-time fused data spatio-temporal model (Fig. 3.4a)	-8.586	0.148	0.007	0.002	59418
Discrete-time fused data spatio-temporal model with weekly support using the temporally aggregated $\tilde{x}(\mathbf{s}, t)$ (Fig. 3.4b)	-8.607	0.150	0.009	0.002	59245
Discrete-time fused data spatio-temporal model with weekly support using COS (Fig. 3.4c)	-8.631	0.151	0.009	0.002	59277
Discrete-time fused data spatio-temporal model with monthly support using the temporally aggregated $\tilde{x}(\mathbf{s}, t)$ (Fig. 3.4d)	-8.638	0.153	0.018	0.003	58923
Discrete-time fused data spatio-temporal model with monthly support using COS (Fig. 3.4e)	-8.786	0.165	0.015	0.002	59127

3.7 Discussion

Ecological processes evolve with time and unfold at vastly different time scales (Carroll et al., 2007). Species-habitat relationships and species abundance are strongly influenced by multiscale ecological processes (Fink et al., 2014). The temporal scale at which species interact with the ecological processes is especially important to study in seasonally fluctuating species populations as they continuously interact with the ecological processes across small to large time scales. Though spatio-temporal point process models are vastly used to model species-habitat relationships and abundance across space and time (Hefley & Hooten, 2016; Renner et al., 2015), the studies seldom discuss the impact of temporal support on the robustness of the models. Our study provides two important contributions; 1) propose a spatio-temporal point process modeling framework using a data fusion approach that enables modeling of the species-habitat relationship and abundance across finer to coarse temporal scales, and thereby 2) study the influence of the temporal support in spatio-temporal point process models to model species- habitat relationships and estimate abundance.

Our study indicates that the parameter estimates associated with the species-habitat relationship (i.e., β_1) is sensitive to the temporal support in the model and may provide biased parameter estimates. Therefore, misspecified temporal support may provide misleading inferences on species-habitat relationships. Moreover, since the estimate for the expected abundance is a function of $\hat{\beta}_1$, misspecified temporal support may yield biased estimates for the estimated expected abundance. Therefore, caution is necessary when defining the temporal support in the model to examine species-habitat relationships and the abundance of a species. Our study indicates that AIC adequately identifies the temporal scale at which the covariate influences species-habitat relationships and abundance. However, under the same temporal support in the model, AIC may not be able to identify whether the covariate that influences species abundance is temporally aggregated or not.

Moreover, we discovered that if the true data generating process is on a high temporal resolution (i.e., covariates drive species abundance in finer temporal scales), there appears to be not much penalty paid for misspecification of the temporal support. In this case, the fitted model with misspecified temporal support still may yield unbiased estimates for species habitat relationships and expected abundance, but with poor efficiency. However, suppose the true data generating process is on coarser temporal resolution (i.e., covariates drive species abundance in coarser temporal scales), a higher penalty is paid for misspecification of the temporal support. In this case, the fitted model with misspecified temporal support may yield biased estimates misleading the inferences on species habitat relationship and expected abundance. Such asymmetry in misspecification of the temporal support suggests that practitioners should pay more attention to the temporal support of the models if they suspect that the covariates drive species abundance in coarser temporal scales.

3.7.1 Future directions

Future directions of our work involves investigating the influence of spatio-temporal support in spatio-temporal autocorrelation. Spatio-temporal autocorrelation may include traditional patterns such as correlated normally distributed effects as well as non-traditional patterns such as discontinuities and abrupt transitions (Mohankumar & Hefley, 2021a). These patterns may operate in multiple spatio-temporal scales, and studying the influence of varying spatial and temporal support on spatio-temporal autocorrelation may provide further insight into understanding species dynamics. In fact, models can be developed to account for the multi-scale spatial dependence by incorporating scientific knowledge of the spatio-temporal data generating process (Hefley et al., 2017b).

Our study investigates the influence of temporal support across continuous and discrete temporal scales using continuous-time and discrete-time intensity functions using temporally aggregated covariates and COS. The continuous intensity function may also be defined by incorporating lagged effects. For example, the continuous intensity function can be defined as, $\lambda(\mathbf{s}, t) = e^{\mathbf{x}(\mathbf{s}, t-\theta)'\boldsymbol{\beta}}$ or $\lambda(\mathbf{s}, t) = \lambda(\mathbf{s}, t-\theta)e^{\mathbf{x}(\mathbf{s}, t)'\boldsymbol{\beta}}$ where the intensity at t may depend on ecological processes at previous states. Furthermore, model-based approaches can be developed to estimate the time lag θ. Future directions may involve including kernel functions to weight the spatio-temporal covariate surface (Hooten & Johnson, 2017; Heaton & Gelfand, 2011, 2012). Such extensions would enable investigating more complex species-habitat relationships and ecological processes.

Acknowledgements

I thank Dr. Trevor Hefley and Dr. Alice Boyle, who are co-authors of the manuscript produced from this chapter. I thank all individuals who contributed to the Grasshopper Sparrows data used in our data example. This material is based upon work supported by the National Science Foundation under Grant No. 1754491 and the Konza Prairie Long-Term Ecological Research (LTER) Grant No. 1440484. This work was permitted by the US Geological Survey's Bird Banding Lab (23836), the Kansas Department of Parks, Wildlife, and Tourism, and the Kansas State University Institutional Animal Care and Use Committee (IACUC) (protocol 4250).

Chapter 4

Conclusion

4.1 Summary of dissertation

Advancing the types of models to model species distributions across space and time involves properly accounting for the limitations associated with the data types used in species distribution studies and addressing limitations in existing modeling approaches that limit the ability to model species distributions reliably. In this dissertation, I studied and addressed some of the most pressing limitations associated with existing modeling approaches for the three most common planned survey data types that are used in species distribution studies: 1) occupancy data, 2) distance sampling data, and 3) capture-recapture data.

In chapter 1, I proposed a hierarchical modeling framework for occupancy data that simultaneously accounts for false absences and accounts for both traditional and non-traditional spatial dependence. Models for occupancy data are used to estimate and map the true presence of a species that typically involves observation errors such as false absences Hepler et al., 2018; Joseph, 2020. Furthermore, researchers often account for spatial dependence in occupancy data by using a correlated, normally distributed site-level random effect (Johnson et al., 2013a), which can account for traditional spatial dependence but might be incapable of modeling non-traditional spatial dependence such as discontinuities and abrupt transitions. This lack of spatial modeling options for occupancy data is in contrast to species distribution models (SDM) that predict the spatial distribution of a species using statistical, and machine learning approaches applied to presence-only, count, and presence-absence data De'ath & Fabricius, 2000; Cutler et al., 2007; Elith et al., 2008. There is a bewildering number of approaches within the SDM literature that are used to model the spatial process. Unfortunately, many of the SDM approaches do not account for false absences in data De'ath & Fabricius, 2000; Cutler et al., 2007; Elith et al., 2008. The proposed approach in this chapter incorporates machine learning approaches into a Bayesian hierarchal modeling framework that can simultaneously account for false absences and account for traditional and non-traditional spatial dependence to estimate and map the true presence of a species reliably. Furthermore, the modeling framework enables any appropriate machine earning approach to be used to model the spatial dependence expanding the types of spatial models practitioners can use for data that fit within the occupancy model framework.

In chapter 2, I propose a fused data modeling approach that combines DS and CR to model species distribution. DS and CR data are high-quality planned survey data (Otis et al., 1978; Burnham et al., 1980; Seber, 1982), but they are seldom used in SDMs due to lack of spatial coverage (McShea et al., 2016). I combine the two data sources using a hierarchical modeling framework that increases spatial coverage, reduces parameter uncertainty, and makes predictions more accurate (Hooten & Hefley, 2019); therefore, it can be used for species distribution modeling. Furthermore, the construction of an adequate fused data SDM for DS and CR data relies upon accounting for missing data issues (Little & Rubin, 2019) unique to each data source. I account for the missing data issues by building models for the missing data mechanism using theory and tools from the missing data literature. I account for two most common missing data issues in DS and CR data: 1) missing individuals that are missing not at random (MNAR) (Little & Rubin, 2019) and 2) partially missing location information. The proposed modeling approach significantly increased the reliability and efficiency of the parameter estimates for species-habitat relationship and expected abundance compared to existing modeling approaches. Furthermore, our modeling framework can be viewed as a unified framework that can be applied to many other data sources (e.g., presence-only data) and a fusion of them addressing critical issues with missing data, advancing the types of models used for species distribution studies.

Finally, in chapter 3, I extend my data fusion approach in chapter 2 to a spatio-temporal modeling framework to investigate the influence of temporal support in spatio-temporal point process models to model species distribution. The temporal dynamics of ecological processes are inherently complex (Fink et al., 2014). Their influence on species-habitat relationships and abundance operates on multiple spatio-temporal scales (Fink et al., 2014); however, the robustness of models to changing temporal scales is rarely studied. The proposed approach in chapter 3 enabled the modeling of the species-habitat relationship and abundance across finer to coarse temporal scales. Therefore the study enables the investigation of the influence of the temporal support in spatio-temporal point process models to model species- habitat relationships and estimate expected abundance.

4.2 Future research

Extensions of the occupancy modeling framework

Although the proposed occupancy modeling framework in chapter 1 is focused on spatial dependence, the model can be extended to use machine learning approaches to capture non-linear and unknown relationships between covariates and the probability of occupancy. Since machine learning approaches are often superior in dealing with a large number of covariates, such extension increases the ability to incorporate large sets of site-level covariates into the model and identify complex relationships Shaby & Fink (2012). The model framework can also be extended to use machine learning approaches to model the probability of detection as a function of predictor variables such as Julian date and observer effort (e.g., similar to the use of cubic splines used by Johnston et al., 2018). Another future direction is that the proposed spatial occupancy model can be extended to account for other types of observation errors in occupancy data, such as false positives (Hooten & Hefley, 2019). This involves modifying the distributional representation for observed data presented in chapter 1.

Account for the confounding of parameters in the detection/capture functions with the parameters in the intensity function

The model framework proposed for the data fusion of DS and CR data in both chapter 2 and chapter 3 does not investigate the confounding between parameters in the detection/capture functions and the underlying intensity function. However, such confounding is possible (Borchers et al., 2006). For example, the underlying intensity and the probability of detection can be a function of the same covariate. Accounting for such confounding of the underlying intensity and the detection/capture probability is an area that needs further research.

Approaches to account for missing spatial covariate values

The proposed model frameworks in both chapter 2 and chapter 3 require spatial covariate values for the entire study region. However, obtaining the spatial covariate values in the entire study region can be trivial in some situations. Some alternate solutions would be to employ an auxiliary model to utilize the available data to predict the spatial covariate values for the entire region and use the predicted values as the input values for the spatial covariate in our models (Hefley et al., 2020). However, further research is needed to account for such missing spatial covariate values explicitly.

Extensions of the fused data framework to identify the factors that influence the probability of detection or capture.

The probability of detection or capture in DS and CR data can be influenced by many covariates, such as the observer's experience level or environmental or geographical features. In the models presented in this dissertation, the probability of detection is defined as only depending on the distance from the transect to the individual, and the probability of capture is defined using a constant parameter. The functions for the probability of detection and capture can be extended to investigate factors that influence the probability of detection or capture.

Investigate the influence of spatio-temporal support in spatio-temporal autocorrelation

The proposed occupancy modeling framework enables the use of machine learning techniques to model complex patterns in spatial dependence. However, the data fusion framework proposed in chapter 2 and chapter 3 does not incorporate a spatial or a spatio-temporal random effect. A bewildering number of approaches within the SDM literature are developed to model the spatial and spatio-temporal autocorrelation (e.g., Chakraborty et al., 2011; Renner et al., 2015; Mohankumar & Hefley, 2021a), which can be used to incorporate a spatial or a spatio-temporal random effect into the proposed fused data modeling framework. However, spatio-temporal autocorrelation may include traditional patterns such as correlated normally distributed effects as well as non-traditional patterns such as discontinuities and abrupt transitions (Mohankumar & Hefley, 2021a) as mentioned in chapter 1. These patterns may operate in multiple spatio-temporal scales, and studying the influence of varying spatial and temporal support on spatio-temporal autocorrelation may provide further insight into understanding species dynamics.

Investigate the influence of temporal support on the models' robustness using kernel functions

The study in this dissertation focuses on investigating the influence of temporal support on the models' robustness across continuous and discrete temporal scales using continuous-time and discrete-time intensity functions. However, the intensity function can be defined using kernel functions to study the influence of more complex species-habitat relationships and ecological processes across time. Kernel functions can be used to weight the spatio-temporal covariate surface (Hooten & Johnson, 2017; Heaton & Gelfand, 2011, 2012) providing further insight into understanding species dynamics.

Bibliography

Aarts, G., Fieberg, J., & Matthiopoulos, J. (2012). Comparative interpretation of count, presence–absence and point methods for species distribution models. *Methods in Ecology and Evolution*, *3*, 177–187.

Albert, J. H., & Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, *88*, 669–679.

Allen, M., Webb, M. H., Alves, F., Heinsohn, R., & Stojanovic, D. (2018). Occupancy patterns of the introduced, predatory sugar glider in Tasmanian forests. *Austral Ecology*, *43*, 470–475.

Andrews, J. E., Brawn, J. D., & Ward, M. P. (2015). When to use social cues: Conspecific attraction at newly created grasslands. *The Condor: Ornithological Applications*, *117*, 297–305.

Araujo, M. B., & Guisan, A. (2006). Five (or so) challenges for species distribution modelling. *Journal of Biogeography*, *33*, 1677–1688.

Borchers, D., Laake, J., Southwell, C., & Paxton, C. (2006). Accommodating unmodeled heterogeneity in double-observer distance sampling surveys. *Biometrics*, *62*, 372–378.

Borchers, D. L., Stevenson, B., Kidney, D., Thomas, L., & Marques, T. A. (2015). A unifying model for capture–recapture and distance sampling surveys of wildlife populations. *Journal of the American Statistical Association*, *110*, 195–204.

Boyle, A. (2019). Cbp01 variable distance line-transect sampling of bird population numbers in different habitats on konza prairie. https://doi.org/10.6073/pasta/053fe6a82e54394a70ff22b4794c0489. Accessed: 2021-12-16.

Boyle, W. A., Shogren, E. H., & Brawn, J. D. (2020). Hygric niches for tropical endotherms. *Trends in Ecology & Evolution*, *35*, 938–952.

Brennan, L. A., & Kuvlesky Jr, W. P. (2005). North american grassland birds: an unfolding conservation crisis? *The Journal of Wildlife Management*, *69*, 1–13.

Buckland, S. T., Anderson, D. R., Burnham, K. P., Laake, J. L., Borchers, D. L., & Thomas, L. (2001). *Introduction to distance sampling: estimating abundance of biological populations*. Oxford, United Kingdom: Oxford University Press.

Buckland, S. T., Anderson, D. R., Burnham, K. P., Laake, J. L., Borchers, D. L., & Thomas, L. (2004). *Advanced distance sampling: estimating abundance of biological populations*. New York: Oxford University Press.

Burnham, K. P., & Anderson, D. R. (1984). The need for distance data in transect counts. *The Journal of Wildlife Management*, *48*, 1248–1254.

Burnham, K. P., Anderson, D. R., & Laake, J. L. (1980). Estimation of density from line transect sampling of biological populations. *Wildlife Monographs*, *72*, 3–202.

Carrijo, T. B., & da Silva, A. R. (2017). Modified Moran's I for small samples. *Geographical Analysis*, *49*, 451–467.

Carroll, S. P., Hendry, A. P., Reznick, D. N., & Fox, C. W. (2007). Evolution on ecological time-scales. *Functional Ecology*, *21*, 387–393.

Chakraborty, A., Gelfand, A. E., Wilson, A. M., Latimer, A. M., & Silander, J. A. (2011). Point pattern modelling for degraded presence-only data over large regions. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *60*, 757–776.

Chandler, R. B., Hepinstall-Cymerman, J., Merker, S., Abernathy-Conners, H., & Cooper, R. J. (2018). Characterizing spatio-temporal variation in survival and recruitment with integrated population models. *The Auk: Ornithological Advances*, *135*, 409–426.

Colwell, R. K., & Rangel, T. F. (2009). Hutchinson's duality: the once and future niche. *Proceedings of the National Academy of Sciences*, *106*, 19651–19658.

Coppedge, B. R., Engle, D. M., Masters, R. E., & Gregory, M. S. (2001). Avian response to landscape change in fragmented southern great plains grasslands. *Ecological Applications*, *11*, 47–59.

Cressie, N. (1991). *Statistics for spatial data*. New York: John Wiley & Sons.

Cressie, N., & Wikle, C. K. (2015). *Statistics for spatio-temporal data*. New Jersey: John Wiley & Sons.

Cutler, D. R., Edwards Jr, T. C., Beard, K. H., Cutler, A., Hess, K. T., Gibson, J., & Lawler, J. J. (2007). Random forests for classification in ecology. *Ecology*, *88*, 2783–2792.

Darwin, C. (1859). *On the origin of species by means of natural selection*. London, United Kingdom: John Murray.

De'ath, G., & Fabricius, K. E. (2000). Classification and regression trees: a powerful yet simple technique for ecological data analysis. *Ecology*, *81*, 3178–3192.

Derksen, D. A., Lafond, G. P., Thomas, A. G., Loeppky, H. A., & Swanton, C. J. (1993). Impact of agronomic practices on weed communities: tillage systems. *Weed Science*, *41*, 409–417.

Dickinson, J. L., Zuckerberg, B., & Bonter, D. N. (2010). Citizen science as an ecological research tool: challenges and benefits. *Annual Review of Ecology, Evolution, and Systematics*, *41*, 149–172.

Diggle, P. J., Besag, J., & Gleaves, J. T. (1976). Statistical analysis of spatial point patterns by means of distance methods. *Biometrics*, *32*, 659–667.

Dorazio, R. M. (2012). Predicting the geographic distribution of a species from presence-only data subject to detection errors. *Biometrics*, *68*, 1303–1312.

Dorazio, R. M. (2014). Accounting for imperfect detection and survey bias in statistical analysis of presence-only data. *Global Ecology and Biogeography*, *23*, 1472–1484.

Dorazio, R. M., & Rodriguez, D. T. (2012). A gibbs sampler for Bayesian analysis of site-occupancy data. *Methods in Ecology and Evolution*, *3*, 1093–1098.

Elith, J., & Leathwick, J. R. (2009). Species distribution models: ecological explanation and prediction across space and time. *Annual Review of Ecology, Evolution, and Systematics*, *40*, 677–697.

Elith, J., Leathwick, J. R., & Hastie, T. (2008). A working guide to boosted regression trees. *Journal of Animal Ecology*, *77*, 802–813.

Farr, M. T. (2021). *Data integration in population and community ecology using hierarchical modeling*. Ph.D. thesis, Michigan State University. https://www.proquest.com/dissertations-theses/data-integration-population-community-ecology/docview/2593648589/se-2?accountid=11789.

Farr, M. T., Green, D. S., Holekamp, K. E., & Zipkin, E. F. (2020). Integrating distance sampling and presence-only data to estimate species abundance. *Ecology*, *102*, e03204.

Fink, D., Damoulas, T., Bruns, N. E., La Sorte, F. A., Hochachka, W. M., Gomes, C. P., & Kelling, S. (2014). Crowdsourcing meets ecology: hemisphere-wide spatiotemporal species distribution models. *AI Magazine*, *35*, 19–30.

Fithian, W., Elith, J., Hastie, T., & Keith, D. A. (2015). Bias correction in species distribution models: pooling survey and collection data for multiple species. *Methods in Ecology and Evolution*, *6*, 424–438.

Fletcher, R. J., Hefley, T. J., Robertson, E. P., Zuckerberg, B., McCleery, R. A., & Dorazio, R. M. (2019). A practical guide for combining data to model species distributions. *Ecology*, *100*, e02710.

Geary, R. C. (1954). The contiguity ratio and statistical mapping. *The Incorporated Statistician*, *5*, 115–146.

Gelfand, A. E., & Schliep, E. M. (2018). Bayesian inference and computing for spatial point patterns. *NSF-CBMS Regional Conference Series in Probability and Statistics*, *10*, 1–125.

Gerber, B. D., Karpanty, S. M., & Kelly, M. J. (2012). Evaluating the potential biases in carnivore capture–recapture studies associated with the use of lure and varying density estimation techniques using photographic-sampling data of the malagasy civet. *Population Ecology*, *54*, 43–54.

Golding, N., & Purse, B. V. (2016). Fast and flexible Bayesian species distribution modelling using Gaussian processes. *Methods in Ecology and Evolution*, *7*, 598–608.

Graham, C. H., Ferrier, S., Huettman, F., Moritz, C., & Peterson, A. T. (2004). New developments in museum-based informatics and applications in biodiversity analysis. *Trends in Ecology & Evolution*, *19*, 497–503.

Guisan, A., & Thuiller, W. (2005). Predicting species distribution: offering more than simple habitat models. *Ecology Letters*, *8*, 993–1009.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: data mining, inference, and prediction*. New York: Springer.

Heaton, M. J., Datta, A., Finley, A. O., Furrer, R., Guinness, J., Guhaniyogi, R., Gerber, F., Gramacy, R. B., Hammerling, D., Katzfuss, M. et al. (2019). A case study competition among methods for analyzing large spatial data. *Journal of Agricultural, Biological and Environmental Statistics*, *24*, 398–425.

Heaton, M. J., & Gelfand, A. E. (2011). Spatial regression using kernel averaged predictors. *Journal of Agricultural, Biological, and Environmental Statistics*, *16*, 233–252.

Heaton, M. J., & Gelfand, A. E. (2012). Kernel averaged predictors for spatio-temporal regression models. *Spatial Statistics*, *2*, 15–32.

Hefley, T. J., Baasch, D. M., Tyre, A. J., & Blankenship, E. E. (2014). Correction of location errors for presence-only species distribution models. *Methods in Ecology and Evolution*, *5*, 207–214.

Hefley, T. J., Baasch, D. M., Tyre, A. J., & Blankenship, E. E. (2015). Use of opportunistic sightings and expert knowledge to predict and compare whooping crane stopover habitat. *Conservation Biology*, *29*, 1337–1346.

Hefley, T. J., Boyle, W. A., & Mohankumar, N. M. (2020). Accounting for location uncertainty in distance sampling data. arXiv:2005.14316.

Hefley, T. J., Brost, B. M., & Hooten, M. B. (2017a). Bias correction of bounded location errors in presence-only data. *Methods in Ecology and Evolution*, *8*, 1566–1573.

Hefley, T. J., & Hooten, M. B. (2016). Hierarchical species distribution models. *Current Landscape Ecology Reports*, *1*, 87–97.

Hefley, T. J., Hooten, M. B., Hanks, E. M., Russell, R. E., & Walsh, D. P. (2017b). Dynamic spatio-temporal models for spatial data. *Spatial Statistics*, *20*, 206–220.

Hefley, T. J., Tyre, A. J., Baasch, D. M., & Blankenship, E. E. (2013). Nondetection sampling bias in marked presence-only data. *Ecology and Evolution*, *3*, 5225–5236.

Heithaus, E. R., Fleming, T. H., & Opler, P. A. (1975). Foraging patterns and resource utilization in seven species of bats in a seasonal tropical forest. *Ecology*, *56*, 841–854.

Hepler, S. A., Erhardt, R., & Anderson, T. M. (2018). Identifying drivers of spatial variation in occupancy with limited replication camera trap data. *Ecology*, *99*, 2152–2158.

Herse, M. R., With, K. A., & Boyle, W. A. (2018). The importance of core habitat for a threatened species in changing landscapes. *Journal of Applied Ecology*, *55*, 2241–2252.

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

Hooten, M. B., & Hefley, T. J. (2019). *Bringing Bayesian models to life*. Florida: Chapman & Hall/CRC Press.

Hooten, M. B., & Hobbs, N. T. (2015). A guide to Bayesian model selection for ecologists. *Ecological Monographs*, *85*, 3–28.

Hooten, M. B., & Johnson, D. S. (2017). Basis function models for animal movement. *Journal of the American Statistical Association*, *112*, 578–589.

Humboldt, A. v., Bonpland, A. et al. (1805). *Analysis of the geography of plants*. Chez Levrault, Scoell et Campagnie, Libraires, Paris.

Hutchinson, R. A., Liu, L.-P., & Dietterich, T. G. (2011). Incorporating boosted regression trees into ecological latent variable models. *Proceedings of the AAAI Conference on Artificial Intelligence*, *25*, 1343–1348.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning*. New York: Springer.

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

Johnson, D. S., Hooten, M. B., & Kuhn, C. E. (2013b). Estimating animal resource selection from telemetry data using point process models. *Journal of Animal Ecology*, *82*, 1155–1164.

Johnson, D. S., Laake, J. L., & Ver Hoef, J. M. (2010). A model-based approach for making ecological inference from distance sampling data. *Biometrics*, *66*, 310–318.

Johnston, A., Fink, D., Hochachka, W. M., & Kelling, S. (2018). Estimates of observer expertise improve species distributions from citizen science data. *Methods in Ecology and Evolution*, *9*, 88–97.

Joseph, M. B. (2020). Neural hierarchical models of ecological populations. *Ecology Letters*, *23*, 734–747.

Kéry, M. (2011). Towards the modelling of true species distributions. *Journal of Biogeography*, *38*, 617–618.

Kéry, M., & Royle, J. A. (2015). *Applied Hierarchical Modeling in Ecology: Analysis of distribution, abundance and species richness in R and BUGS: Volume 1: Prelude and Static Models*. London, United Kingdom: Academic Press.

Knapp, A. K., Briggs, J. M., Hartnett, D. C., & Collins, S. L. (1998). *Grassland dynamics long-term ecological research in tallgrass prairie*. New York: Oxford University Press.

Koshkina, V., Wang, Y., Gordon, A., Dorazio, R. M., White, M., & Stone, L. (2017). Integrated species distribution models: combining presence-background data and site-occupancy data with imperfect detection. *Methods in Ecology and Evolution*, *8*, 420–430.

Le Maho, Y., Whittington, J. D., Hanuise, N., Pereira, L., Boureau, M., Brucker, M., Chatelain, N., Courtecuisse, J., Crenner, F., Friess, B. et al. (2014). Rovers minimize human disturbance in research on wild animals. *Nature Methods*, *11*, 1242–1244.

Lin, Y.-C., Chang, L.-W., Yang, K.-C., Wang, H.-H., Sun, I.-F. et al. (2011). Point patterns of tree distribution determined by habitat heterogeneity and dispersal limitation. *Oecologia*, *165*, 175–184.

Linnaeus, C. (1781). On the increase of the habitable earth. *Amoenitates Academicae*, *2*, 17–27.

Little, R. J. (1992). Regression with missing x's: a review. *Journal of the American statistical association*, *87*, 1227–1237.

Little, R. J., & Rubin, D. B. (2019). *Statistical analysis with missing data*. New Jersey: John Wiley & Sons.

Lomolino, M. V., Riddle, B. R., & Brown, J. H. (2006). *Biogeography*. Sinauer Associates, Inc.

Macías-Duarte, A., Panjabi, A. O., Strasser, E. H., Levandoski, G. J., Ruvalcaba-Ortega, I., Doherty, P. F., & Ortega-Rosas, C. I. (2017). Winter survival of north american grassland birds is driven by weather and grassland condition in the chihuahuan desert. *Journal of Field Ornithology*, *88*, 374–386.

MacKenzie, D. I. (2005). What are the issues with presence-absence data for wildlife managers? *The Journal of Wildlife Management*, *69*, 849–860.

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

Mappes, J., Marples, N., & Endler, J. A. (2005). The complex business of survival by aposematism. *Trends in Ecology & Evolution*, *20*, 598–603.

Martiny, J. B. H., Bohannan, B. J., Brown, J. H., Colwell, R. K., Fuhrman, J. A., Green, J. L., Horner-Devine, M. C., Kane, M., Krumins, J. A., Kuske, C. R. et al. (2006). Microbial biogeography: putting microorganisms on the map. *Nature Reviews Microbiology*, *4*, 102–112.

Mason, A., Richardson, S., Plewis, I., & Best, N. (2012). Strategy for modelling nonrandom missing data mechanisms in observational studies using Bayesian methods. *Journal of Official Statistics*, *28*, 279–302.

Mayor, S. J., Schneider, D. C., Schaefer, J. A., & Mahoney, S. P. (2009). Habitat selection at multiple scales. *Ecoscience*, *16*, 238–247.

McGarigal, K., Wan, H. Y., Zeller, K. A., Timm, B. C., & Cushman, S. A. (2016). Multi-scale habitat selection modeling: a review and outlook. *Landscape Ecology*, *31*, 1161–1175.

McShea, W. J., Forrester, T., Costello, R., He, Z., & Kays, R. (2016). Volunteer-run cameras as distributed sensors for macrosystem mammal research. *Landscape Ecology*, *31*, 55–66.

Michener, W. K., & Jones, M. B. (2012). Ecoinformatics: supporting ecology as a data-intensive science. *Trends in Ecology & Evolution*, *27*, 85–93.

Mohankumar, N. M., & Hefley, T. J. (2021a). Using machine learning to model nontraditional spatial dependence in occupancy data. *Ecology*, *103*, e03563.

Mohankumar, N. M., & Hefley, T. J. (2021b). Using machine learning to model non-traditional spatial dependence in occupancy data. Dryad Digital Repository, https://doi.org/10.5061/dryad.4xgxd259g.

Mohankumar, N. M., Hefley, T. J., Silber, K., & Boyle, W. A. (2022). Data fusion of distance sampling and capture-recapture data. arXiv:2203.03960.

Moran, P. A. (1948). The interpretation of statistical maps. *Journal of the Royal Statistical Society. Series B (Methodological)*, *10*, 243–251.

Moran, P. A. (1950). Notes on continuous stochastic phenomena. *Biometrika*, *37*, 17–23.

Murphy, K. P. (2012). *Machine learning: a probabilistic perspective*. MIT press.

Ostfeld, R. S. (1985). Limiting resources and territoriality in microtine rodents. *The American Naturalist*, *126*, 1–15.

Otis, D. L., Burnham, K. P., White, G. C., & Anderson, D. R. (1978). Statistical inference from capture data on closed animal populations. *Wildlife Monographs*, *62*, 3–135.

Pearce, J. L., & Boyce, M. S. (2006). Modelling distribution and abundance with presence-only data. *Journal of Applied Ecology*, *43*, 405–412.

Phillips, S. J., Anderson, R. P., & Schapire, R. E. (2006). Maximum entropy modeling of species geographic distributions. *Ecological Modelling*, *190*, 231–259.

Pollock, K. H., Nichols, J. D., Brownie, C., & Hines, J. E. (1990). Statistical inference for capture-recapture experiments. *Wildlife Monographs*, *107*, 3–97.

Pressey, R. L., Cabeza, M., Watts, M. E., Cowling, R. M., & Wilson, K. A. (2007). Conservation planning in a changing world. *Trends in Ecology & Evolution*, *22*, 583–592.

Ramon, P., Velazquez, E., Escudero, A., & de la Cruz, M. (2018). Environmental heterogeneity blurs the signature of dispersal syndromes on spatial patterns of woody species in a moist tropical forest. *PloS One*, *13*, e0192341.

Rapacciuolo, G., & Blois, J. L. (2019). Understanding ecological change across large spatial, temporal and taxonomic scales: integrating data and methods in light of theory. *Ecography*, *42*, 1247–1266.

Renner, I. W., Elith, J., Baddeley, A., Fithian, W., Hastie, T., Phillips, S. J., Popovic, G., & Warton, D. I. (2015). Point process models for presence-only analysis. *Methods in Ecology and Evolution*, *6*, 366–379.

Reynolds, A. M., Sword, G. A., Simpson, S. J., & Reynolds, D. R. (2009). Predator percolation, insect outbreaks, and phase polyphenism. *Current Biology*, *19*, 20–24.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*, 581–592.

Seber, G. A. F. (1982). *The estimation of animal abundance and related parameters*. New York: Macmillan.

Shaby, B. A., & Fink, D. (2012). Embedding black-box regression techniques into hierarchical Bayesian models. *Journal of Statistical Computation and Simulation*, *82*, 1753–1766.

Shaffer, J. A., Igl, L. D., Johnson, D. H., Sondreal, M. L., Goldade, C. M., Nenneman, M. P., Wooten, T. L., & Euliss, B. R. (2021). *The effects of management practices on grassland birds—Grasshopper Sparrow (Ammodramus savannarum)*. U.S. Geological Survey. https://doi.org/10.3133/pp1842GG.

Strebel, N., Kéry, M., Guélat, J., & Sattler, T. (2022). Spatiotemporal modelling of abundance from multiple data sources in an integrated spatial distribution model. *Journal of Biogeography*, .

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

Vázquez, F. J., Acea, M. J., & Carballas, T. (1993). Soil microbial populations after wildfire. *FEMS Microbiology Ecology*, *13*, 93–103.

Walker, N. B., Hefley, T. J., & Walsh, D. P. (2020). Bias correction of bounded location error in binary data. *Biometrics*, *76*, 530–539.

Warton, D. I., & Shepherd, L. C. (2010). Poisson point process models solve the" pseudo-absence problem" for presence-only data in ecology. *The Annals of Applied Statistics*, *4*, 1383–1402.

West, A. S., Keyser, P. D., Lituma, C. M., Buehler, D. A., Applegate, R. D., & Morgan, J. (2016). Grasslands bird occupancy of native warm-season grass. *The Journal of Wildlife Management*, *80*, 1081–1090.

Whittle, P. (1954). On stationary processes in the plane. *Biometrika*, *41*, 434–449.

Wikle, C. K. (2019). Comparison of deep neural networks and deep hierarchical models for spatio-temporal data. *Journal of Agricultural, Biological and Environmental Statistics*, *24*, 175–203.

Williams, E. J., & Boyle, W. A. (2018). Patterns and correlates of within-season breeding dispersal: A common strategy in a declining grassland songbird. *The Auk: Ornithological Advances*, *135*, 1–14.

Williams, E. J., & Boyle, W. A. (2019). Causes and consequences of avian within-season dispersal decisions in a dynamic grassland environment. *Animal Behaviour*, *155*, 77–87.

Winnicki, S., Munguía, S., Williams, E., & Boyle, W. (2020). Social interactions do not drive territory aggregation in a grassland songbird. *Ecology*, *101*, e02927.

Worm, B., & Tittensor, D. P. (2018). *A theory of global biodiversity (MPB-60)*. Princeton: Princeton University Press.

Wright, W. J., Irvine, K. M., & Higgs, M. D. (2019). Identifying occupancy model inadequacies: Can residuals separately assess detection and presence? *Ecology*, *100*, e02703.

Zipkin, E. F., Zylstra, E. R., Wright, A. D., Saunders, S. P., Finley, A. O., Dietze, M. C., Itter, M. S., & Tingley, M. W. (2021). Addressing data integration challenges to link ecological processes across scales. *Frontiers in Ecology and the Environment*, *19*, 30–38.

Appendix A

Additional material associated with our spatial modeling framework, synthetic data examples and data examples from chapter 1

A.1 Introduction

This appendix contains an in-depth description and explanation of our spatial occupancy model and the Markov chain Monte Carlo (MCMC) algorithm in Chapter 1. Additionally, this appendix contains the figures and the results associated with scenarios 2-6 in the synthetic data examples discussed in the Chapter 1. We begin by providing additional details about our modeling framework, including the prior distributions. We then introduce the latent variable approach, which enables the construction of an MCMC algorithm that uses closed-form full-conditional distributions. Finally, we describe how to construct an MCMC algorithm that enables the embedding of machine learning approaches within our spatial occupancy model.

A.2 Prior distributions

As described in Chapter 1, we take a Bayesian approach, which requires the specification of prior distributions for unknown parameters in our spatial occupancy model. Assuming p_{ij} in (1.1) in Chapter 1 is constant at all the sites during all sampling periods (i.e. $p_{ij} \equiv p$), we assign the priors $p \sim \text{Beta}(1, 1)$. The prior for $\boldsymbol{\beta}$ in (1.3) of the Chapter 1 is defined as, $\boldsymbol{\beta} \sim \text{N}(0, \sigma_\beta^2 \mathbf{I})$. Implementation of our spatial occupancy model using the MCMC algorithm developed by Shaby & Fink (2012) requires that we assign a prior to a temporary intermediate variable, f_i. We assign the prior for f_i as $f_i \sim \text{N}(0, \sigma_f^2)$. The temporary intermediate variable f_i is explained in-detail in section A.4 in this appendix.

The construction of the MCMC algorithm requires the full-conditional distributions of the parameters. To obtain the full-conditionals, we use the latent normal random variable approach described in the next section.

A.3 Latent normal random variable approach

The latent normal random variable approach enables the Gibbs sampler to be constructed using closed-form full-conditional distributions, as shown by Dorazio & Rodriguez (2012). Using this approach, z_i in (1.2) of Chapter 1 (true presence or absence at the i^{th}site) is expressed as

$$z_i = \begin{cases} 1 & , v_i \geq 0 \\ 0 & , v_i < 0 \end{cases}, \tag{A.1}$$

where v_i is a latent variable (Albert & Chib, 1993; Dorazio & Rodriguez, 2012). If the link function in (1.5) in Chapter 1 is a probit, then the latent variable v_i is distributed as

$$v_i | f(\mathbf{s}_i), \boldsymbol{\beta} \sim \text{N}(\mathbf{x}_i' \boldsymbol{\beta} + f(\mathbf{s}_i), 1). \tag{A.2}$$

Here, $\mathbf{x}_i \equiv (1, x_1, x_2, ..., x_q)'$ and $\boldsymbol{\beta} \equiv (\beta_0, \beta_1, \beta_2, ..., \beta_q)'$ where $x_1, x_2, ..., x_q$ represent the site-level covariates, β_0 represents the intercept parameter, and $\beta_1, \beta_2, ..., \beta_q$ represent the regression coefficients. The function $f(\mathbf{s}_i)$ represent the unknown spatially varying process that depends on the coordinate vector, $\mathbf{s}_i$, of the i^{th} site.

The univariate full-conditional distribution of z_i is

$$[z_i | y_{i\cdot}, \boldsymbol{\beta}, \widehat{f(\mathbf{s}_i)}, p] = \begin{cases} 1 & , y_{i\cdot} = 1 \\ \text{Bernoulli}\left(\frac{\Phi(\mathbf{x}_i'\boldsymbol{\beta} + \widehat{f(\mathbf{s}_i)}) \prod_{j=1}^{J_i}(1-p)}{\Phi(\mathbf{x}_i'\boldsymbol{\beta} + \widehat{f(\mathbf{s}_i)}) \prod_{j=1}^{J_i}(1-p) + 1 - \Phi(\mathbf{x}_i'\boldsymbol{\beta} + \widehat{f(\mathbf{s}_i)})}\right) & , y_{i\cdot} = 0 \end{cases},$$

where, $y_{i\cdot} = \sum_{j=1}^{J_i} y_{ij}$, $\Phi(\cdot)$ denotes the inverse probit link, and $\widehat{f(\mathbf{s}_i)}$ is a scalar. We explain $\widehat{f(\mathbf{s}_i)}$ in section A.4 as this relates to embedding of the machine learning approaches. The univariate full-conditional distribution of p is

$$[p | y_{i\cdot}, z_i] = \text{Beta}(1 + \textstyle\sum_{i=1}^{n} z_i y_{i\cdot}, 1 + \sum_{i=1}^{n} J_i - \sum_{i=1}^{n} z_i y_{i\cdot}).$$

The univariate full-conditional distribution of v_i is

$$[v_i | z_i, \widehat{f(\mathbf{s}_i)}, \boldsymbol{\beta}] = \begin{cases} \text{TN}(\mathbf{x}_i'\boldsymbol{\beta} + \widehat{f(\mathbf{s}_i)}, 1)|_0^{\infty} & , z_i = 1 \\ \text{TN}(\mathbf{x}_i'\boldsymbol{\beta} + \widehat{f(\mathbf{s}_i)}, 1)|_{-\infty}^{0} & , z_i = 0 \end{cases},$$

where, TN denotes a truncated normal distribution. The multivariate full-conditional distribution of $\boldsymbol{\beta}$ is

$$[\boldsymbol{\beta} | \mathbf{v}, \widehat{\mathbf{f}(\mathbf{S})}] = \text{N}\left(\left(\mathbf{X}'\mathbf{X} + \tfrac{1}{\sigma_\beta^2}\mathbf{I}\right)^{-1} \mathbf{X}'(\mathbf{v} - \widehat{\mathbf{f}(\mathbf{S})}), \left(\mathbf{X}'\mathbf{X} + \tfrac{1}{\sigma_\beta^2}\mathbf{I}\right)^{-1}\right),$$

where, $\widehat{\mathbf{f}(\mathbf{S})} \equiv (\widehat{f(\mathbf{s}_1)}, \widehat{f(\mathbf{s}_2)}, ..., \widehat{f(\mathbf{s}_n)})'$, $\mathbf{v} \equiv (v_1, v_2, ..., v_n)'$, and $\mathbf{X} \equiv (\mathbf{x}_1' | \mathbf{x}_2' | \cdots | \mathbf{x}_n')$. See chapter 23 in Hooten & Hefley (2019) for more details on deriving the above full-conditionals.

A.4 Embedding machine learning approaches

As described in Chapter 1, the functional form of $f(\mathbf{s}_i)$ presented in (A.2) is unknown. To approximate the function $f(\mathbf{s}_i)$, we use machine learning approaches within our spatial occupancy model. To accomplish this, Shaby & Fink (2012) developed an approximate Gibbs sampler where in each iteration of the Gibbs sampler, $f(\mathbf{s}_i)$ is replaced with a temporary intermediate variable f_i. The temporary intermediate variable f_i is sampled from the full-conditional distribution

$$[f_i|v_i, \boldsymbol{\beta}] = \text{N}\left(\frac{\sigma_f^2}{\sigma_f^2+1}(v_i - \mathbf{x}_i'\boldsymbol{\beta}), \frac{\sigma_f^2}{\sigma_f^2+1}\right).$$

The sampled f_i enables the use of off-the-shelf software for machine learning approaches. To accomplish this, Shaby & Fink (2012) used the sampled f_i as a response variable and the coordinate vector $\mathbf{s}_i$ as the predictor in a machine learning approach. Then obtain the predicted values $f(\hat{\mathbf{s}}_i)$, that is an approximation of the underlying spatial process $f(\mathbf{s}_i)$. Additional details about this approximate Gibbs sampler is provided in Shaby & Fink (2012).

A.5 Markov chain Monte Carlo Algorithm

This section contains MCMC algorithm and the R code to implement our modeling framework.

Algorithm 1 The MCMC algorithm used to sample from the posterior distributions of our spatial occupancy model. In this algorithm, l is the current iteration and m is the total number of iterations. The $\mathbf{y} \equiv (y_{1\cdot}, y_{2\cdot}, ..., y_{n\cdot})'$ is a $(n\times1)$ vector with i^{th} row representing the number of detections at the i^{th} site, $\mathbf{z} \equiv (z_1, z_2, ..., z_n)'$ represents the true presence or absence at each site, $\mathbf{S} \equiv (\mathbf{s}_1'|\mathbf{s}_2'|\cdots|\mathbf{s}_n')$ is a $(n\times2)$ matrix with i^{th} row representing the coordinate vector of the i^{th} site, and $\mathbf{f} \equiv (f_1, f_2, ..., f_n)'$ represents the vector of temporary intermediate variables discussed in section S4 in this appendix. The $\mathbf{f}(\hat{\mathbf{S}})$ and $\mathbf{v}$ are defined in section A.4 , $\boldsymbol{\beta}$ is defined in section A.3 , and p is defined in section A.2 .

```
Set initial values for $\mathbf{f}, \boldsymbol{\beta}$, and $p$
Set $l$=0
while l < m do
    Sample $\mathbf{z}$ from its full-conditional distribution, $\mathbf{z}^{(l)} \sim [\mathbf{z}^{(l)}|\mathbf{y}, \boldsymbol{\beta}^{(l-1)}, \mathbf{f}(\hat{\mathbf{S}})^{(l-1)}, p^{(l-1)}]$;
    Sample $p$ from its full-conditional distribution, $p^{(l)} \sim [p^{(l)}|\mathbf{y}, \mathbf{z}^{(l-1)}]$;
    Sample $\mathbf{v}$ from its full-conditional distribution, $\mathbf{v}^{(l)} \sim [\mathbf{v}^{(l)}|\mathbf{z}^{(l-1)}, \mathbf{f}(\hat{\mathbf{S}})^{(l-1)}, \boldsymbol{\beta}^{(l-1)}]$;
    Sample $\boldsymbol{\beta}$ from its full-conditional distribution, $\boldsymbol{\beta}^{(l)} \sim [\boldsymbol{\beta}^{(l)}|\mathbf{v}^{(l-1)}, \mathbf{f}(\hat{\mathbf{S}})^{(l-1)}]$;
    Sample the temporary intermediate vector $\mathbf{f}$ from its full-conditional distribution, $\mathbf{f}^{(l)} \sim [\mathbf{f}^{(l)}|\mathbf{v}^{(l-1)}, \boldsymbol{\beta}^{(l-1)}]$;
    Use the sampled $\mathbf{f}^{(l)}$ as a response and the coordinate matrix $\mathbf{S}$ as the predictor in a machine learning approach. Then, obtain the predicted values $\mathbf{f}(\hat{\mathbf{S}})^{(l)} = g(\mathbf{S})$, where $g(\cdot)$ is a machine learning approach;
    Use $\mathbf{f}(\hat{\mathbf{S}})^{(l)}$ to sample from the remaining full-conditional distributions of the parameters.
end while
```

A.6 Model extensions

In this section, we discuss the general framework and the extensions of our model. In Chapter 1, in (1.5), we present an additive model to model the probability of true presence ψ_i using site-level covariates $\mathbf{x}_i$ and the spatial dependence $f(\mathbf{s}_i)$. Our modeling framework can be generalized by writing the effects of site-level covariates using a function $h(\mathbf{x}_i)$, where $h(\mathbf{x}_i)$

can also be modeled using machine learning approaches the same as $f(\mathbf{s}_i)$. The (1.5) in Chapter 1 can be generalized as

$$g(\psi_i) = h(\mathbf{x}_i) + f(\mathbf{s}_i). \tag{A.3}$$

This generalized version (A.3) increases the ability to incorporate large sets of site-level covariates into the model and identify complex relationships using machine learning approaches. Shaby & Fink (2012) provide a thorough explanation of how to incorporate machine learning approaches to model the effect of site-level covariates.

Another extension of our modeling framework is including machine learning approaches to model the probability of detection. In the manuscript, we assumed that the probability of detection does not change across sites. However, we can use machine learning approaches to include informative covariates for detection in to the model such as, Julian date, effort, etc., and identify their complex relationship to the probability of detection. The model can be written as

$$q(p_i) = h_p(\mathbf{x}_i), \tag{A.4}$$

where, $q(\cdot)$ is an appropriate link function and $h_p(\cdot)$ is a function of covariates for detection.

As discussed in the manuscript, our spatial occupancy model can be easily adapted to account for other types of contamination in occupancy data, such as false positives. To account for false positives, we can extend our model framework by modifying (1.1) in Chapter 1 to

$$y_{ij}|z_i, p_{ij1} \sim \begin{cases} \text{Bernoulli}(p_{ij0}) & , z_i = 1 \\ \\ \text{Bernoulli}(p_{ij}) & , z_i = 0 \end{cases}, \tag{A.5}$$

where, p_{ij1} is the probability of detecting at least one individual at i^{th} site during the j^{th} sampling period, and p_{ij0} is the probability of false positives at i^{th} site during the j^{th} sampling period. Hooten & Hefley (2019) provide in detail description of the model specification to account for false positives in occupancy models.

A.7 Figures and results associated with synthetic data examples (scenarios 2-6)

A.7.1 Figures for synthetic data examples

The below figure is associated with scenario 2 in the synthetic data example, where the spatial dependence forms a circle with the probability of occupancy being low in the center and smoothly increasing towards the edge of the circle.

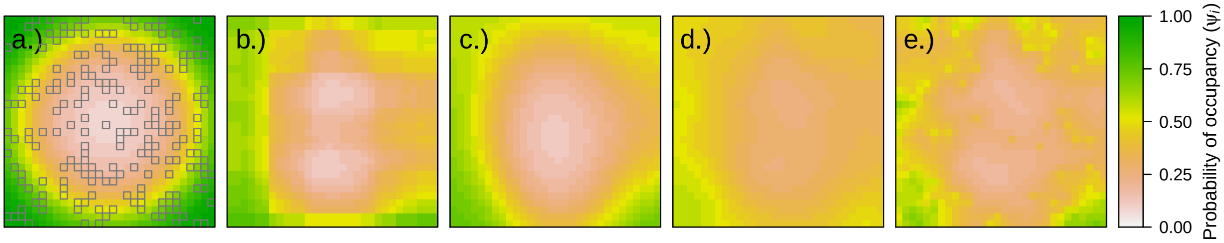

Figure A.1: The probability of occupancy from scenario 2 of the synthetic data example (panel a) and the posterior mean of the probability of occupancy ($\text{E}(\psi_i|\mathbf{y})$) obtained by fitting spatial occupancy models that include an embedded regression tree (panel b), a support vector regression (panel c), a low-rank Gaussian process (panel d), and a Gaussian Markov random field (panel e). The gray squares in the panel show the location of the 200 sampled sites used for model fitting.

The below figure is associated with scenario 3 in the synthetic data example, where the spatial dependence is defined by a cosine function.

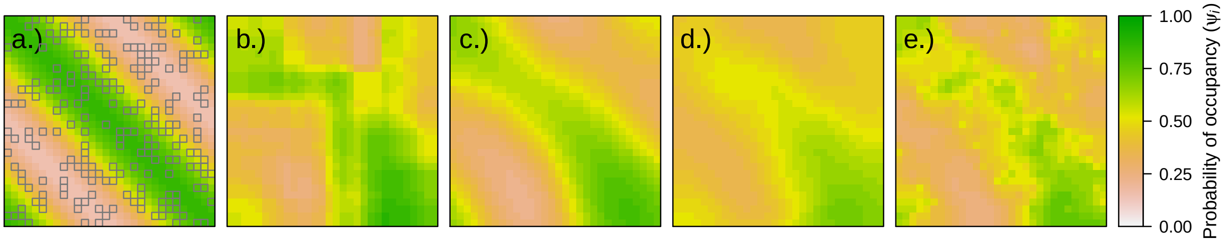

Figure A.2: The probability of occupancy from scenario 3 of the synthetic data example (panel a) and the posterior mean of the probability of occupancy ($\text{E}(\psi_i|\mathbf{y})$) obtained by fitting spatial occupancy models that include an embedded regression tree (panel b), a support vector regression (panel c), a low-rank Gaussian process (panel d), and a Gaussian Markov random field (panel e). The gray squares in the panel show the location of the 200 sampled sites used for model fitting.

The below figure is associated with scenario 4 in the synthetic data example, where the spatial dependence is a normally distributed random effect with a correlation matrix specified by a conditional autoregressive process.

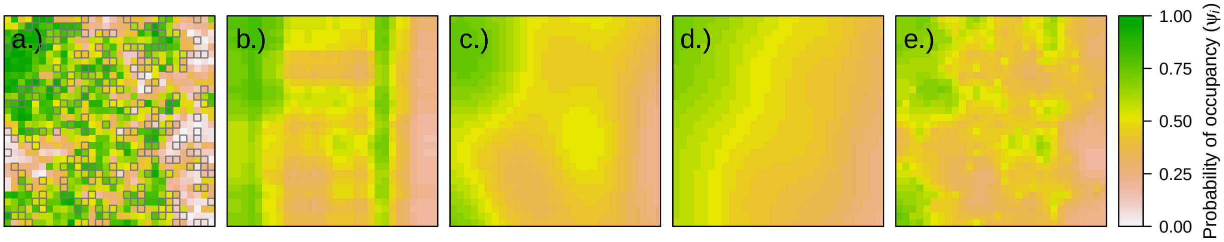

Figure A.3: The probability of occupancy from scenario 4 of the synthetic data example (panel a) and the posterior mean of the probability of occupancy ($\text{E}(\psi_i|\mathbf{y})$) obtained by fitting spatial occupancy models that include an embedded regression tree (panel b), a support vector regression (panel c), a low-rank Gaussian process (panel d), and a Gaussian Markov random field (panel e). The gray squares in the panel show the location of the 200 sampled sites used for model fitting.

The below figure is associated with scenario 5 in the synthetic data example, where the spatial dependence is a normally distributed random effect with a correlation matrix specified by an exponential covariance function.

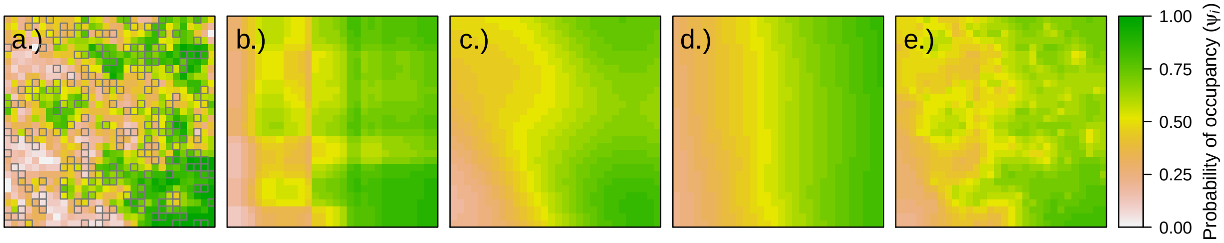

Figure A.4: The probability of occupancy from scenario 5 of the synthetic data example (panel a) and the posterior mean of the probability of occupancy ($\text{E}(\psi_i|\mathbf{y})$) obtained by fitting spatial occupancy models that include an embedded regression tree (panel b), a support vector regression (panel c), a low-rank Gaussian process (panel d), and a Gaussian Markov random field (panel e). The gray squares in the panel show the location of the 200 sampled sites used for model fitting.

The below figure is associated with scenario 6 in the synthetic data example, where the spatial dependence is a normally distributed random effect with a correlation matrix specified by a squared exponential covariance function.

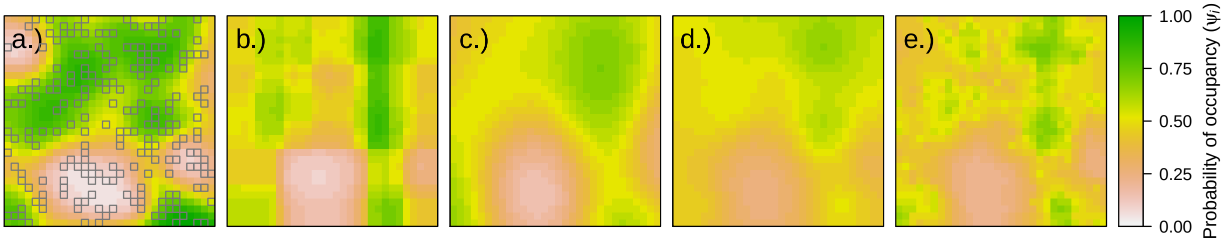

Figure A.5: The probability of occupancy from scenario 6 of the synthetic data example (panel a) and the posterior mean of the probability of occupancy ($\mathrm{E}(\psi_i|\mathbf{y})$) obtained by fitting spatial occupancy models that include an embedded regression tree (panel b), a support vector regression (panel c), a low-rank Gaussian process (panel d), and a Gaussian Markov random field (panel e). The gray squares in the panel show the location of the 200 sampled sites used for model fitting.

A.7.2 Results for synthetic data examples

Table A.1: The -2×log posterior predictive density (i.e., -2×LPPD) for all scenarios. The results are obtained by embedding a regression tree (REG), a support vector regression (SVR), a low-rank Gaussian process (LRGP), and a Gaussian Markov random field (GMRF). We also include the -2× LPPD obtained by fitting the traditional occupancy model without including the spatial effect (NOSP).

Scenario	REG	SVR	LRGP	GMRF	NOSP
Scenario 2	402.60	398.46	424.75	419.82	446.09
Scenario 3	445.96	436.11	448.13	442.04	465.79
Scenario 4	466.40	462.77	466.55	462.30	464.71
Scenario 5	441.05	441.02	445.83	441.83	468.39
Scenario 6	425.89	426.49	434.67	432.21	450.24

A.7.3 Moran's I correlogram plots associated with synthetic data examples

In this section, we obtain the Moran's I correlogram plots to investigate the residual spatial dependence. The below figure shows the Moran's I correlogram plots for scenario 1.

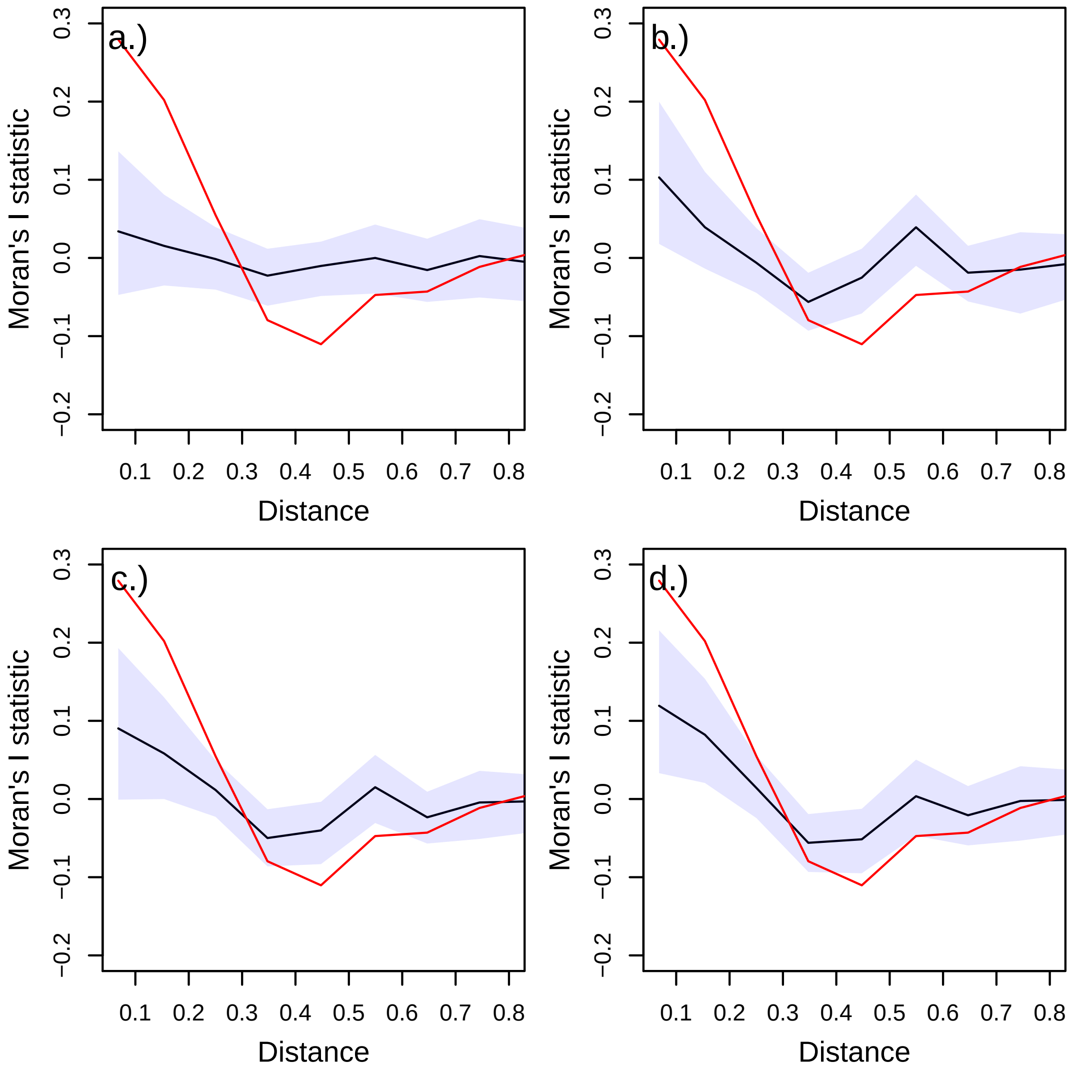

Figure A.6: Moran's I correlogram plots for the fitted occupancy models embedding a) regression trees, b) support vector regression, c) low-rank Gaussian process, and d) Gaussian Markov random field. For each fitted occupancy model, we plot the posterior mean of the Moran's I against the distance class of neighbouring sites (black –) and display the associated 95% credible interval (the band represented by the shaded region). For comparison, we include the posterior mean of the Moran's I that we obtained from the traditional occupancy model fitted without the spatial effect (red –).

The below figure shows the Moran's I correlogram plots for scenario 2.

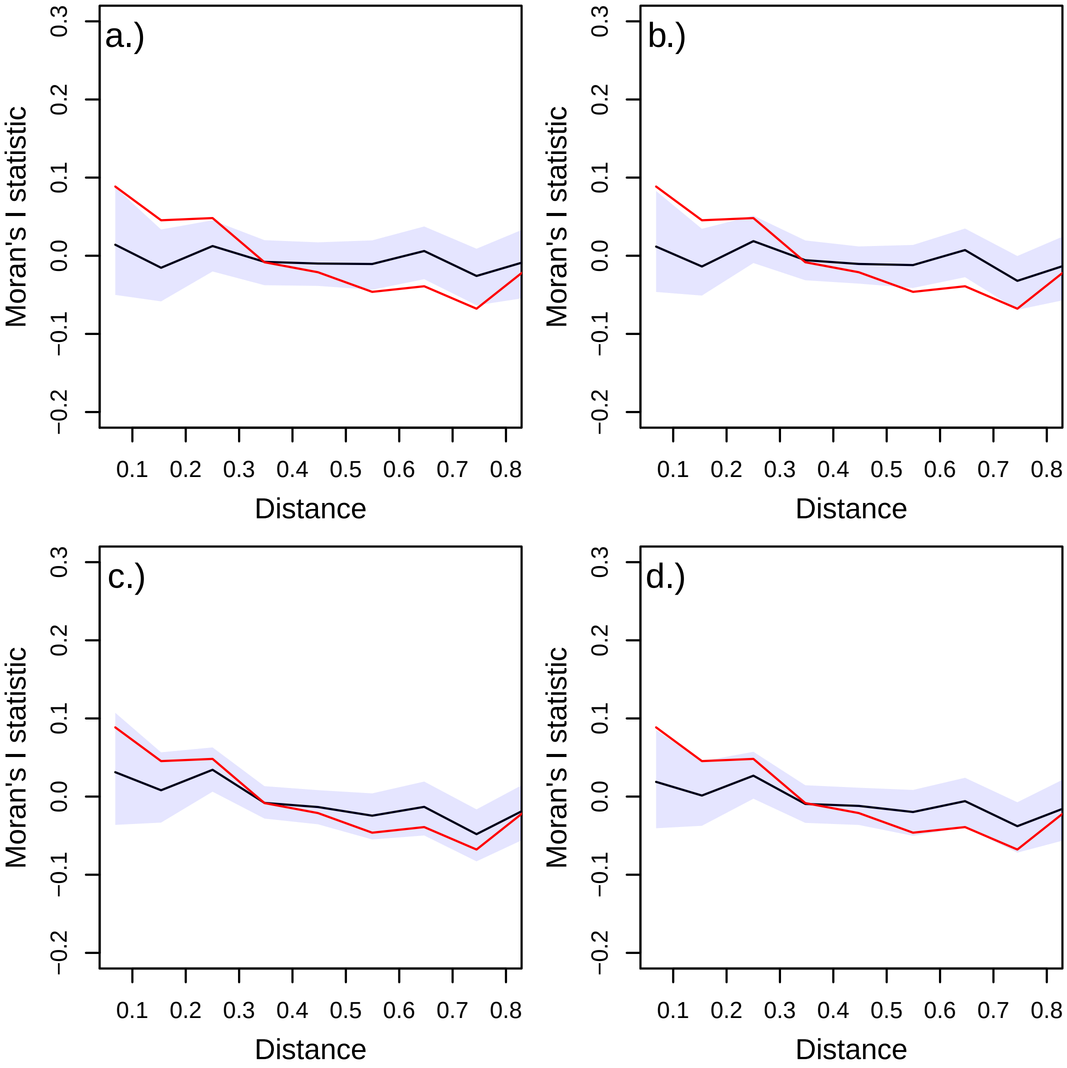

Figure A.7: Moran's I correlogram plots for the fitted occupancy models embedding a) regression trees, b) support vector regression, c) low-rank Gaussian process, and d) Gaussian Markov random field. For each fitted occupancy model, we plot the posterior mean of the Moran's I against the distance class of neighbouring sites (black –) and display the associated 95% credible interval (the band represented by the shaded region). For comparison, we include the posterior mean of the Moran's I that we obtained from the traditional occupancy model fitted without the spatial effect (red –).

The below figure shows the Moran's I correlogram plots for scenario 3.

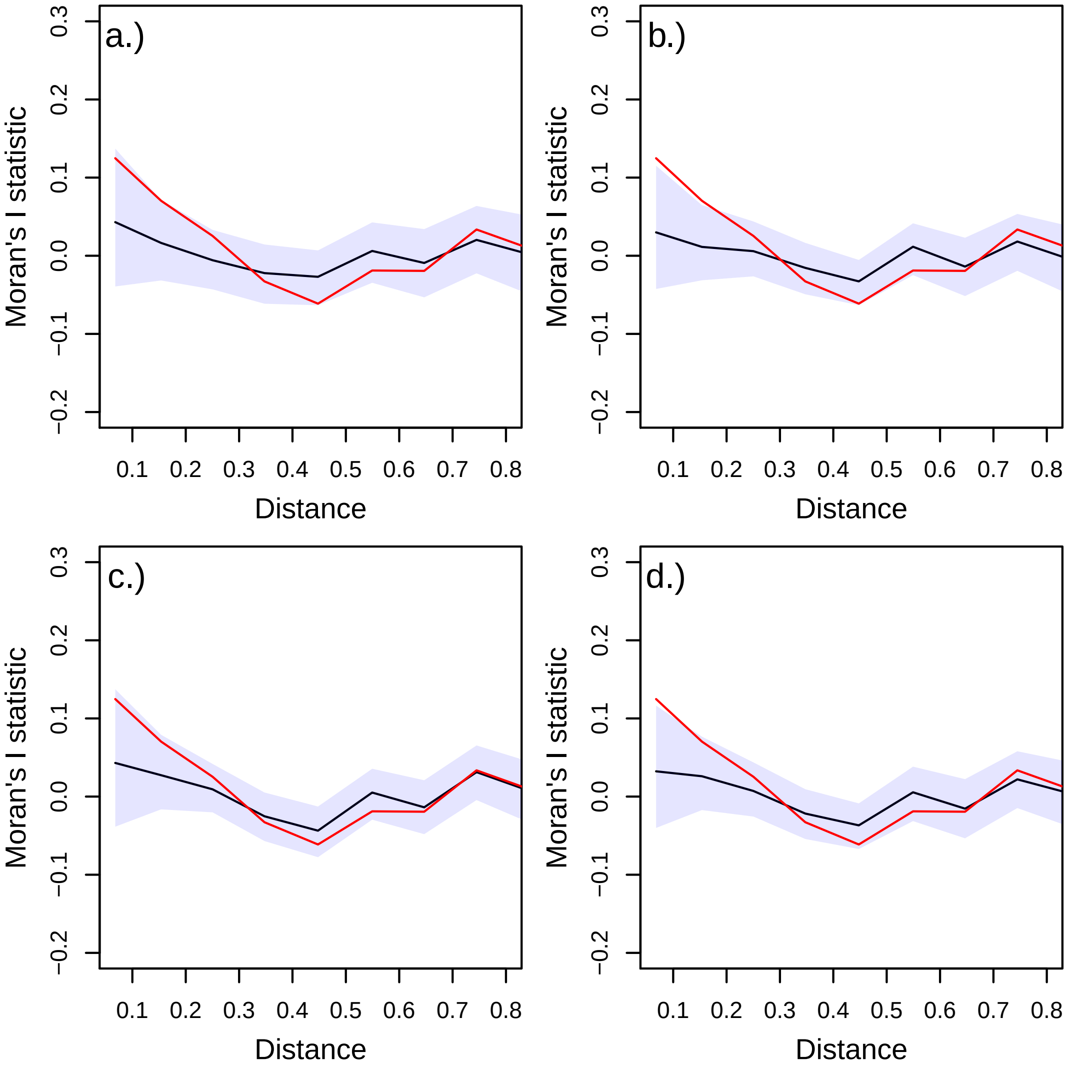

Figure A.8: Moran's I correlogram plots for the fitted occupancy models embedding a) regression trees, b) support vector regression, c) low-rank Gaussian process, and d) Gaussian Markov random field. For each fitted occupancy model, we plot the posterior mean of the Moran's I against the distance class of neighbouring sites (black –) and display the associated 95% credible interval (the band represented by the shaded region). For comparison, we include the posterior mean of the Moran's I that we obtained from the traditional occupancy model fitted without the spatial effect (red –).

The below figure shows the Moran's I correlogram plots for scenario 4.

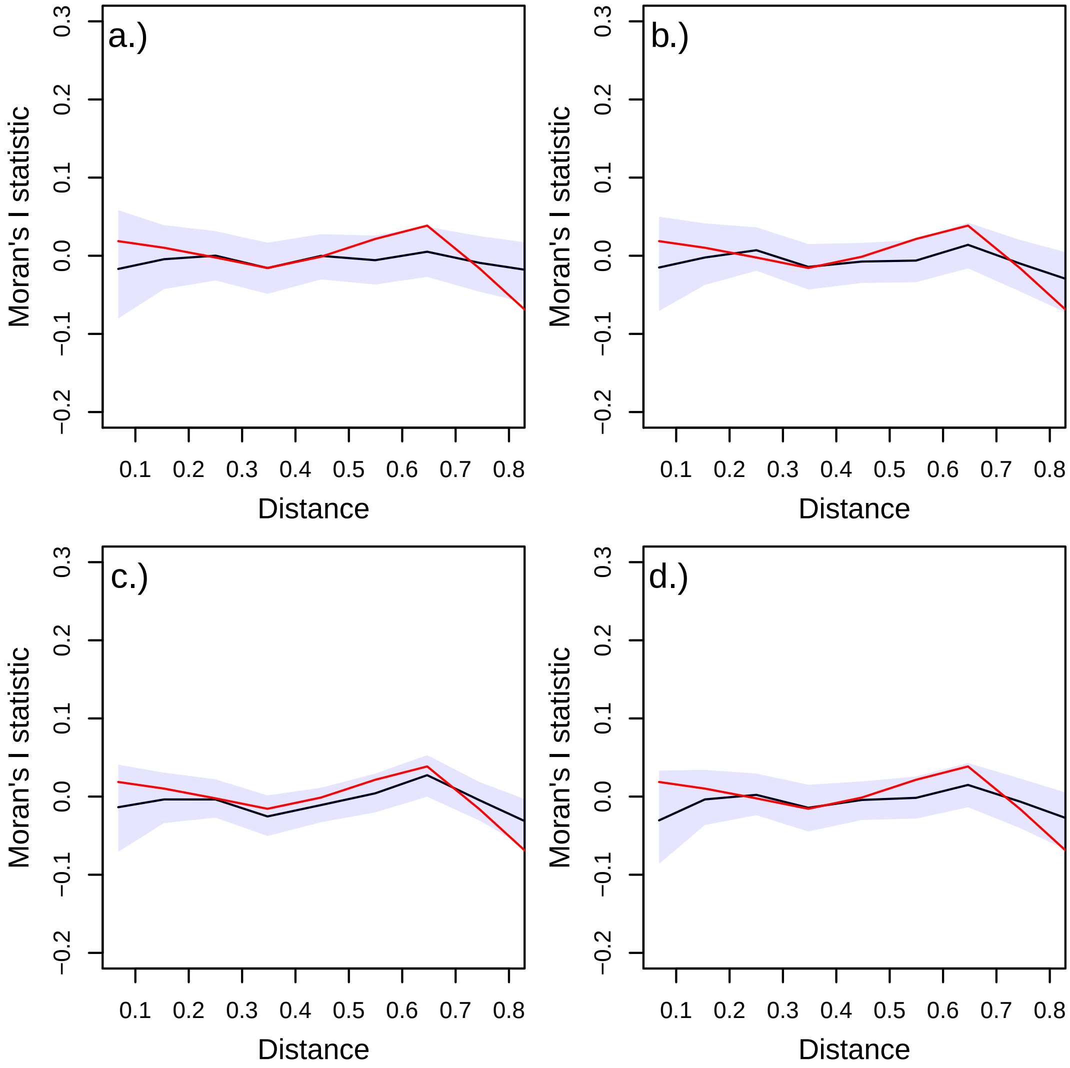

Figure A.9: Moran's I correlogram plots for the fitted occupancy models embedding a) regression trees, b) support vector regression, c) low-rank Gaussian process, and d) Gaussian Markov random field. For each fitted occupancy model, we plot the posterior mean of the Moran's I against the distance class of neighbouring sites (black –) and display the associated 95% credible interval (the band represented by the shaded region). For comparison, we include the posterior mean of the Moran's I that we obtained from the traditional occupancy model fitted without the spatial effect (red –).

The below figure shows the Moran's I correlogram plots for scenario 5.

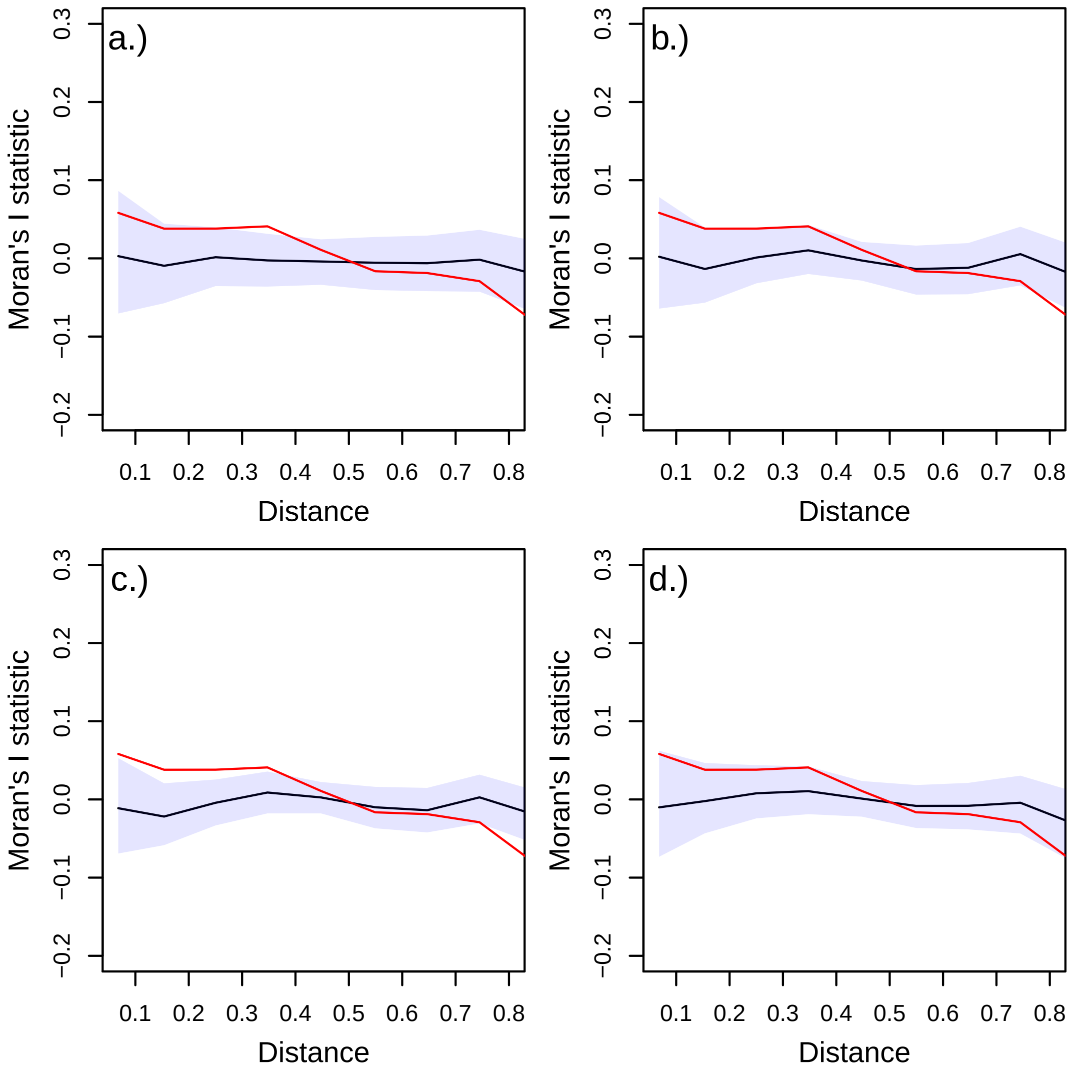

Figure A.10: Moran's I correlogram plots for the fitted occupancy models embedding a) regression trees, b) support vector regression, c) low-rank Gaussian process, and d) Gaussian Markov random field. For each fitted occupancy model, we plot the posterior mean of the Moran's I against the distance class of neighbouring sites (black –) and display the associated 95% credible interval (the band represented by the shaded region). For comparison, we include the posterior mean of the Moran's I that we obtained from the traditional occupancy model fitted without the spatial effect (red –).

The below figure shows the Moran's I correlogram plots for scenario 6.

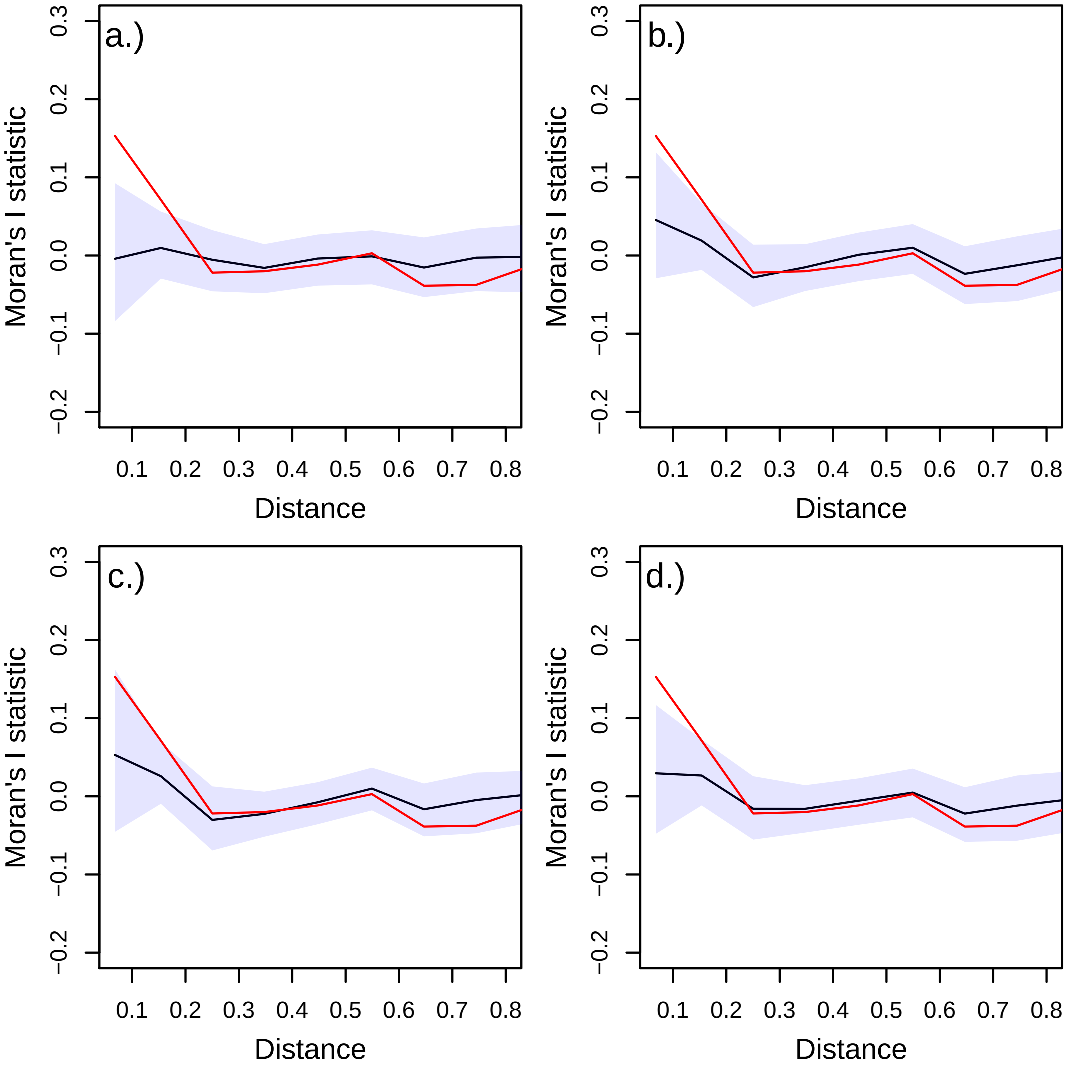

Figure A.11: Moran's I correlogram plots for the fitted occupancy models embedding a) regression trees, b) support vector regression, c) low-rank Gaussian process, and d) Gaussian Markov random field. For each fitted occupancy model, we plot the posterior mean of the Moran's I against the distance class of neighbouring sites (black –) and display the associated 95% credible interval (the band represented by the shaded region). For comparison, we include the posterior mean of the Moran's I that we obtained from the traditional occupancy model fitted without the spatial effect (red –).

A.8 Moran’s I correlogram plots associated with the Thomson’s gazelle data example in Chapter 1

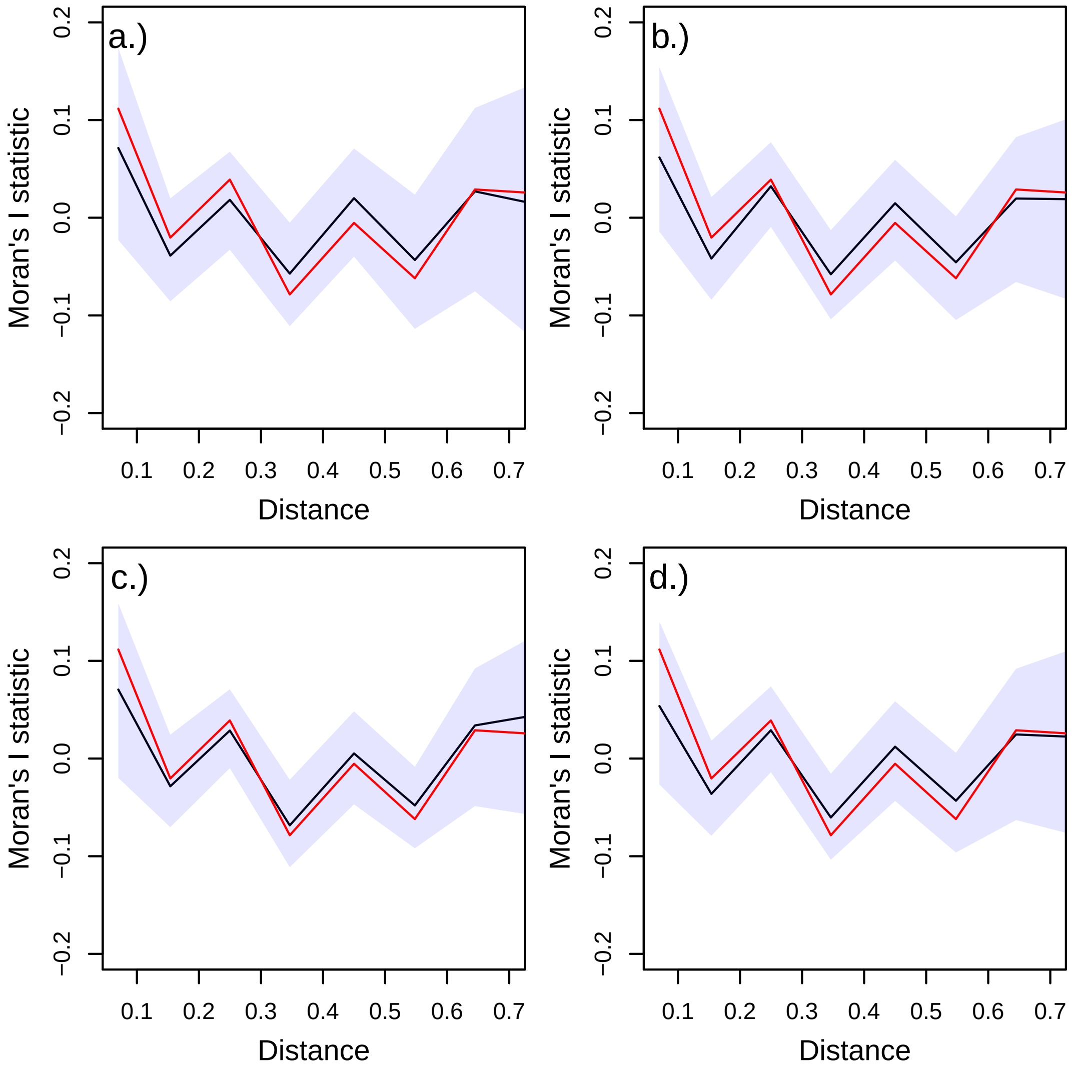

Figure A.12: Moran’s I correlogram plots for the fitted occupancy models embedding a) regression trees, b) support vector regression, c) low-rank Gaussian process, and d) Gaussian Markov random field. For each fitted occupancy model, we plot the posterior mean of the Moran’s I against the distance class of neighbouring sites (black –) and display the associated 95% credible interval (the band represented by the shaded region). For comparison, we include the posterior mean of the Moran’s I that we obtained from the traditional occupancy model fitted without the spatial effect (red –).

A.9 Moran's I correlogram plots associated with the sugar glider data example in Chapter 1

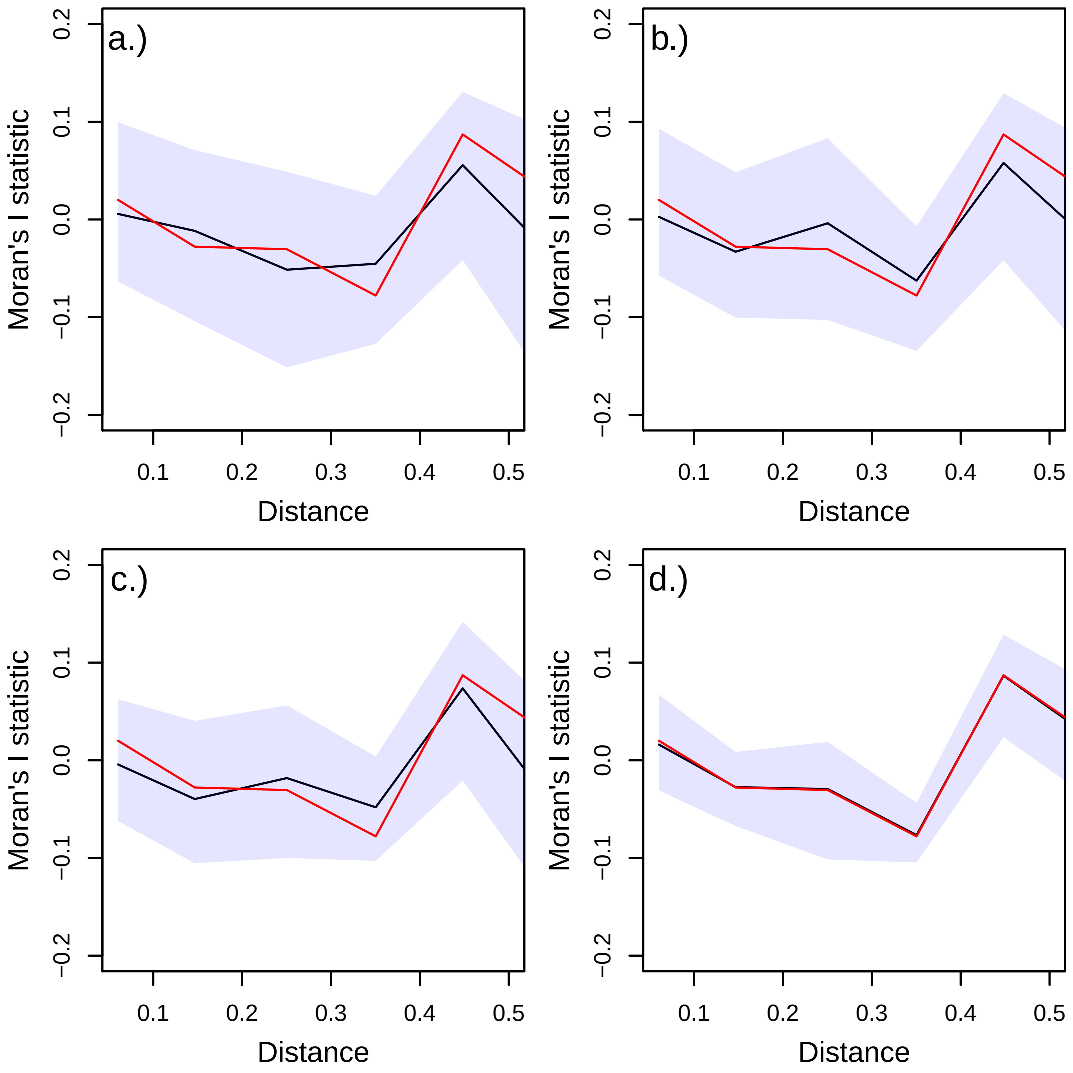

Figure A.13: Moran's I correlogram plots for the fitted occupancy models embedding a) regression trees, b) support vector regression, c) low-rank Gaussian process, and d) Gaussian Markov random field. For each fitted occupancy model, we plot the posterior mean of the Moran's I against the distance class of neighbouring sites (black –) and display the associated 95% credible interval (the band represented by the shaded region). For comparison, we include the posterior mean of the Moran's I that we obtained from the traditional occupancy model fitted without the spatial effect (red –).

Appendix B

Details associated with deriving the models in chapter 2.

B.1 Models accounting for missing individuals that are missing not at random (MNAR)

The construction of the inhomogeneous Poisson point process (IPPP) represented by (2.1) in Chapter 2 involves specifying the PMF for the random number of individuals in the study area $\mathcal{S}$ (i.e., N), and the PDF for a coordinate vector that contains the location of an individual (i.e., $\mathbf{u}_i$) in the study area $\mathcal{S}$. The N follows a Poisson distribution, and the PMF of N can be written as

$$[N|\lambda(\mathbf{s})] = \frac{e^{-\int_{\mathcal{S}} \lambda(\mathbf{s})d\mathbf{s}}(\int_{\mathcal{S}} \lambda(\mathbf{s})d\mathbf{s})^N}{N!}. \tag{B.1}$$

The PDF for the location of the i^{th} individual, $\mathbf{u}_i$ is

$$[\mathbf{u}_i|\lambda(\mathbf{s})] = \frac{\lambda(\mathbf{u}_i)}{\int_S \lambda(\mathbf{s})d\mathbf{s}}. \tag{B.2}$$

To incorporate the missing data mechanism that produce MNAR individuals, we can add a label $m(\mathbf{u}_i)$ to the location of the i^{th} individual representing whether the individual is missing or observed. Gelfand & Schliep (2018) viewed adding a label such as $m(\mathbf{u}_i)$ to a point process as adding an additional coordinate to the points in the point process. That is, the location of the i^{th} individual is represented as a pair $(\mathbf{u}_i, m(\mathbf{u}_i))$. Using the product space representation, $(\mathbf{u}_i, m(\mathbf{u}_i))$ is a point over $\mathcal{S} \times M$, where M is the support for the labels $(m(\mathbf{u}_i) \in M)$. By incorporating the distribution of the $m(\mathbf{u}_i)$ represented by (2.2) in Chapter 2, we can write

$$\begin{aligned}
[\mathbf{u}_i, m(\mathbf{u}_i)|\lambda(\mathbf{s}), q(\mathbf{s}), r(\mathbf{s})] &= [m(\mathbf{u}_i)|\mathbf{u}_i, q(\mathbf{s}), r(\mathbf{s})][\mathbf{u}_i|\lambda(\mathbf{s})] \\
&= \begin{cases} q(\mathbf{u}_i)^{m(\mathbf{u}_i)}(1-q(\mathbf{u}_i))^{1-m(\mathbf{u}_i)} \times \frac{\lambda(\mathbf{u}_i)}{\int_S \lambda(\mathbf{s})d\mathbf{s}} & , \text{if } r(\mathbf{u}_i) = 1 \\ 0 & , \text{if } r(\mathbf{u}_i) = 0 \end{cases} .
\end{aligned} \tag{B.3}$$

Now, using (B.2) and (B.3), we can derive (2.3) in Chapter 2 the PDF for the location of the i^{th} individual conditioned on the label $m(\mathbf{u}_i)$, which is

$$\begin{aligned}
[\mathbf{u}_i|m(\mathbf{u}_i), \lambda(\mathbf{s}), q(\mathbf{s}), r(\mathbf{s})] &= \frac{[\mathbf{u}_i, m(\mathbf{u}_i)|\lambda(\mathbf{s}), q(\mathbf{s}), r(\mathbf{s})]}{\int_{\mathcal{S}}[\mathbf{u}_i, m(\mathbf{u}_i)|\lambda(\mathbf{s}), q(\mathbf{s}), r(\mathbf{s})]d\mathbf{u}_i} \\
&= \begin{cases} \frac{q(\mathbf{u}_i)^{m(\mathbf{u}_i)}(1-q(\mathbf{u}_i))^{1-m(\mathbf{u}_i)}\lambda(\mathbf{u}_i)}{\int_{\mathcal{S}} q(\mathbf{s})^{m(\mathbf{s})}(1-q(\mathbf{s}))^{1-m(\mathbf{s})}\lambda(\mathbf{s})d\mathbf{s}} & , \text{if } r(\mathbf{u}_i) = 1 \\ 0 & , \text{if } r(\mathbf{u}_i) = 0 \end{cases} .
\end{aligned}$$

We assume that if the location of the i^{th} individual $\mathbf{u}_i$ is not in the sampled region within the study area, i^{th} individual is not observed (i.e., $m(\mathbf{u}_i) = 0$). Therefore, $m(\mathbf{u}_i) = 0$ implies $r(\mathbf{u}_i) = 0$ and we can rewrite the PDF for the location of the i^{th} individual conditioned on the label $m(\mathbf{u}_i)$ as

$$[\mathbf{u}_i|m(\mathbf{u}_i),\lambda(\mathbf{s}),q(\mathbf{s}),r(\mathbf{s})] = \begin{cases} \frac{\lambda(\mathbf{u}_i)q(\mathbf{u}_i)}{\int_{\mathcal{S}}\lambda(\mathbf{s})q(\mathbf{s})d\mathbf{s}} & \text{, if } r(\mathbf{u}_i)=1 \;\&\; m(\mathbf{u}_i)=1 \\ \\ 0 & , otherwise \end{cases} . \tag{B.4}$$

The model representation in (B.4) accounts for the missing individuals that are MNAR; however, the model requires complete location information of the individuals (i.e., the exact geographic coordinates of the locations of the individuals). Since distance sampling and capture-recapture data often do not contain complete location information of the individuals, we need to construct models that account for the partially missing location information to represent the observed data.

B.2 Data models accounting for partially missing location information

In (2.4) and (2.5) in Chapter 2, we represent two PDFs for the observed location of the i^{th} individual conditioned on the true location of the individual (i.e., $[\mathbf{y}_i|\mathbf{u}_i]$). The distributional representations in (2.4) and (2.5) account for the partially missing location information of the individuals.

However, in (2.4) and (2.5), the observed location of the i^{th} individual is conditioned on the true location of the individual $\mathbf{u}_i$, but the true location of the individual is of little interest in our study. Therefore, we can integrate the joint likelihood of $\mathbf{y}_i$ and $\mathbf{u}_i$ and remove $\mathbf{u}_i$ from the model as below.

$$[\mathbf{y}_i|m(\mathbf{u}_i),\lambda(\mathbf{s}),q(\mathbf{s}),r(\mathbf{s})] = \int_{\mathcal{S}}[\mathbf{y}_i|\mathbf{u}_i][\mathbf{u}_i|m(\mathbf{u}_i),\lambda(\mathbf{s}),q(\mathbf{s}),r(\mathbf{s})]d\mathbf{u}_i. \tag{B.5}$$

Substituting (B.4) and the PDF $[\mathbf{y}_i|\mathbf{u}_i]$ represented by (2.4) in Chapter 2 we can obtain

$$[\mathbf{y}_i|m(\mathbf{u}_i),\lambda(\mathbf{s}),q(\mathbf{s}),r(\mathbf{s})] = \begin{cases} \int_{\mathcal{S}} \frac{|A_{u_i}|^{-1} I(\mathbf{y}_i \in A_{u_i})\lambda(\mathbf{u}_i)q(\mathbf{u}_i)}{\int_{\mathcal{S}}\lambda(\mathbf{s})q(\mathbf{s})d\mathbf{s}} d\mathbf{u}_i & \text{, if } r(\mathbf{u}_i)=1 \ \&\ m(\mathbf{u}_i)=1 \\ 0 & , otherwise \end{cases}$$

$$= \begin{cases} \frac{\int_{A_{u_i}} |A_{u_i}|^{-1}\lambda(\mathbf{u}_i)q(\mathbf{u}_i)}{\int_{\mathcal{S}}\lambda(\mathbf{s})q(\mathbf{s})d\mathbf{s}} d\mathbf{u}_i & \text{, if } r(\mathbf{u}_i)=1 \ \&\ m(\mathbf{u}_i)=1 \\ 0 & , otherwise \end{cases} . \tag{B.6}$$

Similarly, by substituting the PDF $[\mathbf{y}_i|\mathbf{u}_i]$ represented by (2.5) in Chapter 2 we can obtain

$$[\mathbf{y}_i|m(\mathbf{u}_i),\lambda(\mathbf{s}),q(\mathbf{s}),r(\mathbf{s})] = \begin{cases} \frac{\int_{L_{u_i}} |L_{u_i}|^{-1}\lambda(\mathbf{u}_i)q(\mathbf{u}_i)d\mathbf{u}_i}{\int_{\mathcal{S}}\lambda(\mathbf{s})q(\mathbf{s})d\mathbf{s}} & \text{, if } r(\mathbf{u}_i)=1 \ \&\ m(\mathbf{u}_i)=1 \\ 0 & , otherwise. \end{cases} . \tag{B.7}$$

The (B.6) and (B.7) are equivalent to (2.6) and (2.7) in Chapter 2.

B.3 Data models for the observed data removing missing individuals

To represent the distribution of observed data, the missing individuals should be removed, and the observed individuals should be retained (i.e., $m(\mathbf{u}_i) = 1$). If n is the number of observed individuals in the study area $\mathcal{S}$, the locations of the n individuals out of the total N individuals contain the label $m(\mathbf{u}_i) = 1$. Therefore, n follows a Poisson distribution with the rate parameter $\bar{\lambda}_1 = \int_{\mathcal{S}} \lambda(\mathbf{s}, m(\mathbf{s}) = 1)d\mathbf{s} = \int_{\mathcal{S}} \lambda(\mathbf{s})q(\mathbf{s})I(r(\mathbf{s}) = 1)d\mathbf{s}$. The PMF of n can be written as

$$[n|\lambda(\mathbf{s}),q(\mathbf{s}),r(\mathbf{s})] = \frac{e^{-\int_{\mathcal{S}}\lambda(\mathbf{s})q(\mathbf{s})I(r(\mathbf{s})=1)d\mathbf{s}}(\int_{\mathcal{S}}\lambda(\mathbf{s})q(\mathbf{s})I(r(\mathbf{s})=1)d\mathbf{s})^n}{n!}. \tag{B.8}$$

Combining the Poisson distribution in (B.8) and the PDF in (B.6) conditioned on $m(\mathbf{u}_i)=1$, we can write

$$\begin{aligned}[]
[\mathbf{y}_1,\mathbf{y}_2,...,\mathbf{y}_n,n|\lambda(\mathbf{s}),q(\mathbf{s}),r(\mathbf{s})] &= \frac{e^{-\int_{\mathcal{S}}\lambda(\mathbf{s})q(\mathbf{s})I(r(\mathbf{s})=1)d\mathbf{s}}(\int_{\mathcal{S}}\lambda(\mathbf{s})q(\mathbf{s})I(r(\mathbf{s})=1)d\mathbf{s})^n}{n!}\times \\
&\qquad n!\prod_{i=1}^{n}\frac{\int_{A_{u_i}}|A_{u_i}|^{-1}\lambda(\mathbf{u}_i)q(\mathbf{u}_i)I(r(\mathbf{u}_i)=1)}{\int_{\mathcal{S}}\lambda(\mathbf{s})q(\mathbf{s})I(r(\mathbf{s})=1)d\mathbf{s}} \\
&= e^{-\int_{\mathcal{S}}\lambda(\mathbf{s})q(\mathbf{s})I(r(\mathbf{s})=1)d\mathbf{s}}\times \\
&\qquad \prod_{i=1}^{n}\int_{A_{u_i}}|A_{u_i}|^{-1}\lambda(\mathbf{u}_i)q(\mathbf{u}_i)I(r(\mathbf{u}_i)=1).
\end{aligned} \tag{B.9}$$

Similarly, by combining the Poisson distribution in (B.8) and the PDF in (B.7) conditioned on $m(\mathbf{u}_i)=1$, we can write

$$\begin{aligned}[]
[\mathbf{y}_1,\mathbf{y}_2,...,\mathbf{y}_n,n|\lambda(\mathbf{s}),q(\mathbf{s}),r(\mathbf{s})] &= e^{-\int_{\mathcal{S}}\lambda(\mathbf{s})q(\mathbf{s})I(r(\mathbf{s})=1)d\mathbf{s}}\times \\
&\qquad \prod_{i=1}^{n}\int_{L_{u_i}}|L_{u_i}|^{-1}\lambda(\mathbf{u}_i)q(\mathbf{u}_i)I(r(\mathbf{u}_i)=1).
\end{aligned} \tag{B.10}$$

The joint distributions in (B.9) and (B.10) are the two distributional representations for the observed data used in our approaches that is based on the assumption that the observed individual locations are conditionally independent given $\lambda(\mathbf{s}), q(\mathbf{s})$, and $r(\mathbf{s})$. By representing distance sampling and capture-recapture data using (B.9) and (B.10) and by obtaining the joint likelihood, the proposed two fused data SDMs in (2.8) and (2.9) in Chapter 2 are obtained.