Statistical methods for modeling infectious disease outbreaks using disease surveillance data

by

Nelson B. Walker

B.S., Brigham Young University, 2014

M.S., Kansas State University, 2018

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

2021

Abstract

Disease surveillance data are commonly used by epidemiologists, veterinary and plant pathologists, and wildlife and plant ecologists to identify, understand, mitigate, and prevent the spread of infectious disease. We develop three statistical methods that may be applied to spatio-temporal disease surveillance data to understand different aspects of an infectious disease outbreak.

First, we develop a method that provides individual-level inference on spatial covariates despite using several types of spatially aggregated binary disease surveillance data. Our method provides individual-level inference on spatial covariates by applying a series of transformations, including a change of support, to a bivariate point process model. The series of transformations preserves the convenient interpretation of desirable binary regression models that are commonly applied to individual-level disease surveillance data. Using a simulation experiment, we compare the performance of the proposed method under varying types of spatial aggregation against the performance of standard approaches using the original individual-level data. We illustrate our method by modeling individual-level probability of infection using a disease surveillance data set that has been aggregated to protect several at-risk or threatened species of bats in the northeastern U.S.

Second, we develop a staged approximate Bayesian model averaging (SABMA) method to estimate the spatio-temporal origins of an epidemic. Specifically, we estimate the number, locations, and times that a pathogen was introduced into a population using spatio-temporal binary disease surveillance data. We employ an ensemble of simple ecological diffusion processes to model the spatio-temporal spread of the pathogen from multiple locations. We study the statistical properties of the SABMA method, in terms of credible interval coverage for parameters and out-of-sample prediction performance, using a simulation experiment.

We then apply our SABMA method to two sets of binary disease surveillance data in white-tailed deer (*Odocoileus virginianus*); the first in the lower peninsula of Michigan in the U.S., and the second in southern Wisconsin and northern Illinois in the U.S.

Third, we develop a Bayesian hierarchical mixture of ecological diffusion models (BHMEDM) that provides inference on the number, locations, and times of pathogen introduction during an epidemic, using spatio-temporal binary disease surveillance data. Our model incorporates a mixture of ecological diffusion processes that account for both the growth and diffusion of the pathogen. As part of the hierarchical framework, we obtain inference on the spatio-temporal process that produced the pathogen introductions, and predict where new pathogen introductions are likely to occur in the future. We demonstrate the BHMEDM using binary disease surveillance data in white-tailed deer from southern Wisconsin and northern Illinois in the U.S.

Statistical methods for modeling infectious disease outbreaks using disease surveillance data

by

Nelson B. Walker

B.S., Brigham Young University, 2014

M.S., Kansas State University, 2018

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

2021

Approved by:

Major Professor
Dr. Trevor Hefley

Copyright

Abstract

Disease surveillance data are commonly used by epidemiologists, veterinary and plant pathologists, and wildlife and plant ecologists to identify, understand, mitigate, and prevent the spread of infectious disease. We develop three statistical methods that may be applied to spatio-temporal disease surveillance data to understand different aspects of an infectious disease outbreak.

First, we develop a method that provides individual-level inference on spatial covariates despite using several types of spatially aggregated binary disease surveillance data. Our method provides individual-level inference on spatial covariates by applying a series of transformations, including a change of support, to a bivariate point process model. The series of transformations preserves the convenient interpretation of desirable binary regression models that are commonly applied to individual-level disease surveillance data. Using a simulation experiment, we compare the performance of the proposed method under varying types of spatial aggregation against the performance of standard approaches using the original individual-level data. We illustrate our method by modeling individual-level probability of infection using a disease surveillance data set that has been aggregated to protect several at-risk or threatened species of bats in the northeastern U.S.

Second, we develop a staged approximate Bayesian model averaging (SABMA) method to estimate the spatio-temporal origins of an epidemic. Specifically, we estimate the number, locations, and times that a pathogen was introduced into a population using spatio-temporal binary disease surveillance data. We employ an ensemble of simple ecological diffusion processes to model the spatio-temporal spread of the pathogen from multiple locations. We study the statistical properties of the SABMA method, in terms of credible interval coverage for parameters and out-of-sample prediction performance, using a simulation experiment.

We then apply our SABMA method to two sets of binary disease surveillance data in white-tailed deer (*Odocoileus virginianus*); the first in the lower peninsula of Michigan in the U.S., and the second in southern Wisconsin and northern Illinois in the U.S.

Third, we develop a Bayesian hierarchical mixture of ecological diffusion models (BHMEDM) that provides inference on the number, locations, and times of pathogen introduction during an epidemic, using spatio-temporal binary disease surveillance data. Our model incorporates a mixture of ecological diffusion processes that account for both the growth and diffusion of the pathogen. As part of the hierarchical framework, we obtain inference on the spatio-temporal process that produced the pathogen introductions, and predict where new pathogen introductions are likely to occur in the future. We demonstrate the BHMEDM using binary disease surveillance data in white-tailed deer from southern Wisconsin and northern Illinois in the U.S.

Table of Contents

List of Figures

List of Tables

Acknowledgments

My Doctor of Philosophy program is the single most challenging and intellectually stimulating educational endeavor I have undertaken in my life. I owe a great debt to many for the support and encouragement I have received, including my immediate family, major professor, advisory committee, departmental faculty, collaborators, classmates, and friends. It truly "takes a village" (Clinton, 2006). I first and foremost acknowledge the love and support of my wife, Kelsee; she is a companion for all seasons. Thanks go to Gwen and Milo for joining us in the adventure. My heartfelt thanks go to my major professor, Dr. Trevor Hefley, for seeing potential in me as I started graduate school and acting as a fantastic mentor. I thank those that served on my advisory committee for any duration, Drs. Nora Bello, Gyuhyeong Goh, Michael Sanderson, and Christopher Vahl, for their interest, collaboration, encouragement, and rigor. I acknowledge the faculty and support staff in the statistics department at Kansas State University for their instruction, encouragement, and service. They created an environment that made my graduate education possible. I thank my collaborators at the National Wildlife Health Center and University of Wisconsin — Drs. Anne Ballmann, Ian McGahan, Robin Russell, and Daniel Walsh — for providing opportunities, interesting problems to solve, and invaluable subject matter expertise. I likewise thank my collaborators at the various Departments of Natural Resources of Illinois, Iowa, Michigan, Minnesota, and Wisconsin. Special thanks go to fellow students in Dr. Hefley's research group, Congxing Zhu, Haoyu Zhang, Liying Jin, and Narmadha "Meenu" Mohankumar. I acknowledge and express thanks as well for research funding support from the US Geological survey through USGS G18AC00317 and USGS G16AC00413, from the Kansas State University Graduate School via the Timothy R. Donoghue Graduate Scholarship Program, and from Dr. Lolafaye Coyne and the Kansas State University Department of Statistics via the Lolafaye Coyne Graduate Research Scholarship.

Dedication

For Kelsee, Gwen, and Milo; hope for a world with fewer infectious diseases.

Introduction

According to the World Health Organization, three of the ten leading causes of death worldwide in 2020 stemmed from broad categories of infectious disease (lower respiratory infections, neonatal conditions, and diarrheal diseases; World Health Organization, 2020). It was estimated that these disease categories and other infectious diseases yearly take the lives of at least 6.1 millions of people around the world and negatively impact global economic output (World Health Organization, 2020). For example, Fan et al. (2018) estimated that the yearly influenza pandemic alone costs the world economy approximately $500 billion in lost economic output and life. By contrast, the current COVID-19 pandemic is estimated to have taken the lives of approximately 4.8 million people (as of October 4, 2021; World Health Organization, 2021). The equivalent of nearly 16 trillion U.S. dollars was provided globally in 2020 to soften the economic disruption, while the global economy shrunk by 3.5% (Yeyati and Filippini, 2021). However, these statistics fail to encompass the human misery caused by illness and death of loved ones and caregivers. For example, Hillis et al. (2021) estimated that as of April 30, 2021 at least 1.56 million children globally (likely much higher) have lost a primary or secondary care-giver to COVID-19. During the pandemic, mental health has declined, particularly among those with lower socioeconomic status, due to social isolation, job loss/economic uncertainty, and other factors (Graham, 2020; Cost et al., 2021; Panchal et al., 2021; Giuntella et al., 2021). These sobering facts highlight the importance of health care and public health monitoring measures to identify and contain infectious disease outbreaks at an early stage. Collecting and analyzing disease surveillance data is one way that public health researchers identify factors that contributed to an outbreak. Public health officials can then advocate for policies that reduce the risk for similar outbreaks.

An early example of data collection and analysis that helped identify, contain, and prevent infectious disease outbreaks comes from the 1854 cholera epidemic in the Soho area of London, England. This epidemic is at least partially notable because of its virulence - it

resulted in the deaths of over 500 individuals in ten days. As the outbreak unfolded, a local surgeon, Dr. John Snow, collected data by interviewing inhabitants and plotted the number and locations of cholera cases on a street map of the area (Chave, 1958; Centers for Disease Control and Prevention, 2012). Dr. Snow's analysis led him to hypothesize that the outbreak was caused by contaminated water from the Broad Street pump. As the outbreak waned, Dr. Snow met with the local governing board and recommended that the pump handle be removed (Chave, 1958). Though incredulous, the local board acquiesced. Later that year, a local priest, Reverend Henry Whitehead, conducted an independent investigation of the epidemic among his parishioners in an attempt to disprove Dr. Snow's hypothesis. At the conclusion of his investigation, Reverend Whitehead concurred with Dr. Snow's belief that the outbreak was likely caused by contaminated water from the pump. Further, he concluded that the source of the contamination was likely the contents of a sick child's diaper that had been washed into a leaky cistern just a few feet from the Broad Street pump. Whitehead later credited deactivating the Broad Street pump with preventing a second wave of cholera in the same neighborhood soon after the first wave had concluded (Chave, 1958). While Dr. Snow's theory about the spread of cholera via contaminated water was not immediately accepted, his efforts are remarkable, in part, for his use of spatial analysis (Chave, 1958; Centers for Disease Control and Prevention, 2012). These events are also instructive because the location of an outbreak or pathogen introduction can be discerned separately from the source of the pathogen. In this case, the location of the cholera pathogen introduction for the neighborhood was the cistern and the adjacent pump, while the originating source of the pathogen in the neighborhood was the sick child. In the subsequent 167 years, the field of epidemiology has evolved into a mature science with a rich knowledge base in infectious disease. Likewise, fields such as wildlife disease ecology and animal pathology have appeared and matured. In modern times, the collection and analysis of disease surveillance data has become essential to quickly identify and mitigate disease outbreaks, particularly in the age of emerging zoonotic diseases like COVID-19 (Simonsen et al., 2016; Watsa and Wildlife Disease Surveillance Focus Group, 2020; Ibrahim, 2020; Budd et al., 2020).

Disease surveillance data come in several forms with various spatio-temporal resolutions,

from counts of individuals (human, animal, or plant) that were diagnosed with a disease (within a geopolitical area over a time period), to individual diagnostic test results (at precise times and locations). Whole volumes have been written about how to effectively establish and execute disease surveillance programs for human, animal, and plant health (e.g., Salman, 2003; Institute of Medicine, 2007; Lee et al., 2010; M'ikanatha et al., 2013; World Organization for Animal Health, 2021; also see Artois et al., 2009 for a chapter about wildlife disease surveillance). Disease surveillance programs exist at the local, state/provincial, national, and international levels. At the national level within the U.S., disease surveillance data are collected by a number of agencies and departments, including the U.S. Department of Agriculture's Animal and Plant Health Inspection Service (APHIS; livestock and plant), the Centers for Disease Control and Prevention (human), the U.S. Food and Drug Administration (human), the Department of Defense (human; United States Government Accountability Office, 2003), and the U.S. Geological Survey (wildlife). At the international level, infectious disease surveillance data is collected by the World Health Organization and the World Organization for Animal Health.

The prudence of collecting infectious disease surveillance data in humans, livestock, and plants is well-recognized. Conducting surveillance for infectious disease in wildlife populations is also prudent in several respects. First, altruistically, it is wise to protect environmental and ecosystem health (Wilcox et al., 2012). Infectious disease can ravage local species populations, impact the health of the ecosystem, and reduce biodiversity (Wilcox et al., 2012). In fact, the Endangered Species Act in the U.S. (16 U.S.C. Ch. 35) codifies the protection of endangered species, including from the ravages of infectious disease. Second, infectious disease outbreaks in farmed and hunted wildlife can have a negative impact on local economies in the form of lost hunting revenue (Narrod et al., 2012; Barratt et al., 2019; Erickson et al., 2019). Third, pathogens in wildlife or livestock may jump between species and infect humans (Aguirre et al., 2012). Examples of this include avian influenza, swine flu, Nipah virus, Middle East respiratory syndrome, and Ebola (Rohr et al., 2019). In fact, it is estimated that up to 76% of emerging infectious diseases in humans are zoonotic (Rohr et al., 2019).

The U.S. Geological Survey National Wildlife Health Center (NWHC) in the U.S. is one of the primary research institutions for tracking and studying infectious disease in wildlife. Examples of diseases that have been of interest include: avian influenza (birds), chronic wasting disease (CWD; deer and other cervids), West Nile virus (e.g., birds, humans), and white-nose syndrome (WNS; bats). The disease surveillance data that motivate the methodological developments in this dissertation came from partnering with researchers at the NWHC and numerous state agencies. In particular, we were granted access to surveillance data on WNS in multiple species of bats and CWD in white-tailed deer (*Odocoileus virginianus*). While the data examples in this dissertation are wildlife-centric, the methods that are developed and presented may be applied to plant and human infectious disease surveillance data as well.

The objective of my research was to expand on previous work by Hefley et al. (2017c) to model the dynamics of CWD in Wisconsin. Additionally, a major effort would focus on forecasting how CWD will spread in the upper Midwestern U.S. states. The process of attaining these goals was broken into several steps, the first few of which were the focus of my master's research (Walker, 2018; Walker et al., 2020) and PhD dissertation.

The first challenge was that the CWD surveillance data suffered from location error. Location error occurs when the recorded location of an observation is different from its true location. In the case of the CWD surveillance data, the location of each tested deer was recorded as the centroid of the section of land that the deer occupied (the area of each section was ≈ 2.59 km^2, according to the Public Land Survey System). Without accounting for this location error, commonly used binary regression models for disease risk factor analyses would provide biased inference on parameters associated with spatial covariates. Walker et al. (2020) developed a method that applied a change of support (COS) transformation to account for location error and obtain unbiased inference on spatial covariates.

Chapter 1 of this dissertation was published as Walker et al. (2021). This chapter extended the ideas presented in Walker et al. (2020) to enable individual-level spatial inference on various types of aggregated disease surveillance data. Chapter 1 acknowledged that disease surveillance data are often aggregated to protect privacy. The chapter identified

several types of aggregated data and detailed how a COS transformation could be used to obtain several distributional results. Models based on these distributional results enabled individual-level spatial inference that would have otherwise been impossible to obtain from aggregated data.

The second challenge was to extend the capability developed by Hefley et al. (2017b) and showcased in Hefley et al. (2017c) and Hefley et al. (2020). Hefley et al. (2017c) modeled the spatio-temporal dynamics of the growth and spread of CWD across southwestern Wisconsin in the U.S. Hefley et al. (2017c) assumed that a single location and time of pathogen introduction was responsible for the spread of the pathogen. Likewise, Hefley et al. (2020) modeled the spatio-temporal dynamics of the growth and spread of the causative pathogen for WNS across the eastern U.S., assuming a single location and time of pathogen introduction. Unlike Hefley et al. (2017c), however, Hefley et al. (2020) also estimated the location and time of pathogen introduction. Holistically, disease surveillance data of both CWD in Wisconsin (and the surrounding states) and WNS in the continental U.S. have suggested that the respective pathogens were introduced at multiple locations and times. Hence, a single introduction model would be inadequate for holistically modeling the spread of either pathogen. Methods would need to be developed that could both estimate the number, locations, and times that the pathogen was introduced and also estimate the diffusion and growth dynamics of the pathogen. Once these and other modeling capabilities were developed, they could be combined for a future multi-state analysis of CWD surveillance data.

Chapter 2 and chapter 3 of this dissertation were developed somewhat concurrently to ensure success and provide options for a future multi-state CWD data analysis. The purpose of chapter 2 was to first examine whether it was possible to estimate the number, locations, and times of pathogen introduction and tackle the associated trans-dimensional estimation problem. This research effort resulted in an approximate Bayesian model-averaged method for estimating the number, locations, and times of pathogen introduction using an ensemble of simple ecological diffusion processes. Chapter 3 tackled the additional problem of obtaining inference on the spatio-temporal process associated with the number, locations, and times of pathogen introduction. Ultimately, in chapter 3 the trans-dimensional estimation

problem was re-framed and addressed using tools from mixture model analysis and the missing data literature. An additional goal of obtaining inference on the spatio-temporal process associated with the number, locations, and times of pathogen introduction was addressed using a point process model. Partway through development of chapter 3, a collaborator and applied mathematician, Dr. Ian McGahan, introduced an approximate analytical solution to the ecological diffusion partial differential equation (PDE) used by Hefley et al. (2017c) and Hefley et al. (2020). Dr. McGahan's addition was crucial because the method in chapter 3 could be adapted to estimate the spatio-temporal diffusion and growth dynamics of the pathogen.

The research in this dissertation represents a considerable contribution to my original objectives by enabling inference on the spread, growth, and spatial distribution of the causative pathogen for CWD in the upper Midwestern U.S. The work of integrating the methods from Walker et al. (2020), chapter 1 (Walker et al., 2021), and chapter 3 in a multi-state CWD data analysis will be accomplished at a later time.

Chapter 1

Recovering Individual-level Spatial Inference from Aggregated Binary Data.

1.1 Abstract

Binary regression models are commonly used in disciplines such as epidemiology and ecology to determine how spatial covariates influence individuals. In many studies, binary data are shared in a spatially aggregated form to protect privacy. For example, rather than reporting the location and result for each individual that was tested for a disease, researchers may report that a disease was detected or not detected within geopolitical units. Often, the spatial aggregation process obscures the values of response variables, spatial covariates, and locations of each individual, which makes recovering individual-level inference difficult. We show that applying a series of transformations, including a change of support, to a bivariate point process model allows researchers to recover individual-level inference for spatial covariates from spatially aggregated binary data. The series of transformations preserves the convenient interpretation of desirable binary regression models that are commonly applied to individual-level data. Using a simulation experiment, we compare the performance of our proposed

method under varying types of spatial aggregation against the performance of standard approaches using the original individual-level data. We illustrate our method by modeling individual-level probability of infection using a data set that has been aggregated to protect several at-risk or threatened species of bats. Our simulation experiment and data illustration demonstrate the utility of the proposed method when access to original non-aggregated data is impractical or prohibited. This chapter was published as an article in *Spatial Statistics* as Walker et al. (2021).

1.2 Introduction

Spatially referenced binary data are among the most common types of data that enable inference about spatial covariates. Scientists and policy makers are often interested in understanding how spatial covariates influence the probability of a binary outcome, such as whether a plant or animal tests positive or negative for a disease. Sometimes spatial binary data are aggregated to protect privacy. For example, wild plants and animals are protected by law (e.g., threatened or endangered species under the U.S. Endangered Species Act (ESA) of 1973). As a result, spatially referenced binary data involving protected plants and animals may be reported in aggregate to reduce the potential for human contact (e.g., tourism, vandalism, and theft). The aggregation process can make individual-level inference difficult to obtain for spatial covariates because the original values of the binary responses, locations, and spatial covariates cannot be recovered.

An example where spatial binary data are aggregated is a disease surveillance study for white-nose syndrome (WNS), which is caused by the fungal pathogen *P. destructans*. In a disease surveillance study, binary observations are collected on individual bats found within geopolitical areas (counties). However, the observations are aggregated to the county-level when making them accessible to researchers and the public in accordance with federal law and to protect the wildlife (see **Figure 1.1**). The map in **Figure 1.1** indicates which counties in the northeastern United States contained individual bats that were tested and which counties had at least one diagnosed case of WNS from 2008-2012. When the individual test results are

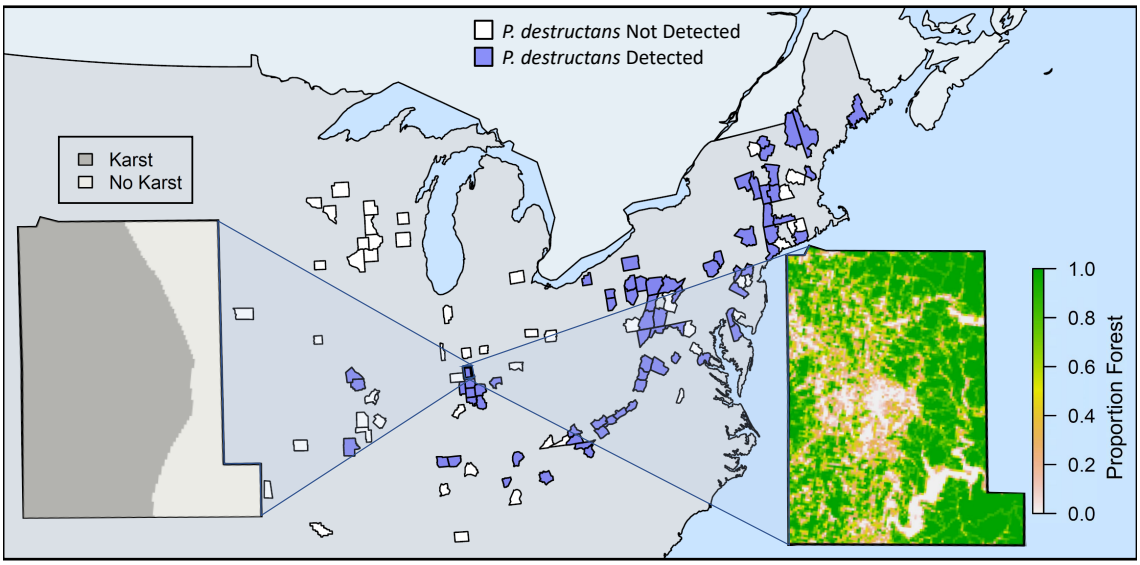

Figure 1.1: The motivating data set shows which counties contained bats that were individually tested for *P. destructans*, the causative agent of white-nose syndrome, within the northeastern United States from 2008-2012. The counties that contained at least one bat that tested positive for *P. destructans* are shown in purple fill while counties with no positive bats are shown in white fill. The covariates 'proportion of land classified as forest' (inset right) and 'presence of karst' (inset left) from Monroe county, Indiana, USA (outlined in bold black). Karst is a type of landscape characterized by caves and sinkholes that can provide habitat to cave-hibernating bats. Spatially referenced wildlife data are often accessible to researchers in aggregated form to reduce the potential for human contact. When binary data within a county are aggregated into an indicator that denotes whether the county contained at least one sampled bat that tested positive, individual-level spatial covariates and inference cannot be obtained.

aggregated as shown in **Figure 1.1**, it can be difficult to recover the original individual-level inference for spatial covariates because the original values of the binary response, location, and spatial covariates for each observation are unknown. For these types of data, researchers commonly resort to fitting regression models to the aggregated data and may interpret the areal-level inference about spatial covariates as if it was obtained from a model that was fit to individual-level data, which is a well-documented ecological fallacy (Piantadosi et al., 1988; Gotway and Young, 2002).

Univariate point process-based methods have traditionally formed the backbone of efforts to make individual-level inference on spatially aggregated data (e.g., Bradley et al., 2016;

Hefley et al., 2017a; Taylor et al., 2018; Gelfand and Shirota, 2019). Perhaps less common, bivariate point process models enable individual-level inference on spatially aggregated data where the non-aggregated data consist of binary marks at specific locations (Diggle et al., 2010a; Chang et al., 2015; Wang et al., 2017; Johnson et al., 2019; Walker et al., 2020). For binary data, these methods are capable of recovering individual-level inference on spatial covariates under varying types of spatial aggregation (see **Table 1.2** and **Figure 1.2**). For example, when the individual-level binary data are aggregated over areal units into separate counts of the number of observations with a specific binary mark, the methods by Wang et al. (2017), Johnson et al. (2019), and Walker et al. (2020) can be used to recover individual-level inference for spatial covariates (see **Table 1.1**, Type C). When at least some of the binary data are aggregated into counts (e.g., number of observations with a mark of zero) and the rest of the data are not aggregated, the methods from Diggle et al. (2010a), Chang et al. (2015), and Walker et al. (2020) can be used to recover individual-level inference for spatial covariates (see **Table 1.1**, Type B).

Aside from Type B and C data, we have identified two additional types of aggregated data that appear in practice. First, when the data are aggregated into counts of the total number of observations in areal units and also aggregated into binary indicators that denote whether at least one observation in the areal unit had a mark of one, the existing methods are insufficient to recover individual-level inference on spatial covariates (see **Table 1.1** and **Figure 1.2**, Type D). Likewise, to the best of our knowledge, no methods exist to recover individual-level inference on spatial covariates when the aggregated data consist only of the binary indicators over areal units (see **Table 1.1** and **Figure 1.2**, Type E). This is unfortunate because, presumably, data categorized as Type D or E are more likely to be accessible when compared to data of Type B or C. We hypothesize that Type D and Type E data would be more accessible because Type D and E are a degraded form of Types A-C data and offer a higher degree of privacy protection. Thus, Type D and E aggregated data are an untapped data source for individual-level inference. For example, the disease surveillance example from **Figure 1.1** may be classified as Type E data.

The contribution of this chapter is to enable individual-level inference for spatial co-

Table 1.1: Different types of aggregation or privacy protection for spatially referenced binary data, along with their relative information content, and references that successfully recover individual-level inference for each type of data. See **Figure 1.2** for visualizations of the aggregation types.

Type	Aggregation/ Privacy Protection	Information Content	Example References That Enable Individual-level Inference
A	None	Complete	Diggle and Giorgi (2019)
B	Some non-aggregated data w/ subregion counts of ones or zeros	High	Diggle et al. (2010a) Chang et al. (2015) Walker et al. (2020)
C	Subregion counts of ones and zeros	High	Wang et al. (2017) Johnson et al. (2019) Walker et al. (2020)
D	Total subregion counts and subregion indicator of ones	Medium	**No methods currently exist**
E	Subregion indicator of ones or zeros	Low	**No methods currently exist**

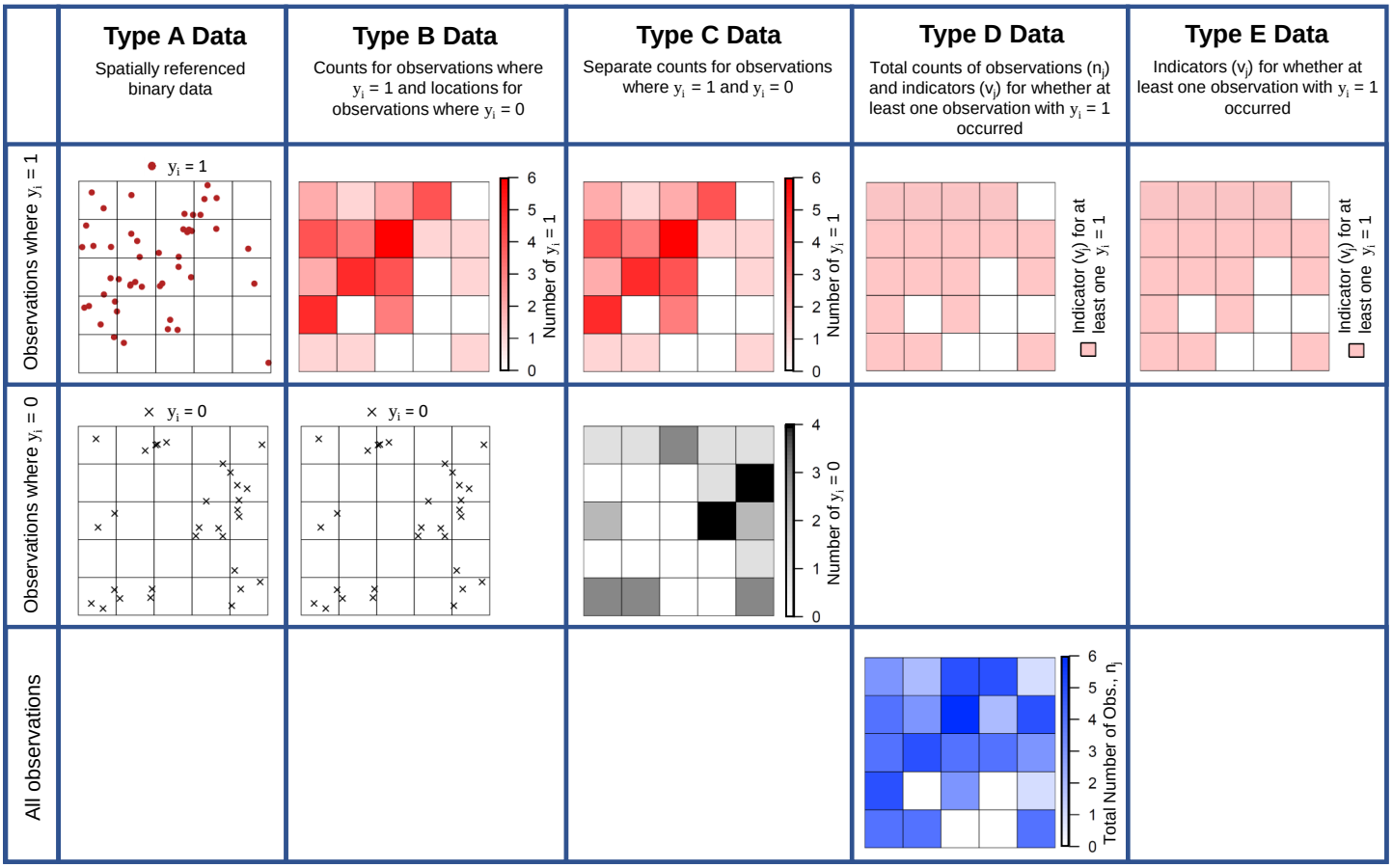

Figure 1.2: Graphical representations of the types of aggregation for spatially referenced binary data found in **Table 1.1**. The data set shown under Type A is progressively aggregated across sub-regions, starting from the exact locations of all observations (Type A data) and ending with binary indicators (Type E data). We define y_i as the binary mark associated with the i^{th} spatially referenced observation. For the j^{th} subregion, we define n_j as the total number of observations contained therein. We also define v_j as a binary indicator that at least one observation with $y_i = 1$ occurred in the j^{th} subregion.

variates from Type D and E aggregated binary data. We accomplish this by transforming the bivariate inhomogeneous Poisson point process (BIPPP) regression model and applying several distributional results. Importantly, and following Walker et al. (2020), the proposed methods preserve the interpretation of commonly used binary regression methods (e.g., logistic and probit regression). Thus the proposed methods are easy to interpret and are widely applicable to aggregated binary data.

The remainder of this chapter proceeds as follows: In the Methods Section, we review regression models for binary data, including the BIPPP. We then present several distributional results for the transformed BIPPP that may be used to recover individual-level spatial inference under various types of aggregation. In the Simulation Experiment Section, we evaluate and compare the proposed models to traditional approaches for the analysis of spatial binary data (e.g., logistic regression) using a simulation study. In the Application Section, we apply our proposed regression models to a data example from wildlife disease surveillance where the aggregated data result in a binary indicator for each geopolitical unit. Finally, in the Discussion Section, we identify potential modifications and model comparisons that practitioners may consider.

1.3 Methods

1.3.1 Binary Regression

Binary regression is arguably one of the most popular types of regression models and can be written as

$$y_i \sim \text{Bernoulli}(p_i), \tag{1.1}$$

$$g(p_i) = \beta_0 + \mathbf{x}_i'\boldsymbol{\beta}, \tag{1.2}$$

where y_i is the i^{th} binary response from $\mathbf{y} \equiv (y_1, y_2, \ldots, y_n)'$, n is the number of observations, p_i is the probability that $y_i = 1$, and $g(\cdot)$ is an appropriate link function (e.g., logit or

probit). Additionally, β_0 is an intercept, $\mathbf{x}_i \equiv (x_1, x_2, \ldots, x_q)'$ is a vector of q covariates, and $\boldsymbol{\beta} \equiv (\beta_1, \beta_2, \ldots, \beta_q)'$ is a vector of q regression coefficients. Regression models like (1.1-1.2) are often used to model spatial binary data (e.g., Gelfand and Schliep, 2018; Diggle and Giorgi, 2019). In the case that (1.2) includes spatial covariates $\mathbf{x}(\mathbf{s})$, then p_i becomes a spatially varying function such that

$$g(p(\mathbf{s})) = \beta_0 + \mathbf{x}(\mathbf{s})'\boldsymbol{\beta}, \tag{1.3}$$

where $\mathbf{s} \equiv (s_1, s_2)'$ is a coordinate vector within the study area $\mathcal{S}$ (i.e., $\mathbf{s} \subseteq \mathcal{S}$). In what follows, we specify $g(\cdot)$ using the logit link function, however, as with any binary regression model, the choice is flexible.

A similar spatial binary regression model to (1.1) and (1.3) that incorporates the locations of n observations in a study area $\mathcal{S} \subset \mathbb{R}^2$, is the bivariate point process (Gelfand and Schliep, 2018). Perhaps the most common type of point process used for binary data is the bivariate inhomogeneous Poisson point process (BIPPP; Gelfand and Schliep, 2018). The BIPPP is a joint distribution composed of a Poisson probability mass function that generates n, a location density that generates the coordinates of each observation, $\mathbf{u}_i$, and the Bernoulli probability mass function from (1.1) that generates binary outcomes, $\mathbf{y}$, called marks (Gelfand and Schliep, 2018). The BIPPP can be written as:

$$\begin{aligned} f(n, \mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_n, \mathbf{y}|\lambda, p) = & \frac{e^{-(\int_{\mathcal{S}} \lambda(\mathbf{s})d\mathbf{s})}(\int_{\mathcal{S}} \lambda(\mathbf{s})d\mathbf{s})^n}{n!} \times \\ & \prod_{i=1}^{n} \frac{\lambda(\mathbf{u}_i)}{\int_{\mathcal{S}} \lambda(\mathbf{s})d\mathbf{s}} \; p(\mathbf{u}_i)^{y_i}(1 - p(\mathbf{u}_i))^{1-y_i}, \end{aligned} \tag{1.4}$$

where $\lambda(\cdot)$ is a spatially varying thinned intensity function that captures both the distribution of bats and the sampling process (Gelfand and Shirota, 2019). The function $p(\cdot)$ is identical to (1.3) and may be viewed as a classification function because it relates a binary mark to each of n locations. For example, in our motivating data set, the binary marks represent test results for individual bats that tested positive (i.e., $y_i = 1$) or negative ($y_i = 0$) for *P. destructans*, the causative agent of WNS. We note that the BIPPP offers no obvious

advantage for spatial binary data over the model formed from (1.1) and (1.3) unless the binary observations are spatially aggregated, the locations of the observations are obscured by location error (e.g., Walker et al., 2020), or the observations are collected via preferential sampling (e.g., Diggle et al., 2010b).

In many applications, researchers often specify $\lambda(\cdot)$ using

$$\log(\lambda(\mathbf{s})) = \alpha_0 + \mathbf{z}(\mathbf{s})'\boldsymbol{\alpha}\,, \tag{1.5}$$

where α_0 is an intercept, $\mathbf{z}(\mathbf{s}) \equiv (z(\mathbf{s})_1, z(\mathbf{s})_2, \ldots, z(\mathbf{s})_r)'$ is a vector of r spatial covariates, and $\boldsymbol{\alpha} \equiv (\alpha_1, \alpha_2, \ldots, \alpha_r)'$ is a vector of r regression coefficients (Gelfand and Schliep, 2018). Some situations may require an alternative, and potentially more flexible, specification in (1.5). For example, a Gaussian process could be added to (1.5) by way of a spatial random effect (Gelfand and Schliep, 2018). We focus on a log-linear specification for $\lambda(\cdot)$ because the specification is reasonable for our motivating data set and because we can more easily discover parameter identifiability issues.

1.3.2 Change of Support and Distributional Results

While the distributions from (1.1) and (1.4) are appropriate for spatially referenced binary data, they are inadequate when the spatial binary data are aggregated (see **Table 1.1**). In what follows, we outline several transformations of the BIPPP that result in distributions that match the distributional attributes of aggregated spatial binary data of Types C, D, and E (see **Table 1.1** and **Figure 1.2**).

The transformation of a spatial process from continuous to areal support is called a change of support (COS). To implement a COS, the study area $\mathcal{S}$ is partitioned into J non-overlapping subregions, $\mathcal{A}_1, \mathcal{A}_2, \ldots, \mathcal{A}_J$, such that $\mathcal{S} = \cup_{j=1}^{J} \mathcal{A}_j$. The partition is determined by how the data were aggregated. For example, our motivating data set reported the county that each bat was sampled from in the northeastern United States (see **Figure 1.1**). Thus, $\mathcal{S}$ is defined by the combined area of the counties that contained sampled bats and the partition is defined by the boundaries of the counties which contained the bats.

If we know the number of observations with a mark of one (n_{1j}) and a mark of zero (n_{0j}) contained within the j^{th} subregion (see **Table 1.1** and **Figure 1.2**, Type C data), a result of applying the COS to the BIPPP is n_{1j} and n_{0j} are Poisson random variables distributed as follows (Gelfand and Schliep, 2018):

$$n_{1j} \sim \text{Pois}(\int_{\mathcal{A}_j} \lambda(\mathbf{s})p(\mathbf{s})d\mathbf{s}), \tag{1.6}$$

$$n_{0j} \sim \text{Pois}(\int_{\mathcal{A}_j} \lambda(\mathbf{s})(1-p(\mathbf{s}))d\mathbf{s}). \tag{1.7}$$

The joint distribution of n_{1j} and n_{0j} is an appropriate density for binary data that have been aggregated into counts and results in a regression model that recovers individual-level inference on spatial covariates. Effectively, this models two point patterns, with intensities $\lambda(\mathbf{s})p(\mathbf{s})$ and $\lambda(\mathbf{s})(1-p(\mathbf{s}))$, for presence and absence of a mark. Wang et al. (2017) and Walker et al. (2020) both used this type of binary regression model to make individual-level inference from aggregated binary data using spatial covariates. Similar to (1.6-1.7), the number of observations in the j^{th} subregion, $n_j = n_{1j} + n_{0j}$, is also a Poisson random variable (Cressie and Wikle, 2011, p. 207),

$$n_j \sim \text{Pois}(\int_{\mathcal{A}_j} \lambda(\mathbf{s})d\mathbf{s}). \tag{1.8}$$

1.3.3 Proposed Change-of-Support based Methods

In some cases, we may have access to n_j (e.g., the total number of individuals tested within each county) and a binary indicator $v_j = \text{I}(n_{1j} > 0)$ for each subregion (see **Table 1.1** and **Figure 1.2**, Type D data). In our motivating data set, $v_j = 1$ indicates that the j^{th} county contains at least one sampled bat that tested positive for the pathogen, and $v_j = 0$ indicates that all of the sampled bats tested negative in the county. Conditioning v_j on n_j, we obtain the following density:

$$v_j | n_j \sim \text{Bern}(1-(1-\tilde{p}_j)^{n_j}), \tag{1.9}$$

where

$$\tilde{p}_j = \frac{\int_{\mathcal{A}_j} \lambda(\mathbf{s}) p(\mathbf{s}) d\mathbf{s}}{\int_{\mathcal{A}_j} \lambda(\mathbf{s}) d\mathbf{s}}. \tag{1.10}$$

The conditional distribution of v_j given n_j is an appropriate density for binary data that have been aggregated into Type D data. The joint density of (1.8) and (1.9) can also be used to construct a regression model for Type D aggregated binary data. Models based on (1.9) or the joint distribution of (1.8) and (1.9) are a novel development because both can recover individual-level inference on spatial covariates from Type D aggregated data (see **Table 1.1**).

Under the form of aggregation in Type E data, we may assume only v_j is given for each subregion (see **Table 1.1** and **Figure 1.2**). The data generated by the indicator function follow a Bernoulli distribution and is given as follows:

$$v_j \sim \text{Bern}(1 - e^{-\int_{A_j} \lambda(\mathbf{s}) p(\mathbf{s}) d\mathbf{s}}). \tag{1.11}$$

A model for Type E data based on (1.11) is also a novel development, as the model is capable of recovering individual-level inference on spatial covariates from Type E aggregated data.

1.3.4 Parameter Identifiability

The distributions presented in Section 1.3.3 form the basis for regression models that recover individual-level spatial inference from various types of aggregated binary data (see **Table 1.1** and **Figure 1.2**). Like all binary regression models and point process models, the proposed transformed BIPPP models may have parameter identifiability issues (e.g., complete separation; Hefley and Hooten, 2015) when sample size is small or the data contain little information (e.g., a very large number of zeros).

1.3.5 Model Implementation

We use the Nelder-Mead algorithm in the program R to numerically minimize the negative log-likelihoods for the densities introduced in this chapter and simultaneously estimate all parameters (R Core Team, 2021). Evaluating the negative log-likelihood functions requires approximating the integrals contained therein. We approximate the integrals using simple quadrature for ease of implementation (e.g., $\int_{\mathcal{A}_j} \lambda(\mathbf{s})d\mathbf{s} \approx \sum_{k=1}^{K} |W| * \lambda(\mathbf{s}_k)$, where $\lambda(\mathbf{s}_k)$ is the value of $\lambda(\mathbf{s})$ at the k^{th} quadrature point and $|W|$ is the area of a grid cell that is both a subset of $\mathcal{A}_j$ and approximated by a quadrature point). For all model parameters, we approximate variances by inverting the Hessian matrix and then construct Wald-type confidence intervals (CIs).

1.4 Simulation Experiment

We conducted a simulation experiment to compare the performance of our proposed models, using different types of aggregated binary data, to traditional models for non-aggregated binary data (e.g., logistic regression). We simulated data using a unit square study area, $\mathcal{S} = [0, 1] \times [0, 1]$, that was divided into 400 regular grid cells (subregions), such that $\mathcal{S} = \cup_{j=1}^{400} \mathcal{A}_j$ and $|\mathcal{A}_j| = \frac{1}{400}$. We generated spatial covariates, $x(\mathbf{s})$ and $z(\mathbf{s})$, and simulated the locations and binary marks of observations from a BIPPP where the intensity function was $\log(\lambda(\mathbf{s})) = \alpha_0 + \alpha_1 z(\mathbf{s})$ and the classification function was $\text{logit}(p(\mathbf{s})) = \beta_0 + \beta_1 x(\mathbf{s})$. We focused on and compared estimates of β_1 among five models because β_1 is highly affected by aggregation and inference on the slope parameters of the classification function is likely to be the focus of many applied studies (Walker et al., 2020). We accomplished the comparison of estimates of β_1 by assessing bias, coverage probabilities (CPs), and relative efficiency for estimates of β_1 among the following five scenarios:

1. A traditional logistic regression model from (1.1) and (1.3) fit to non-aggregated data (see **Table 1.1** and **Figure 1.2**, Type A);

2. A joint model for n_{1j} and n_{0j} that is specified by combining the distributions in (1.6)

and (1.7; see **Table 1.1** and **Figure 1.2**, Type C);

3. A joint model for v_j and n_j that is specified by combining the distributions in (1.8) and (1.9; see **Table 1.1** and **Figure 1.2**, Type D);

4. The conditional model for v_j given n_j from (1.9; see **Table 1.1** and **Figure 1.2**, Type D);

5. The Bernoulli model for v_j from (1.11; see **Table 1.1** and **Figure 1.2**, Type E).

We simulated 1000 data sets from four different settings using a combination of two factors: covariate equivalence ($x(\mathbf{s}) = z(\mathbf{s})$ vs. $x(\mathrm{s}) \neq z(\mathbf{s})$); and average sample size (small vs. large). Thus our simulation experiment uses a total of 4,000 simulated data sets and realizations of $z(\mathbf{s})$ and $x(\mathbf{s})$. Each simulated data set was aggregated to fit each data type in scenarios 2-5. We drew each spatial covariate realization from a low-rank Gaussian process (Higdon, 2002) on a 200×200 grid with knots at every fourth grid cell to reduce computation time. We chose parameter values of $\alpha_1 = 1$, and $\beta_1 = 1$ for all settings. We chose values for α_0 and β_0 for each setting such that the average sample size per subregion was either 10 or 50 (small vs. large) and the proportion of subregions that contained a binary mark of one was approximately constant across all settings. The values of α_0 and β_0 in settings 1-4 were $7.800, 9.410, 7.820, 9.405$ and $-5.500, -7.070, -4.750, -6.350$, respectively.

We fit the model in scenario one (i.e., traditional logistic regression) using the `glm` function in R to obtain the maximum likelihood estimates (MLEs) of β_0 and β_1. We fit the models in scenarios two through five as described in Section 1.3.5. For each model and setting, we calculated and compared the CPs from the 95% Wald-type CIs for β_1. We also constructed box plots comparing the distribution of $\hat{\beta}_1$ obtained from the 1000 data sets for each scenario and setting. We calculated the standard deviation of the empirical distribution of the 1000 estimates of β_1 in each scenario. We then calculated the relative efficiency of $\hat{\beta}_1$ for scenarios two through five by dividing the standard deviation of the distribution of $\hat{\beta}_1$ for the respective scenario by that of scenario one. Lastly, we calculated the mean squared predictive error (MSPE) in the estimated intensity and probability surfaces for each of the models in scenarios

two through five. However, we only calculated the MSPE for the estimated probability surface for the model in scenario one.

When binary data are generated according to a BIPPP and then spatially aggregated, we expect to obtain unbiased estimates in scenarios two, three, four, and five. Of the proposed models based on the distributional results presented in Sections 1.3.2-1.3.3, we expect that the model for scenario two will have the highest relative efficiency among all settings covered by the experiment, followed by the models from scenarios three, four, and five. We expect the MSPE of the estimated intensity and probability surfaces to be smallest for the model in scenario two, followed by three, four, and five. We provide annotated R code capable of reproducing the simulation experiment in the `simulation.R` file in the supporting information for Walker et al. (2021).

1.4.1 Simulation Results

In our simulation experiment, we crossed two factors (average sample size per subregion and covariate equivalence) with two levels each. With our choices of α_0, the average numbers of observations within each grid cell were about 10.2 and 50.1 for small and large sample settings, respectively. With our choices of β_0 for each setting, we maintained a proportion of approximately 0.11 of grid-cells that contained a binary mark of one (see **Table 1.2**).

As expected, under the model and data in scenario one (traditional logistic regression with no data aggregation), the MLEs for β_1 appear to be unbiased for all settings and had CPs between 0.945 and 0.951. Under the model and data in scenario two (joint distribution of n_{1j} and n_{0j}) the MLEs for β_1 appear to be unbiased for all settings in the simulation study (see **Figure 1.3** for graphical comparisons of estimates and Appendix A for additional plots and summaries). The CPs for $\hat{\beta}_1$, in scenario two, were between 0.94 and 0.961 for all settings. Additionally, the relative efficiency of $\hat{\beta}_1$, obtained from scenario two, ranged from about 1.1 (settings 1, 2) to about 1.2 (setting 3). The CPs obtained for scenarios one and two, and efficiencies for scenario two, are available in **Table 1.2**.

Under the model and data in scenario three (joint distribution of v_j and n_j) the MLEs

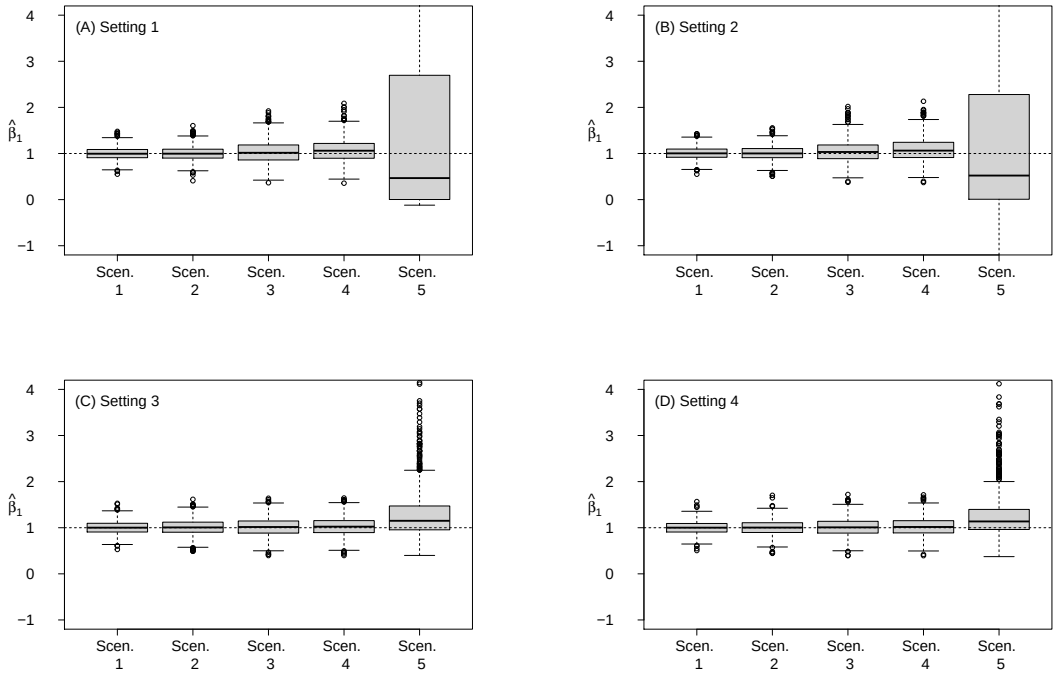

Figure 1.3: Panels (A) and (B) show box plots of results from small and large average sample size simulation experiment settings where $x(\mathbf{s}) = z(\mathbf{s})$. Panels (C) and (D) show small and large sample size simulation experiments where $x(\mathbf{s}) \neq z(\mathbf{s})$. We show maximum likelihood estimates of β_1 obtained using five different models (each under a different data aggregation scenario), which included: **Scen. 1)** logistic regression with no data aggregation (Type A data); **Scen. 2)** a joint model for n_{1j} and n_{0j} where binary data were aggregated into counts for each subregion (Type C data); **Scen. 3)** a joint model for v_j and n_j using data aggregated into a count and indicator variable for each subregion (Type D data); **Scen. 4)** a conditional model for v_j given n_j using data aggregated into a count and indicator variable for each subregion (Type D data); **Scen. 5)** a Bernoulli model for v_j using data aggregated into an indicator variable for each subregion (Type E data). Each of the four panels used 1,000 simulated data sets, and each panel shows the true value of $\beta_1 = 1$ (dotted line). The distribution of $\hat{\beta}_1$ from scenario five (Bernoulli model) was such that some estimates fell outside the upper bounds of the plots. Each box plot shows (from bottom to top) the lower bound of 1.5 times the inter-quartile range, the 25^{th} percentile, the median, the 75^{th} percentile, and the upper bound of 1.5 times the inter-quartile range. See **Table 1.2** for a summary of all settings.

for β_1 appear to be unbiased for all settings in the simulation study (see **Figure 1.3**). The CPs for $\hat{\beta}_1$, in scenario three, were between 0.946 and 0.964 for all settings. Additionally, the relative efficiency of $\hat{\beta}_1$, obtained from scenario three, ranged from about 1.4 (setting 4) to about 1.8 (setting 2). The CPs and efficiencies obtained for scenario three are available in **Table 1.2**.

Under the model and data in scenario four (conditional distribution of v_j given n_j) the MLEs for β_1 appear to be unbiased for all settings in the simulation study (see **Figure 1.3**). The CPs for $\hat{\beta}_1$, in scenario four, were between 0.919 and 0.943 for all settings. Additionally, the relative efficiency of $\hat{\beta}_1$, obtained from scenario four, ranged from about 1.4 (setting 4) to about 1.9 (setting 2). Finally, under the model and data in scenario five (Bernoulli distribution of v_j), the MLEs for β_1 were weakly identifiable with efficiencies of $\hat{\beta}_1$ ranging from about 13.1 (setting 4) to over 18,000 (setting 3) and CPs between 0.819 and 0.956. The CPs and efficiencies obtained for scenarios four and five are available in **Table 1.2**.

As expected, the MSPE of the estimated probability surfaces was smallest for the model in scenario one, followed by two, three, four, and five across all settings. In general, the MSPE of the estimated intensity surfaces were smallest for the model in scenario two, followed by three, four, and five. Plots showing the distributions of the MSPE for the estimated intensity and probability surfaces among each of the scenarios for all settings are given in Appendix A.

1.5 Application

1.5.1 Disease Risk Factor Analysis

The distributional results outlined in the Methods Section are useful for disease risk factor analyses when data have been spatially aggregated. Using the transformed distributions enables researchers to recover individual-level inference about how spatial covariates influence the probability of infection. We illustrate our proposed methods using disease surveillance data collected to understand and manage the spread of white-nose syndrome (WNS). As

Table 1.2: Results from our simulation experiment using two sample sizes (small vs. large) and two levels of covariate equivalence ($x(\mathbf{s}) = z(\mathbf{s})$ vs. $x(\mathbf{s}) \neq z(\mathbf{s})$). For each setting, we report the average number of observations within each grid cell ($\bar{n}_j$), the average number per grid cell that had a mark of one ($\bar{n}_{1j}$) and zero ($\bar{n}_{0j}$), and the average proportion of grid cells that contained an observation with a mark of one ($\bar{v} = \frac{1}{1000}\frac{1}{400}\sum_{sim=1}^{1000}\sum_{j=1}^{400} \mathrm{I}(n_{1j} > 0)$) from 1,000 simulated data sets. We show the relative efficiency (Eff.) for estimating β_1 and the 95% CI coverage probability (CP) for each of the proposed models (using appropriate types of aggregated data). We also report the 95% CI CP for logistic regression using the exact locations of observations. We calculate the relative efficiency for each proposed model as the ratio of the standard deviation of the empirical distribution of β_1 from the respective proposed model against that of logistic regression.

Setting	Covariate Equivalence ($x(\mathbf{s}) = z(\mathbf{s})$)	Sample Size	$\bar{n}_j$	$\bar{n}_{1j}$	$\bar{n}_{0j}$	$\bar{v}$	CP Scen. 1	CP Scen. 2	CP Scen. 3	CP Scen. 4	CP Scen. 5	Eff. Scen. 2	Eff. Scen. 3	Eff. Scen. 4	Eff. Scen. 5
1	Yes	Small	10.1	0.18	9.89	0.11	0.951	0.960	0.964	0.919	0.875	1.12	1.75	1.81	1,606
2	Yes	Large	50.3	0.19	50.1	0.11	0.950	0.961	0.958	0.932	0.819	1.12	1.78	1.86	248.4
3	No	Small	10.2	0.14	10.1	0.11	0.945	0.940	0.946	0.922	0.956	1.22	1.41	1.43	18,516
4	No	Large	49.9	0.14	49.8	0.11	0.951	0.955	0.959	0.943	0.940	1.15	1.37	1.39	13.13

previously mentioned, WNS is a fungal disease caused by the pathogen *P. destructans* that threatens several North American species of bats (Ingersoll et al., 2016). The earliest documentation of the disease in North America was in 2006 based on photographic evidence from Howes Cave, near Albany, New York (Blehert et al., 2009; Frick et al., 2010; Hefley et al., 2020). The pathogen, *P. destructans*, has since spread throughout the eastern and midwestern United States resulting in high mortality rates among several species of cave-hibernating bats. Surveillance for *P. destructans* in the United States began in 2007 using a combination of passive and active surveillance methods. During 2007–2012, samples were obtained from individual bats associated with morbidity or mortality investigations occurring year-round at underground hibernacula or on the above-ground landscape. An individual sample consisted of a bat carcass, biopsies of wing skin, or tape lifts of fungal growth on the muzzle. A small number of individual samples were also obtained from target species (including *Myotis* spp., *Perimyotis subflavus*, and *Eptesicus fuscus*) that were admitted to rehabilitation facilities or state diagnostic laboratories for rabies testing from approximately December to May. A positive or negative diagnosis of WNS in individual bats was determined by observing characteristic histopathologic lesions in skin tissues using light microscopy (Meteyer et al., 2009). A diagnosis of 'suspect WNS' was assigned to individuals with clinical signs suggestive of the disease that had ambiguous skin histopathology or that had the causative agent (*P. destructans*) detected by fungal culture, fungal tape lift, or polymerase chain reaction in the absence of available or definitive histopathology and regardless of observed clinical signs (Lorch et al., 2010). We treated 'suspect WNS' diagnoses as positive cases for our analysis.

We illustrate our modeling approach using a subset of the WNS surveillance data collected during 2008–2012 that included individual samples of little brown bats (*Myotis lucifugus*), big brown bats (*Eptesicus fuscus*), northern long-eared bats (*Myotis septentrionalis*), and tri-colored bats (*Perimyotis subflavus*). This resulted in a total of 428 samples with 226 positive or suspected positive cases of WNS (Ballmann et al., 2021). As a result of the data collection process, the study area $\mathcal{S}$ was defined as the 120 counties that contained at least one bat that was tested for WNS between 2008 and 2012. The resulting study area collectively covered approximately 195,000 km^2. We note that this number reflects the sum

of the areas of the included counties rather than the area of the northeastern United States. To comply with the Endangered Species Act and protect the bats and their environment, the locations of the tested bats were recorded as the respective county centroids and thus suffered from bounded location error (*sensu* Walker et al., 2020). As bounded location error is equivalent to aggregation in this instance, the original data are Type C and require an appropriate model (i.e. the joint model for n_{1j} and n_{0j} from (1.6-1.7)) to obtain bias corrected individual-level inference. As Type C data can be further aggregated to become Type D and E, the WNS data are well-positioned to demonstrate our proposed models.

We were interested in two spatial covariates when we evaluated our proposed models. The first spatial covariate was 'presence of karst' (karst), a type of landscape characterized by cave formation. Therefore, the presence of karst in any particular area serves as a plausible surrogate covariate for the presence or absence of caves where bats might congregate (Medellin et al., 2017). The second spatial covariate was 'proportion of land classified as forest' (forest) and was calculated from the 2011 National Land Cover Database by determining what proportion of land within each 300m $\times$ 300m grid cell in the study area was composed of any kind of forest (Homer et al., 2015). The forest covariate is notable because the proportion of the immediate vicinity that is covered in forest may be an ecologically relevant predictor for the presence of WNS (Jachowski et al., 2014).

We fit each of four regression models that enable individual-level spatial inference from aggregated binary data (i.e., the joint model for n_{1j} and n_{0j} from (1.6-1.7); the joint model for v_j and n_j from (1.8) and (1.9); the conditional model for v_j given n_j from (1.9); and the Bernoulli model for v_j from (1.11)) to the WNS data set under the types of aggregation introduced in **Table 1.1** (Types C, D, and E). We incorporated the spatial covariate 'presence of karst' in the thinned intensity function, $\lambda(\mathbf{s})$, of the proposed transformed models and we included 'proportion of land classified as forest' (forest) as the spatial covariate in $p(\mathbf{s})$ in the transformed models.

We also fit three logistic regression models to the Type E aggregated WNS data, consisting of indicator variables (see **Table 1.1**, Type E). These three models represent the approach some researchers resort to when attempting to make individual-level inference

from aggregated data. The first model that was fit to Type E data used the value of the forest covariate from the centroid of each county (Areal County Centroid), while the second model used the average of the forest covariate for each county (Areal County Average). The third logistic regression model that was fit to Type E data used the average of the forest covariate across areas in each respective county where karst landscape was present (Areal % Forest in Karst).

We fit the regression models that enable individual-level spatial inference from aggregated binary data as outlined in Section 1.3.5 using the program R. We used the `glm` function in the program R to fit the specified logistic regression models (R Core Team, 2021). Numerically optimizing the likelihood functions for the proposed regression models each required approximately one and a half hours on a standard desktop computer. We compare MLEs and Wald-type 95% CIs among the proposed regression models and we provide the MLEs and Wald-type 95% CIs for the three logistic regression models fit to Type E data as a reference. We provide annotated R code capable of reproducing the disease risk factor analysis in the `wns.R` file in the supporting information for Walker et al. (2021) and in Ballmann et al. (2021).

1.5.2 Results

Our results show that the proposed regression models give similar inference to each other regardless of the type of data or level of aggregation, as long as the appropriate model is used (see **Figure 1.4** for comparisons and Appendix A for additional plots). The joint model for n_{1j} and n_{0j} from (1.6-1.7) provided the most precise estimates and matched the distribution of the available WNS data. As a result, the joint model for n_{1j} and n_{0j} provides the most efficient individual-level inference among the proposed models. This is unsurprising because the data, which are Type C, contain the most information (see **Table 1.1**).

The results for the logistic regression models fit to Type E data differed among themselves substantially, although the 95% CIs for $\hat{\beta}_{forest}$ overlapped between two pairs of the three models. While it would be tempting to compare the results from the logistic regression

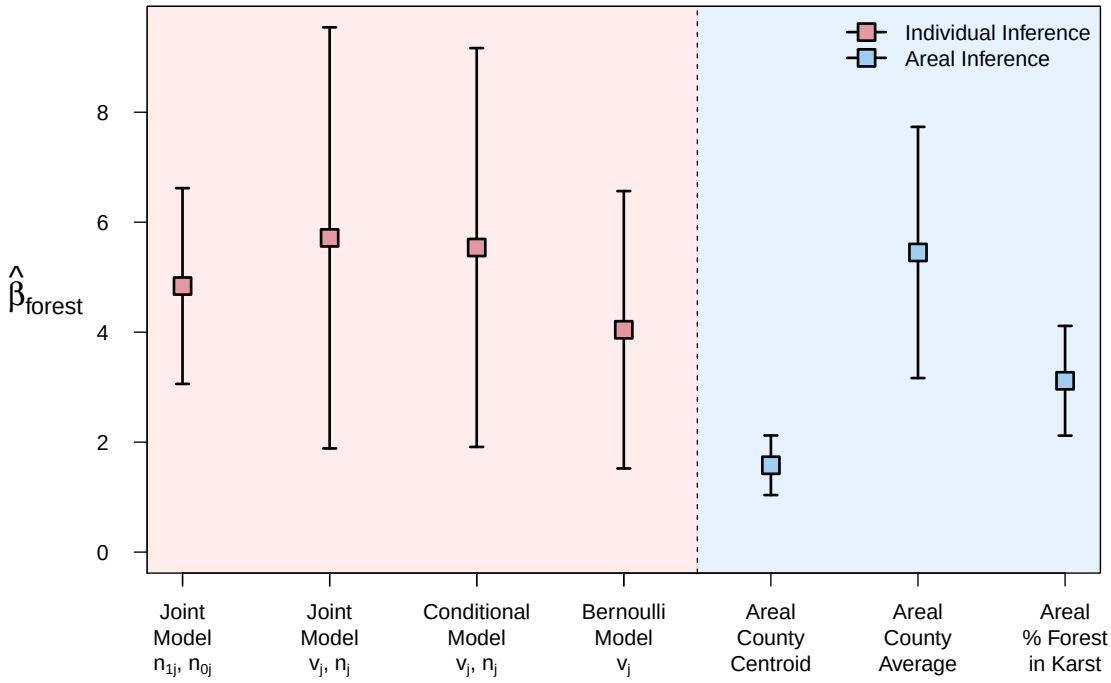

Figure 1.4: Binary regression model coefficient estimates and 95% CIs for the spatial covariate 'proportion of land classified as forest' (forest) that affects the probability of *P. destructans* infection for cave-hibernating bats in the northeastern United States (see **Figure 1.1** for visual). Estimates were obtained from the joint model for n_{1j} and n_{0j} in (1.6-1.7), the joint model for v_j and n_j in (1.8) and (1.9), the conditional model for v_j given n_j in (1.9), and the Bernoulli model for v_j in (1.11) that were fit using the respective data types. Here, n_{1j} is the number of observations in the j^{th} county that tested positive or suspect positive for WNS, n_{0j} is the number of observations in the j^{th} county that tested negative, n_j is the total number of observations in the j^{th} county, and $v_j = \text{I}(n_{1j} > 0)$. Also, using data that consists of the binary indicators (v_j), we give the areal-level results for logistic regression models that have the covariates of county centroid value of forest (Areal County Centroid), county averaged forest (Areal County Average), and county averaged forest in karst landscape (Areal % Forest in Karst). We delineate which models can recover individual-level inference (pink) and which are suited to areal-level inference (blue). For each model, we give the coefficient estimate (box) followed by the 95% CI limits (whisker ends).

models fit to Type E data against the models that produce individual-level inference, it would be fallacious to do so (Piantadosi et al., 1988; Gotway and Young, 2002).

1.6 Discussion

Our results demonstrated that models based on the proposed distributional results were capable of recovering individual-level inference on spatial covariates from aggregated binary data. As the degree of data aggregation increases, from Type C data to Type E, the relative efficiency of slope parameter estimates and intercept estimates decreases (see Appendix A for additional results). Further, the probability of obtaining extreme values of coefficient estimates and standard errors from the proposed models increases as aggregation increases from Type A to Type E data. However, even without more specific information than an indicator variable (i.e., Type E data) for each county, our results show that it may be possible to recover individual-level inference.

In many situations, such as our WNS surveillance data, data curators will be unable to release the exact locations of binary data (i.e., Type A data). Likewise, there will be many situations where data curators may be unwilling or unable to release Type C aggregated data because the data contain too much specific information to adequately protect privacy. The next level of privacy protection that enables individual-level inference comes from releasing the number of observations in each subregion (n_j) and an indicator variable for each subregion ($v_j = \mathrm{I}(n_{1j} > 0)$). Releasing n_j and v_j would provide the data required to fit models based on (1.9) and the joint density of (1.8) and (1.9). We note that inference from the joint model for v_j and n_j is usually preferable in practice if n_{1j} and n_{0j} are unavailable. This is because parameter estimates from the joint model for v_j and n_j are more efficient than that of the conditional model. The model for Type E data based on (1.11) has an increased probability of providing extreme coefficient estimates and large or infinite standard errors for some situations (similar to complete separation in binary regression models). However, if auxiliary information is available about $\lambda(\cdot)$ (e.g., the sampling design for the study or a point estimate for $\lambda(\cdot)$), models based on (1.11) would have a higher probability of being useful

(i.e., estimates may not be extreme and confidence intervals may be of reasonable width). In general, if individual-level inference is required, we recommend that practitioners fit the appropriate model for the type of aggregated data that is available to them. If the standard errors are large for the parameters of interest in the appropriate model, we recommend applying standard techniques to address complete separation (e.g., a Firth correction; Firth, 1993).

Two issues linger from our disease risk factor analysis. First, in some disease risk factor analyses there may be a need to account for spatial correlation among the responses. A spatial random effect may be added to the models proposed in this chapter, either in the specification for $\lambda(\mathbf{s})$, or $p(\mathbf{s})$, or both (e.g., Diggle et al., 1998), as follows:

$$\log(\lambda(\mathbf{s})) = \alpha_0 + \mathbf{z}(\mathbf{s})'\boldsymbol{\alpha} + \eta(\mathbf{s}), \tag{1.12}$$

$$\text{logit}(p(\mathbf{s})) = \beta_0 + \mathbf{x}(\mathbf{s})'\boldsymbol{\beta} + \gamma(\mathbf{s}), \tag{1.13}$$

where each value of $\eta(\mathbf{s})$ and $\gamma(\mathbf{s})$ is assumed to follow a multivariate normal distribution, as follows:

$$\begin{bmatrix} \eta(\mathbf{s}_1) \\ \vdots \\ \eta(\mathbf{s}_n) \\ \gamma(\mathbf{s}_1) \\ \vdots \\ \gamma(\mathbf{s}_n) \end{bmatrix} \sim \text{N}\left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_\eta & \boldsymbol{\Sigma}_{\eta\gamma} \\ \boldsymbol{\Sigma}_{\gamma\eta} & \boldsymbol{\Sigma}_\gamma \end{bmatrix} \right). \tag{1.14}$$

Here, $\boldsymbol{\Sigma}_\eta$ and $\boldsymbol{\Sigma}_\gamma$ are block diagonal components of the covariance matrix and $\boldsymbol{\Sigma}_{\eta\gamma} = \boldsymbol{\Sigma}'_{\gamma\eta}$ is an $n \times n$ block of zeros. Although practitioners could perform standard visual model checking procedures (e.g., semivariogram) to determine if spatial auto-correlation occurs in either the location data or the binary marks, we are unaware of how these techniques could be applied to aggregated data. Instead, we recommend that practitioners fit the proposed

models with a spatial random effect(s), and then again without, and perform model selection (Burnham and Anderson, 2002).

The second common issue for disease risk factor analyses is that collection of opportunistic disease surveillance data is often and likely the result of preferential sampling. Preferential sampling arises if $\eta(\mathbf{s})$ and$\gamma(\mathbf{s})$ from (12-14) are correlated, or when the off-diagonal blocks of the covariance matrix are non-zero. Including a spatial random effect is therefore a straightforward way to account for preferential sampling that may be present when using any of the models included in this chapter (Diggle et al., 2010b). Adapting assumptions 1-3 from Diggle et al. (2010b) to our notation from (4) and assuming that (4) is specified with spatial random effects:

1. $\eta(\mathbf{s}) \sim \mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma}_\eta)$, where $\eta(\mathbf{s})$ is a spatial random effect assumed to follow a multivariate normal distribution and $\mathbf{s}$ is the coordinate vector in the study area $\mathcal{S}$ (i.e., $\mathbf{s} \subseteq \mathcal{S}$).

2. $\mathbf{U} \sim \mathrm{IPPP}(\lambda(\mathbf{s}))$ where $\mathbf{U} \equiv (\mathbf{u}_1, \mathbf{u}_2, ..., \mathbf{u}_n)'$ is a matrix of locations for the tested bats generated from an inhomogeneous Poisson point process with $\log(\lambda(\mathbf{s})) = \alpha_0 + \mathbf{z}(\mathbf{s})'\boldsymbol{\alpha} + \theta\eta(\mathbf{s})$ and θ as a scaling parameter.

3. $y_i \sim \mathrm{Bern}(p(\mathbf{u}_i))$, where y_i is the i^{th} observation, $\mathbf{u}_i$ is the location of the i^{th} bat, and $g(p(\mathbf{s})) = \beta_0 + \mathbf{x}(\mathbf{s})'\boldsymbol{\beta} + \eta(\mathbf{s})$. For our purposes, $g(\cdot)$ is the logit link.

Following Diggle et al. (2010b), the model specified in items 1-3 accounts for preferential sampling.

Lastly, non-spatial individual-level covariates (e.g., sex or age) can be included in models for Type B and C data (e.g., Walker et al., 2020). However, due to the constraints inherent in the aggregation process for Type D and E data, it is not likely that non-spatial, individual-level covariates would be available. A future contribution might incorporate non-spatial, aggregated individual-level covariates (e.g., average age of tested individuals in a county) into the proposed transformed models for data Types D and E. Furthermore, Taylor et al. (2018) and Heaton et al. (2020) showed it may be possible to relax the assumption of a discretized partition of the study area that normally applies to models that include a COS

transformation. Relaxing this assumption would accommodate overlapping and uncertain subregion boundaries.

Acknowledgements

We thank all state, federal and other partners for submitting samples and the USGS National Wildlife Health Center (Madison) for processing the samples. We thank the associate editor and two anonymous referees from *Spatial Statistics* for their valuable feedback. We likewise thank Dr. Kathi Irvine for her comments via a Fundamental Science Practices (FSP) review. We acknowledge support for this research from USGS G18AC00317 and G16AC00413. Any use of trade, firm, or product names is for descriptive purposes only and does not imply endorsement by the U.S. Government. Declaration of Interest: None.

Supporting Information

The `.R` files referenced in Sections 1.4 and 1.5 are available within the `Recovering Inference.zip` from the Supplementary Material associated with Walker et al. (2021). The disease surveillance data used in this chapter are available in the Supplementary Material and from the data release Ballmann et al. (2021). These data were provided by the U.S. Geological Survey, National Wildlife Health Center from a database that is continuously updated (accessed on Aug 22, 2019). Updated versions of the data may be requested from Anne Ballmann (aballmann@usgs.gov) with the permission of the National Wildlife Health Center and contributing partner agencies.

Chapter 2

A Staged Approximate Bayesian Model Averaging Method for Estimating the Number, Locations, and Times of Introduction for a Novel Pathogen.

2.1 Abstract

Disease surveillance data are an important resource for epidemiologists, animal and plant pathologists, and ecologists to identify, understand, and mitigate the spread of infectious disease in human, livestock, wildlife, and plant populations. While methods exist to identify likely sources of a pathogen in simple circumstances or forecast the spread of an infectious disease in broad terms, few methods explicitly estimate the spatio-temporal origins of a pathogen (i.e., the locations and times that a pathogen was introduced) and the number of pathogen introductions in a widespread epidemic. We outline a likelihood based method and a staged approximate Bayesian model averaging (SABMA) method to obtain this inference

under an ensemble model of simple ecological diffusion processes for the spread of a pathogen through a population. We evaluate the predictive performance and credible interval coverage of the SABMA method compared to the inference and predictive ability of the likelihood-based method via a simulation experiment. We then apply the likelihood-based and SABMA methods to two sets of disease surveillance data on chronic wasting disease (CWD) in white-tailed deer (*Odocoileus virginianus*); the first in the lower peninsula of Michigan in the U.S., and the second in southern Wisconsin and northern Illinois in the U.S.

2.2 Introduction

In epidemiology, animal and plant pathology, and disease ecology, a central concern of researchers is to identify and understand outbreaks of infectious disease. The collection and analysis of spatial and spatio-temporal disease surveillance data, related to the presence and prevalence of infectious disease, provide essential insight into the dynamics, severity, and causes of infectious disease outbreaks (Lee et al., 2010). Specifically, researchers may use disease surveillance data to identify clusters of infectious disease cases, determine the severity of an outbreak, and identify high-risk populations. Further, disease surveillance data may also be used to inform public health measures to mitigate the spread of the pathogen, assess the effectiveness of interventions, and form and test hypotheses about risk factors associated with the start and growth of an outbreak (Garcia-Abreu et al., 2002). In this chapter, we focus on this last aspect by estimating the spatio-temporal origins of an outbreak and predicting the spread of the pathogen.

Many non-spatial methods are capable of modeling and forecasting the temporal spread of infectious disease (e.g., susceptible, infectious, recovered (SIR) compartmental models; Vynnycky and White, 2010; Adivar and Selin Selen, 2013). Additionally, many methods are capable of identifying the likely source or timing of an outbreak (e.g., case-control analysis and epidemic curve analysis; National Research Council, 2009; Borgan et al., 2018; Egan and Hall, 2015). However, efforts to estimate or obtain inference on the spatio-temporal origins (i.e., the location and time of pathogen introduction) of an outbreak have focused

on two main areas: geographic clustering or profiling, and Bayesian spatio-temporal modeling. Geographic clustering or profiling identifies high-likelihood areas of pathogen origin without estimating the time of pathogen introduction from temporal data (LeComber et al., 2011; Mohler and Short, 2012; Stevenson et al., 2012; Verity et al., 2014). Bayesian spatio-temporal models often employ differential or partial differential equations (PDEs) to mechanistically account for the spread of the pathogen that caused an outbreak For example, Legrand et al. (2009) and Hefley et al. (2020) developed and implemented different Bayesian spatio-temporal dynamic models that could simultaneously account for the dynamic process of a pathogen invasion and estimate the initial location and time of introduction for the pathogen. However, many infectious disease outbreaks are not well explained using a model that assumes a single pathogen introduction (e.g., Levy et al., 2011; Verity et al., 2014; Kissler et al., 2019). While Hefley et al. (2020) noted that their proposed model could be modified by including mixtures of diffusion models to estimate multiple locations and times of introduction, Hefley et al. (2020) assumed the number of introductions was known and stopped short of implementing their idea for their preferred dynamic spatio-temporal model.

The main technical problem that hampered Hefley et al. (2020) was the infeasibility of sampling a random number of three-dimensional vectors (locations and times of introduction). In contrast, Verity et al. (2014) developed a method in the geographic clustering literature to estimate the number and locations of introduction while ignoring time, and Levy et al. (2011) developed a distance-based Bayesian clustering method to select the number of introductions and estimate the locations and times of introduction. Levy et al. (2011) did not, however, employ a diffusion PDE as Legrand et al. (2009) and Hefley et al. (2020) had done, nor did Levy et al. (2011) account for uncertainty in the number of introductions that best fit the data. The purpose of this chapter is to develop a method that estimates the number, locations, and times that a pathogen was introduced into a population using an ensemble of PDE-based diffusion processes to account for the spread of the pathogen. Our method should also be capable of predicting the individual-level probability of being infected at any location and time, and provide a measure of uncertainty in the estimated number of pathogen introductions.

The remainder of the chapter proceeds as follows: First, we introduce a simple diffusion process for pathogen spread and an ensemble model framework that is capable of modeling the number, locations, and times of introduction. We then outline a staged approximate Bayesian model-averaging (SABMA) method that provides approximate Bayesian inference on the number, locations, and times of pathogen introduction. Next, we conduct a simulation experiment to evaluate the performance of our SABMA method, present results from the simulation experiment, and apply our method to two ecological disease surveillance data examples. Finally, we discuss the ramifications of our simulation experiment and data examples and suggest future lines of research.

2.3 Methods

2.3.1 Ecological Diffusion From One Introduction

An ecological diffusion process is a partial differential equation (PDE) that can be used to describe the movement of biotic entities from an initial high concentration to a dispersed low concentration. Researchers have previously used ecological diffusion PDEs to model the spread of infectious agents (Garlick et al., 2011, 2014; Hefley et al., 2017b,c, 2020). The PDE for ecological diffusion originating from a single location is expressed as:

$$\frac{\partial}{\partial t}u(\mathbf{s},t) = \left(\frac{\partial^2}{\partial s_1^2} + \frac{\partial^2}{\partial s_2^2}\right)[\mu(\mathbf{s})u(\mathbf{s},t)] \tag{2.1}$$

where $u(\mathbf{s},t)$ is the intensity of pathogen particles at any given location $\mathbf{s} \equiv (s_1, s_2)' \subseteq \mathcal{S}$ and time $t \subseteq \mathcal{T}$. We define $\mu(\mathbf{s})$ as the diffusion coefficient that determines the rate of spread for any location. Integrating the pathogen intensity $u(\mathbf{s},t)$ over a given spatial area provides the expected pathogen concentration within that area at that time.

The pathogen intensity $u(\mathbf{s},t)$ at any given location and time is unobserved. However, binary disease surveillance data are commonly collected from individuals and contain information about the latent pathogen intensity. Let y_i be the i^{th} binary observation ($i = 1, 2, ..., n$), where $y_i = 1$ denotes the presence of the pathogen in that individual and $y_i = 0$ otherwise.

If we assume that y_i is related to the latent pathogen intensity $u(\mathbf{s}_i, t_i)$ at the i^{th} location and time through some link function $g(\cdot)$, we define a statistical model for binary spatio-temporal disease surveillance data based on (2.1) as follows:

$$y_i \sim \text{Bern}(p(\mathbf{s}_i, t_i)) \tag{2.2}$$

$$g(p(\mathbf{s}_i, t_i)) = u(\mathbf{s}_i, t_i) \tag{2.3}$$

$$\frac{\partial}{\partial t} u(\mathbf{s}, t) = \left(\frac{\partial^2}{\partial s_1^2} + \frac{\partial^2}{\partial s_2^2} \right) [\mu(\mathbf{s}) u(\mathbf{s}, t)]. \tag{2.4}$$

Under the ecological diffusion process and Dirichlet boundary conditions, researchers may use disease surveillance data collected over space and time to estimate $\mu(\mathbf{s})$ and then back-cast and forecast to estimate the spatial distribution of the pathogen at earlier and later time points, respectively. Assuming that the pathogen was introduced at an exact location and time, back-casting could enable the estimation of the location of pathogen introduction, $\boldsymbol{\omega} \equiv (\omega_1, \omega_2)' \subseteq \mathcal{S}$, the time of introduction, t_0, and the initial pathogen intensity, θ, similar to Legrand et al. (2009) and Hefley et al. (2020).

2.3.2 An Ensemble of Ecological Diffusion Processes for Multiple Introductions

While some infectious disease outbreaks are well explained using a model that assumes a single pathogen introduction (e.g., white-nose syndrome in bats in the eastern U.S.; Drees et al., 2017), other disease outbreaks may require a more complex modeling method that assumes multiple pathogen introductions at different times within the study area $\mathcal{S}$ (e.g., chagas disease and influenza; Levy et al., 2011; Kissler et al., 2019). In a multiple introduction scenario, the total pathogen intensity $u(\mathbf{s}_i, t_i)$ at the location and time of the i^{th} observation may be viewed as a sum of J component pathogen intensities:

$$u(\mathbf{s}_i, t_i) = \sum_{j=1}^{J} u_j(\mathbf{s}_i, t_i) \tag{2.5}$$

where $u_j(\mathbf{s}_i, t_i)$ represents the pathogen intensity at the i^{th} location and time that was contributed by the j^{th} introduction ($j = 1, 2, ..., J$; Hefley et al., 2020). Each of the J component pathogen intensities arises from an ecological diffusion process with unique initial conditions:

$$u_j(\mathbf{s}, t_{0_j}) = \begin{cases} \theta_j & \text{if } \mathbf{s} = \boldsymbol{\omega}_j \\ 0 & \text{if } \mathbf{s} \neq \boldsymbol{\omega}_j \end{cases}, \tag{2.6}$$

where t_{0_j} is the time of the j^{th} introduction, θ_j is the intensity of pathogen released at the j^{th} introduction, and $\boldsymbol{\omega}_j$ is the location of the j^{th} introduction. Following Hefley et al. (2020), we assume the diffusion rate is constant (i.e., $\mu(\mathbf{s}) = \mu$) and specify Dirichlet boundary conditions. As a result, the analytical solution to the j^{th} ecological diffusion PDE at an arbitrary time t and location $\mathbf{s}$ is (Pielou, 1969, Ch 11, Eq. 11.3):

$$u_j(\mathbf{s}, t) = \frac{\theta_j}{4\pi\mu(t - t_{0_j})} \exp\left\{ \frac{-||\mathbf{s} - \boldsymbol{\omega}_j||^2}{4\mu(t - t_{0_j})} \right\}, \tag{2.7}$$

where $||\mathbf{s} - \boldsymbol{\omega}_j||^2$ the squared Euclidean distance between any location $\mathbf{s}$ and the j^{th} location of introduction $\boldsymbol{\omega}_j$. The ensemble ecological diffusion model then becomes:

$$y_i \sim \text{Bernoulli}(p_i) \tag{2.8}$$

$$g(p(\mathbf{s}_i, t_i)) = u(\mathbf{s}_i, t_i) \tag{2.9}$$

$$u(\mathbf{s}_i, t_i) = \sum_{j=1}^{J} u_j(\mathbf{s}_i, t_i) \tag{2.10}$$

$$u_j(\mathbf{s}_i, t_i) = \frac{\theta_j}{4\pi\mu(t_i - t_{0_j})} \exp\left\{ \frac{-||\mathbf{s}_i - \boldsymbol{\omega}_j||^2}{4\mu(t_i - t_{0_j})} \right\}. \tag{2.11}$$

Within this specification, the unknown parameters are $\boldsymbol{\Phi} \equiv (\mu, \boldsymbol{\omega}_1, ..., \boldsymbol{\omega}_J, t_{0_1}, ..., t_{0_J}, \theta_1, ..., \theta_J)'$ and J, where $\boldsymbol{\Phi}$: $\mu \subset \mathcal{R}^+$, $\boldsymbol{\omega}_1, ..., \boldsymbol{\omega}_J \subseteq \mathcal{S}$, $t_{0_1}, ..., t_{0_J} \subset \mathcal{T} \subset \mathcal{R}$, $\theta_1, ..., \theta_J \subset \mathcal{R}^+$, and J is a positive integer (i.e., $J = 1, 2, ...$). The log-likelihood for the ensemble model is as follows:

$$\log f(\mathbf{y}|\mathbf{\Phi}, J) = \sum_{i=1}^{n}[y_i \log\left(g^{-1}\left(\sum_{j=1}^{J} \frac{\theta_j}{4\pi\mu(t_i - t_{0_j})} \exp\left\{\frac{-||\mathbf{s}_i - \boldsymbol{\omega}_j||^2}{4\mu(t_i - t_{0_j})}\right\}\right)\right) + (1 - y_i)\log\left(1 - g^{-1}\left(\sum_{j=1}^{J} \frac{\theta_j}{4\pi\mu(t_i - t_{0_j})} \exp\left\{\frac{-||\mathbf{s}_i - \boldsymbol{\omega}_j||^2}{4\mu(t_i - t_{0_j})}\right\}\right)\right)], \quad (2.12)$$

where $\mathbf{y} = (y_1, y_2, ..., y_n)'$.

2.3.3 Fitting the Ensemble Model

The ensemble model in (2.8-2.11) may be fit in a frequentist or Bayesian paradigm to obtain inference on the parameters of interest. Maximum likelihood estimates (MLEs) for the parameters $\mathbf{\Phi}$ may be obtained if the true parameter values are not close to the boundary of the parameter space (e.g., if $\boldsymbol{\omega}_j$ is not close to the boundary of $\mathcal{S}$; Marchand and Strawderman, 2004). We can obtain MLEs using common numerical optimization methods, such as the `BFGS` algorithm from the `optim` function in the program R (Nocedal and Wright, 2006; R Core Team, 2021). After obtaining MLEs, we may approximate the variances of the parameter estimates $\hat{\mathbf{\Phi}}$ and obtain Wald-type confidence intervals by inverting the Hessian matrix. However, in some circumstances, such as fitting a mis-specified model or a model with weakly identifiable parameters, MLEs may be unavailable or the Hessian may be singular (Albert and Anderson, 1984). Moreover, a maximum likelihood estimation approach for the model in (2.8-2.11) must assume one of the following: 1) J is known; 2) J can be selected using a model selection criterion (e.g., Bayesian information criterion); 3) J can be estimated using a mixed integer optimization method (see Kronqvist et al., 2019 for a review of different methods). To the best of our knowledge, mixed integer optimization methods do not provide a measure of uncertainty associated with the MLE of J. Not accounting for uncertainty in the selection or estimation of J may lead to underestimating the uncertainty associated with the other parameter estimates $\hat{\mathbf{\Phi}}$ (Madigan and Raftery, 1994).

Under the Bayesian paradigm, classical Markov chain Monte Carlo (MCMC) methods such as the Metropolis-Hastings algorithm or Gibbs sampler are unfortunately difficult to apply because the number of parameters may be large (as J increases), closed-form solutions for

the full conditional posterior distributions are unavailable, and estimating all parameters is a challenging trans-dimensional problem. The trans-dimensional nature of estimating parameters in (2.8-2.11) exists because for every one unit increase in the value of J, the number of parameters to be estimated increases by four (θ_{j+1}, t_{0j+1}, and $\boldsymbol{\omega}_{j+1} \equiv (\omega_{1,j+1}, \omega_{2,j+1})'$). Thus, sampling from the posterior distribution of J changes the dimension and meaning of the parameters in the joint posterior distribution $p(\boldsymbol{\Phi}, J|\mathbf{y})$. We address this trans-dimensional difficulty by obtaining inference from $p(\boldsymbol{\Phi}, J|\mathbf{y})$ using a staged approximate Bayesian model averaging (SABMA) method. In the first stage, we sample from $p(J|\mathbf{y})$ using the Markov chain Monte Carlo model composition (MC^3) method (Madigan and York, 1995). We employ a Laplace approximation to the marginal likelihood (LAML) for each candidate value of J as part of the MC^3 algorithm. In the second stage, we obtain approximate samples from $p(\boldsymbol{\Phi}|\mathbf{y}, J)$ using the weighted Bayesian bootstrap (WBB; Newton and Raftery, 1994; Newton et al., 2021). As $p(\boldsymbol{\Phi}, J|\mathbf{y}) = p(\boldsymbol{\Phi}|\mathbf{y}, J)p(J|\mathbf{y})$, our staged approach provides valid approximate inference on $p(\boldsymbol{\Phi}, J|\mathbf{y})$. Additionally, employing a Bayesian model averaging method may improve the quality of inference on $\boldsymbol{\Phi}$ and the predictive performance of the model beyond what is available with the likelihood-based method, as measured by the logarithmic scoring rule (Madigan and Raftery, 1994; Gneiting and Raftery, 2007).

2.3.4 Markov Chain Monte Carlo Model Composition (MC^3)

In the first stage of the SABMA method, we employ the MC^3 method to obtain the posterior distribution $p(J|\mathbf{y})$. The MC^3 method is used in Bayesian model averaging to calculate the posterior probability of candidate models given the data (Madigan and York, 1995). It is thus possible to obtain a Markov chain for J ($J^{(r)}$: $r = 1, 2, ..., m$) with the stationary distribution $p(J|\mathbf{y})$. Let $\text{nbd}(J^{(r)})$ be the set of models with different values of J in the neighborhood for the current model $J^{(r)}$, then:

$$\text{nbd}(J^{(r)}) = \begin{cases} \{1, 2, 3, 4, 5\} & \text{if } J^{(r)} \leq 2 \\ \{J^{(r)} - 2, J^{(r)} - 1, J^{(r)}, J^{(r)} + 1, J^{(r)} + 2\} & \text{if } J^{(r)} \geq 3, \end{cases} \tag{2.13}$$

and the MC^3 algorithm is given as (Madigan and York, 1995):

1. Define J' by randomly selecting one value of J from nbd$(J^{(r)})$;

2. Calculate

$$h = \frac{\#\text{nbd}(J^{(r)})f(J'|\mathbf{y})}{\#\text{nbd}(J')f(J^{(r)}|\mathbf{y})}, \tag{2.14}$$

   where $\#\text{nbd}(\cdot)$ is the number of candidate models in the neighborhood of $J^{(r)}$ or J'. Based on how (2.13) was specified, $\#\text{nbd}(\cdot) = 5$.

3. Generate $v \sim \text{Uniform}(0, 1)$ and update

$$J^{(r+1)} = \begin{cases} J' & \text{if } v \leq h \\ J^{(r)} & \text{otherwise.} \end{cases} \tag{2.15}$$

The last component required to obtain a Markov chain $J^{(r)}$ is to specify a form for $f(J|\mathbf{y})$.

Laplace Approximation to the Marginal Likelihood

We now specify a form for $f(J|\mathbf{y})$ using the Laplace approximation to the marginal likelihood (LAML). The LAML is a popular approximate method for Bayesian model selection and model averaging. Let $\{M_1, M_2, ...\} \subset \mathcal{M}$ be the set of candidate models where each model assumes a different value for $J = 1, 2, ...$ The marginal likelihood, $p(\mathbf{y}|M_J)$, for a given model M_J is defined as:

$$p(\mathbf{y}|M_J) = \int_{\boldsymbol{\phi}_{M_J}} p(\mathbf{y}|\boldsymbol{\phi}_{M_J})p(\boldsymbol{\phi}_{M_J})d\boldsymbol{\phi}_{M_J}, \tag{2.16}$$

where $\boldsymbol{\phi}_{M_J}$ is the vector of parameters in the model M_J. If we assume a uniform prior on J (i.e., $p(J) \propto 1$ for all $J = 1, 2, ...$), the posterior mode of $p(J|\mathbf{y})$ is approximated by the following form of the LAML:

$$LAML(M_J) = -2\log p(\mathbf{y}|\hat{\boldsymbol{\phi}}_{M_J}) + k_{M_J}\log n, \tag{2.17}$$

where M_J is the given model, and $p(\mathbf{y}|\hat{\boldsymbol{\phi}}_{M_J})$ is the likelihood for the given model evaluated at the vector of MLEs $\hat{\boldsymbol{\phi}}_{M_J}$ of the parameters of interest. Additionally, $\boldsymbol{\phi}_{M_J} \equiv (\phi_1, \phi_2, ...\phi_{k_{M_J}})'$ is the vector of parameters for the given model M_J and k_{M_J} is the number of parameters in model M_J. Therefore, the explicit form of $f(J|\mathbf{y})$ from (2.14) is:

$$f(J|\mathbf{y}) \propto \exp\{-\frac{1}{2}LAML(M_J)\}. \tag{2.18}$$

2.3.5 The Weighted Bayesian Bootstrap

In the second stage of the SABMA method, we address the problem of obtaining approximate draws from $p(\boldsymbol{\Phi}|\mathbf{y}, J)$. Rather than using classical MCMC methods like the Metropolis-Hastings algorithm or a Gibbs sampler, we choose an approximate Bayesian method, the weighted Bayesian bootstrap (WBB), to fit the ensemble model and obtain inference (Newton and Raftery, 1994; Newton et al., 2021). The WBB is ideal for our purposes because the WBB requires the repeated optimization of the weighted likelihood that can be easily done in parallel, the WBB is compatible with flat priors, and the WBB is simple to implement (Newton and Raftery, 1994; Newton et al., 2021). Further, the WBB adapts well to changing the number of model parameters in $\boldsymbol{\Phi}$ between draws according to the posterior distribution of J, thus solving the trans-dimensionality problem.

To sample from the approximate posterior distribution $p(\boldsymbol{\Phi}|\mathbf{y}, J)$, we scale the likelihood for each observation by a random weight, $w_i \sim \text{Exp}(1)$. The resulting randomly weighted posterior distribution is proportional to the likelihood times flat prior distributions. One draw $\boldsymbol{\Phi}^{(r)}$ from the posterior distribution $p(\boldsymbol{\Phi}|\mathbf{y}, J)$ is generated as follows:

1. Sample $\mathbf{w}= (w_1, ..., w_n)'$ where $w_i \sim \text{Exp}(1)$ for the i^{th} element in $\mathbf{w}$;
2. Solve $\boldsymbol{\Phi}^{(r)} = \arg\max_{\boldsymbol{\Phi}} \sum_{i=1}^{n} w_i \log(p(y_i|\boldsymbol{\Phi}, J))$.

We complete step 2 on (2.12) using the `BFGS` method from the `optim` function in R. We

repeat the above procedure m times ($r = 1, ..., m$) to obtain m draws from the approximate posterior distribution of $p(\mathbf{\Phi}|\mathbf{y}, J)$.

2.4 Simulation

We conducted a simulation experiment to show the validity of our SABMA method for obtaining an approximation of $p(\mathbf{\Phi}, J|\mathbf{y})$. We show the validity of the SABMA method by evaluating out-of-sample prediction performance and credible interval coverage for parameters in the ensemble model in (2.8-2.11). We also compare the prediction performance and coverage probabilities obtained by the SABMA method against the prediction performance and confidence interval coverage obtained using a likelihood-based model fitting and selection method. We first define a study area and describe how locations and times of pathogen introduction were simulated. Next, we describe how the training and prediction sets of binary spatio-temporal disease surveillance data were simulated. We then describe how we implemented the SABMA and likelihood-based methods to fit the ensemble model and how we measured the predictive performance of the models. Lastly, we present the results of the simulation experiment.

We defined a unit square study area, $\mathcal{S} = [0, 1] \times [0, 1]$, with a circular sampling field (for locations of pathogen introduction) centered at $(0.5, 0.5)'$ with a radius of 0.45. Locations of pathogen introduction could be sampled within the circle with a constant probability of $\frac{1}{\pi(0.45)^2}$ and zero outside the circle. We simulated 1,000 scenarios, where each scenario had a pre-specified number of pathogen introductions J, from one to five, with each value of J assigned to 200 scenarios. We simulated the coordinates of the location of the first introduction from a uniform distribution over the sampling field. Where applicable, the location of the second introduction was drawn from the uniform distribution over the sampling field, except for a 0.1 radius circle cutout centered at the location of the first introduction. Likewise, where applicable, the location of the third introduction was simulated from the same uniform sampling field, except for two 0.1 radius circle cutouts centered at the locations of the first and second introductions. This pattern was followed for the locations of the fourth and

fifth introductions whenever applicable. The times of the five potential introductions were drawn from scaled beta distributions and constrained such that $t_{0_1} < t_{0_2} < t_{0_3} < t_{0_4} < t_{0_5}$, as follows:

$$\text{If } J \geq 1, \quad t_{0_1} \sim 30 * \text{Beta}(2, 15); \tag{2.19}$$

$$\text{If } J \geq 2, \quad t_{0_2} = \begin{cases} t^*_{0_2} & \text{if } t^*_{0_2} > t_{0_1} \\ t_{0_1} + 1.1 & \text{if } t^*_{0_2} < t_{0_1}; \end{cases} \tag{2.20}$$

$$\text{If } J \geq 3, \quad t_{0_3} = \begin{cases} t^*_{0_3} & \text{if } t^*_{0_3} > t_{0_2} \\ t_{0_2} + 1.1 & \text{if } t^*_{0_3} < t_{0_2}; \end{cases} \tag{2.21}$$

$$\text{If } J \geq 4, \quad t_{0_4} = \begin{cases} t^*_{0_4} & \text{if } t^*_{0_4} > t_{0_3} \\ t_{0_3} + 1.1 & \text{if } t^*_{0_4} < t_{0_3}; \end{cases} \tag{2.22}$$

$$\text{If } J = 5, \quad t_{0_5} = \begin{cases} t^*_{0_5} & \text{if } t^*_{0_5} > t_{0_4} \\ t_{0_4} + 1.1 & \text{if } t^*_{0_5} < t_{0_4}; \end{cases} \tag{2.23}$$

where

$$\text{if } J \geq 2, \; t^*_{0_2} \sim 30 * \text{Beta}(8, 20); \tag{2.24}$$

$$\text{If } J \geq 3, \; t^*_{0_3} \sim 30 * \text{Beta}(20, 20); \tag{2.25}$$

$$\text{If } J \geq 4, \; t^*_{0_4} \sim 30 * \text{Beta}(20, 8); \tag{2.26}$$

$$\text{If } J = 5, \; t^*_{0_5} \sim 30 * \text{Beta}(15, 2). \tag{2.27}$$

In each scenario, we randomly sampled the locations and times of $n = 2,000$ training observations and $n^* = 400$ prediction observations. We chose $n = 2,000$, rather than a larger number, to keep the simulation computation time at a reasonable level for 1,000 scenarios. We chose $n^* = 400$ because it is common and feasible for practitioners to set a similarly-sized fraction of their data as a prediction set. The locations of observations

were sampled with a constant probability of $\frac{1}{\pi(0.45)^2}$ inside the sampling circle (centered at $(0.5, 0.5)'$ with a radius of 0.45) and a probability of zero outside the circle. We assigned a sampling time to each sampled location from both data sets, drawn from a discrete uniform distribution over $t = \{12, 13, ..., 47, 48\}$. We finished constructing the training data and prediction data using (2.8-2.11) as the generative model for the binary marks $\mathbf{y}$ (training data) and $\mathbf{y}^* \equiv (y_1^*, y_2^*, ..., y_{n^*}^*)$ (prediction data), respectively. In this chapter, we follow Hefley et al. (2020) in using the standard log-normal cumulative distribution function as $g^{-1}(\cdot)$. We also specified $\theta = \theta_j = 1000$ for $j = 1, ..., J$, and $\mu = 0.0001$ for all scenarios. Thus, y_i is the i^{th} binary mark associated with the i^{th} location and time, $\mathbf{s}_i$ and t_i, in the training data set. Likewise, y_i^* is the i^{th} binary mark associated with the i^{th} location and time, $\mathbf{s}_i^*$ and t_i^*, in the prediction data set. **Figure 2.1** shows a plot of a partial example training data set with observations that were sampled at $t = 12, 24, 36, 48$.

Once the data were generated, we fit the model in (2.8-2.11) twice each for $j = 1, 2, 3, 4, 5, 6, 7$. The first set of models were fit using the likelihood-based method with BIC for model selection. The second set of models was fit using our SABMA method. Both model fitting methods involved the `BFGS` algorithm from `optim` function in R. We assumed starting values of $\theta_j = 1000$ for $j = 1, 2, ...7$ and $\mu = 0.0001$. Since practitioners may be able to guess reasonable starting values for $t_{0_1}, ..., t_{0_7}$ based on data, we attempted to recreate this 'guessing' behavior by drawing starting values for $t_{0_1}, ..., t_{0_7}$ from (N($t_{0_1}, 2^2$),...,N($t_{0_7}, 2^2$)). Additionally, we employed K-means clustering on the locations of observations where $y_i = 1$ to obtain starting values for the location of introduction parameters $\boldsymbol{\omega}_1, ..., \boldsymbol{\omega}_7$.

To apply the likelihood-based estimation method for each scenario, we obtained the MLEs $\hat{\boldsymbol{\Phi}}$ from each of seven models using the `BFGS` algorithm from the `optim` function in R. We then used the MLEs of the parameters for each model to calculate the BIC value for each of the seven models. We selected the model with the lowest BIC value to obtain our estimate of the number of introductions, $\hat{J}$. We estimated the variances of $\hat{\boldsymbol{\Phi}}$, given the selected model, and obtained Wald-type 95% confidence intervals for $\hat{\boldsymbol{\Phi}}$ by inverting the Hessian matrix. We then calculated the 0.95 confidence interval coverage probabilities for $\boldsymbol{\Phi}$.

To apply the SABMA method, we first retrieved the MLEs that were obtained for each

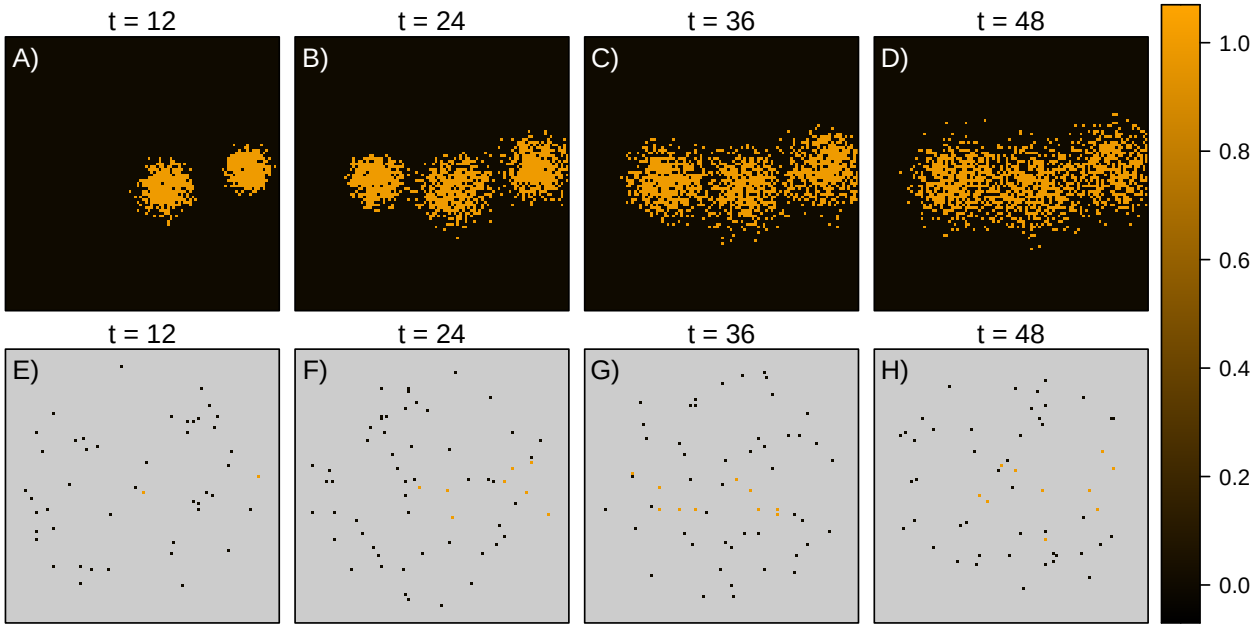

Figure 2.1: Panels A-D: Plots showing the introduction and diffusion of pathogen particles across a study area, as evidenced by locations of simulated individuals marked in orange that are positive for a pathogen. The black represents the locations of simulated individuals that do not have the pathogen. Individuals at every location in the study area were tested across thirty-seven time points, although only $t = 12, 24, 36$, and 48 are shown. Two introductions occurred before time $t = 12$ and a third introduction occurred between $t = 12$ and $t = 24$. **Panels E-H**: An example simulated data set showing binary marks associated with individuals at randomly sampled locations and times ($t = 12, 24, 36$, and 48 are shown). An orange dot shows that an individual has the pathogen, and a black dot shows that an individual does not have the pathogen. The top and bottom plots shared the same locations and times of pathogen introduction.

of the seven models when employing the likelihood-based method. We used the MLEs to calculate the LAML for each of the seven models. We then used the MC^3 algorithm to obtain 4,000 draws from $p(J|\mathbf{y})$ in the first stage. The first 2,000 draws were discarded as burn in, leaving 2,000 draws. In the second stage we used the remaining MC^3 draws to specify which models would be fit via the WBB. We then employed the WBB to obtain 2,000 draws from the approximate posterior distributions $p(\mathbf{\Phi}|\mathbf{y}, J)$. We obtained 95% credible intervals and calculated the 0.95 credible interval coverage probabilities.

2.4.1 Evaluating Predictive Performance

For each of 1,000 scenarios, we evaluated the predictive performance of the ensemble model fit using the SABMA method versus the likelihood-based method. Predictive performance was evaluated using the logarithmic scoring rule on out-of-sample prediction data (LSR; Gneiting and Raftery, 2007). We selected the LSR because it is a strictly proper rule and because the LSR matches the likelihood employed when fitting the models and generating the data (Gneiting and Raftery, 2007). Cases of disease were comparatively rare in our simulation experiment and data examples. To evaluate the quality of prediction from the two models, recall that $\mathbf{y}^* \equiv (y_1^*, y_2^*, ..., y_{n^*}^*)'$ is the prediction set of observations that is generated from the true model. Each prediction observation is associated with a location $\mathbf{s}_i^*$ and time t_i^*. The formula for the LSR of the BIC selected model is equivalent to the log-likelihood for the Bernoulli distribution as follows:

$$LSR_{BIC} = \sum_{i=1}^{n^*} y_i^* \log(g^{-1}(u(\mathbf{s}_i^*, t_i^*))) + (1 - y_i^*) \log(1 - g^{-1}(u(\mathbf{s}_i^*, t_i^*))), \tag{2.28}$$

where $g^{-1}(u(\mathbf{s}_i^*, t_i^*))$ is the predicted probability of $y_i^* = 1$, given the selected value of J and the MLEs of $\mathbf{\Phi}$.

We define the formula for the LSR for m WBB draws ($r = 1, 2, ..., m$) from the approximate posterior distributions of $p(\mathbf{\Phi}|\mathbf{y}, J)$ as follows:

$$LSR_{WBB} = \sum_{i=1}^{n^*} y_i^* \log(\frac{1}{m}\sum_{r=1}^{m}\left\{g^{-1}(u(\mathbf{s}_i^*, t_i^*))^{(r)}\right\}) + (1 - y_i^*)\log(1 - \frac{1}{m}\sum_{r=1}^{m}\left\{g^{-1}(u(\mathbf{s}_i^*, t_i^*))^{(r)}\right\}), \quad (2.29)$$

where $g^{-1}(u(\mathbf{s}_i^*, t_i^*))^{(r)}$ is the predicted probability that $y_i^* = 1$, given the r^{th} draw from the approximate posterior distribution $p(\mathbf{\Phi}, J|\mathbf{y})$. When using the LSR to compare the predictive performance of models, the better model is identified by the LSR value that is least negative.

For each simulated scenario we also compared how the SABMA and likelihood-based methods performed in estimating or selecting the true number of introductions J. We tabulated the number of the 2,000 MC^3 draws that correctly identified J for each true value of J, separately. A classification success for the MC^3 algorithm was declared if the mode of the MC^3 draws was the true value of J for a particular data set. We also tabulated the number of scenarios in which BIC correctly identified the true value of J, separately. The simulation experiment required approximately seventy-five hours on a rented 96-core AWS elastic computing server for 1,000 scenarios.

2.4.2 Results

After generating 1,000 simulated data sets and fitting the various ensemble models using the SABMA and likelihood-based methods, the model with the correct value of J was selected by BIC and estimated by MC^3 in 88% of the scenarios (see **Tables 2.1-2.2** for selection and estimation breakdowns). The MC^3 credible interval coverage for J was 0.892. We note that the likelihood-based method did not produce a coverage probability for J. The average LSR_{BIC} was approximately -149.85 with a standard error of approximately 2.76. The average LSR_{WBB} was approximately -140.40 with a standard error of approximately 1.65. Therefore, the SABMA ensemble model performed better than the BIC selected ensemble model in terms of out-of-sample prediction using the LSR. The collective coverage of the credible intervals for all parameters in $\mathbf{\Phi}$ from the SABMA method was approximately 0.926, while the collective coverage probabilities of the confidence intervals for the parameters in the

likelihood-based method was approximately 0.882.

We note that 469 of the scenarios contained abnormalities where at least one BIC value was an `NA`. In scenarios where the BIC value from one or more models was an `NA`, we removed those models from consideration and conducted model selection using the remaining criterion values. This approach follows an ad hoc procedure that practitioners might use. We also note that the Hessian on the BIC-selected ensemble model was singular for twenty-three of the 1000 data sets, and therefore confidence intervals were unavailable. When calculating confidence interval coverage probabilities, we treated these cases as if the confidence intervals did not cover the true values. Additionally, when the WBB produced a draw from $p(\mathbf{\Phi}|\mathbf{y}, J)$ that was invalid because the `BFGS` algorithm did not converge, the draw was excluded from further calculations. Lastly, where applicable, the confidence intervals for t_{0_2}, t_{0_3}, t_{0_4}, and t_{0_5} from the BIC-selected ensemble model were obtained using the delta method from the `car` package in the program R (Fox et al., 2020; R Core Team, 2021). The delta method was necessary because extensive experimentation by the authors found that estimating the change in time between introductions resulted in more accurate estimation of J and the times of introduction, as opposed to estimating the time of introduction for all introductions directly.

Table 2.1: Results for BIC-selected value of J compared to the true value of J for each scenario. The generative model for each scenario contained between one to five introductions. We fit models to each data set that assumed anywhere from one to seven introductions. The bold numbers (across the diagonal from left to right) show the number of times that the model with the true number of introductions was correctly selected ($\frac{197+189+175+157+162}{1000} \times 100 =$ 88.0% correctly selected).

Model Choice for	True Number of Introductions in the Generative Model				
Number of Introductions	1	2	3	4	5
1	**197**	1	0	0	0
2	3	**189**	5	0	0
3	0	9	**175**	8	1
4	0	1	18	**157**	8
5	0	0	2	30	**162**
6	0	0	0	3	21
7	0	0	0	2	8

Table 2.2: Results for the MC^3 estimated and classified value of J compared to the true value of J for each scenario. The generative model for each scenario contained between one to five introductions. We fit models to each data set that assumed anywhere from one to seven introductions. We tabulated the number of the 2,000 MC^3 draws that were associated with each value of J separately. The classified value of J was determined by the mode of the MC^3 draws. The bold numbers (across the diagonal from left to right) show the number of times that the model with the true number of introductions was correctly estimated ($\frac{197+189+175+157+162}{1000} \times 100 = 88.0\%$ correctly estimated).

Model Choice for	True Number of Introductions in the Generative Model				
Number of Introductions	1	2	3	4	5
1	**197**	1	0	0	1
2	3	**189**	5	1	0
3	0	9	**175**	10	5
4	0	1	18	**157**	14
5	0	0	2	30	**162**
6	0	0	0	2	18
7	0	0	0	0	0

2.5 Michigan Data Example

We illustrate the utility of the SABMA and likelihood-based methods with an exploratory analysis of binary spatio-temporal surveillance data for CWD in white-tailed deer, collected in the lower peninsula of Michigan in the U.S. Our purpose is to obtain inference about the number, locations, and times that the pathogenic prion was introduced in the study area. We hypothesize that cases of CWD in the lower peninsula of Michigan in the mid-2010's are the result of approximately three to four separate pathogen introductions in the vicinity.

Chronic wasting disease (CWD) is an invariably fatal transmissible spongiform encephalopathy that affects cervids (e.g., elk, deer). First discovered in captive deer populations in Colorado, USA in the 1960s, it has spread to at least 26 U.S. states and can be found in five additional countries (Rivera et al., 2019). The causative prion has been found to spread by contact between deer (including between carcasses and live individuals) via saliva, urine, feces, and blood, and has been found to persist in the environment on vegetation and soil (Rivera et al., 2019). Thus, transmission may occur directly between individuals or indirectly through the environment.

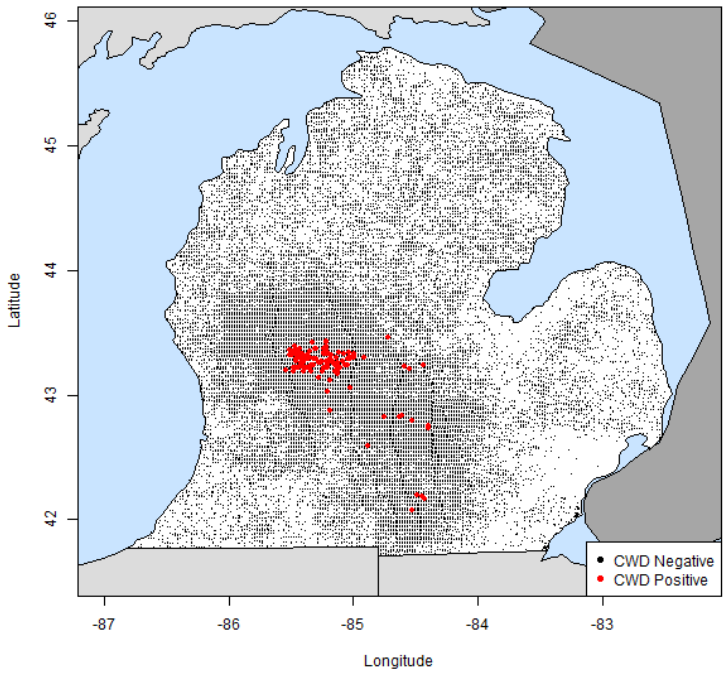

Figure 2.2: Plot of the lower peninsula of Michigan in the U.S. with the approximate locations of deer that tested positive for CWD (red) and negative for CWD (black) from 2002 - 2020.

The Michigan Department of Natural Resources has collected surveillance data on CWD since 2002. We defined our study area as the lower peninsula of Michigan covering approximately 106,000 km^2 that contained the primary outbreaks of CWD in Michigan. We restricted our analysis to data that were collected from 2002 (when surveillance for CWD began in Michigan) through the beginning of 2020. Including observations from these years allowed us to capture the initial diffusion dynamics of the early outbreaks. We then randomly split the data into a training set (80,401 obs. with 154 positive cases) and a prediction set (16,080 obs. with 22 positive cases). In all, our analysis included 96,481 deer, of which 176 tested positive for CWD (see **Figure 2.2** for a plot of the study area and data).

After examining the data, we determined that $J = 1, 2, ..., 10$ was a reasonable range for the possible values of J. To apply the likelihood-based method, we fit ten models of the form following (2.8-2.11) with each assuming a different value of J. We followed Hefley et al. (2020) in using the standard log-normal cumulative distribution function as $g^{-1}(\cdot)$. We fit

each model using the `BFGS` algorithm in the `optim` function in R. We chose starting values for locations of introduction using K-means clustering on the locations of observations where $y_i = 1$. We chose starting values for times of introduction using the earliest date of a CWD case within each cluster, subtracted by three years. We made this time adjustment based on the assumption that CWD was present several years before it was detected. We then calculated the BIC value for each candidate model and selected the model with the lowest BIC value. To apply the SABMA method, we first retrieved the MLEs that were obtained for the ten candidate ensemble models when employing the likelihood-based method. We then calculated the LAML for each of the ten candidate ensemble models and employed the MC^3 algorithm to draw 2,000 samples from $p(J|\mathbf{y})$. For this application, we redefined the neighborhood from (2.13) as:

$$\text{nbd}(J^{(r)}) = \begin{cases} \{1,2,3,4\} & \text{if } J^{(r)} = 1 \\ \{1,2,3,4,5\} & \text{if } J^{(r)} = 2 \\ \{1,2,3,4,5,6\} & \text{if } J^{(r)} = 3 \\ \{J^{(r)}-3, J^{(r)}-2, J^{(r)}-1, J^{(r)}, J^{(r)}+1, J^{(r)}+2, J^{(r)}+3\} & \text{if } 4 \leq J^{(r)} \leq 7 \\ \{5,6,7,8,9,10\} & \text{if } J^{(r)} = 8 \\ \{6,7,8,9,10\} & \text{if } J^{(r)} = 9 \\ \{7,8,9,10\} & \text{if } J^{(r)} = 10. \end{cases} \tag{2.30}$$

In due course, we obtained 2,000 approximate draws from $p(\mathbf{\Phi}, |\mathbf{y}, J)$ using the WBB. Finally, we used the LSR to compare the predictive performance of the ensemble model that was fitted using the SABMA method against the predictive performance of the BIC-selected ensemble model.

2.5.1 Results

The result of the SABMA method was a point-mass distribution for $p(J|\mathbf{y})$ at $J = 4$. The mean times of introduction were: 1991.829 (October 30, 1991), 2015.663 (August 31, 2015),

Table 2.3: Inference for each parameter in $\mathbf{\Phi}$ obtained using the SABMA and likelihood-based methods. We first present posterior summaries, composed of posterior means and 95% credible intervals, for parameters in the ensemble model obtained using the SABMA method. We then present MLEs and upper and lower limits of 95% confidence intervals for the same parameters obtained using the likelihood-based method.

	SABMA			Likelihood		
Parameter	Mean	2.5%	97.5%	MLE	2.5% CL	97.5% CL
μ	0.00219	0.00072	0.00317	0.00236	0.00173	0.00323
ω_{11} (Long.)	-84.598	-84.765	-84.517	-84.592	-84.654	-84.530
ω_{12} (Lat.)	42.936	42.787	43.085	42.943	42.857	43.030
t_{0_1}	1991.829	1808.234	2004.892	2004.802	2004.538	2005.066
θ_1	28,348	15,769	93,507	23,978	17,401	33,040
ω_{21} (Long.)	-85.468	-85.499	-85.430	-85.475	-85.508	-85.443
ω_{22} (Lat.)	43.271	43.246	43.294	43.270	43.243	43.298
t_{0_2}	2015.663	2011.976	2016.850	2016.079	2015.222	2016.937
θ_2	14,111	10,219	18,516	13,673	9,586	19,503
ω_{31} (Long.)	-85.133	-85.179	-85.079	-85.138	-85.181	-85.096
ω_{32} (Lat.)	43.248	43.217	43.277	43.247	43.216	43.277
t_{0_3}	2012.365	2011.835	2014.675	2012.086	2008.599	2015.573
θ_3	22,820	8,750	30,649	25,713	18,796	35,176
ω_{41} (Long.)	-84.477	-84.508	-84.437	-84.476	-84.537	-84.414
ω_{42} (Lat.)	42.169	42.121	42.205	42.168	42.104	42.231
t_{0_4}	2017.246	2015.980	2018.345	2017.184	2015.331	2019.037
θ_4	5,518	1,943	9,008	6,330	2,809	14,262

2012.365 (May 13, 2012), and 2017.246 (March 31, 2017). We provide posterior summaries (posterior mean and credible intervals) for all parameters in $\mathbf{\Phi}$ in **Table 2.3**. Additionally, we provide a plot of posterior inference on $\boldsymbol{\omega}_1, \boldsymbol{\omega}_2, \boldsymbol{\omega}_3$, and $\boldsymbol{\omega}_4$ within the lower peninsula of Michigan (see **Figure 2.3**).

The likelihood-based method selected the ensemble model where $J = 4$. The MLEs of times of introduction were: 2004.802 (October 20, 2004), 2016.079 (January 30, 2016), 2012.086 (February 1, 2012), and 2017.184 (March 9, 2017). We provide the MLEs and 95% Wald-type confidence intervals from the BIC-selected ensemble model in **Table 2.3**. Additionally, we provide a plot of the confidence regions for $\boldsymbol{\omega}_1, \boldsymbol{\omega}_2, ..., \boldsymbol{\omega}_7$ within the study area (see **Figure 2.3**). We note that the confidence intervals for t_{0_2}, t_{0_3}, and t_{0_4} were obtained using the delta method from the `car` package in the program R (Fox et al., 2020; R Core Team, 2021). The delta method was necessary because extensive experimentation by the

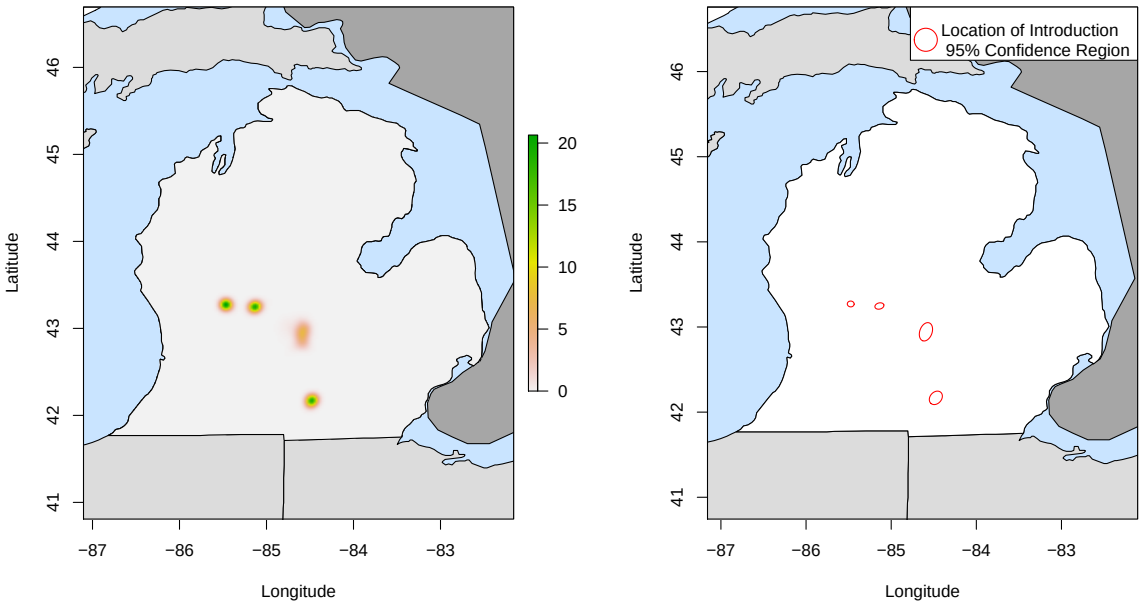

Figure 2.3: Kernel density plot (left) of $p(\boldsymbol{\omega}_1, \boldsymbol{\omega}_2, \boldsymbol{\omega}_3, \boldsymbol{\omega}_4 | \mathbf{y}, J = 4)$ within the lower peninsula of Michigan in the U.S. The frequentist 95% confidence regions (right) for $\boldsymbol{\omega}_1, \boldsymbol{\omega}_2, ..., \boldsymbol{\omega}_4$ within the lower peninsula of Michigan in the U.S.

authors found that estimating the change in time between introductions resulted in more accurate estimation of J and the times of introduction, as opposed to estimating the time of introduction for all introductions directly. We found that the likelihood-based method performed slightly better than the SABMA method in terms of out-of-sample predictive performance using the LSR (-116.6055 vs. -116.8642).

2.6 Wisconsin and Illinois Data Example

We now apply the SABMA and likelihood-based methods to a more difficult scenario of analyzing binary spatio-temporal disease surveillance data for chronic wasting disease (CWD) in white-tailed deer, collected in southern Wisconsin and northern Illinois. Like the Michigan data example, we complete an exploratory analysis. The Wisconsin and Illinois Departments of Natural Resources have collected disease surveillance data on CWD since 2001 and 2002, respectively. Obtaining sensible inference in this data scenario is more challenging because disease surveillance data was not widely collected in either Wisconsin or Illinois before the

first cases of CWD were detected. Therefore, much less information is available from the surveillance data on when and where the causative pathogen appeared. Nevertheless, our purpose is to conduct an exploratory analysis to obtain inference about the number, locations, and times of pathogenic prion introduction in the study area. We hypothesize that cases of CWD in southern Wisconsin and northern Illinois are the result of approximately three to six separate pathogen introductions in the vicinity.

We defined a study area in southern Wisconsin and northern Illinois covering approximately 34,500 km^2 that contained the initial outbreaks of CWD in both states. We restricted our analysis to data that were collected from 2001 (when CWD was first discovered in Wisconsin) through 2006. Including observations from these years allowed us to capture the initial diffusion dynamics of the early outbreaks in both states, while balancing computational considerations. We then randomly split the data into a training set (75,392 obs. with 799 positive cases) and a prediction set (15,078 obs. with 159 positive cases). In all, our analysis included 90,470 deer, of which 958 tested positive for CWD (see **Figure 2.4** for a plot of the study area and data).

After examining the data, we determined that $J = 1, 2, ..., 15$ was a reasonable range for the possible values of J. To apply the likelihood-based method, we fit fifteen models of the form following (2.8-2.11) with each assuming a different value of J. We followed Hefley et al. (2020) in using the standard log-normal cumulative distribution function as $g^{-1}(\cdot)$. We fit each model using the `BFGS` algorithm in the `optim` function in R. We chose starting values for locations of introduction using K-means clustering on the locations of observations where $y_i = 1$. We chose starting values for times of introduction using the earliest date of a CWD case within each cluster, subtracted by three years. We made this time adjustment based on the assumption that CWD was present several years before it was detected. We then calculated the BIC value for each candidate model and selected the model with the lowest BIC value. To apply the SABMA method, we first retrieved the MLEs that were obtained for the fifteen candidate ensemble models when employing the likelihood-based method. We then calculated the LAML for each of the fifteen candidate ensemble models and employed the MC^3 algorithm to draw 2,000 samples from $p(J|\mathbf{y})$. For this application, we redefined

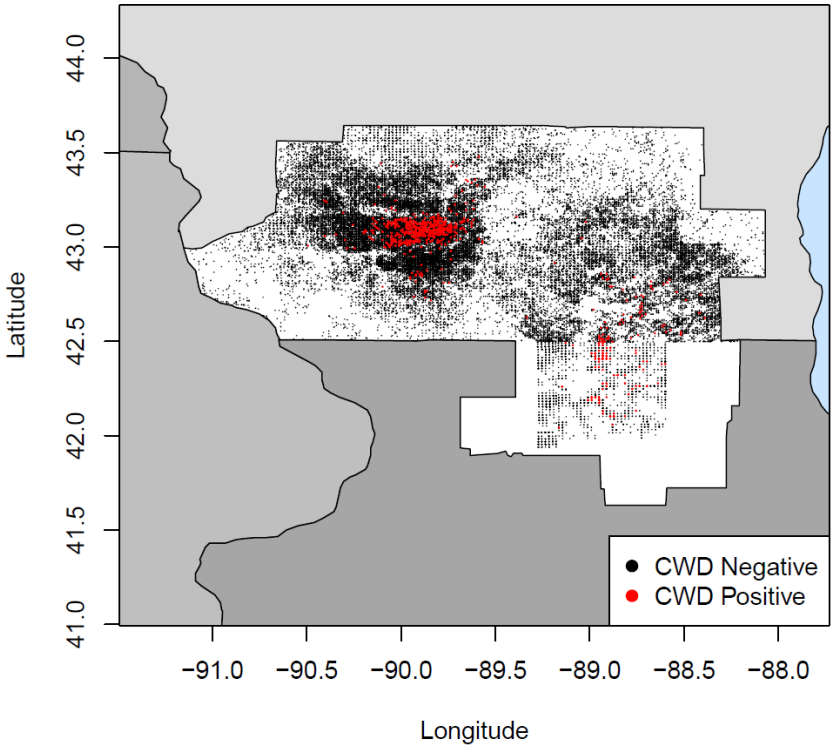

Figure 2.4: Plot of the study area in southern Wisconsin and northern Illinois in the U.S. with the approximate locations of deer that tested positive for CWD (red) and negative for CWD (black) from 2001 – 2006.

the neighborhood from (2.13) as:

$$\text{nbd}(J^{(r)}) = \begin{cases} \{1,2,3,4\} & \text{if } J^{(r)} = 1 \\ \{1,2,3,4,5\} & \text{if } J^{(r)} = 2 \\ \{1,2,3,4,5,6\} & \text{if } J^{(r)} = 3 \\ \{J^{(r)}-3, J^{(r)}-2, J^{(r)}-1, J^{(r)}, J^{(r)}+1, J^{(r)}+2, J^{(r)}+3\} & \text{if } 4 \leq J^{(r)} \leq 12 \\ \{10,11,12,13,14,15\} & \text{if } J^{(r)} = 13 \\ \{11,12,13,14,15\} & \text{if } J^{(r)} = 14 \\ \{12,13,14,15\} & \text{if } J^{(r)} = 15. \end{cases} \tag{2.31}$$

In due course, we obtained 2,000 approximate draws from $p(\mathbf{\Phi}, |\mathbf{y}, J)$ using the WBB. Finally, we used the LSR to compare the predictive performance of the ensemble model that was fitted using the SABMA method against the predictive performance of the BIC-selected ensemble model.

2.6.1 Results

The result of the SABMA method was a discrete distribution for $p(J|\mathbf{y})$ such that $Pr(J = 7|\mathbf{y}) = 0.9675$ and $Pr(J = 8|\mathbf{y}) = 0.0325$. The mean times of introduction for $J = 7$ were: 1985.889 (November 21, 1985), 1894.146 (February 23, 1894), 1994.221 (March 22, 1994), 1994.781 (October 13, 1994), 1872.770 (October 8, 1872), 1648.640 (August 22, 1648), and 1392.451 (June 14, 1392). The mean times of introduction for $J = 8$ were: 1983.975 (December 22, 1983), 1988.736 (September 26, 1988), 1817.557 (July 23, 1917), 1979.195 (March 13, 1979), 1980.596 (August 5, 1980), 1732.342 (May 5, 1732), 1764.451 (June 13, 1764), and 1798.512 (July 6, 1798). We provide posterior summaries (posterior mean and credible intervals) for all parameters in $\mathbf{\Phi}$ in **Table 2.4**. Additionally, we provide a plot of posterior inference on $p(\boldsymbol{\omega}_1, \boldsymbol{\omega}_2, ..., \boldsymbol{\omega}_J|\mathbf{y}, J)$ within the study area (see **Figure 2.5**).

The likelihood-based method selected the ensemble model where $J = 7$. The MLEs of times of introduction were: 1988.062 (January 1, 1988), 1925.671 (September 2, 1925),

Table 2.4: Posterior summaries, composed of posterior means and 95% credible intervals, for parameters in the ensemble model fit using the SABMA method. We first present summaries for WBB draws corresponding to $J = 7$, followed by summaries for WBB draws where $J = 8$

	$J = 7$			$J = 8$		
Parameter	Mean	2.5%	97.5%	Mean	2.5%	97.5%
μ	0.00014	0.00005	0.00025	0.00012	0.00006	0.00018
ω_{11} (Long.)	-89.943	-90.027	-89.907	-89.929	-89.988	-89.902
ω_{12} (Lat.)	43.089	43.079	43.100	43.084	43.076	43.093
t_{0_1}	1985.899	1966.176	1996.530	1983.975	1964.355	1996.894
θ_1	4,485	2,165	6,330	4,733	2,193	7,378
ω_{21} (Long.)	-89.555	-89.661	-89.434	-89.778	-89.808	-89.747
ω_{22} (Lat.)	43.260	43.148	43.383	43.110	43.102	43.121
t_{0_2}	1894.146	1636.453	1994.800	1988.736	1970.814	1998.469
θ_2	16,311	3,464	36,677	3,644	1,703	6,209
ω_{31} (Long.)	-89.789	-89.843	-89.764	-90.158	-90.319	-90.082
ω_{32} (Lat.)	43.107	43.098	43.116	43.101	43.072	43.149
t_{0_3}	1994.221	1981.283	1999.830	1817.557	1680.769	1906.675
θ_3	2,615	1,614	4,804	23,344	16,985	36,559
ω_{41} (Long.)	-88.926	-88.955	-88.898	-89.876	-89.901	-89.836
ω_{42} (Lat.)	42.429	42.394	42.475	42.771	42.740	42.801
t_{0_4}	1994.781	1973.806	2002.925	1979.195	1956.822	1992.862
θ_4	2,020	435	4,835	4,345	2,406	6,677
ω_{51} (Long.)	-88.894	-88.974	-88.824	-89.663	-89.728	-89.626
ω_{52} (Lat.)	42.118	41.994	42.228	43.382	43.246	43.446
t_{0_5}	1872.770	1554.685	1998.009	1980.596	1926.103	2006.080
θ_5	23,072	3,048	59,876	4,627	1,448	9,783
ω_{61} (Long.)	-88.775	-88.821	-88.724	-88.756	-88.825	-88.660
ω_{62} (Lat.)	42.571	42.477	42.671	42.620	42.493	42.741
t_{0_6}	1648.640	1188.930	1884.732	1732.342	1451.360	1844.861
θ_6	73,768	44,351	100,489	51,798	22,826	90,008
ω_{71} (Long.)	-90.089	-90.194	-89.963	-89.464	-89.627	-89.317
ω_{72} (Lat.)	42.953	42.749	43.071	43.127	42.974	43.218
t_{0_7}	1392.451	765.367	1743.294	1764.451	1503.463	1969.954
θ_7	75,756	39,381	135,598	28,787	6,379	43,539
ω_{81} (Long.)				-88.901	-88.981	-88.852
ω_{82} (Lat.)				42.227	41.952	42.346
t_{0_8}				1798.512	1617.372	1935.709
θ_8				42,399	18,281	64,677

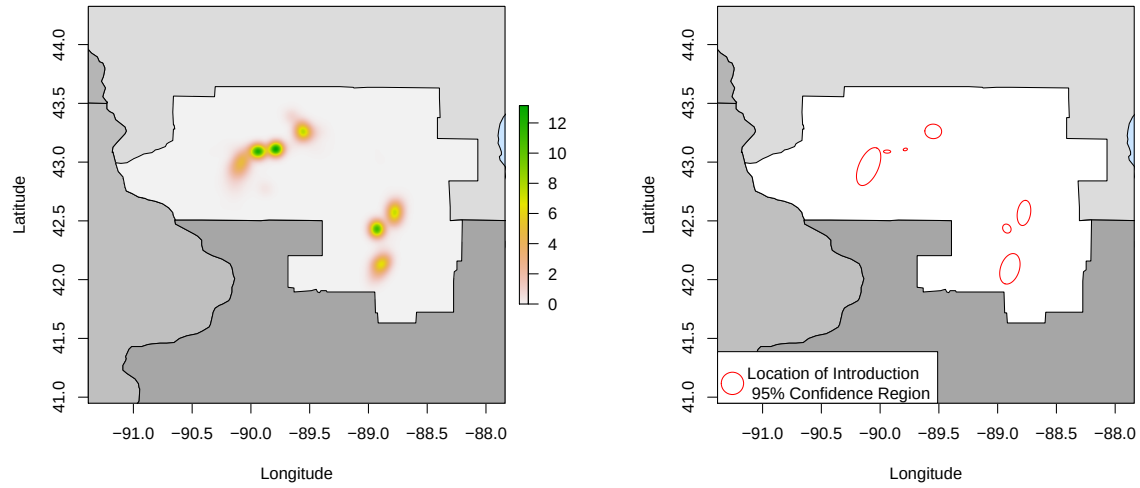

Figure 2.5: Kernel density plot (left) of $p(\boldsymbol{\omega}_1, \boldsymbol{\omega}_2, ..., \boldsymbol{\omega}_J|\mathbf{y}, J)$ within the study area in southern Wisconsin and northern Illinois in the U.S. The 95% confidence regions (right) for $\boldsymbol{\omega}_1, \boldsymbol{\omega}_2, ..., \boldsymbol{\omega}_7$ within the study area in southern Wisconsin and northern Illinois in the U.S.

1996.582 (August 1, 1996), 1995.669 (September 2, 1995), 1902.046 (January 17, 1902), 1710.411 (May 31, 1710), and 1485.846 (November 5, 1485). We provide the MLEs and 95% Wald-type confidence intervals from the BIC-selected ensemble model in **Table 2.5**. Additionally, we provide a plot of the confidence regions for $\boldsymbol{\omega}_1, \boldsymbol{\omega}_2, ..., \boldsymbol{\omega}_7$ within the study area (see **Figure 2.5**). We note that the confidence intervals for t_{0_2}, t_{0_3},...,t_{0_7} were obtained using the delta method from the `car` package in the program R (Fox et al., 2020; R Core Team, 2021). The delta method was necessary because extensive experimentation by the authors found that estimating the change in time between introductions resulted in more accurate estimation of J and the times of introduction, as opposed to estimating the time of introduction for all introductions directly. We found that the SABMA method performed better than the likelihood-based method in terms of out-of-sample prediction performance using the LSR (-710.8041 vs. -714.0254).

Table 2.5: Inference for each parameter in $\mathbf{\Phi}$ in the ensemble model from the likelihood-based method. We present the MLE and upper and lower limits of the respective 95% confidence interval for each parameter.

Parameter	MLE	2.5% CL	97.5% CL
μ	0.00015	0.00006	0.00035
ω_{11} (Long.)	-89.941	-89.965	-89.916
ω_{12} (Lat.)	43.089	43.079	43.100
t_{0_1}	1988.062	1975.094	2001.031
θ_1	4,753	3,256	6,937
ω_{21} (Long.)	-89.548	-89.605	-89.491
ω_{22} (Lat.)	43.261	43.212	43.311
t_{0_2}	1925.671	1843.770	2007.572
θ_2	15,050	8,426	26,882
ω_{31} (Long.)	-89.785	-89.799	-89.771
ω_{32} (Lat.)	43.108	43.099	43.117
t_{0_3}	1996.582	1990.390	2002.774
θ_3	2,403	1,667	3,465
ω_{41} (Long.)	-88.922	-88.951	-88.893
ω_{42} (Lat.)	42.434	42.404	42.465
t_{0_4}	1995.669	1985.116	2006.222
θ_4	2,111	1,012	4,407
ω_{51} (Long.)	-88.896	-88.965	-88.826
ω_{52} (Lat.)	42.092	41.988	42.196
t_{0_5}	1902.046	1789.647	2014.445
θ_5	22,292	8,570	57,876
ω_{61} (Long.)	-88.776	-88.823	-88.730
ω_{62} (Lat.)	42.568	42.482	42.653
t_{0_6}	1710.411	1412.809	2008.014
θ_6	77,312	54,283	110,109
ω_{71} (Long.)	-90.098	-90.181	-90.015
ω_{72} (Lat.)	42.962	42.833	43.091
t_{0_7}	1485.846	980.371	1991.321
θ_7	74,298	37,021	149,110

2.7 Discussion

We have proposed two methods to fit an ensemble model and obtain inference on the number, locations, and times of introduction of a novel pathogen using binary spatio-temporal disease surveillance data. We evaluated the two methods of fitting the ensemble model by conducting a simulation experiment and an exploratory analysis of two CWD surveillance data sets. Our simulation experiment demonstrated that the model fit using our SABMA method performed similarly to traditional BIC-based model selection in estimating the number of introductions, and performed better than the BIC-selected models in estimating the locations and times of introduction. The simulation experiment also showed that the model fit using the SABMA method was better in terms of predictive performance according to the LSR. While the likelihood-based method produced an ensemble model that was marginally better at estimation and prediction in a well-behaved disease surveillance example from Michigan, the sensibility of inference from the likelihood-based ensemble model began to break down in the data example from Wisconsin and Illinois. Particularly, the upper confidence limits on some times of introduction from the likelihood-based method were past 2006, when the last observation was collected. In contrast, the SABMA method provided upper credible interval limits on most of the times of introduction that were sensible, given the difficulties of the Wisconsin/Illinois data set. However, both the likelihood-based and SABMA methods produced some unreasonable MLE/posterior mean values and wide confidence or credible intervals on times of introduction. Despite this, the SABMA ensemble model fit to the Wisconsin/Illinois data set was preferred to the likelihood-based method in terms of out-of-sample predictive performance according to the LSR. It is notable that the likelihood-based method performs relatively well in simulation and the Michigan data example, and that the likelihood-based method may be considered an early step in implementing the SABMA method.

Many compartment-based infectious disease forecasting models explicitly account for the incubation time of the pathogen (between exposure and when clinical symptoms present themselves (Vynnycky and White, 2010). The minimum incubation time for CWD in deer

in an experimental setting was about fifteen months (Williams et al., 2002; Williams and Miller, 2002). When estimating the locations and times that the pathogen was introduced, we did not explicitly account for the incubation period (and the possible symptomatic period) between the time that a deer was infected and the time that the deer tested positive for CWD. Our ensemble model could be adjusted to account for the uncertainty due to these lag times and would likely result in estimates of pathogen introduction times that are earlier. Previous work has also shown that the susceptibility of deer to the pathogen may be affected by individual-level covariates (e.g., age, sex; Heisey et al., 2010). Future work may append a susceptibility component to the total pathogen intensity to account for differences in susceptibility among individual deer.

Due to data privacy concerns, disease surveillance data may only be available in an aggregated form within study area subregions and segments of time (e.g., counts of individuals with positive and negative test results). If the aggregated form of the data matches any of Data Types B-E from Walker et al. (2021) or Chapter 1, the distributional results from Walker et al. (2021) or Chapter 1 may be adapted to the components of the ensemble model in (2.8-2.11) to estimate J and $\mathbf{\Phi}$. For example, suppose the data are counts of individuals that tested positive (n_{1l}) and negative (n_{0l}) for an infectious disease in the l^{th} space-time cube composed of a subregion and time interval ($\mathcal{A}_l \times \mathcal{T}_l$) for $l = 1, 2, ..., L$. The joint model for these counts (Type C aggregated data) can be adapted from Walker et al. (2021) and (2.8-2.11) as follows:

$$n_{1l} \sim \text{Pois}\left(\int_{\mathcal{A}_l}\int_{\mathcal{T}_l} \lambda(\mathbf{s},t)p(\mathbf{s},t)dtd\mathbf{s}\right), \tag{2.32}$$

$$n_{0l} \sim \text{Pois}\left(\int_{\mathcal{A}_l}\int_{\mathcal{T}_l} \lambda(\mathbf{s},t)(1-p(\mathbf{s},t))dtd\mathbf{s}\right), \tag{2.33}$$

$$g(p(\mathbf{s},t)) = u(\mathbf{s},t) \tag{2.34}$$

$$u(\mathbf{s},t) = \sum_{j=1}^{J} u_j(\mathbf{s},t) \tag{2.35}$$

$$u_j(\mathbf{s},t) = \frac{\theta_j}{4\pi\mu(t-t_{0_j})}\exp\left\{\frac{-||\mathbf{s}-\boldsymbol{\omega}_j||^2}{4\mu(t-t_{0_j})}\right\}, \tag{2.36}$$

where $\lambda(\mathbf{s}, t)$ is the sampling intensity that influenced how many individuals were tested within each space-time cube. If $\lambda(\mathbf{s}, t)$ is known, it may be substituted directly. Alternatively, $\lambda(\mathbf{s}, t)$ may be specified using a function of spatial covariates and parameters, and the parameters may be jointly estimated with $\mathbf{\Phi}$. A common specification for the sampling intensity is $\log(\lambda(\mathbf{s}, t)) = \beta_0 + \mathbf{x}(\mathbf{s}, t)'\boldsymbol{\beta}$. Here, $\mathbf{x}(\mathbf{s}, t)$ is a vector of spatial covariates that may vary over time, $\boldsymbol{\beta}$ is a corresponding vector of regression parameters, and β_0 is an intercept parameter.

The approach from this chapter uses an analytical solution (2.7) to the simple ecological diffusion PDE (2.1) under certain simplifying assumptions. As it has been said, "all models are wrong, but some are useful" (Box and Draper, 1987). We therefore acknowledge that the constant diffusion assumption required for (2.7) may be too strong for some scenarios, particularly when applied to data collected long after an initial outbreak or when inhomogenous diffusion is suspected. Additionally, (2.1) is incapable of accounting for growth in the pathogen intensity over time, as the pathogen reproduces and sheds from its host. As such, future work will address the problem of obtaining inference on the number, locations, and times of introduction as well as inhomogenous pathogen diffusion and growth dynamics. As our method stops short of attempting to draw inference about contributing spatial factors that may influence where pathogen introductions occur, future work may also address this challenge.

Acknowledgements

We thank the Illinois, Michigan, and Wisconsin Departments of Natural Resources for obtaining deer tissue samples and the hunters in Illinois, Michigan, and Wisconsin who provided them. We acknowledge support for this research from USGS G18AC00317 and G16AC00413. Any use of trade, firm, or product names is for descriptive purposes only and does not imply endorsement by the U.S. Government.

Chapter 3

Predicting the Risk of Novel Pathogen Introductions from Disease Surveillance Data

3.1 Abstract

In the course of an infectious disease outbreak, researchers often must estimate or infer the source of the causative pathogen, the risk factors associated with the spread and growth of the pathogen, and risk factors that may be associated with new outbreaks. Because the exact time and location of introduction for the pathogen is usually unobserved, these questions must be addressed using incomplete or indirect data, such as disease surveillance data. We introduce a Bayesian hierarchical mixture of ecological diffusion models (BHMEDM), for disease surveillance data, that estimates parameters associated with the dynamic process of a pathogen diffusing and multiplying through a study area from multiple initial locations. We address several computational challenges and provide inference for the number, locations, and times of introduction of the pathogen into a population. We also obtain inference on the spatio-temporal process associated with the pathogen introductions and predict where new introductions are likely to occur in the future. We apply this method to disease surveillance

data for chronic wasting disease in white-tailed deer (*Odocoileus virginianus*) from southern Wisconsin and northern Illinois in the United States.

3.2 Introduction

Scientists and practitioners working in areas such as public health, animal and plant pathology and disease ecology are often concerned with identifying and studying outbreaks of infectious disease. One way that outbreaks are identified and studied is through collecting disease surveillance data. Disease surveillance data, often consisting of individual or aggregated test results for a pathogen, contain information about the presence and prevalence of infectious disease. When analyzed, disease surveillance data provide clues about the origins and causes of an outbreak, and inform current understanding of how pathogens diffuse and grow through a population or ecosystem (Lee et al., 2010). In this chapter, we again focus on studying the initiation and progression of infectious disease outbreaks. However, we reframe and address the trans-dimensional estimation problem from Section 2.3.3 using tools and ideas from the mixture model, point process, and missing data literature. Incorporating tools and ideas from these areas in our estimation framework allow us to make predictions about where the pathogen is likely to be introduced in the future (which we did not attempt in Chapter 2).

Non-spatial methods, such as compartmental models, are commonly used to forecast the temporal spread of a pathogen through a population (Vynnycky and White, 2010; Adivar and Selin Selen, 2013). Additionally, non-spatially explicit methods such as case-control and epidemic curve analysis are commonly used to identify the likely physical source of a pathogen and the time of pathogen introduction in infectious disease outbreaks (Borgan et al., 2018; Egan and Hall, 2015). In contrast, methods that explicitly model the spatio-temporal dynamics of an outbreak are less frequently employed (e.g., using partial differential equations or agent-based models; Legrand et al., 2009; Garlick et al., 2014; Hefley et al., 2017c; Hefley et al., 2020; Banks and Hooten, 2021). Likewise, methods that explicitly estimate the spatio-temporal origins of an outbreak, while forecasting the spread of the

pathogen, are much less common than non-spatially explicit methods. For example, the spatial, but not temporal, origins of an outbreak can be identified using geographic clustering or profiling (LeComber et al., 2011; Mohler and Short, 2012; Stevenson et al., 2012; Verity et al., 2014). Levy et al. (2011) estimated the locations and times that a pathogen was introduced into a population using a distance-based epicenter regression model, but they did not forecast the spatio-temporal spread of the pathogen. To the best of our knowledge, only Legrand et al. (2009), Hefley et al. (2020), and Chapter 2 of this dissertation estimate the location and time of pathogen introduction while forecasting the spread of the pathogen. Of these, only Chapter 2 of this dissertation accounts for and estimates more than one location and time of pathogen introduction. However, the method in Chapter 2 relied on a simplified representation of homogeneous ecological diffusion with no pathogen growth that may be unrealistic in practice.

Predicting where novel pathogens may be introduced into a population in the future is a difficult problem, at least partially because pathogens can adapt in unexpected ways to form a niche in otherwise inhospitable new environs (Roy et al., 2017). It is well recognized that the movements of plants, animals, and people are large drivers for the risk of a novel pathogen introduction (Fèvre et al., 2006; Santini et al., 2018; Gottwald et al., 2019; Kraemer et al., 2019). Efforts to predict the risk of pathogen introductions have relied on both mechanistic and phenomenological models to account for the movement of animals or people. For example, Gottwald et al. (2019) developed a method that produced census-tract level risk predictions for a pathogen introduction across the U.S. The Gottwald et al. (2019) method relied on pathogen distribution and prevalence data from countries where the pathogen was already present, data about travel between those countries and the U.S., and data about travel between U.S. census tracts. Likewise, Oleson and Wikle (2013) developed a dynamic spatio-temporal model that produced county-level risk predictions for the introduction of avian influenza into poultry farms in the U.S. by migrating waterfowl. The Oleson and Wikle (2013) model relied on counts of migrating waterfowl across the U.S. over time and the number of poultry farms in each U.S. county. The resulting spatio-temporal model predicted the varying weekly county-level risk of pathogen introduction throughout the waterfowl

migration season. Methods that relied on spatial or environmental covariates (e.g., elevation variability, forest density, and climate) have also been used to predict the risk of other spatially-referenced events (e.g., tornado touchdowns; Karpman et al., 2013) or invasive species introductions (e.g., Asian giant hornet queen wintering and hive sites; Norderud et al., 2021). In particular, Karpman et al. (2013) modeled the locations of tornado touchdowns and spatially predicted the future touchdown risk in the eastern U.S. using an inhomogeneous Poisson point process.

To the best of our knowledge, no methods exist that provide a unified framework to estimate the origins of an infectious disease outbreak, predict the growth and spread of the pathogen, and predict where new introductions are likely to occur in the future. This chapter seeks to fill that gap in the literature. In this chapter, we propose a Bayesian hierarchical mixture of ecological diffusion models (BHMEDM) with a spatio-temporal inhomogeneous Poisson point process (IPPP) component. The BHMEDM incorporates partial differential equations (PDEs) that represent ecological diffusion processes with exponential growth. Including these PDEs make the BHMEDM capable of providing predictive inference on the growth and diffusion of a pathogen through a study area over time. The mixture aspect of the BHMEDM enables inference on the number, locations, and times that a pathogen was introduced, and the spatio-temporal IPPP component of the model provides inference on the spatio-temporal process for the locations and times of pathogen introduction. The IPPP component also enables prediction for where new introductions are likely to occur in the future. We then apply the BHMEDM to binary spatio-temporal disease surveillance data of chronic wasting disease (CWD) in white-tailed deer (*Odocoileus virginianus*) in southern Wisconsin and northern Illinois in the U.S. Importantly, we obtain predictive inference about where new pathogen introductions are likely to occur in a broader region that includes much of Illinois, Iowa, Michigan, Minnesota, and Wisconsin.

3.3 Methods

3.3.1 Sum of Ecological Diffusion with Exponential Growth PDEs

An ecological diffusion with exponential growth process is a PDE that describes the movement of biotic entities across a region of interest over time from an initial intensity at some location or area, while accounting for growth in intensity due to favorable reproductive conditions. Researchers have previously used ecological diffusion PDEs to model the spread of pathogens from a single location of introduction (Garlick et al., 2011; Garlick et al., 2014; Hefley et al., 2017b; Hefley et al., 2017c; Hefley et al., 2020). However, numerically solving the PDEs associated with these models has been sufficiently difficult that researchers have historically been limited to modeling epidemics caused by a single pathogen introduction. We present a sum of ecological diffusion PDEs that accounts for a pathogen spreading through a population from multiple locations. We also present several techniques that reduce the computational burden and make the model more practical to fit.

Let $\mathbf{s} \equiv (s_1, s_2)'$ be the coordinates of any given point within a two dimensional study area $\mathcal{S}$ and let t signify time in some interval $\mathcal{T}$. Assuming that pathogens may be introduced to a population at multiple times and locations, the unobserved total pathogen intensity $u(\mathbf{s}, t)$ at any location and time can be viewed as a sum of J component pathogen intensities. These component pathogen intensities, in turn, are modeled by individual PDEs that represent ecological diffusion with exponential growth. The sum of J PDEs for ecological diffusion with exponential growth can be expressed as (Hefley et al., 2017c; Hefley et al., 2020; Chapter 2):

$$u(\mathbf{s}, t) = \sum_{j=1}^{J} u_j(\mathbf{s}, t), \tag{3.1}$$

$$\frac{\partial}{\partial t} u_j(\mathbf{s}, t) = \left(\frac{\partial^2}{\partial s_1^2} + \frac{\partial^2}{\partial s_2^2} \right) [\mu(\mathbf{s}) u_j(\mathbf{s}, t)] + \lambda(\mathbf{s}) u_j(\mathbf{s}, t), \tag{3.2}$$

where $\mu(\mathbf{s})$ is a common spatially varying diffusion coefficient that determines the rate of spread for any location, and $\gamma(\mathbf{s})$ is a commonly shared growth rate for the pathogen at any location. Integrating $u(\mathbf{s}, t)$ over any spatial area provides the expected pathogen concen-

tration within that area at that time.

To initiate this group of ecological diffusion processes, J initial intensities $u_j(\mathbf{s}, t_{0_j})$ (for $j = 1, 2, ..., J$) must be defined across the study area, where t_{0_j} is the time that the j^{th} introduction occurred and $\boldsymbol{\omega}_j \equiv (\omega_{1,j}, \omega_{2,j})'$ is the vector of coordinates for the center of the initial intensity. Following Hefley et al. (2017c) and Williams et al. (2017), we choose bivariate Gaussian kernel initial conditions:

$$c_j(\mathbf{s}, t_{0_j}) = \frac{\theta_j}{\sqrt{4\pi^2\sigma_1^2\sigma_2^2}} e^{\{\frac{-(s_1-\omega_{1,j})^2}{2\sigma_1^2}\}} e^{\{\frac{-(s_2-\omega_{2,j})^2}{2\sigma_2^2}\}}, \tag{3.3}$$

where $u_j(\mathbf{s}, t_{0_j}) = \frac{c_j(\mathbf{s}, t_{0_j})}{\mu(\mathbf{s})}$ and θ_j is a scaling factor for the initial pathogen intensity of the j^{th} introduction. We define $\sigma^2 = \sigma_1^2 = \sigma_2^2$ as the common variance of the Gaussian kernel initial conditions.

Computational Challenges with Solving PDEs

Numerically solving PDEs over space and time is a significant computational challenge that prevents the wide-spread use of PDEs in statistical applications. Typically, to obtain a value from the PDE at a particular location and time, practitioners must numerically approximate the solution to the PDE (e.g., using the finite difference method). Additional computational tools, such as homogenization, have made PDE-based models more accessible for statistical applications (Garlick et al., 2011; Hooten et al., 2013). Broadly speaking, homogenization is a harmonic mean-based form of upscaling that improves the efficiency and stability of an algorithm for solving a PDE (Hooten et al., 2013). For example, Hefley et al. (2017c) used the homogenization technique and the finite difference method to approximate the ecological diffusion PDE in (3.2) for a single pathogen introduction. Following Hooten et al. (2013) and Hefley et al. (2017c), the homogenized PDE is given as:

$$\frac{\partial}{\partial t} c_j(\mathbf{s}, t) = \bar{\mu}(\mathbf{s}) \left(\frac{\partial^2}{\partial s_1^2} + \frac{\partial^2}{\partial s_2^2} \right) c_j(\mathbf{s}, t) + \bar{\lambda}(\mathbf{s}) c_j(\mathbf{s}, t), \tag{3.4}$$

where the homogenized diffusion rate $\bar{\mu}(\mathbf{s})$ is:

$$\bar{\mu}(\mathbf{s}) = \left(\frac{1}{|\mathcal{A}|} \int_{\mathcal{A}} \frac{1}{\mu(\mathbf{s})} d\mathbf{s} \right)^{-1}, \tag{3.5}$$

and the the homogenized growth rate $\lambda(\mathbf{s})$ is:

$$\bar{\lambda}(\mathbf{s}) = \frac{\bar{\mu}(\mathbf{s})}{|\mathcal{A}|} \int_{\mathcal{A}} \frac{\lambda(\mathbf{s})}{\mu(\mathbf{s})} d\mathbf{s}. \tag{3.6}$$

We define $c_j(\mathbf{s}, t)$ as the homogenized pathogen intensity for the j^{th} introduction and $u_j(\mathbf{s}, t) \approx \frac{c_j(\mathbf{s},t)}{\mu(\mathbf{s})}$ at any location and time after t_{0_j}.

Unfortunately, certain computational barriers persist in numerically solving a sum of PDEs in (3.1-3.3). For example, the stability of the finite difference approximation is dependent upon the level of temporal resolution in the model. Additionally, the computational complexity of the mathematical model in (3.1-3.3) renders it inappropriate when the spread of a pathogen is known to be from multiple initial locations that are separated in space and time. We address these computational issues in subsequent sections.

Analytical Solution to Homogenized Ecological Diffusion

The homogenized PDE in (3.4) has been central to several statistical applications (Hooten et al., 2013; Williams et al., 2017; Hefley et al., 2017b,c; Lu et al., 2020; Hefley et al., 2020). However, numerically solving (3.4) is computationally intensive and, in some cases, prohibitive. We now introduce an analytical solution to the homogenized ecological diffusion PDE in (3.4) that was obtained through collaboration with Dr. Ian McGahan at the University of Wisconsin. The analytical solution provides an approximation to the PDE in (3.2) for any location and time in the study domain and results in substantial computational savings when obtaining $u_j(\mathbf{s}, t)$. To obtain the analytical solution, we first transform both the initial conditions from (3.3) and the homogenized PDE from (3.4) into the Fourier parameter space. Second, we solve the resulting ordinary differential equation. Lastly, we back-transform the solution to the ordinary differential equation into real space. The result is an analytical solution to the homogenized PDE for any location and time after the initial time t_{0_j}, as

follows (Logan, 2014):

$$u_j(\mathbf{s},t) = \frac{1}{\mu(\mathbf{s})}\frac{\theta_j}{2\pi(\sigma^2+2\bar{\mu}(\mathbf{s})(t-t_{0_j}))}e^{\bar{\lambda}(\mathbf{s})(t-t_{0_j})}e^{\frac{-||\mathbf{s}-\boldsymbol{\omega}_j||^2}{2(\sigma^2+2\bar{\mu}(\mathbf{s})(t-t_{0_j}))}}, \tag{3.7}$$

where $||\mathbf{s}-\boldsymbol{\omega}||^2$ is the squared Euclidean distance between any location $\mathbf{s}$ and the location of introduction $\boldsymbol{\omega}_j$. We can directly substitute the analytical solution in (3.7) for (3.2) when specifying (3.1-3.3). As a result, we may obtain the pathogen intensity at any location and time much faster and easier than numerically solving (3.2) or even (3.4). We derive the analytical solution from (3.7) in Appendix B.

3.3.2 A Bayesian Hierarchical Mixture of Ecological Diffusion Models

We introduce a Bayesian hierarchical mixture of ecological diffusion models (BHMEDM), based on (3.1, 3.3, 3.7), that may be fit to binary spatio-temporal disease surveillance data. In the context of disease surveillance, the pathogen intensity at any given location and time is unobserved. Rather, practitioners may have binary data from individuals that denote the presence or absence of the pathogen. Thus, the i^{th} observation is a binary mark reported for an individual at a specific location $\mathbf{s}_i$ and time t_i, such that $y_i = 1$ (for $i = 1, ..., n$) denotes the presence of the pathogen in that individual and $y_i = 0$ otherwise. These binary observations may be viewed as Bernoulli random variables, each dependent on the latent total pathogen intensity $u(\mathbf{s}_i, t_i)$ and the deterministic approximate PDE analytical solution in (3.7) as follows:

$$y_i \sim \text{Bern}(p(\mathbf{s}_i, t_i)), \tag{3.8}$$

$$p(\mathbf{s}_i, t_i) = g^{-1}(u(\mathbf{s}_i, t_i)e^{\mathbf{b}_i'\boldsymbol{\tau}}), \tag{3.9}$$

$$u(\mathbf{s}_i, t_i) = \sum_{j=1}^{J} u_j(\mathbf{s}_i, t_i), \tag{3.10}$$

$$u_j(\mathbf{s}_i, t_i) = \frac{1}{\mu(\mathbf{s}_i)} \frac{\theta_j}{2\pi(\sigma^2 + 2\bar{\mu}(\mathbf{s}_i)(t_i - t_{0_j}))} e^{\bar{\lambda}(\mathbf{s}_i)(t_i - t_{0_j})} e^{\frac{-||\mathbf{s}_i - \boldsymbol{\omega}_j||^2}{2(\sigma^2 + 2\bar{\mu}(\mathbf{s}_i)(t_i - t_{0_j}))}}, \tag{3.11}$$

where $p(\mathbf{s}_i, t_i)$ is the probability that $y_i = 1$ and $g^{-1}(\cdot)$ is a suitable inverse-link function (e.g., standard log-normal cumulative distribution function). We define $e^{\mathbf{b}_i'\boldsymbol{\tau}}$ as a susceptibility factor that accounts for differences in susceptibility to the pathogen among individuals. This susceptibility factor is dependent on individual-level attributes (e.g., sex), denoted by the covariate vector $\mathbf{b}_i$ for the i^{th} observation, and a vector of susceptibility parameters $\boldsymbol{\tau}$.

As before, the unobserved total pathogen intensity $u(\mathbf{s}_i, t_i)$ at the i^{th} location and time can be viewed as a sum of J component pathogen intensities. Thus, we assume that cases of an infectious disease within a population are the result of one or more pathogen introductions across a study area. Particularly, the location-level risk associated with having the pathogen is cumulative, according to how many pathogen introductions occurred beforehand as well as the locations of the introductions relative to the locations of sampled individuals. This corresponds to the idea that an individual may be exposed to the pathogen from one or more sources with varying degrees of exposure from each source at any given time. The common diffusion and growth terms across the J homogenized PDEs are represented as:

$$\log(\mu(\mathbf{s})) = \alpha_0 + \mathbf{z}(\mathbf{s})'\boldsymbol{\alpha}, \tag{3.12}$$

$$\lambda(\mathbf{s}) = \gamma_0 + \mathbf{w}(\mathbf{s})'\boldsymbol{\gamma}. \tag{3.13}$$

As specified, $\mu(\mathbf{s})$ and $\lambda(\mathbf{s})$ are dependent on log-linear and linear models, respectively. We define α_0 as an intercept parameter, $\mathbf{z}(\mathbf{s})$ as a vector of spatial covariates, and $\boldsymbol{\alpha}$ as the corresponding vector of slope parameters that influence the diffusion of the pathogen.

Likewise, we define γ_0 as an intercept parameter, $\mathbf{w}(\mathbf{s})$ as a vector of spatial covariates, and $\boldsymbol{\gamma}$ as the corresponding vector of slope parameters that influence the growth of the pathogen.

Predicting where pathogen introductions are likely to occur in the future is difficult (Roy et al., 2017). In some cases, data may be available about the movements of plants, animals, or people to inform a model and predict the risk of a pathogen introduction (Oleson and Wikle, 2013; Gottwald et al., 2019). In other circumstances, environmental or spatial covariates may act as surrogates for the underlying spatio-temporal process that influences the introduction of a novel pathogen. We propose to obtain predictive inference on this spatio-temporal process by specifying a spatio-temporal inhomogeneous Poisson point process (IPPP) as a joint model for J, $\boldsymbol{\Omega} \equiv (\boldsymbol{\omega}_1, \boldsymbol{\omega}_2, ..., \boldsymbol{\omega}_J)'$, and $\mathbf{t}_0 \equiv (t_{0_1}, t_{0_2}, ...t_{0_J})'$ in the initial conditions (3.3). Inference on the hyper-parameters in the IPPP is the key to overall predictive inference on where introductions are likely to occur in a broader spatial domain. We define the spatio-temporal IPPP as follows (Gelfand and Schliep, 2018):

$$p(J, \boldsymbol{\Omega}, \mathbf{t}_0|\Lambda) = \frac{e^{-\int_{\mathcal{S}}\int_{\mathcal{T}}\Lambda(\mathbf{s})dtd\mathbf{s}}(\int_{\mathcal{S}}\int_{\mathcal{T}}\Lambda(\mathbf{s},t)dtd\mathbf{s})^J}{J!}\prod_{j=1}^{J}\frac{\Lambda(\boldsymbol{\omega}_j, t_{0_j})}{\int_{\mathcal{S}}\int_{\mathcal{T}}\Lambda(\mathbf{s},t)dtd\mathbf{s}}, \tag{3.14}$$

where $\Lambda(\mathbf{s}, t)$ is the spatio-temporally varying intensity function and $\mathcal{T}$ is the temporal domain. The intensity may be specified as $\log(\Lambda(\mathbf{s}, t)) = \beta_0 + \mathbf{x}(\mathbf{s})'\boldsymbol{\beta}$, where $\mathbf{x}(\mathbf{s})$ is a vector of spatial covariates and $\boldsymbol{\beta}$ is a corresponding vector of regression parameters.

While inference on J, $\boldsymbol{\Omega}$, and $\mathbf{t}_0$ may be interesting, $p(\beta_0, \boldsymbol{\beta}|J, \boldsymbol{\Omega}, \mathbf{t}_0)$ and the posterior predictive distribution of the IPPP, $p(\tilde{J}, \tilde{\boldsymbol{\Omega}}, \tilde{\mathbf{t}}_0|J, \boldsymbol{\Omega}, \mathbf{t}_0)$, provide predictive inference on where introductions are likely to occur. If $\mathcal{S} \subset \tilde{\mathcal{S}}$, $\boldsymbol{\Omega} \subset \mathcal{S}$, and $\tilde{\boldsymbol{\Omega}} \subset \tilde{\mathcal{S}}$, we may draw from the posterior predictive distribution for the number, locations, and times of new pathogen introductions, $p(\tilde{J}, \tilde{\boldsymbol{\Omega}}, \tilde{\mathbf{t}}_0|J, \boldsymbol{\Omega}, \mathbf{t}_0)$ (Hooten and Hefley, 2019):

$$p(\tilde{J}, \tilde{\boldsymbol{\Omega}}, \tilde{\mathbf{t}}_0|J, \boldsymbol{\Omega}, \mathbf{t}_0) = \int_{\beta_0}\int_{\boldsymbol{\beta}} p(\tilde{J}, \tilde{\boldsymbol{\Omega}}, \tilde{\mathbf{t}}_0, \beta_0, \boldsymbol{\beta}|J, \boldsymbol{\Omega}, \mathbf{t}_0)d\boldsymbol{\beta}d\beta_0 \tag{3.15}$$

$$= \int_{\beta_0}\int_{\boldsymbol{\beta}} p(\tilde{J}, \tilde{\boldsymbol{\Omega}}, \tilde{\mathbf{t}}_0|\beta_0, \boldsymbol{\beta}, J, \boldsymbol{\Omega}, \mathbf{t}_0)p(\beta_0, \boldsymbol{\beta}|J, \boldsymbol{\Omega}, \mathbf{t}_0)d\boldsymbol{\beta}d\beta_0, \tag{3.16}$$

where $\tilde{J}$, $\tilde{\boldsymbol{\Omega}}$, and $\tilde{\mathbf{t}}_0$ are the number, locations, and times of pathogen introduction that are

likely to occur in the larger study area $\tilde{\mathcal{S}}$, given the spatio-temporal IPPP. Thus, inference on the spatio-temporal IPPP process, including from the posterior predictive distribution, may inform public health, agricultural, and wildlife management policy to reduce the risk of new pathogen introductions.

3.3.3 Model Fitting

BHMEDM

Ordinarily, one might fit the BHMEDM using an MCMC algorithm and use a Metropolis-Hastings algorithm to sample J, $\mathbf{\Omega}$, and $\mathbf{t}_0$. However, fitting the BHMEDM to obtain inference on J, $\mathbf{\Omega}$, and $\mathbf{t}_0$ is computationally difficult because J is a discrete, unknown parameter that presents a trans-dimensional estimation problem. That is, as J increases by one, the number of parameters that must be estimated increases by three ($t_{0_{J+1}}$ and the two coordinates in $\boldsymbol{\omega}_{J+1}$). Thus, as currently specified, tuning an MCMC algorithm to fit the BHMEDM is difficult, if not impossible. In the remainder of this section, we turn to ideas from the mixture model and missing date literature to overcome these trans-dimensionality and model tuning challenges. Specifically, we will define an over-parameterized finite mixture model of ecological diffusion processes (Rousseau and Mengersen, 2011), where the locations and times of pathogen introduction have been fixed. We will then rely on a form of indicator variable selection to eliminate superfluous pathogen introductions (Rousseau and Mengersen, 2011). Finally, we explain how to use Bayesian imputation to obtain inference from the spatio-temporal IPPP component of the model (Scharf et al., 2017).

We now define the over-parameterized mixture model. Let J^* to be an arbitrarily large integer such that $J < J^*$. In this context, J is the unknown but true number of introductions, as before. We note that computational feasibility may limit the size of J^*. Fixing J^* in this way results in an over-parameterized number of pathogen introductions, which we call pseudo-introductions (Rousseau and Mengersen, 2011). Pseudo-introductions are composed of the matrix of potential locations of introduction $\mathbf{\Omega}^* \equiv (\boldsymbol{\omega}_1^*, \boldsymbol{\omega}_2^*, ..., \boldsymbol{\omega}_{J^*}^*)'$ and the corresponding potential times of introduction $\mathbf{t}_0^* \equiv (t_{0_1}^*, t_{0_2}^*, ..., t_{0_{J^*}}^*)$. We will define how the

pseudo-introductions are selected shortly. With this modification, the mixture of component pathogen intensities in (3.1) may be re-expressed as:

$$u(\mathbf{s}_i, t_i) = \sum_{j=1}^{J^*} v_j u_j(\mathbf{s}_i, t_i), \tag{3.17}$$

where v_j is a binary indicator that denotes whether the j^{th} component pathogen intensity contributes to the total pathogen intensity. Thus, when $v_j = 1$, $u_j(\mathbf{s}_i, t_i)$ denotes the pathogen intensity at the i^{th} location and time that is contributed by the j^{th} pseudo-introduction. As before, the PDE for the j^{th} ecological diffusion model that provides $u_j(\mathbf{s}_i, t_i)$ is approximated by the analytical solution to the homogenized PDE in (3.7)

After J^* has been pre-chosen, we then use expert elicitation to select where and when pathogen introductions were likely to have occurred as well as stochastically define additional pseudo-introductions in the vicinity of the elicited locations and times. Alternatively, we may randomly draw the location and time of pseudo-introductions from a uniform space-time cube or polyhedron defined in the study area and temporal domain ($\mathcal{S} \times \mathcal{T}$). The extent of this space-time cube or polyhedron may be informed by the data or be defined through expert elicitation. After J^* pseudo-introductions are defined, we consider the selected $t^*_{0_j}$ and $\boldsymbol{\omega}^*_j$ values to be fixed for $j = 1, 2, ..., J^*$. We then rely upon the posterior inference from the binary indicators, $\mathbf{v} \equiv (v_1, v_2, ..., v_{J^*})'$ to "empty" the model of superfluous pseudo-introductions and provide a measure of uncertainty in the locations and times of the true pathogen introductions (Rousseau and Mengersen, 2011; Thompson et al., 2017). Thus, pseudo-introductions (the pairing of a pre-specified location and time) and the corresponding ecological diffusion processes are components of the BHMEDM that are then added or removed by indicator variables. We can now easily fit the BHMEDM in (3.8-3.9, 3.17, 3.11-3.13) using an MCMC algorithm (see Appendix C).

Inference about the number, locations, and times of pathogen introduction are available as derived quantities through the posterior distributions of v_j (i.e., $\mathbf{v}$) and the J^* pre-specified pseudo-introduction locations and times. We define $J^{**k} = \sum_{j=1}^{J*} v_j^k$ as the derived k^{th} posterior draw from the distribution of the indicator variables that represents inference on the

number of pathogen introductions. We define $\boldsymbol{\Omega}^{**k} \equiv (\boldsymbol{\omega}_1^{**k}, \boldsymbol{\omega}_2^{**k}, ..., \boldsymbol{\omega}_{J^{**k}}^{**k})'$ as the k^{th} derived draw from the distribution of the indicator variables and associated pseudo-introduction locations that represents inference on the locations of pathogen introduction. Likewise, we define $\mathbf{t}_0^{**k} \equiv (t_{0_1}^{**k}, t_{0_2}^{**k}, ..., t_{0_{J^{**k}}}^{**k})'$ as the corresponding k^{th} derived draw from the distribution of the indicator variables and associated pseudo-introduction times that represent inference on the distribution of the times of pathogen introduction. If the pseudo-introductions have been correctly specified, then $(J^{**k}, \boldsymbol{\Omega}^{**k}, \mathbf{t}_0^{**k})'$ represents a draw from $p(J, \boldsymbol{\Omega}, \mathbf{t}_0|\mathbf{y}, ...)$. However, if a subset of the pseudo-introductions are mis-specified, then at best $(J^{**k}, \boldsymbol{\Omega}^{**k}, \mathbf{t}_0^{**k})'$ represents a draw from a distribution that approximates $p(J, \boldsymbol{\Omega}, \mathbf{t}_0|\mathbf{y}, ...)$.

Understanding that the set of pseudo-introductions may be mis-specified, we employ Bayesian imputation and substitute J^{**k}, $\boldsymbol{\Omega}^{**k}$, and $\mathbf{t}_0^{**k}$ for J, $\boldsymbol{\Omega}$ and $\mathbf{t}_0$ in 3.14 within the k^{th} MCMC iteration to obtain inference on the parameters of the IPPP, β_0 and $\boldsymbol{\beta}$ (Scharf et al., 2017). Let $\underline{J}^{**}$, $\underline{\boldsymbol{\Omega}}^{**}$, and $\underline{\mathbf{t}}_0^{**}$ be the sets of J^{**k}, $\boldsymbol{\Omega}^{**k}$, and $\mathbf{t}_0^{**}$, respectively, for all k. We may then obtain draws from $p(\beta_0, \boldsymbol{\beta}|\underline{J}^{**}, \underline{\boldsymbol{\Omega}}^{**}, \underline{\mathbf{t}}_0^{**})$. If $\mathcal{S} \subset \tilde{\mathcal{S}}$, $\boldsymbol{\Omega} \subset \mathcal{S}$, and $\tilde{\boldsymbol{\Omega}} \subset \tilde{\mathcal{S}}$, we may draw from the posterior predictive distribution for new pathogen introductions, $p(\tilde{J}, \tilde{\boldsymbol{\Omega}}, \tilde{\mathbf{t}}_0|\underline{J}^{**}, \underline{\boldsymbol{\Omega}}^{**}, \underline{\mathbf{t}}_0^{**})$ using (3.15-3.16; Hooten and Hefley, 2019).

3.4 Wisconsin and Illinois Data Example

We illustrate the utility of our BHMEDM with an exploratory analysis of spatio-temporal disease surveillance data for chronic wasting disease (CWD) in white-tailed deer collected in southern Wisconsin and northern Illinois in the U.S. Our purpose is two-fold: first, to model the influence of spatial, ecologically relevant covariates on the diffusion and growth dynamics of CWD in white-tailed deer; second, to obtain inference on the spatio-temporal process for the pathogen introductions and predict where new introductions are likely to occur in the future. We hypothesized that cases of CWD in Wisconsin and Illinois within the first few years of surveillance data collection are the result of three to six separate pathogen introductions in the vicinity.

Chronic wasting disease (CWD) is an invariably fatal transmissible spongiform encephalopa-

thy that affects cervids (e.g., elk, deer). First discovered in captive deer populations in Colorado, U.S. in the 1960s, it has spread to at least 26 U.S. states and can be found in five additional countries (Rivera et al., 2019). The causative prion has been found to spread by contact between deer (including between carcasses and live individuals) via saliva, urine, feces, and blood, and has been found to persist in the environment on vegetation and soil (Rivera et al., 2019). Thus, transmission may occur directly between individuals or indirectly through the environment.

Modeling the disease dynamics of infectious diseases like CWD may be complicated for at least two reasons. First, the dynamic spread of CWD may be linked to both migrations and animal movement within the home range (often dependent on landscape characteristics). Second, modeling efforts should account for pathogen reservoirs in the environment (also possibly dependent on landscape characteristics; Rivera et al., 2019). For example, deer may be more likely to travel quickly through landscape that is open, such as pasture or cropland, and travel slowly or linger in forested areas. Additionally, there is evidence that certain land cover types and soil components act as reservoirs for the prion, and hence allow it to persist in the environment (Rivera et al., 2019). When the prion persists in the environment it can multiply because it may be picked up by an uninfected deer, replicate within that deer, and then return to the environment in greater numbers as it is shed by the deer.

Modeling the spatio-temporal process for novel pathogen introductions provides another layer of difficulty. For example, when an infected deer travels long distances to choose a new home range, certain ecologically related spatial covariates (e.g., forest and crop density, and water availability) may influence the choice of home range, and therefore the locale of the resulting novel pathogen introduction. Other spatial factors like distance to nearest deer farm or distance to nearest highway can also impact how humans might facilitate the novel introduction of the pathogen through transporting infected captive animals or discarding infected carcasses.

The Wisconsin and Illinois Departments of Natural Resources have collected disease surveillance data on CWD since 2001 and 2002, respectively. We defined a study area $\mathcal{S}$ in

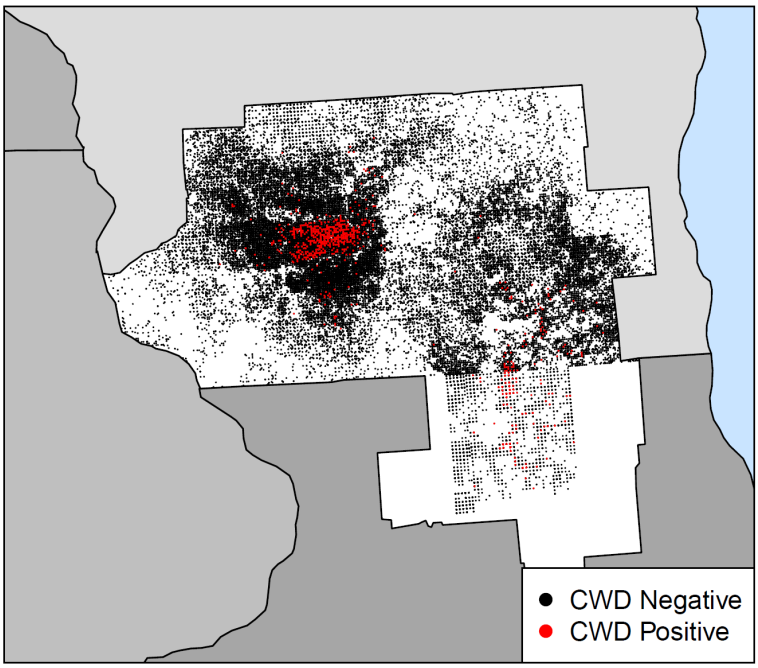

Figure 3.1: The study area in southern Wisconsin and northern Illinois in the U.S. with plotted locations of deer that tested positive for CWD (red) and negative for CWD (black) between 2001 and 2006.

southern Wisconsin and northern Illinois covering approximately 34,500 km^2 that contained the initial outbreaks of CWD in both states. We restricted our analysis to data that were collected from 2001 (when CWD was first discovered in Wisconsin) through 2006. Including observations from these years allowed us to capture the initial diffusion and growth dynamics of the early outbreaks in both states, while balancing computational considerations. In all, our analysis included 90,467 deer, of which 958 tested positive for CWD (see **Figure 3.1** for a plot of the study area and data).

As forest and human development land cover, or lack thereof, are thought to influence the diffusion of the pathogen, we included forest density and development density as spatial covariates in the diffusion component of the BHMEDM (Rivera et al., 2019). The forest and development density covariates were obtained from the 2001 National Land Cover Database (NLCD) by calculating the percentage of land within 300m $\times$300m grid cells classified as forest and developed by humans, respectively (Homer et al., 2007). We also included a spatial indicator variable (east vs. west) in the diffusion component of the BHMEDM that

differentiates between the east and west sides of the study area. We include this spatial indicator to check whether the east and west parts of the study area that contain CWD cases are influenced by slightly different diffusion and growth processes.

We included three spatial covariates related to soils across the study area in the growth component of the BHMEDM: percent clay, percent organic carbon content, and cation exchange capacity. Clay content has been shown to affect the persistence of pathogenic prions in the soil (Walter et al., 2011; Dorak et al., 2017; Rivera et al., 2019), while cation exchange capacity and organic carbon content of soils may also be of interest. The clay, organic carbon, and cation exchange covariates were obtained from the ISRIC SoilGrids database using the program QGIS (Poggio et al., 2021; QGIS Development Team, 2021). Following Hefley et al. (2017c), we included the forest and development density spatial covariates in the growth component of the BHMEDM. We also included the east vs west indicator variable. All spatial covariates in the BHMEDM, except for the east vs. west covariate, were centered and scaled. Then, all spatial covariates were homogenized from a resolution of 300m $\times$ 300m to a resolution of $4,500\text{m} \times 4,500\text{m}$ while fitting the BHMEDM (original resolution shown in **Figure 3.2**). Lastly, as sex of deer is associated with the susceptibility of individual deer to developing CWD, we included the sex of the deer in the susceptibility factor of the BHMEDM. We employ descriptive notation for the parameters in the BHMEDM to aid the reader (i.e., $\alpha_{forest}, \alpha_{development}, \alpha_{evw}, \gamma_{clay}, \gamma_{cec}, \gamma_{socc}, \gamma_{forest}, \gamma_{development}, \gamma_{evw}$, and τ_{sex}).

We must pre-specify potential locations and times of pseudo-introduction to fit the BHMEDM model. We drew the locations of forty pseudo-introductions from two separate bivariate uniform distributions defined on two regions of the study area that were identified by the data as having the majority of the CWD cases. We drew twenty pseudo-introductions for each region. We drew the times of pseudo-introduction in decimal years from a uniform distribution with lower and upper bounds of 1994.000 and 2004.000, respectively. The range of drawn dates for the pseudo introductions was (1994.003, 2003.670), in decimal years.

We specified priors for the parameters in the BHMEDM as follows: $q_j \sim \text{Beta}(0.5, 0.5)$, $(\alpha_0, \boldsymbol{\alpha})' \sim \text{MVN}(\mathbf{0}, 10^6\mathbf{I})$, $(\gamma_0, \boldsymbol{\gamma})' \sim \text{MVN}(\mathbf{0}, 10^6\mathbf{I})$, and $(\log(\theta), \tau_{sex})' \sim \text{MVN}((36, 0)', \left(\begin{smallmatrix} 2 & 0 \\ 0 & 1 \end{smallmatrix}\right))$. We fit the BHMEDM using **Algorithm 1** in Appendix C. We used Bayesian imputation

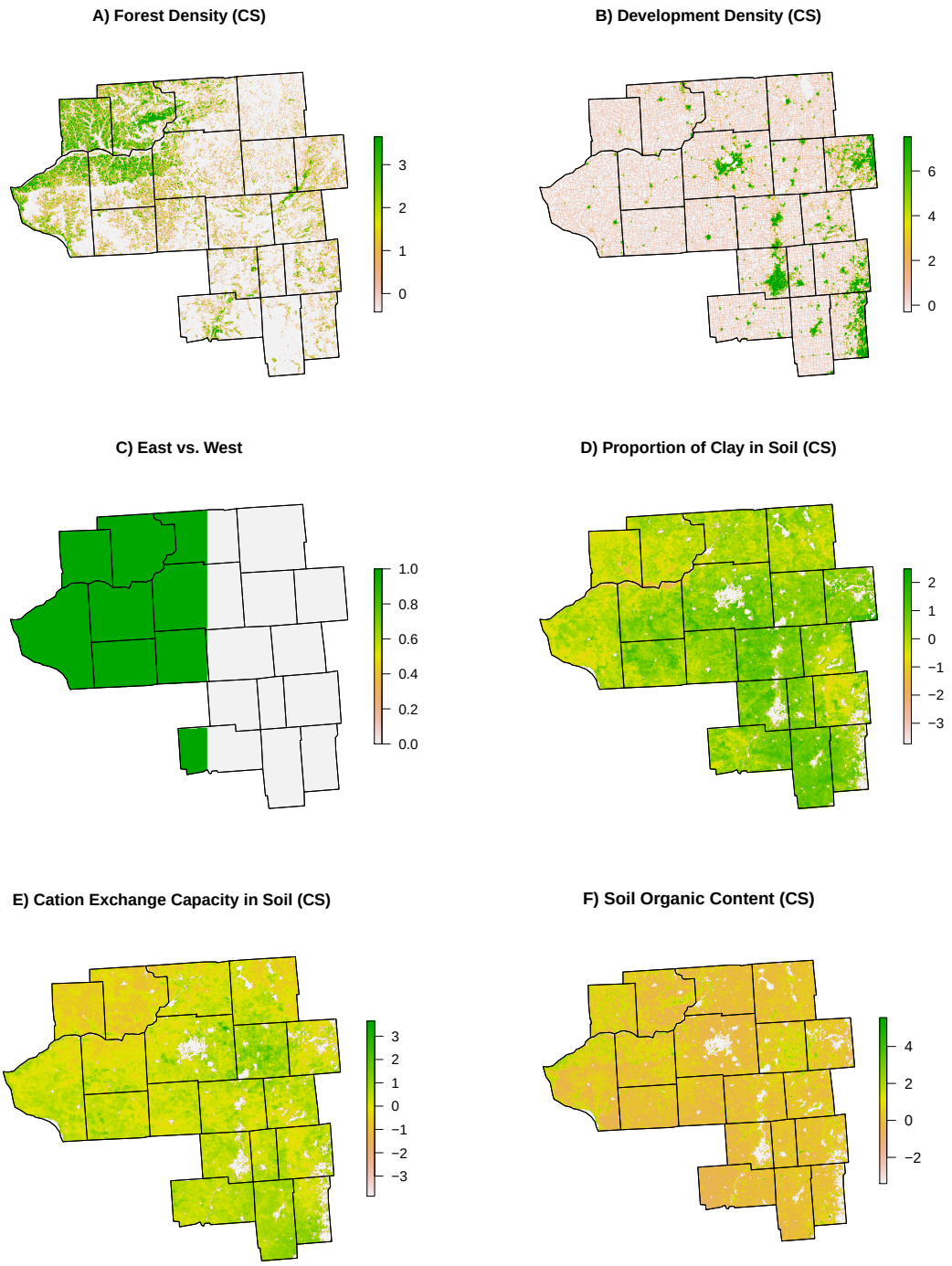

Figure 3.2: Centered and scaled (CS) spatial covariates used in the BHMEDM to obtain inference on the growth and diffusion of the causative prion across the study area. The forest density (A), development density (B), and east vs. west indicator (C) spatial covariates were included in the diffusion term of the BHMEDM. The clay density (D), cation exchange capacity density (E), and oil organic carbon content (F) spatial covariates were included in the growth term of the BHMEDM. The forest density, development density, and east vs. west indicator spatial covariates were also included in the growth term of the BHMEDM. The east vs. west indicator spatial covariate was not centered and scaled.

and an MCMC algorithm to fit the IPPP component of the model from (3.14) by substituting J^{**k}, $\mathbf{\Omega}^{**k}$, and $\mathbf{t}_0^{**k}$ for J, $\mathbf{\Omega}$, and $\mathbf{t}_0$ at the k^{th} iteration. We specified the vector $\mathbf{x}(\mathbf{s})$ in the intensity $\Lambda(\mathbf{s}, t)$ as the centered and scaled transformations of distance to nearest highway, water density, crop density, and forest density (all shown in **Figure 3.3**). Like the forest and development spatial covariates, the crop and water density spatial covariates were derived from the 2001 NLCD as the percentage of land cover within 300m $\times$300m grid cells classified as crops or water, respectively. We specified priors for the parameters in the IPPP as follows: $(\beta_0, \boldsymbol{\beta}) \sim \text{MVN}(\mathbf{0}, 10^6\mathbf{I})$. For both the BHMEDM and IPPP component, we employed 50,000 MCMC iterations and an adaptive Metropolis-Hastings algorithm to tune the proposal distributions (Roberts and Rosenthal, 2007; Roberts and Rosenthal, 2009). We discarded the first 8,000 MCMC iterations as burn-in. We calculated the predicted probability of CWD infection for female and male deer across the study area between 2001 and 2008. After employing Bayesian imputation to fit the IPPP component, we obtained 42,000 draws from the posterior predictive distribution of the IPPP across most of the northern Midwest U.S. (where $\tilde{\mathcal{S}}$ included much of Illinois, Iowa, Michigan, Minnesota, and Wisconsin). We subsequently obtained the mean number of introductions likely to occur in each county across the predicted region between January 1, 2004 and December 31, 2013.

3.5 Results

An increase in the forest density covariate was associated with a decrease in the diffusion rate of the pathogen. The development density and east vs. west (evw) covariates were not significant to the diffusion of the pathogen, according to the associated parameter 95% credible intervals. Increases in the percent cation exchange capacity of the soil and forest density covariates were associated with an increased growth rate. The clay concentration and organic carbon content of the soil covariates were found to not influence the growth rate of the pathogen, according to the associated parameter 95% credible intervals. Likewise, the development density and east vs. west indicator covariates did not influence the growth rate. We provide posterior summaries (mean values and credible intervals) for α_0, α_{forest}, $\alpha_{development}$,

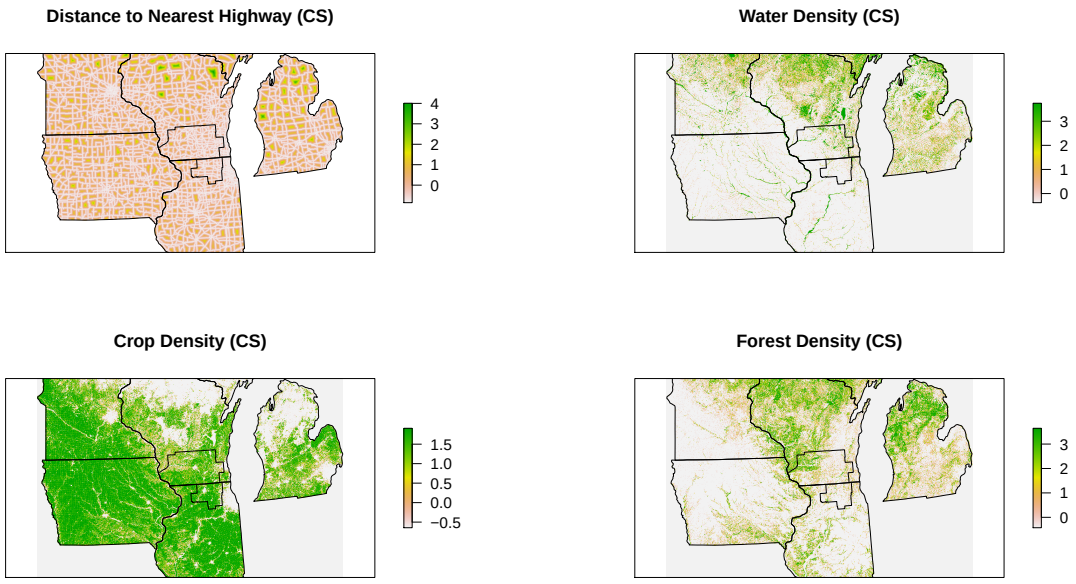

Figure 3.3: Centered and scaled (CS) spatial covariates of interest across Illinois, Iowa, Michigan, Minnesota, and Wisconsin. When fitting an IPPP to Bayesian imputed pathogen introductions, the spatial covariates were used from within the study area in southern Wisconsin and northern Illinois (outlined in black). When drawing from the posterior predictive distribution of the IPPP, the spatial covariates were used within the broader region of most of Illinois, Iowa, Michigan, Minnesota, and Wisconsin.

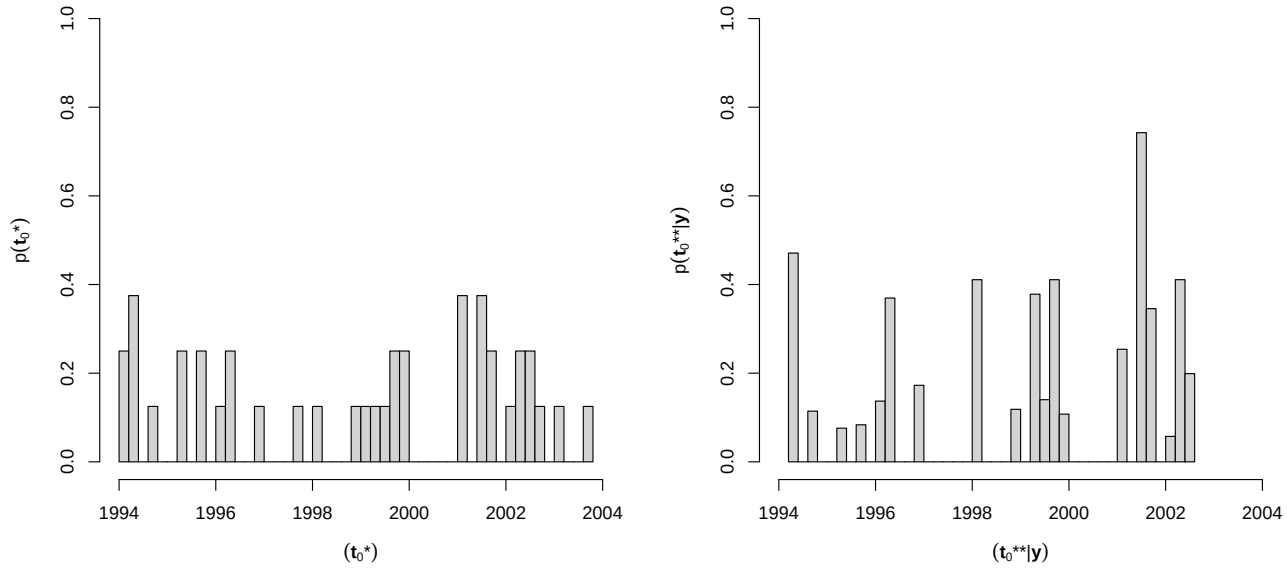

Figure 3.4: Histograms of the distribution of pre-specified pseudo-introduction times (left panel) and posterior-selected pseudo-introduction times (right panel) in decimal years, within the study area.

α_{evw}, γ_0, γ_{clay}, γ_{cec}, γ_{socc}, γ_{forest}, $\gamma_{development}$, γ_{evw}, $\log(\theta)$, and τ_{sex} from the BHMEDM in **Table 3.1**. The derived distribution of J^{**} had a mean of 12.1 introductions and a 95% credible interval from 10 to 15. The times associated with the posterior-selected pseudo-introductions ranged between 1994.347 (May 7, 1994) and 2002.538 (July 16, 2002). We provide a comparison of the distribution of pre-specified times of introduction and posterior-selected times of introduction in **Figure 3.4**. Plots that compare the distribution of pre-specified locations of pseudo-introduction with that of the posterior-selected locations of pseudo-introduction are provided in **Figure 3.5**. The results of calculating the predicted probability of CWD infection for female and male deer across the study area from 2001 to 2008 are provided in **Figure 3.6**.

We provide posterior summaries (mean values and credible intervals) for β_0, $\beta_{highway}$, β_{water}, β_{crop}, and β_{forest} from the IPPP in **Table 3.1**. While the 95% credible intervals for the slope parameters all included zero (except for the water density covariate), the placement and skew of each posterior distribution, relative to zero, provided information that was

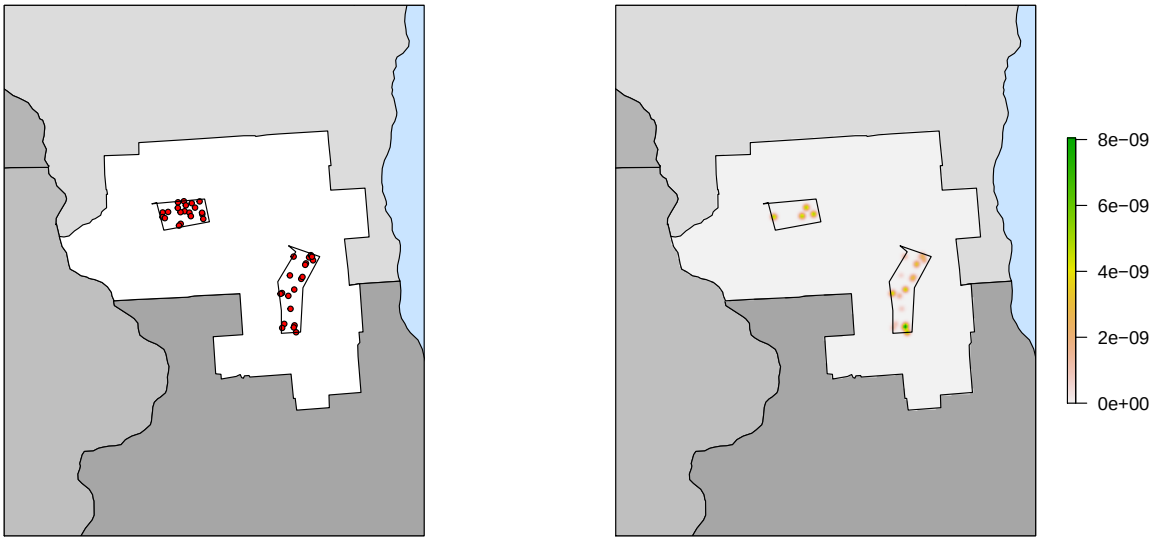

Figure 3.5: Pre-specified pseudo-introduction locations (left panel) and posterior-selected pseudo-introduction locations (right panel) within the study area in southern Wisconsin and northern Illinois. The two regions of the study area are outlined in black that were identified by the data as containing the majority of the CWD cases. Twenty pseudo-introductions were randomly drawn from each region.

evident via the posterior predictive distribution summaries. We obtained 42,000 draws from the posterior predictive distribution of the IPPP across most of the northern Midwest U.S. (Illinois, Iowa, Michigan, Minnesota, and Wisconsin) and calculated the mean number of introductions likely to occur in each county across the region between January 1, 2004 and December 31, 2013 (see **Figure 3.7**).

3.6 Discussion

In this chapter, we introduced a BHMEDM that included an IPPP component and applied the BHMEDM in an exploratory analysis of CWD surveillance data. With the BHMEDM, we obtained inference on the diffusion and growth dynamics for the causative pathogen of an infectious disease. We also obtain Bayesian inference on the number, locations, and times that a pathogen was introduced into a population. From the IPPP component of the model, we obtained inference on the spatio-temporal process for the number and locations

Table 3.1: Posterior summaries for parameters in the Bayesian hierarchical mixture of ecological diffusion models (BHMEDM) and inhomogeneous Poisson point process (IPPP) component (obtained using Bayesian imputation).

			Quantile		
Model	Parameter	Mean	2.5%	50%	97.5%
BHMEDM	α_0	17.455	17.216	17.461	17.667
	α_{forest}	-0.028	-0.055	-0.028	-0.002
	$\alpha_{development}$	0.012	-0.051	0.011	0.080
	α_{evw}	0.064	-0.140	0.064	0.268
	γ_0	0.154	0.099	0.156	0.205
	γ_{clay}	-0.0003	-0.015	-0.0003	0.015
	γ_{cec}	0.027	0.004	0.027	0.051
	γ_{socc}	-0.017	-0.037	-0.017	0.002
	γ_{forest}	0.033	0.021	0.032	0.045
	$\gamma_{development}$	0.004	-0.018	0.004	0.026
	γ_{evw}	0.009	-0.041	0.007	0.066
	$\log(\theta)$	34.887	34.593	34.904	35.110
	τ_{sex}	0.212	0.160	0.212	0.266
IPPP	β_0	-70.368	-510.229	-11.401	2.576
	$\beta_{highway}$	-0.313	-3.293	-0.212	1.993
	β_{water}	-190.639	-1352.265	-35.488	-0.326
	β_{crop}	0.379	-1.129	0.246	2.928
	β_{forest}	0.033	-1.389	-0.012	1.743

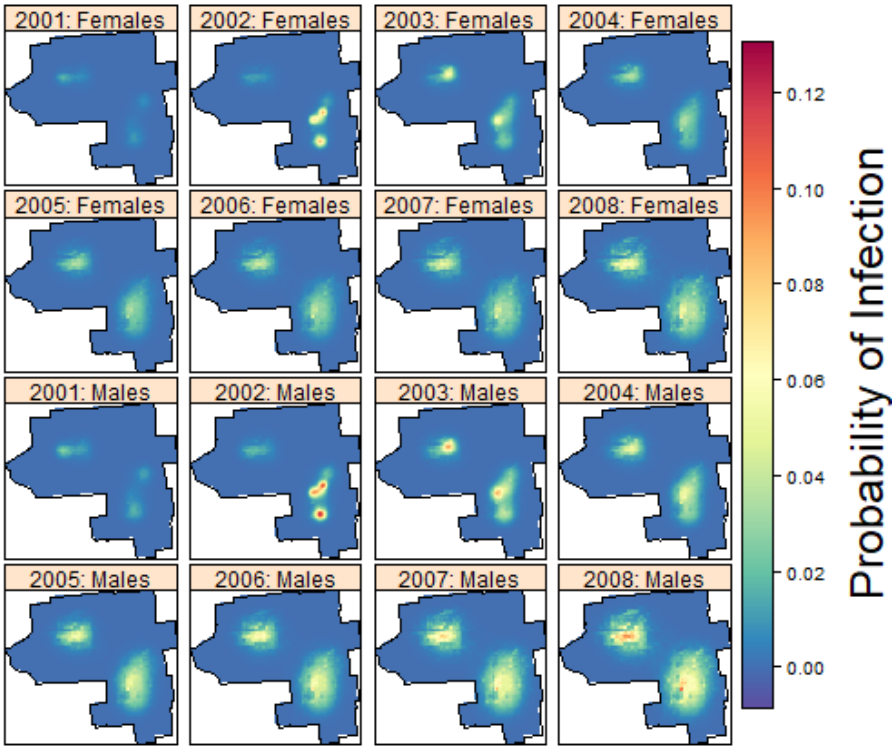

Figure 3.6: Plots of the predicted probability of infection for CWD broken down by sex of deer, across the study area from 2001 - 2008. The predicted probabilities of infection from 2007 - 2008 are forecasts because we only included disease surveillance data from 2001 - 2006 when fitting the BHMEDM.

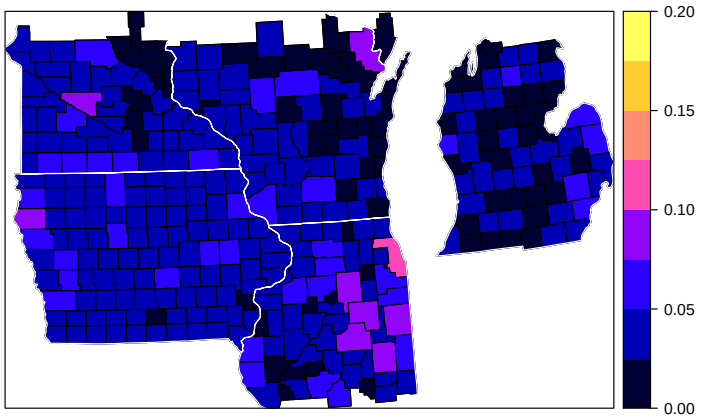

Figure 3.7: Plot of the mean number of pathogen introductions expected within each county across most of the northern Midwest U.S. (Illinois, Iowa, Michigan, Minnesota, and Wisconsin) in the time between January 1, 2004 and December 31, 2013, given the number, locations, and times of selected pseudo-introductions in the study area (southern Wisconsin and northern Illinois).

of pathogen introduction. We also used the IPPP to obtain predictive inference for which counties in the upper mid-western U.S. may have an increased risk for novel introductions. Like the ensemble model in Chapter 2 of this dissertation, the BHMEDM may be modified to accommodate any type of aggregated binary data from Chapter 1.

Because J^* must be finite, the uncertainty related to inference from the posterior distribution of $\mathbf{v}$ and the associated set of selected pseudo-introductions J^{**}, $\underline{\boldsymbol{\Omega}}^{**}$, and $\mathbf{t}_0^{**}$ is limited by the number, locations, and times of the pre-specified pseudo-introductions. That is, the degree of uncertainty in the locations and times of pathogen introductions is dependent on how the locations and times of pseudo-introductions were specified and the distance between each pseudo-introduction in space and time. Verity et al. (2014) and Thompson et al. (2017) sought to resolve this problem by specifying a Dirichlet process model with an infinite number of locations and times of introduction. While instructive, the Dirichlet process model is impractical for the large number of observations in the CWD data set. Thus, we address the behavior of the BHMEDM using a heuristic related to spatial infill asymptotics and using ideas from over-fitted or over-parameterized mixture models.

Infill asymptotics are concerned with the consistency of parameter estimates within a spatio-temporal model, particularly as the number of observations increases and the distance

between observations in space and time decreases in a fixed study area (see Stein, 2010, for an overview of spatial asymptotics). A heuristic, related to spatio-temporal infill of pseudo-introductions, is that we may conceivably obtain more valid inference on the spatio-temporal IPPP component of the model as J^* increases and the spatio-temporal distance between pseudo-introductions decreases to zero.

In infinite mixture model theory, if an infinitely dense grid of pseudo-introductions is specified, the selected pseudo-introductions create a type of space-time cloud around the true introduction locations and times and provide a measure of uncertainty regarding the true locations and times of introduction (Thompson et al., 2017). In the case of over-fitted or over-parameterized mixture models, where $J^* > J$, Neal (2000) acknowledged that properly eliminating superfluous model components can be a challenging technical problem. However, Rousseau and Mengersen (2011) showed that certain prior specifications will asymptotically empty the model of superfluous pseudo-introductions.

Several features of the BHMEDM are important to note. First, inference from the IPPP, including from the posterior predictive distribution of introductions, is highly dependent upon the locations of the pseudo-introductions that were pre-specified for the BHMEDM. This dependence is reduced if the resolution is low for the spatial covariates in the IPPP. One way to assess this dependence is to fit the BHMEDM multiple times with separate sets of pre-specified pseudo-introductions and compare inference from the IPPP among the model fits. Additionally, one may perform Bayesian model averaging across the separate BHMEDM fits to account for the uncertainty in the process of pre-specifying pseudo-introductions. On the other hand, the practitioner may assess how the selected pseudo-introductions affect inference from the IPPP by comparing inference from the IPPP fit to the selected pseudo-introductions (J^{**}, $\mathbf{\Omega}^{**}$, and $\mathbf{t}_0^{**}$), versus inference from the IPPP fit to the set of pre-specified pseudo-introductions (J^*, $\mathbf{\Omega}^*$, and $\mathbf{t}_0^*$).

Second, we note that the spatial covariate effect sizes from the diffusion and growth components of the BHMEDM tend to attenuate as more pseudo introductions are pre-specified in the mixture. This phenomenon is an artifact of decreasing the spatio-temporal distance between pseudo-introductions as J^* increases. Specifically, as more pseudo-introductions

are selected from a neighborhood, the diffusion and growth terms of the BHMEDM are less integral to properly fitting the data.

Third, practitioners should carefully choose the hyper-parameters on the beta hyper-prior for the Bernoulli distribution of v_j. In our disease surveillance data example we chose hyper-parameters equal to 0.5 to produce a Jeffreys hyper-prior. A hyper-prior that tends to move $P(v_j = 1)$ close to zero for $j = 1, ..., J^*$ should properly empty pseudo-introductions from the BHMEDM (Rousseau and Mengersen, 2011).

Acknowledgements

We thank the Illinois and Wisconsin Departments of Natural Resources for obtaining deer tissue samples and the hunters in Illinois and Wisconsin who provided them. We acknowledge support for this research from USGS G18AC00317 and G16AC00413. Any use of trade, firm, or product names is for descriptive purposes only and does not imply endorsement by the U.S. Government.

Bibliography

Adivar, B. and Selin Selen, E. (2013). Review of research studies on population specific epidemic disasters. *Disaster Prevention and Management*, 22(3):243–264.

Aguirre, A., Ostfeld, R., and Daszak, P., editors (2012). *New Directions in Conservation Medicine: Applied Cases of Ecological Health*. Oxford University Press, Inc, New York.

Albert, A. and Anderson, J. A. (1984). On the existence of maximum likelihood estimates in logistic regression models. *Biometrika*, 71(1):1–10.

Artois, M., Bengis, R., Delahay, R. J., Duchêne, M., Duff, J. P., Ferroglio, E., Gortazar, C., Hutchings, M. R., Kock, R. A., Leighton, F. A., Mrner, T., and Smith, G. C. (2009). Wildlife disease surveillance and monitoring. In Delahay, R. J., Smith, G. C., and Hutchings, M. R., editors, *Management of Disease in Wild Mammals*, pages 187–213. Springer, Tokyo, New York.

Ballmann, A., Russell, R., Walsh, D., Walker, N., and Hefley, T. (2021). *Pseudogymnoascus destructans* detections by US county (2008-2012). Web. https://doi.org/10.5066/P9XUPDIB.

Banks, D. L. and Hooten, M. B. (2021). Statistical challenges in agent-based modeling. *The American Statistician*, 75(3):235–242.

Barratt, A. S., Rich, K. M., Eze, J. I., Porphyre, T., Gunn, G. J., and Stott, A. W. (2019). Framework for estimating indirect costs in animal health using time series analysis. *Frontiers in Veterinary Science*, 6:190.

Blehert, D. S., Hicks, A. C., Behr, M., Meteyer, C. U., Berlowski-Zier, B. M., Buckles, E. L., Coleman, J. T. H., Darling, S. R., Gargas, A., Niver, R., Okoniewski, J. C., Rudd,

R. J., and Stone, W. B. (2009). Bat white-nose syndrome: An emerging fungal pathogen? *Science*, 323(5911):227.

Borgan, Ø., Breslow, N. E., Chatterjee, N., Gail, M. H., Scott, A., and Wild, C. J. (2018). *Handbook of Statistical Methods for Case-Control Studies*. CRC Press, Taylor & Francis Group, Boca Raton, Florida.

Box, G. and Draper, N. (1987). *Empirical Model-Building and Response Surfaces*. Wiley, New York.

Bradley, J. R., Wikle, C. K., and Holan, S. H. (2016). Bayesian spatial change of support for count-valued survey data with application to the American community survey. *Journal of the American Statistical Association*, 111(514):472–487.

Budd, J., Miller, B. S., Manning, E. M., Lampos, V., Zhuang, M., Edelstein, M., Rees, G., Emery, V. C., Stevens, M. M., Keegan, N., Short, M. J., Pillay, D., Manley, E., Cox, I. J., Heymann, D., Johnson, A. M., and McKendry, R. A. (2020). Digital technologies in the public-health response to COVID-19. *Nature Medicine*, 26:1183–1192.

Burnham, K. P. and Anderson, D. R. (2002). *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*. Springer-Verlag, New York.

Centers for Disease Control and Prevention (2012). Principles of epidemiology in public health practice: An introduction to applied epidemiology and biostatistics. Print, Web. https://www.cdc.gov/csels/dsepd/ss1978/SS1978.pdf.

Chang, X., Waagepetersen, R., Yu, H., Ma, X., Holford, T. R., Wang, R., and Guan, Y. (2015). Disease risk estimation by combining case-control data with aggregated information on the population at risk. *Biometrics*, 71(1):114–121.

Chave, S. P. W. (1958). Henry Whitehead and cholera in Broad Street. *Medical History*, 2(2):92–108.

Clinton, H. R. (2006). *It Takes a Village*. Simon & Schuster, New York.

Cost, K. T., Crosbie, J., Anagnostou, E., Birken, C. S., Charach, A., Monga, S., Kelley, E., Nicolson, R., Maguire, J. L., Burton, C. L., Schachar, R. J., Arnold, P. D., and Korczak, D. J. (2021). Mostly worse, occasionally better: Impact of COVID-19 pandemic on the mental health of Canadian children and adolescents. *European Child & Adolescent Psychiatry*, pages 1–14.

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

Diggle, P. J. and Giorgi, E. (2019). *Model-based Geostatistics for Global Public Health*. CRC Press, Taylor & Francis Group, Boca Raton, Florida.

Diggle, P. J., Guan, Y., Hart, A. C., Paize, F., and Stanton, M. (2010a). Estimating individual-level risk in spatial epidemiology using spatially aggregated information on the population at risk. *Journal of the American Statistical Association*, 105(492):1394–1402.

Diggle, P. J., Menezes, R., and Su, T. (2010b). Geostatistical inference under preferential sampling. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 59(2):191–232.

Diggle, P. J., Tawn, J. A., and Moyeed, R. A. (1998). Model-based geostatistics. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 47(3):299–350.

Dorak, S. J., Green, M. L., Wander, M. M., Ruiz, M. O., Buhnerkempe, M. G., Tian, T., Novakofski, J. E., and Mateus-Pinilla, N. E. (2017). Clay content and pH: Soil characteristic associations with the persistent presence of chronic wasting disease in northern Illinois. *Scientific Reports*, 7:18062.

Drees, K. P., Lorch, J. M., Puechmaille, S. J., Parise, K. L., Wibbelt, G., Hoyt, J. R., Sun, K., Jargalsaikhan, A., Dalannast, M., Palmer, J. M., Lindner, D. L., Kilpatrick, A. M., Pearson, T., Keim, P. S., Blehert, D. S., and Foster, J. T. (2017). Phylogenetics of a fungal invasion: Origins and widespread dispersal of white-nose syndrome. *mBio*, 8(6):e01941–17.

Egan, J. R. and Hall, I. M. (2015). A review of back-calculation techniques and their potential to inform mitigation strategies with application to non-transmissible acute infectious diseases. *Journal of The Royal Society Interface*, 12(106):20150096.

Erickson, D., Reeling, C., and Lee, J. G. (2019). The effect of chronic wasting disease on resident deer hunting permit demand in Wisconsin. *Animals*, 9(12):1096.

Fan, V. Y., Jamison, D. T., and Summers, L. H. (2018). Pandemic risk: How large are the expected losses? *Bulletin of the World Health Organization*, 96(2):129–134.

Fèvre, E. M., Bronsvoort, B. M., Hamilton, K. A., and Cleaveland, S. (2006). Animal movements and the spread of infectious diseases. *Trends in Microbiology*, 14(3):125–131.

Firth, D. (1993). Bias reduction of maximum likelihood estimates. *Biometrika*, 80(1):27–38.

Fox, J., Weisberg, S., Price, B., Adler, D., Bates, D., Baud-Bovy, G., Bolker, B., Ellison, S., Firth, D., Friendly, M., Gorjanc, G., Graves, S., Heiberger, R., Krivitsky, P., Laboissiere, R., Maechler, M., Monette, G., Murdoch, D., Nilsson, H., Ogle, D., Ripley, B., Venables, W., Walker, S., Winsemius, D., Zeileis, A., and R-Core (2020). car: Companion to applied regression (version 3.0-8). Web. https://CRAN.R-project.org/package=car.

Frick, W. F., Pollock, J. F., Hicks, A. C., Langwig, K. E., Reynolds, D. S., Turner, G. G., Butchkoski, C. M., and Kunz, T. H. (2010). An emerging disease causes regional population collapse of a common North American bat species. *Science*, 329(5992):679–682.

Garcia-Abreu, A., Halperin, W., and Daniel, I. (2002). *Public Health Surveillance Toolkit: A guide for busy task managers.* World Bank, Washington, DC.

Garlick, M. J., Powell, J. A., Hooten, M. B., and MacFarlane, L. R. (2014). Homogenization, sex, and differential motility predict spread of chronic wasting disease in mule deer in southern Utah. *Journal of Mathematical Biology*, 69(2):369–399.

Garlick, M. J., Powell, J. A., Hooten, M. B., and McFarlane, L. R. (2011). Homogenization

of large-scale movement models in ecology. *Bulletin of Mathematical Biology*, 73(9):2088–2108.

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

Gelfand, A. E. and Shirota, S. (2019). Preferential sampling for presence/absence data and for fusion of presence/absence data with presence-only data. *Ecological Monographs*, 89(3):e01372.

Giuntella, O., Hyde, K., Saccardo, S., and Sadoff, S. (2021). Lifestyle and mental health disruptions during COVID-19. *Proceedings of the National Academy of Sciences*, 118(9):e2016632118.

Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378.

Gottwald, T., Luo, W., Posny, D., Riley, T., and Louws, F. (2019). A probabilistic census-travel model to predict introduction sites of exotic plant, animal and human pathogens. *Philosophical Transactions of the Royal Society B*, 374(1776):20180260.

Gotway, C. A. and Young, L. J. (2002). Combining incompatible spatial data. *Journal of the American Statistical Association*, 97(458):632–648.

Graham, C. (2020). The human costs of the pandemic: Is it time to prioritize well-being? Web. https://www.brookings.edu/research/the-human-costs-of-the-pandemic-is-it-time-to-prioritize-well-being/.

Heaton, M. J., Berrett, C., Pugh, S., Evans, A., and Sloan, C. (2020). Modeling bronchiolitis incidence proportions in the presence of spatio-temporal uncertainty. *Journal of the American Statistical Association*, 115(529):66–78.

Hefley, T. J., Brost, B. M., and Hooten, M. B. (2017a). Bias correction of bounded location errors in presence-only data. *Methods in Ecology and Evolution*, 8(11):1566–1573.

Hefley, T. J. and Hooten, M. B. (2015). On the existence of maximum likelihood estimates for presence-only data. *Methods in Ecology and Evolution*, 6(6):648–655.

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

Hefley, T. J., Hooten, M. B., Russell, R. E., Walsh, D. P., and Powell, J. A. (2017c). When mechanism matters: Bayesian forecasting using models of ecological diffusion. *Ecology Letters*, 20(5):640–650.

Hefley, T. J., Russell, R. E., Ballmann, A. E., and Zhang, H. (2020). When and where: Estimating the date and location of introduction for exotic pests and pathogens. Web. https://arxiv.org/pdf/2006.16982v1.

Heisey, D. M., Osnas, E. E., Cross, P. C., Joly, D. O., Langenberg, J. A., and Miller, M. W. (2010). Linking process to pattern: estimating spatiotemporal dynamics of a wildlife epidemic from cross-sectional data. *Ecological Monographs*, 80(2):221–240.

Higdon, D. (2002). Space and space-time modeling using process convolutions. In Anderson, C. W., Barnett, V., Chatwin, P. C., and El-Shaarawi, A. H., editors, *Quantitative Methods for Current Environmental Issues*, pages 37–56. Springer, London.

Hillis, S. D., Unwin, H. J. T., Chen, Y., Cluver, L., Sherr, L., Goldman, P. S., Ratmann, O., Donnelly, C. A., Bhatt, S., Villaveces, A., Butchart, A., Bachman, G., Rawlings, L., Green, P., Nelson, C. A., and Flaxman, S. (2021). Global minimum estimates of children affected by COVID-19-associated orphanhood and deaths of caregivers: A modelling study. *The Lancet*, 398(10298):391–402.

Homer, C. G., Dewitz, J. A., Fry, J. A., Coan, M. J., Hossain, S. M. N., Larson, C. R., Herold, N., McKerrow, A., Van Driel, J. N., and Wickham, J. (2007). Completion of the 2001 National Land Cover Database for the conterminous United States. *Photogrammetric Engineering and Remote Sensing*, 73(4):337–341.

Homer, C. G., Dewitz, J. A., Yang, L., Jin, S., Danielson, P., Xian, G., Coulston, J., Herold, N., Wickham, J., and Megown, K. (2015). Completion of the 2011 National Land Cover Database for the conterminous United States - representing a decade of land cover change information. *Photogrammetric Engineering and Remote Sensing*, 81(5):345–354.

Hooten, M. B., Garlick, M. J., and Powell, J. A. (2013). Computationally efficient statistical differential equation modeling using homogenization. *Journal of Agricultural, Biological, and Environmental Statistics*, 18(3):405–428.

Hooten, M. B. and Hefley, T. (2019). *Bringing Bayesian Models to Life*. CRC Press, Taylor & Francis Group, Boca Raton, Florida.

Ibrahim, N. K. (2020). Epidemiologic surveillance for controlling Covid-19 pandemic: Types, challenges and implications. *Journal of Infection and Public Health*, 13(11):1630–1638.

Ingersoll, T. E., Sewall, B. J., and Amelon, S. K. (2016). Effects of white-nose syndrome on regional population patterns of 3 hibernating bat species. *Conservation Biology*, 30(5):1048–1059.

Institute of Medicine (2007). *Global Infectious Disease Surveillance and Detection: Assessing the Challenges - Finding Solutions: Workshop Summary*. National Academies Press, Washington D.C.

Jachowski, D. S., Johnson, J. B., Dobony, C. A., Edwards, J. W., and Ford, W. M. (2014). Space use and resource selection by foraging Indiana bats at the northern edge of their distribution. *Endangered Species Research*, 24(2):149–157.

Johnson, O., Diggle, P., and Giorgi, E. (2019). A spatially discrete approximation to log-Gaussian Cox processes for modelling aggregated disease count data. *Statistics in Medicine*, 38(24):4871–4887.

Karpman, D., Ferreira, M. A., and Wikle, C. K. (2013). A point process model for tornado report climatology. *Stat*, 2(1):1–8.

Kissler, S. M., Gog, J. R., Viboud, C., Charu, V., Bjørnstad, O. N., Simonsen, L., and Grenfell, B. T. (2019). Geographic transmission hubs of the 2009 influenza pandemic in the United States. *Epidemics*, 26:86–94.

Kraemer, M. U. G., Golding, N., Bisanzio, D., Bhatt, S., Pigott, D. M., Ray, S. E., Brady, O. J., Brownstein, J. S., Faria, N. R., Cummings, D. A. T., Pybus, O. G., Smith, D. L., Tatem, A. J., Hay, S. I., and Reiner, R. C. (2019). Utilizing general human movement models to predict the spread of emerging infectious diseases in resource poor settings. *Scientific Reports*, 9(1):5151.

Kronqvist, J., Bernal, D. E., Lundell, A., and Grossmann, I. E. (2019). A review and comparison of solvers for convex MINLP. *Optimization and Engineering*, 20(2):397–455.

LeComber, S. C., Rossmo, D. K., Hassan, A. N., Fuller, D. O., and Beier, J. C. (2011). Geographic profiling as a novel spatial tool for targeting infectious disease control. *International Journal of Health Geographics*, 10(1):35.

Lee, L. M., Teutsch, S. M., Thacker, S. B., and St. Louis, M. E., editors (2010). *Principles & Practice of Public Health Surveillance*. Oxford University Press, New York.

Legrand, J., Egan, J. R., Hall, I. M., Cauchemez, S., Leach, S., and Ferguson, N. M. (2009). Estimating the location and spatial extent of a covert anthrax release. *PLoS Computational Biology*, 5(1):e1000356.

Levy, M. Z., Small, D. S., Vilhena, D. A., Bowman, N. M., Kawai, V., Cornejo del Carpio, J. G., Cordova-Benzaquen, E., Gilman, R. H., Bern, C., and Plotkin, J. B. (2011). Retracing micro-epidemics of chagas disease using epicenter regression. *PLoS Computational Biology*, 7(9):e1002146.

Logan, J. D. (2014). *Applied Partial Differential Equations*. Springer International Publishing, New York.

Lorch, J. M., Gargas, A., Meteyer, C. U., Berlowski-Zier, B. M., Green, D. E., Shearn-Bochsler, V., Thomas, N. J., and Blehert, D. S. (2010). Rapid polymerase chain reaction

diagnosis of white-nose syndrome in bats. *Journal of Veterinary Diagnostic Investigation*, 22(2):224–230.

Lu, X., Williams, P. J., Hooten, M. B., Powell, J. A., Womble, J. N., and Bower, M. R. (2020). Nonlinear reaction-diffusion process models improve inference for population dynamics. *Environmetrics*, 31(3):e2604.

Madigan, D. and Raftery, A. E. (1994). Model selection and accounting for model uncertainty in graphical models using Occam's window. *Journal of the American Statistical Association*, 89(428):1535–1546.

Madigan, D. and York, J. (1995). Bayesian graphical models for discrete data. *International Statistical Review*, 63(2):215–232.

Marchand, E. and Strawderman, W. E. (2004). Estimation in restricted parameter spaces: A review. *Lecture Notes - Monograph Series*, 45:21–44.

Medellin, R. A., Wiederholt, R., and Lopez-Hoffman, L. (2017). Conservation relevance of bat caves for biodiversity and ecosystem services. *Biological Conservation*, 211:45–50.

Meteyer, C. U., Buckles, E. L., Blehert, D. S., Hicks, A. C., Green, D. E., Shearn-Bochsler, V., Thomas, N. J., Gargas, A., and Behr, M. J. (2009). Histopathologic criteria to confirm white-nose syndrome in bats. *Journal of Veterinary Diagnostic Investigation*, 21(4):411–414.

M'ikanatha, N. M., Lynfield, R., Van Beneden, C. A., and de Valk, H., editors (2013). *Infectious Disease Surveillance*. Wiley-Blackwell, Chicester, England.

Mohler, G. O. and Short, M. B. (2012). Geographic profiling from kinetic models of criminal behavior. *SIAM Journal on Applied Mathematics*, 72(1):163–180.

Narrod, C., Zinsstag, J., and Tiongco, M. (2012). A one health framework for estimating the economic costs of zoonotic diseases on society. *EcoHealth*, 9(2):150–162.

National Research Council (2009). *Letter Report on the Review of the Food Safety and Inspection Service Proposed Risk-Based Approach to and Application of Public-Health Attribution*. National Academies Press, Washington, D.C.

Neal, R. M. (2000). Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, 9(2):249–265.

Newton, M. A., Polson, N. G., and Xu, J. (2021). Weighted Bayesian bootstrap for scalable posterior distributions. *Canadian Journal of Statistics*, 49(2):421–437.

Newton, M. A. and Raftery, A. E. (1994). Approximate Bayesian inference with the weighted likelihood bootstrap. *Journal of the Royal Statistical Society: Series B (Methodological)*, 56(1):3–26.

Nocedal, J. and Wright, S. J. (2006). *Numerical Optimization*. Springer-Verlag, New York.

Norderud, E. D., Powell, S. L., and Peterson, R. K. D. (2021). Risk assessment for the establishment of *Vespa mandarinia* (Hymenoptera: Vespidae) in the Pacific Northwest, United States. *Journal of Insect Science*, 21(4):1–14.

Oleson, J. J. and Wikle, C. K. (2013). Predicting infectious disease outbreak risk via migratory waterfowl vectors. *Journal of Applied Statistics*, 40(3):656–673.

Panchal, N., Kamal, R., Cox, C., and Garfield, R. (2021). The implications of COVID-19 for mental health and substance use. Kaiser Family Foundation. Web. https://www.kff.org/coronavirus-covid-19/issue-brief/the-implications-of-covid-19-for-mental-health-and-substance-use/.

Piantadosi, S., Byar, D. P., and Green, S. B. (1988). The ecological fallacy. *American Journal of Epidemiology*, 127(5):893–904.

Pielou, E. C. (1969). *An Introduction to Mathematical Ecology*. Wiley-Interscience, New York.

Poggio, L., de Sousa, L. M., Batjes, N. H., Heuvelink, G. B. M., Kempen, B., Ribeiro, E., and Rossiter, D. (2021). SoilGrids 2.0: Producing soil information for the globe with quantified spatial uncertainty. *SOIL*, 7(1):217–240.

QGIS Development Team (2021). QGIS geographic information system. Web. http://qgis.osgeo.org.

R Core Team (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rivera, N. A., Brandt, A. L., Novakofski, J. E., and Mateus-Pinilla, N. E. (2019). Chronic wasting disease in cervids: Prevalence, impact and management strategies. *Veterinary Medicine (Auckland, N. Z.)*, 10:123–139.

Roberts, G. O. and Rosenthal, J. S. (2007). Coupling and ergodicity of adaptive Markov chain Monte Carlo algorithms. *Journal of Applied Probability*, 44(2):458–475.

Roberts, G. O. and Rosenthal, J. S. (2009). Examples of adaptive MCMC. *Journal of Computational and Graphical Statistics*, 18(2):349–367.

Rohr, J. R., Barrett, C. B., Civitello, D. J., Craft, M. E., Delius, B., DeLeo, G. A., Hudson, P. J., Jouanard, N., Nguyen, K. H., Ostfeld, R. S., Remais, J. V., Riveau, G., Sokolow, S. H., and Tilman, D. (2019). Emerging human infectious diseases and the links to global food production. *Nature Sustainability*, 2(6):445–456.

Rousseau, J. and Mengersen, K. (2011). Asymptotic behaviour of the posterior distribution in overfitted mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(5):689–710.

Roy, H. E., Hesketh, H., Purse, B. V., Eilenberg, J., Santini, A., Scalera, R., Stentiford, G. D., Ariaens, T., Bacela-Spychalska, K., Bass, D., Beckmann, K. M., Bessell, P., Bojko, J., Booy, O., Cardoso, A. C., Essl, F., Groom, Q., Harrower, C., Kleespies, R., Martinou, A. F., van Oers, M. M., Peeler, E. J., Pergl, J., Rabitsch, W., Roques, A., Schaffner, F.,

Schindler, S., Schmidt, B. R., Schönrogge, K., Smith, J., Solarz, W., Stewart, A., Stroo, A., Tricarico, E., Turvey, K. M. A., Vannini, A., Vilà, M., Woodward, S., Wynns, A. A., and Dunn, A. M. (2017). Alien pathogens on the horizon: Opportunities for predicting their threat to wildlife. *Conservation Letters*, 10(4):477–484.

Salman, M., editor (2003). *Animal Disease Surveillance Survey Systems: Methods and Applications*. Iowa State Press, Ames, Iowa.

Santini, A., Liebhold, A., Migliorini, D., and Woodward, S. (2018). Tracing the role of human civilization in the globalization of plant pathogens. *The ISME Journal*, 12(3):647–652.

Scharf, H., Hooten, M. B., and Johnson, D. S. (2017). Imputation approaches for animal movement modeling. *Journal of Agricultural, Biological and Environmental Statistics*, 22(3):335–352.

Simonsen, L., Gog, J. R., Olson, D., and Viboud, C. (2016). Infectious disease surveillance in the big data era: Towards faster and locally relevant systems. *Journal of Infectious Diseases*, 214(S4):S380–S385.

Stein, M. (2010). Asymptotics for spatial processes. In Gelfand, A., Diggle, P. J., Fuentes, M., and Guttorp, P., editors, *Handbook of Spatial Statistics*. CRC Press, Boca Raton, Florida.

Stevenson, M. D., Rossmo, D. K., Knell, R. J., and Le Comber, S. C. (2012). Geographic profiling as a novel spatial tool for targeting the control of invasive species. *Ecography*, 35(8):704–715.

Taylor, B. M., Andrade-Pacheco, R., and Sturrock, H. J. W. (2018). Continuous inference for aggregated point process data. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 181(4):1125–1150.

Thompson, J., Finnie, T., Hall, I., and Dobrinkova, N. (2017). Catching clouds: Simultaneous optimization of the parameters of biological agent plumes using Dirichlet processes to

best estimate infection source location. In *Proceedings of the 2017 Federated Conference on Computer Science and Information Systems*. IEEE.

United States Government Accountability Office (2003). Emerging infectious diseases: Review of state and federal disease surveillance efforts. Web. https://www.gao.gov/assets/gao-04-877.pdf.

Verity, R., Stevenson, M. D., Rossmo, D. K., Nichols, R. A., and Le Comber, S. C. (2014). Spatial targeting of infectious disease control: Identifying multiple, unknown sources. *Methods in Ecology and Evolution*, 5(7):647–655.

Vynnycky, E. and White, R. D. (2010). *An Introduction to Infectious Disease Modelling*. Oxford University Press, New York.

Walker, N. B. (2018). Bias correction of bounded location errors in binary data. K-Rex, Web. http://hdl.handle.net/2097/39089.

Walker, N. B., Hefley, T. J., Ballmann, A. E., Russell, R. E., and Walsh, D. P. (2021). Recovering individual-level spatial inference from aggregated binary data. *Spatial Statistics*, 44:100514.

Walker, N. B., Hefley, T. J., and Walsh, D. P. (2020). Bias correction of bounded location error in binary data. *Biometrics*, 76(2):530–539.

Walter, W. D., Walsh, D. P., Farnsworth, M. L., Winkelman, D. L., and Miller, M. W. (2011). Soil clay content underlies prion infection odds. *Nature Communications*, 2:200.

Wang, F., Wang, J., Gelfand, A., and Li, F. (2017). Accommodating the ecological fallacy in disease mapping in the absence of individual exposures. *Statistics in Medicine*, 36(30):4930–4942.

Watsa, M. and Wildlife Disease Surveillance Focus Group (2020). Rigorous wildlife disease surveillance. *Science*, 369(6500):145–147.

Wilcox, B. A., Aguirre, A. A., and Horwitz, P. (2012). Ecohealth: Connecting ecology, health and sustainability. In Aguirre, A. A., Ostfeld, R., and Daszak, P., editors, *New Directions in Conservation Medicine: Applied Cases of Ecological Health*. Oxford University Press, Inc, New York.

Williams, E. S. and Miller, M. W. (2002). Chronic wasting disease in deer and elk in North America. *Revue Scientifique et Technique*, 21(2):305–316.

Williams, E. S., Miller, M. W., Kreeger, T. J., Kahn, R. H., and Thorne, E. T. (2002). Chronic wasting disease of deer and elk: A review with recommendations for management. *The Journal of Wildlife Management*, 66(3):551–563.

Williams, P. J., Hooten, M. B., Womble, J. N., Esslinger, G. G., Bower, M. R., and Hefley, T. J. (2017). An integrated data model to estimate spatiotemporal occupancy, abundance, and colonization dynamics. *Ecology*, 98(2):328–336.

World Health Organization (2020). The top 10 causes of death. Web. https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death. Accessed: 2021-07-29.

World Health Organization (2021). WHO coronavirus (COVID-19) dashboard. Web. https://covid19.who.int/. Assessed: 2021-10-4.

World Organization for Animal Health (2021). Terrestrial animal health code. Web. https://www.oie.int/en/what-we-do/standards/codes-and-manuals/terrestrial-code-online-access/.

Yeyati, E. L. and Filippini, F. (2021). Social and economic impact of COVID-19. Web. https://www.brookings.edu/wp-content/uploads/2021/06/Social-and-economic-impact-COVID.pdf.

Appendix A

Additional Details of Simulation Experiment and Data Example From Chapter 1

A.1 Introduction

This appendix provides additional results for the simulation experiment and additional information about the white-nose syndrome data example presented in Chapter 1. This appendix was originally published as Supporting Material for Walker et al. (2021). Figures and tables are labeled identically to the chapter (e.g., **Figure 1.1**, etc). The R code that reproduces the simulation experiment and the figures is found in the `simulation.R` file in the Supporting Material associated with Walker et al. (2021). The R code that reproduces the results from the data example and the associated figures is found in the `wns.R` file, also in the Supporting Material associated with Walker et al. (2021).

A.2 Simulation Experiment

We conducted a simulation experiment to compare performance of our proposed models, using different types of aggregated binary data, to traditional models for non-aggregated binary data (e.g., logistic regression). We generated simulated data using a unit square study area, $\mathcal{S} = [0,1] \times [0,1]$, that was divided into 400 regular grid cells (subregions), such that $\mathcal{S} = \cup_{j=1}^{400} \mathcal{A}_j$ and $|\mathcal{A}_j| = \frac{1}{400}$. We generated spatial covariates, $x(\mathbf{s})$ and $z(\mathbf{s})$, and simulated the locations and binary marks of observations from a BIPPP where the intensity function was $\log(\lambda(\mathbf{s})) = \alpha_0 + \alpha_1 z(\mathbf{s})$ and the classification function was $\text{logit}(p(\mathbf{s})) = \beta_0 + \beta_1 x(\mathbf{s})$. We focused on and compared estimates of β_0 and β_1 among five models because β_0 and β_1 are highly affected by aggregation and inference on the parameters of the classification function is likely to be the focus of many applied studies (Walker et al., 2020). We accomplished the comparison of estimates of β_0 and β_1 by assessing bias, coverage probabilities, and relative efficiency for estimates of β_0 and β_1 among the following five scenarios:

1. A traditional logistic regression model from (1.1) and (1.3) fit to non-aggregated data (see **Table 1.1** and **Figure 1.2**, Type A);

2. A joint model for n_{1j} and n_{0j} that is specified by combining the distributions in (1.6) and (1.7; see **Table 1.1** and **Figure 1.2**, Type C);

3. A joint model for v_j and n_j that is specified by combining the distributions in (1.8) and (1.9; see **Table 1.1** and **Figure 1.2**, Type D);

4. The conditional model for v_j given n_j from (1.9; see **Table 1.1** and **Figure 1.2**, Type D);

5. The Bernoulli model for v_j from (1.11; see **Table 1.1** and **Figure 1.2**, Type E).

We simulated 1000 data sets from four different settings using a combination of two factors: covariate equivalence ($x(\mathbf{s}) = z(\mathbf{s})$ vs. $x(\mathbf{s}) \neq z(\mathbf{s})$); and average sample size (small vs. large). Thus our simulation experiment uses a total of 4000 simulated data sets and realizations of $z(\mathbf{s})$ and $x(\mathbf{s})$. Each simulated data set was aggregated to fit each data type

in scenarios 2-5. We drew each spatial covariate realization from a low-rank Gaussian process (Higdon, 2002) on a 200×200 grid with knots at every fourth grid cell to reduce computation time. We chose parameter values of $\alpha_1 = 1$, and $\beta_1 = 1$ for all settings. We chose values for α_0 and β_0 for each setting such that the average sample size per subregion was either 10 or 50 (small vs. large) and the proportion of subregions that contained a binary mark of one was approximately constant across all settings (see **Table A.1**).

Setting	Covariate Equivalence $x(\mathbf{s}) = z(\mathbf{s})$	Average Sample Size	α_0	β_0
1	Yes	Small	7.800	-5.500
2	Yes	Large	9.410	-7.070
3	No	Small	7.820	-4.750
4	No	Large	9.405	-6.350

Table A.1: Settings from the simulation experiment using covariate equivalence ($x(\mathbf{s}) = z(\mathbf{s})$ vs. $x(\mathbf{s}) \neq z(\mathbf{s})$ and two average sample sizes (small vs. large), along with the values for α_0 and β_0 that were used when simulating data.

We fit the model in scenario one (i.e., traditional logistic regression) using the `glm` function in R to obtain the maximum likelihood estimates (MLEs) of β_0 (R Core Team, 2021). We fit the models in scenarios two through five as described in Section 1.3.5. For each model and setting, we calculated and compared the coverage probabilities (CPs) from the 95% Wald-type CIs for β_0. We also constructed box plots comparing the distribution of $\hat{\beta}_0$ obtained from the 1000 data sets for each scenario and setting. We calculated the standard deviation of the empirical distribution of 1000 estimates of β_0 in each scenario. We then calculated the relative efficiency of $\hat{\beta}_0$ for scenarios two through five by dividing the standard deviation of the distribution of $\hat{\beta}_0$ for the respective scenario by that of scenario one. Additionally, we calculated the mean squared predictive error (MSPE) in the estimated intensity and probability surfaces for each of the models in scenarios two through five. However, we only calculated the MSPE for the estimated probability surface for the model in scenario one. Lastly, we produce four data sets (each data set generated from a different simulation setting) and fit each of the models for Type C-E data in the appropriate data scenario. We then produce plots of the estimated intensity and probability surfaces obtained from each of the

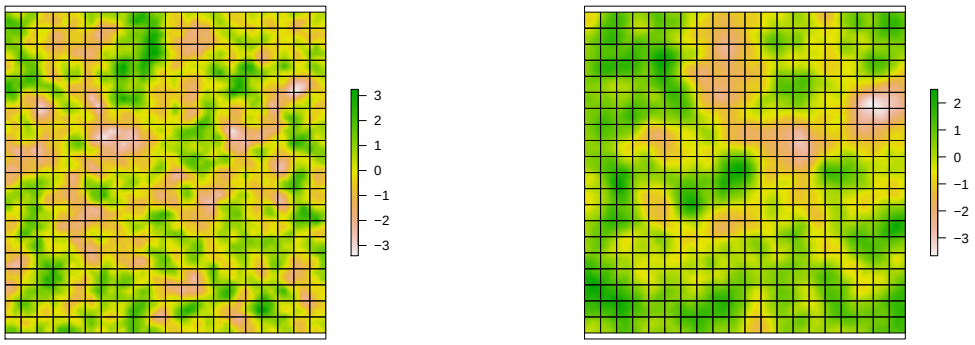

Figure A.1: Two plots showing a realization of the spatial covariates that were used for $x(\mathbf{s})$ (left) and $z(\mathbf{s})$ (right). In all cases $x(\mathbf{s})$ and $z(\mathbf{s})$ are spatially correlated and drawn from a low-rank Gaussian process on a 200×200 grid with knots at every fourth grid cell. The grid in each plot shows the partition of the study area into the 400 subregions over which the data were aggregated.

models for Types C-E data (for a total of thirty-two estimated surfaces).

When binary data are generated according to a BIPPP and then spatially aggregated, we expect to obtain unbiased estimates in scenarios two, three, four, and five. Of the models based on the distributional results presented in Sections 1.3.2-1.3.3, we expect that the model for scenario two will have the highest relative efficiency among all settings covered by the experiment, followed by the models from scenarios three, four, and five. We expect the MSPE of the estimated intensity and probability surfaces to be smallest for the model in scenario two, followed by three, four, and five. The detailed R code capable of reproducing the simulation experiment may be found in the `simulation.R` file in the supporting information associated with Walker et al. (2021). Plots showing results from all settings in the simulation experiment are provided in this appendix.

A.2.1 Spatial Covariates

The spatial covariates used for all simulated data sets in the settings outlined above are similar in characteristics to the covariates shown in **Figure A.1**.

A.3 Results

This section contains the supplemental results from the simulation experiment. We first provide a summary of the results for β_0 in **Table A.2**. The subsequent figures allow the reader to empirically evaluate bias and relative efficiency among all five data scenarios for the four settings. **Figure A.2** contains comparisons of the estimates of β_0. In Section A.3.1 we provide results that address the question of how well the estimated intensity and probability surfaces match the true intensity and probability surfaces, using mean squared predictive error (MSPE). We accomplish this by first providing graphical summaries of mean squared predictive error (MSPE) for all scenarios, data sets, and settings in **Figures A.3-A.4**. We then provide and evaluate plots of example estimated intensity and probability surfaces obtained from each of the models for Types C-E data using an example of covariates and data from each of the four settings.

In our simulation experiment, we crossed two factors (average sample size per subregion and covariate equivalence) with two levels each. With our choices of α_0, the average numbers of observations within each grid cell were 10.2 and 50.1 for small and large sample settings, respectively. With our choices of β_0 for each setting, we maintained a proportion of approximately 0.11 of grid-cells that contained a binary mark of one (see **Table A.2**).

As expected, under the model and data in scenario one (traditional logistic regression with no data aggregation), the MLEs for β_0 appear to be unbiased for all settings and had CPs between 0.950 and 0.956. Under the model and data in scenario two (joint distribution of n_{1j} and n_{0j}) the MLEs for β_0 appear to be unbiased for all settings in the simulation study (see **Figure A.2** for graphical comparisons of estimates). The CPs for $\hat{\beta}_0$, in scenario two, were between 0.948 and 0.957 for all settings. Additionally, the relative efficiency of $\hat{\beta}_0$, obtained from scenario two, ranged from about 1.06 (settings 4) to about 1.11 (setting 3). The coverage probabilities obtained for scenarios one and two, and efficiencies for scenario two, are available in **Table A.2**.

Under the model and data in scenario three (joint distribution of v_j and n_j) the MLEs for β_0 appear to be unbiased for all settings in the simulation study (see **Figure A.2**). The

CPs for $\hat{\beta}_0$, in scenario three, were between 0.944 and 0.969 for all settings. Additionally, the relative efficiency of $\hat{\beta}_0$, obtained from scenario three, ranged from about 1.14 (setting 4) to about 1.39 (setting 2). The coverage probabilities and efficiencies obtained for scenario three are available in **Table A.2**.

Under the model and data in scenario four (conditional distribution of v_j given n_j) the MLEs for β_0 appear to be unbiased for all settings in the simulation study (see **Figure A.2**). The CPs for $\hat{\beta}_0$, in scenario four, were between 0.905 and 0.938 for all settings. Additionally, the relative efficiency of $\hat{\beta}_0$, obtained from scenario four, ranged from about 1.16 (setting 4) to about 1.66 (setting 1). Finally, under the model and data in scenario five (Bernoulli distribution of v_j), the MLEs for β_0 were weakly identifiable with efficiencies of $\hat{\beta}_0$ ranging from about 31.8 (setting 4) to over 3,900 (setting 3) and CPs between 0.510 and 0.777. The coverage probabilities and efficiencies obtained for scenarios four and five are available in **Table A.2**.

Setting	Covariate Equivalence ($x(\mathbf{s}) = z(\mathbf{s})$)	Sample Size	CP Scen. 1	CP Scen. 2	CP Scen. 3	CP Scen. 4	CP Scen. 5	Eff. Scen. 2	Eff. Scen. 3	Eff. Scen. 4	Eff. Scen. 5
1	Yes	Small	0.950	0.956	0.969	0.905	0.605	1.10	1.37	1.66	653.0
2	Yes	Large	0.956	0.957	0.960	0.910	0.510	1.10	1.39	1.64	248.7
3	No	Small	0.951	0.948	0.944	0.915	0.777	1.11	1.20	1.22	3,929
4	No	Large	0.953	0.957	0.955	0.938	0.663	1.06	1.14	1.16	31.78

Table A.2: Results from our simulation experiment using two sample sizes (small vs. large) and two levels of covariate equivalence ($x(\mathbf{s}) = z(\mathbf{s})$ vs. $x(\mathbf{s}) \neq z(\mathbf{s})$). For each setting, we show the relative efficiency (Eff.) for estimating β_0 and the 95% CI coverage probability (CP) for each of the models that were based on the proposed distributional results (using appropriate types of aggregated data) for 1,000 simulated data sets. We also report the 95% CI coverage probabilities for logistic regression using the exact locations of observations. We calculate the relative efficiency for each model as the ratio of the standard deviation of the empirical distribution of β_0 from the respective model against that of logistic regression.

A.3.1 MSPE of Estimated Risk and Intensity Surfaces and Example Estimated Surfaces

We evaluate the performance of the models for Type A and C-E data in estimating the risk and intensity surfaces ($\lambda(\cdot)$ and $p(\cdot)$) using mean squared predictive error (MSPE). The

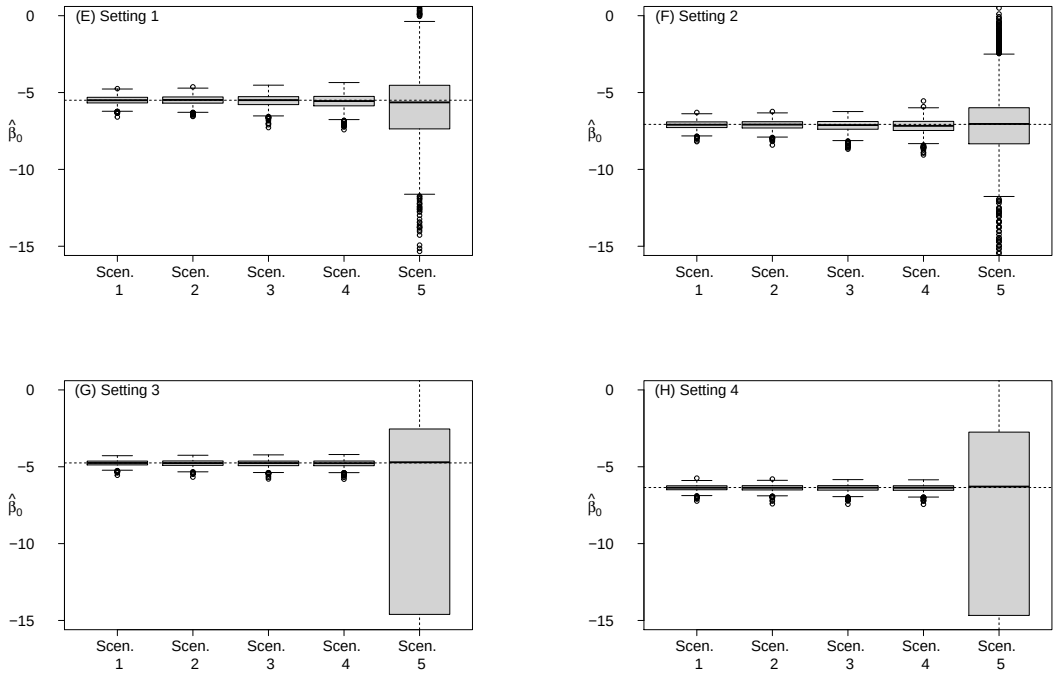

Figure A.2: Panels (E) and (F) show box plots of results for β_0 from small and large average sample size simulation experiment settings where $x(\mathbf{s}) = z(\mathbf{s})$. Panels (G) and (H) show small and large sample size simulation experiment settings where $x(\mathbf{s}) \neq z(\mathbf{s})$. We show estimates of β_0 obtained using five different models (each under a different data aggregation scenario), which included: Scen. 1) logistic regression with no data aggregation (Type A data); Scen. 2) a joint model for n_{1j} and n_{0j} where binary data were aggregated into counts for each subregion (Type C data); Scen. 3) a joint model for v_j and n_j using data aggregated into a count and indicator variable for each subregion (Type D data); Scen. 4) a conditional model for v_j using data aggregated into an indicator variable for each subregion (Type E data). Each of the four panels used 1,000 simulated data sets, and each panel shows the true value of β_0 (dotted line). The distribution of $\hat{\beta}_0$ from scenario five (Bernoulli model) was such that some estimates fell outside the upper bounds of the plots. Each box plot shows (from bottom to top) the lower bound of 1.5 times the inter-quartile range, the 25^{th} percentile, the median, the 75^{th} percentile, and the upper bound of 1.5 times the inter-quartile range. See **Table A.2** for a summary of all settings.

estimated surfaces were produced using the inverse link functions from the paper and linear combination of the relevant coefficient estimates and spatial covariates $x(\mathbf{s})$ and $z(\mathbf{s})$:

$$\lambda(\mathbf{s}) = e^{\alpha_0 + z(\mathbf{s})\alpha_1},$$

$$p(\mathbf{s}) = \frac{e^{\beta_0 + x(\mathbf{s})\beta_1}}{1 + e^{\beta_0 + x(\mathbf{s})\beta_1}}.$$

The estimated surfaces were then compared against the true surfaces (calculated using the true parameter values) at all locations in the unit study area using MSPE. In the context of our simulation, 'all locations' refers to all 40,000 grid cell centroids given by our definition of the covariates $x(\mathbf{s})$ and $x(\mathbf{s})$ (each covariate defined on a 200×200 grid). **Figures A.3-A.4** contain the graphical comparisons of MSPE for the estimated intensity and risk surfaces among the models for Types A and C-E data in all four settings. In addition, we provide example estimated risk and intensity surface maps ($\lambda(\cdot)$ and $p(\cdot)$) from a simulated data set from each of the four simulation settings for each of the following models: the joint model for Type C data, the joint model for Type D data, the conditional model for Type D data, and the Bernoulli model for Type E data. The maps were produced using the inverse link functions from the paper and linear combination of the relevant coefficient estimates and spatial covariates $x(\mathbf{s})$ and $z(\mathbf{s})$. **Figures A.5-A.12** contain the estimated $\lambda(\mathbf{s})$ and $p(\mathbf{s})$ surfaces from all four models from Settings 1-4. We provide **Table A.3** as a quick reference for locating the figures for the intensity and risk surfaces for each simulation setting.

Setting	Figures for Estimated $\lambda(\mathbf{s})$ Surface	Figure for Estimated $p(\mathbf{s})$ Surfaces
1	**Figure A.5**	**Figure A.6**
2	**Figure A.7**	**Figure A.8**
3	**Figure A.9**	**Figure A.10**
4	**Figure A.11**	**Figure A.12**

Table A.3: Figure titles for the figures that contain example estimated intensity ($\lambda(\mathbf{s})$) and risk surfaces ($p(\mathbf{s})$) for each simulation setting.

The plots of MSPE across models and settings in **Figures A.3-A.4** and the plots of the estimated surfaces $\lambda(\mathbf{s})$ and $p(\mathbf{s})$ in **Figures A.5-A.12** highlight that as the degree of

data aggregation increases, from Type C data to Type E, the relative efficiency of parameter estimates decreases (in terms of the precision of estimated parameters for a given number of sampled individuals). Further, the probability of obtaining extreme values of coefficient estimates and standard errors (and by extension, intensity and probability surfaces) from the proposed models increases as aggregation increases from Type C to Type E data due to identifiability issues. As a result, we see weaker performance (as measured by MSPE) in estimating intensity and risk surfaces among the conditional and Bernoulli models for Type D and E data.

A.4 Disease Risk Factor Analysis Data and Figures

This section contains the estimated risk and intensity maps ($\lambda(\cdot)$ and $p(\cdot)$) from the risk factor analysis for each of the following models: the joint model for Type C data, the joint model for Type D data, the conditional model for Type D data, and the Bernoulli model for Type E data. The maps were produced using the inverse link functions from the paper and linear combination of the relevant coefficient estimates and spatial covariates:

$$\lambda(\mathbf{s}) = e^{\alpha_0 + z(\mathbf{s})\alpha_{karst}},$$

$$p(\mathbf{s}) = \frac{e^{\beta_0 + x(\mathbf{s})\beta_{forest}}}{1 + e^{\beta_0 + x(\mathbf{s})\beta_{forest}}},$$

where $z(\mathbf{s})$ is spatial covariate that takes a value of 1 wherever karst landscape is present, and 0 everywhere else. Additionally, $x(\mathbf{s})$ is a spatial covariate for proportion of land classified as forest, that takes a value between 0 and 1 inclusive. **Figure A.13** contains the estimated $\lambda(\mathbf{s})$ surfaces from all four models, while **Figure A.14** shows the estimated $p(\mathbf{s})$ surfaces for the same models. The portrayals of $\widehat{\lambda(\mathbf{s})}$ and $\widehat{p(\mathbf{s})}$ in **Figures A.13-A.14** highlight that as the degree of data aggregation increases, from Type C data to Type E, the relative efficiency of parameter estimates and the unreported intercept estimates decreases (in terms of the precision of estimated parameters for a given number of sampled individuals). Further, the probability of obtaining extreme values of coefficient estimates and standard errors from the

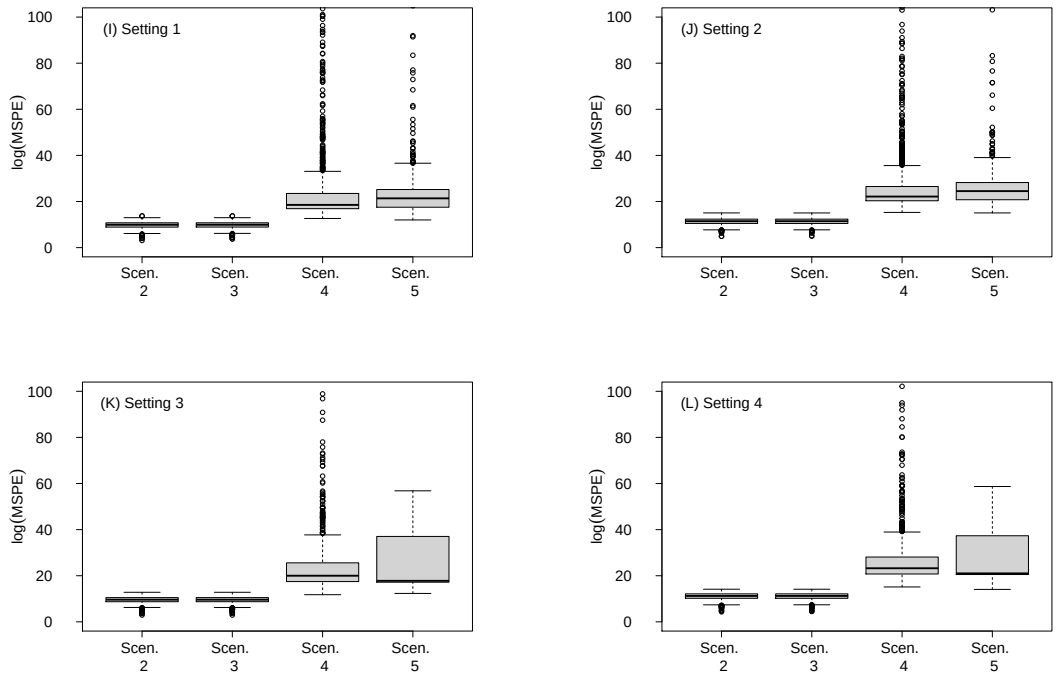

Figure A.3: Box plots of the log transformed mean squared predictive errors (MSPE) of the estimated intensity surfaces $\widehat{\lambda(\mathbf{s})}$ from each data scenario and data set in each setting. Panels (I) and (J) show the log transformed MSPE obtained from small and large average sample size simulation experiment settings where $x(\mathbf{s}) = z(\mathbf{s})$. Panels (K) and (L) show the log transformed MSPE from small and large sample size simulation experiment settings where $x(\mathbf{s}) \neq z(\mathbf{s})$. We show the log MSPE calculated using estimates from four different models (each under a different data aggregation scenario), which included: Scen. 2) a joint model for n_{1j} and n_{0j} where binary data were aggregated into counts for each subregion (Type C data); Scen. 3) a joint model for v_j and n_j using data aggregated into a count and indicator variable for each subregion (Type D data); Scen. 4) a conditional model for v_j using data aggregated into an indicator variable for each subregion (Type E data). A smaller log MSPE is indicative of a estimated intensity surface that is closer to the true intensity surface. Each of the four panels used 1,000 simulated data sets. The distributions of the log MSPE from scenarios four and five (Conditional and Bernoulli models) were such that some estimates fell outside the upper bounds of the plots. Each box plot shows (from bottom to top) the lower bound of 1.5 times the inter-quartile range, the 25^{th} percentile, the median, the 75^{th} percentile, and the upper bound of 1.5 times the inter-quartile range.

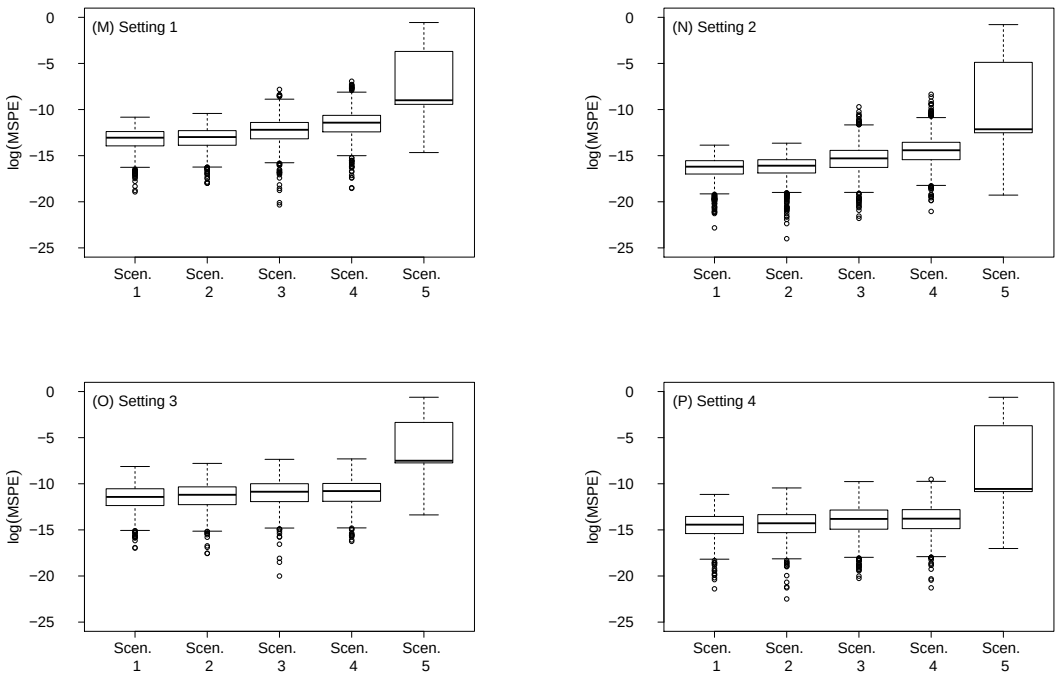

Figure A.4: Box plots of the mean squared predictive errors (MSPE) of the estimated probability surfaces $\widehat{p(\mathbf{s})}$ from each data scenario and data set in each setting. Panels (M) and (N) show the log transformed MSPE obtained from small and large average sample size simulation experiment settings where $x(\mathbf{s}) = z(\mathbf{s})$. Panels (O) and (P) show the log transformed MSPE from small and large sample size simulation experiment settings where $x(\mathbf{s}) \neq z(\mathbf{s})$. We show the MSPE calculated using estimates from five different models (each under a different data aggregation scenario), which included: Scen. 1) logistic regression with no data aggregation (Type A data); Scen. 2) a joint model for n_{1j} and n_{0j} where binary data were aggregated into counts for each subregion (Type C data); Scen. 3) a joint model for v_j and n_j using data aggregated into a count and indicator variable for each subregion (Type D data); Scen. 4) a conditional model for v_j using data aggregated into an indicator variable for each subregion (Type E data). A smaller MSPE is indicative of a estimated probability surface that is closer to the true probability surface. Each of the four panels used 1,000 simulated data sets. Each box plot shows (from bottom to top) the lower bound of 1.5 times the inter-quartile range, the 25^{th} percentile, the median, the 75^{th} percentile, and the upper bound of 1.5 times the inter-quartile range.

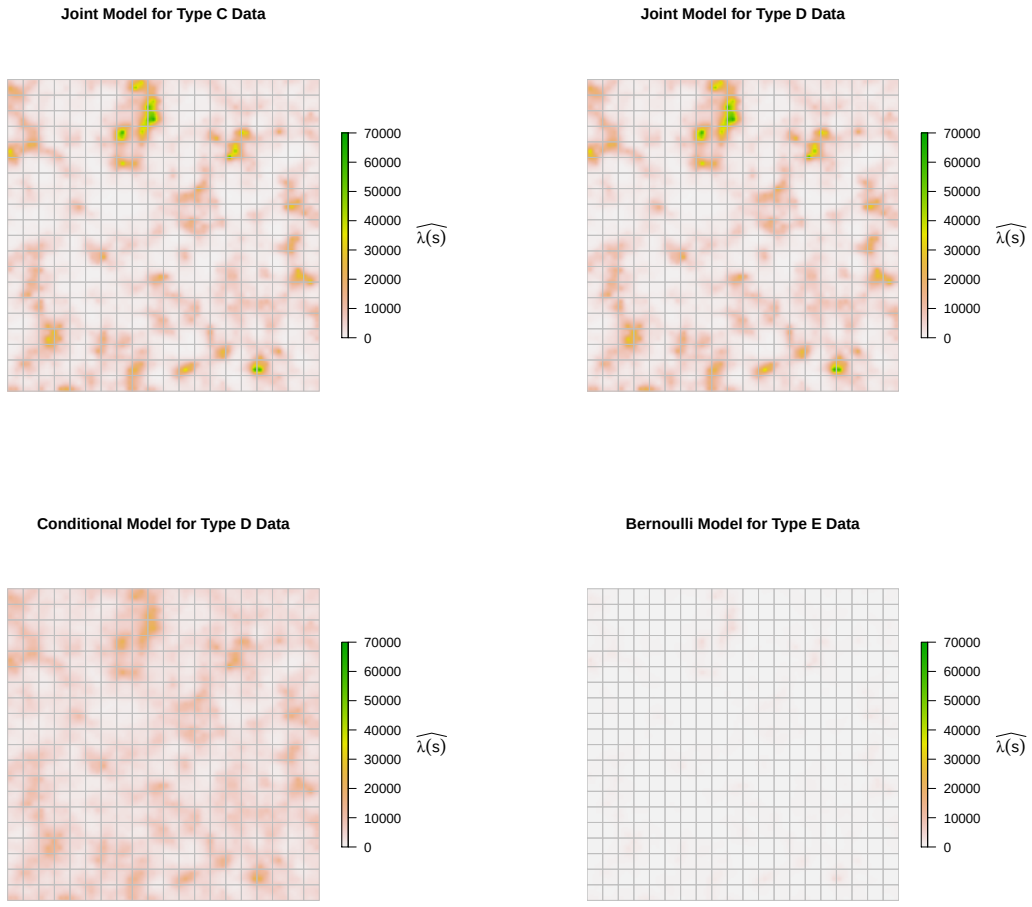

Figure A.5: An example estimated $\lambda(\mathbf{s})$ surface across the simulated unit study area from Setting 1, obtained by fitting the following models to an example data set: 1) Joint model for Type C data (top left); 2) Joint model for Type D data (top right); 3) Conditional model for Type D data (bottom left) 4) Bernoulli model for Type E data (bottom right).

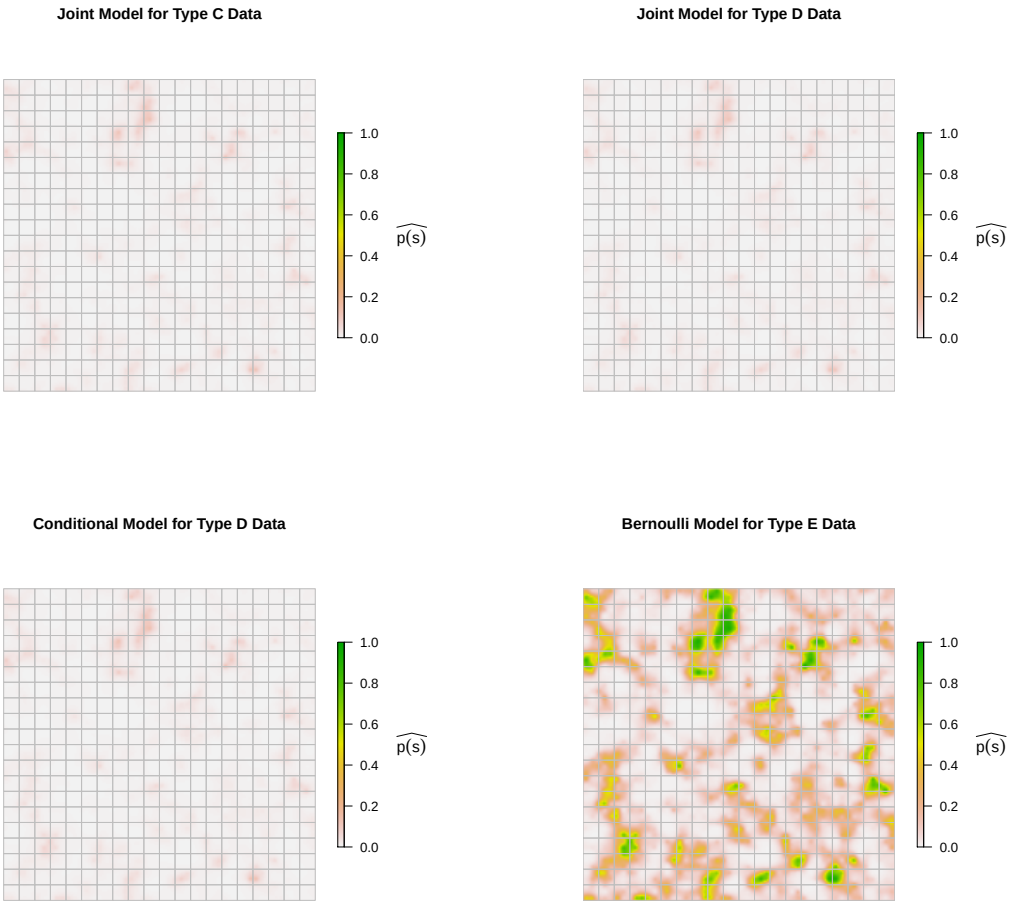

Figure A.6: An example estimated $p(\mathbf{s})$ surface across the simulated unit study area from Setting 1, obtained by fitting the following models to an example data set: 1) Joint model for Type C data (top left); 2) Joint model for Type D data (top right); 3) Conditional model for Type D data (bottom left) 4) Bernoulli model for Type E data (bottom right).

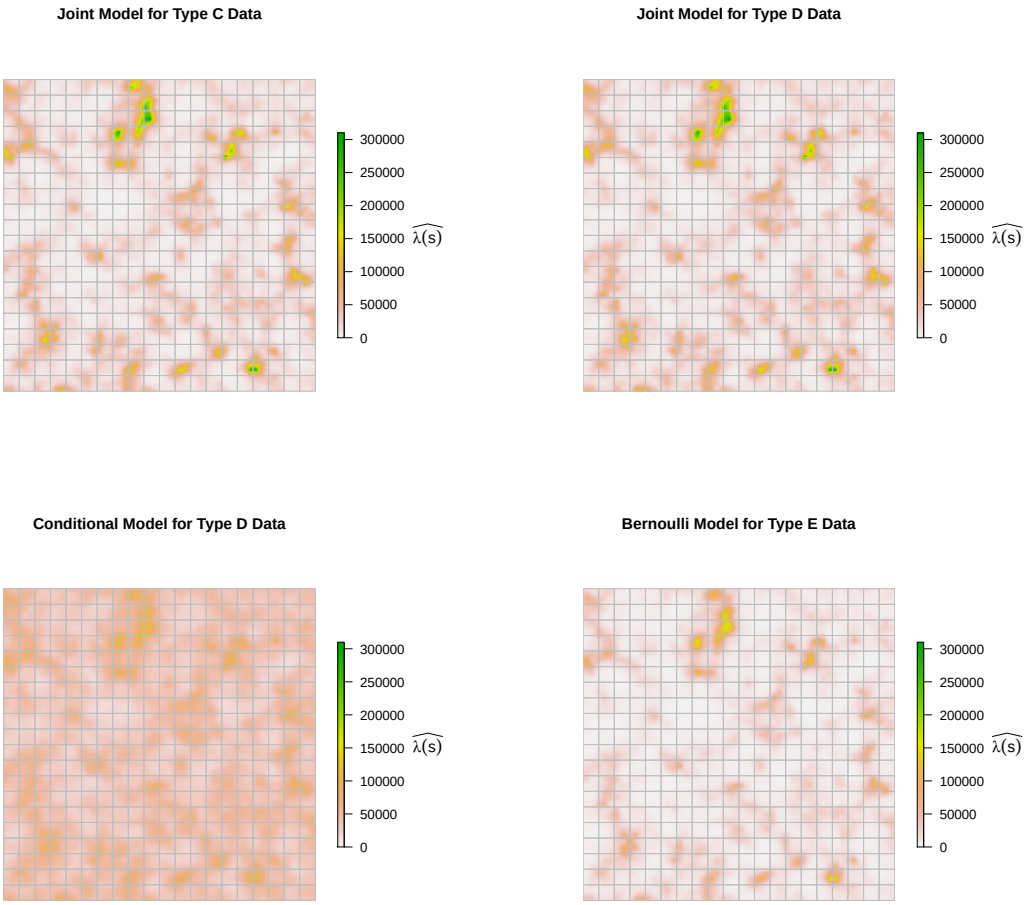

Figure A.7: An example estimated $\lambda(\mathbf{s})$ surface across the simulated unit study area from Setting 2, obtained by fitting the following models to an example data set: 1) Joint model for Type C data (top left); 2) Joint model for Type D data (top right); 3) Conditional model for Type D data (bottom left) 4) Bernoulli model for Type E data (bottom right).

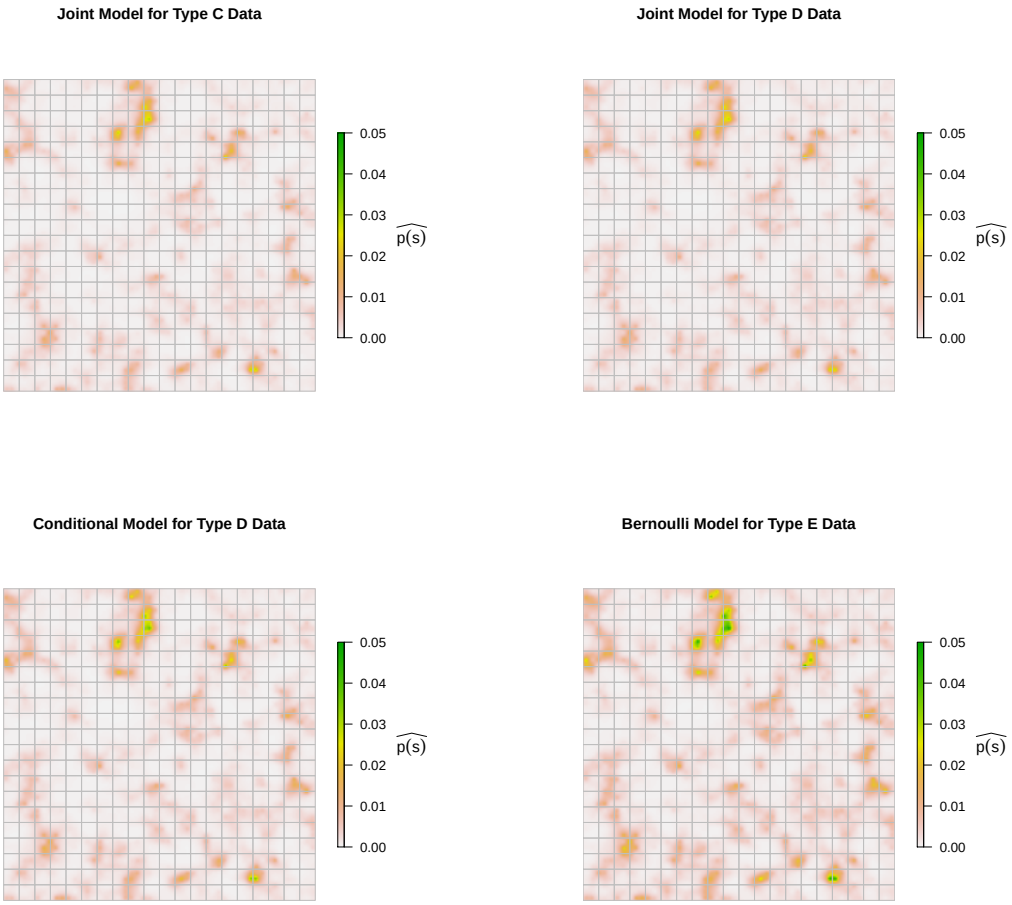

Figure A.8: An example estimated $p(\mathbf{s})$ surface across the simulated unit study area from Setting 2, obtained by fitting the following models to an example data set: 1) Joint model for Type C data (top left); 2) Joint model for Type D data (top right); 3) Conditional model for Type D data (bottom left) 4) Bernoulli model for Type E data (bottom right).

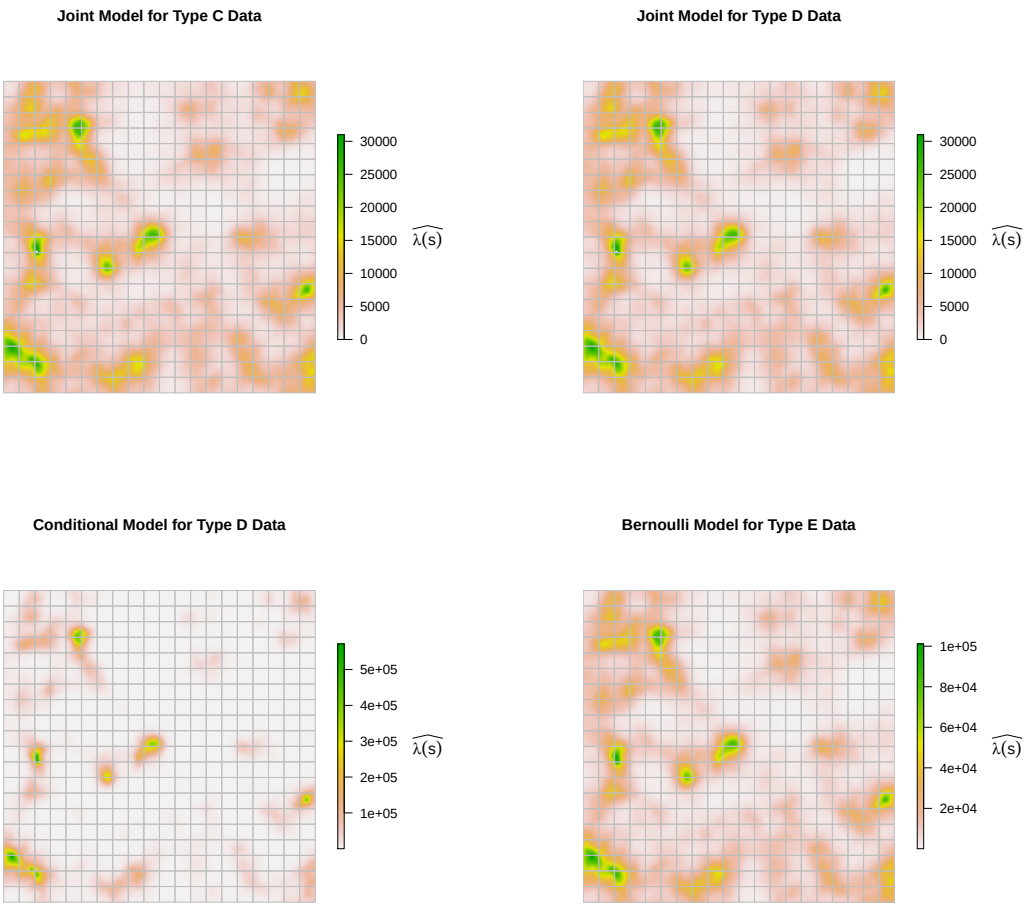

Figure A.9: An example estimated $\lambda(\mathbf{s})$ surface across the simulated unit study area from Setting 3, obtained by fitting the following models to an example data set: 1) Joint model for Type C data (top left); 2) Joint model for Type D data (top right); 3) Conditional model for Type D data (bottom left) 4) Bernoulli model for Type E data (bottom right). We note that the scales of the legends on the four plots are different from each other.

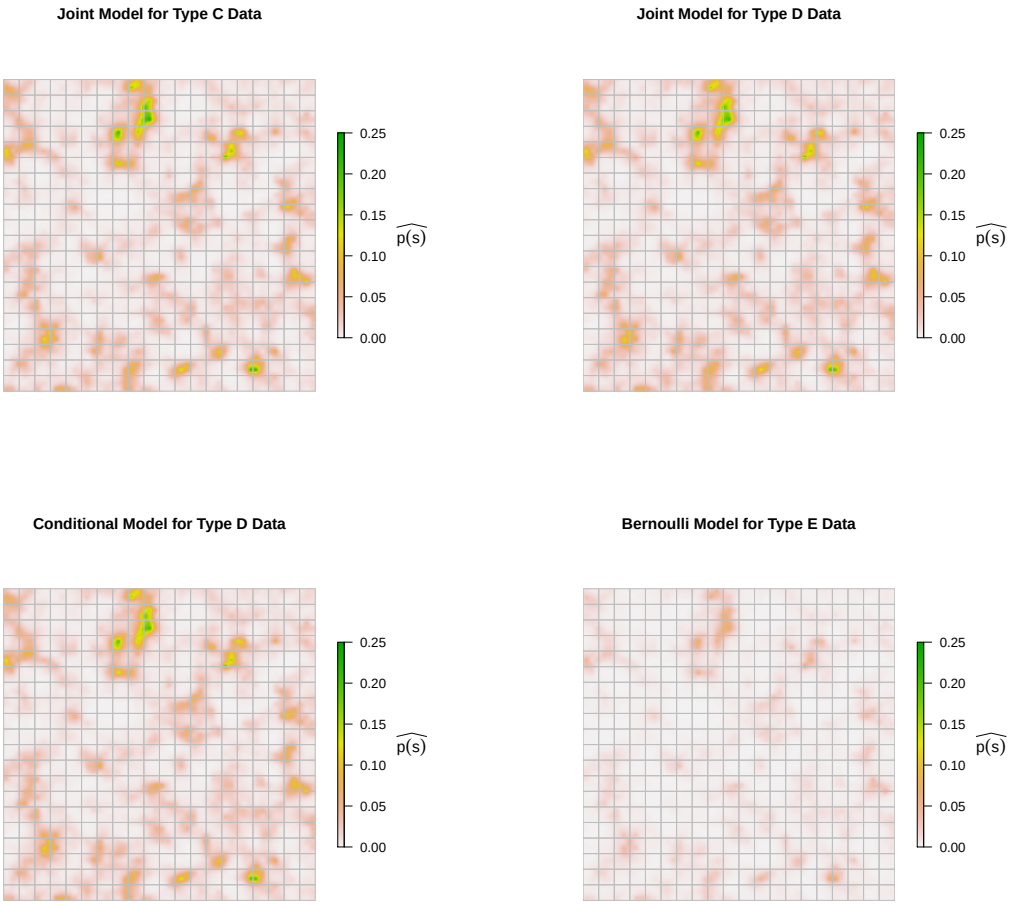

Figure A.10: An example estimated $p(\mathbf{s})$ surface across the simulated unit study area from Setting 3, obtained by fitting the following models to an example data set: 1) Joint model for Type C data (top left); 2) Joint model for Type D data (top right); 3) Conditional model for Type D data (bottom left) 4) Bernoulli model for Type E data (bottom right).

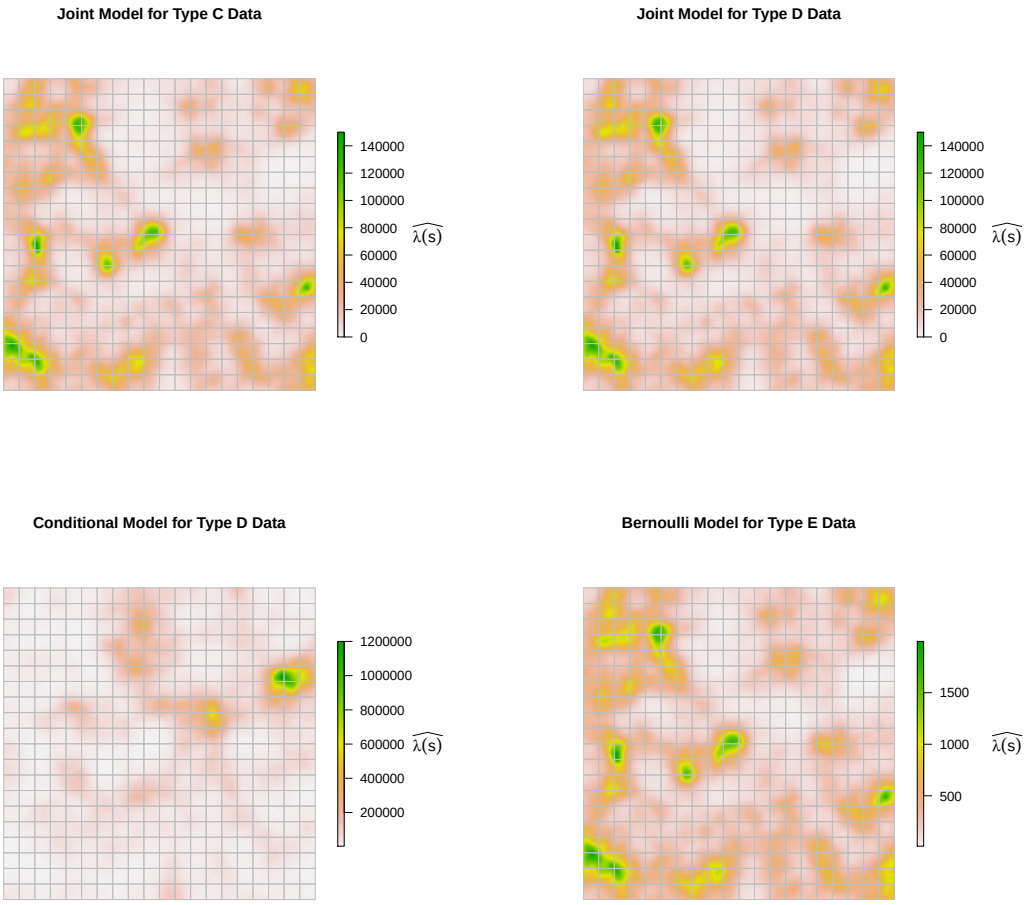

Figure A.11: An example estimated $\lambda(\mathbf{s})$ surface across the simulated unit study area from Setting 4, obtained by fitting the following models to an example data set: 1) Joint model for Type C data (top left); 2) Joint model for Type D data (top right); 3) Conditional model for Type D data (bottom left) 4) Bernoulli model for Type E data (bottom right). The surface produced by the Conditional model for Type D data (bottom left) is starkly different because the sign of the estimate of the slope parameter α_1 was negative. We note that the scales of the legends on the four plots are different from each other.

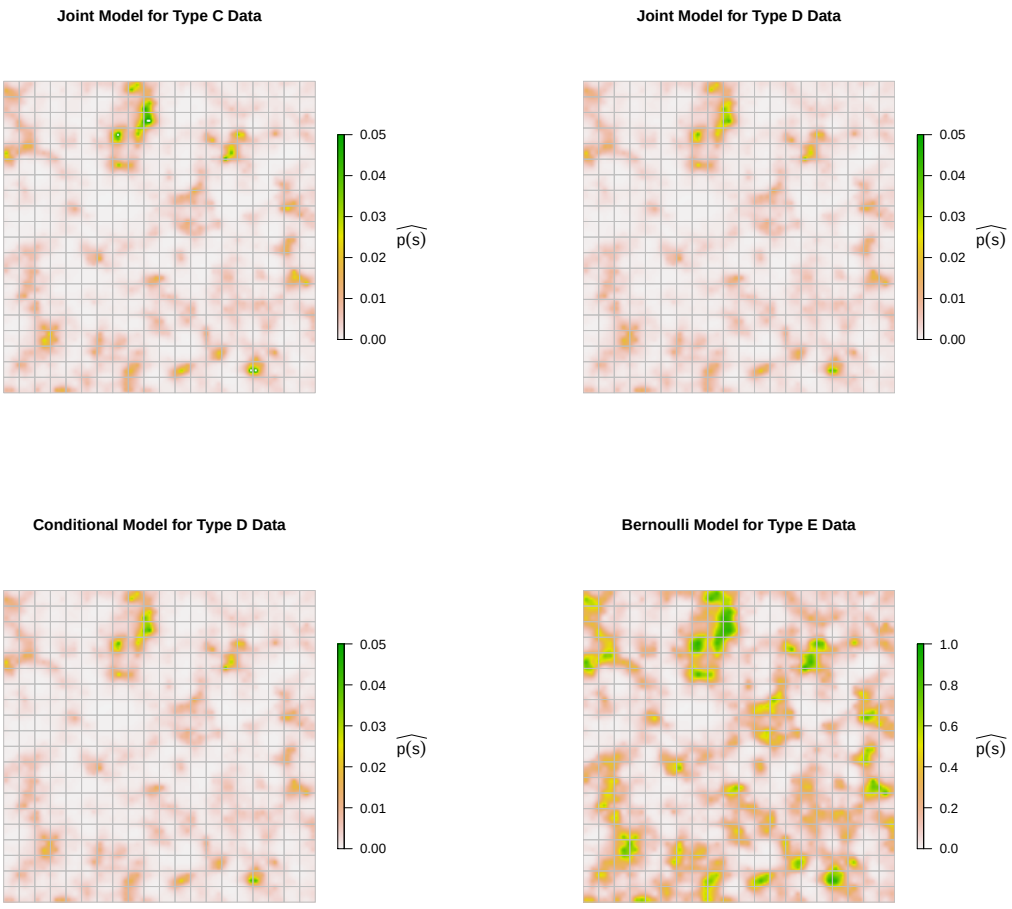

Figure A.12: An example estimated $p(\mathbf{s})$ surface across the simulated unit study area from Setting 4, obtained by fitting the following models to an example data set: 1) Joint model for Type C data (top left); 2) Joint model for Type D data (top right); 3) Conditional model for Type D data (bottom left) 4) Bernoulli model for Type E data (bottom right). We note that the scales of the legends on the four plots are different from each other.

proposed models increases as aggregation increases from Type C to Type E data due to identifiability issues.

We took several steps to ensure that the aberrant results for the Bernoulli model for Type E data were the result of identifiability issues in the intercept terms for the model, $\hat{\alpha}_0$ and $\hat{\beta}_0$. First, we determined through a sensitivity analysis that the parameter estimates from the Bernoulli model were not sensitive to the parameter starting values that we provided to the `optim` function in the program R. Next, we confirmed that $\hat{\alpha}_0$ and $\hat{\beta}_0$ from the Bernoulli model were highly correlated, using the `cov2cor` function on the inverse of the Hessian matrix in R. Third, we refit a constrained Bernoulli model using a fixed value for $\alpha_0 = 5.9631673$ (the MLE of α_0 from the joint model of n_{1j} and n_{0j}). This fixed value for α_0 was reasonable because the joint model for n_{1j} and n_{0j} (Type C data) provides the most precise inference. Refitting the Bernoulli model with this constraint on α_0 produced inference that was more reasonable (see **Figure A.15**). The sensitivity analysis, the correlated estimates, and the improved performance of the constrained Bernoulli model all confirm that the unconstrained Bernoulli model for Type E data has a higher probability of extreme coefficient estimates and standard errors. Additionally, we demonstrate that incorporating auxiliary information about $\lambda(\mathbf{s})$ into the Bernoulli model improves the quality of inference.

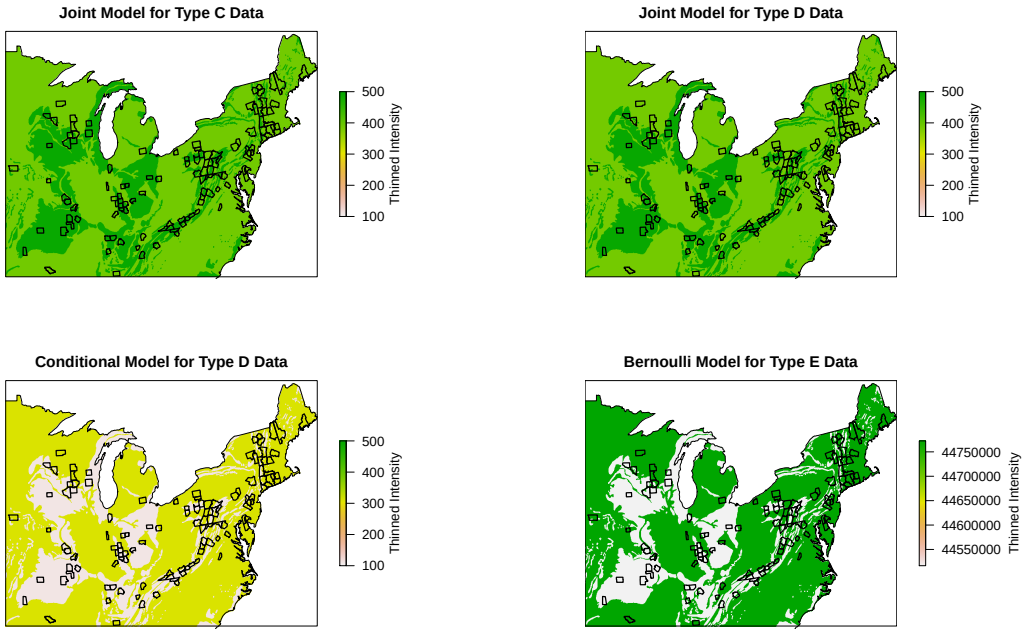

Figure A.13: The estimated $\lambda(\mathbf{s})$ surface across the northeastern United States obtained from the following models: 1) Joint model for Type C data (top left); 2) Joint model for Type D data (top right); 3) Conditional model for Type D data (bottom left) 4) Bernoulli model for Type E data (bottom right). The estimates for α_1 had a negative sign from the conditional model for Type D data and the Bernoulli model.

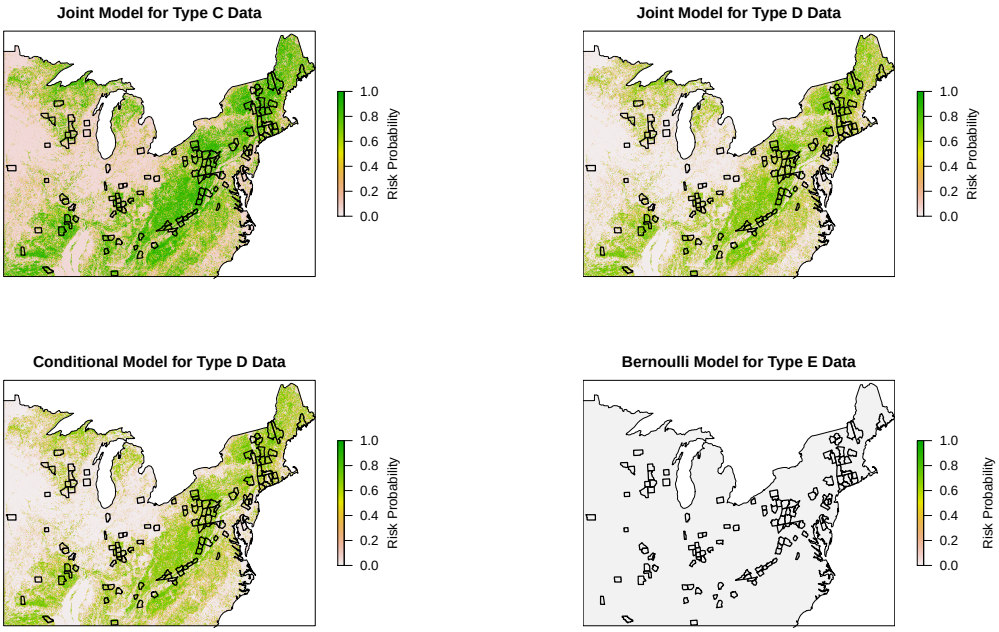

Figure A.14: The estimated $p(\mathbf{s})$ surface across the northeastern United States obtained from the following models: 1) Joint model for Type C data (top left); 2) Joint model for Type D data (top right); 3) Conditional model for Type D data (bottom left) 4) Bernoulli model for Type E data (bottom right). Risk probabilities differed significantly for the Bernoulli model because the Bernoulli model provided a large negatively-signed estimate for β_0 in comparison to the other models, due to identifiability issues in α_0 and β_0.

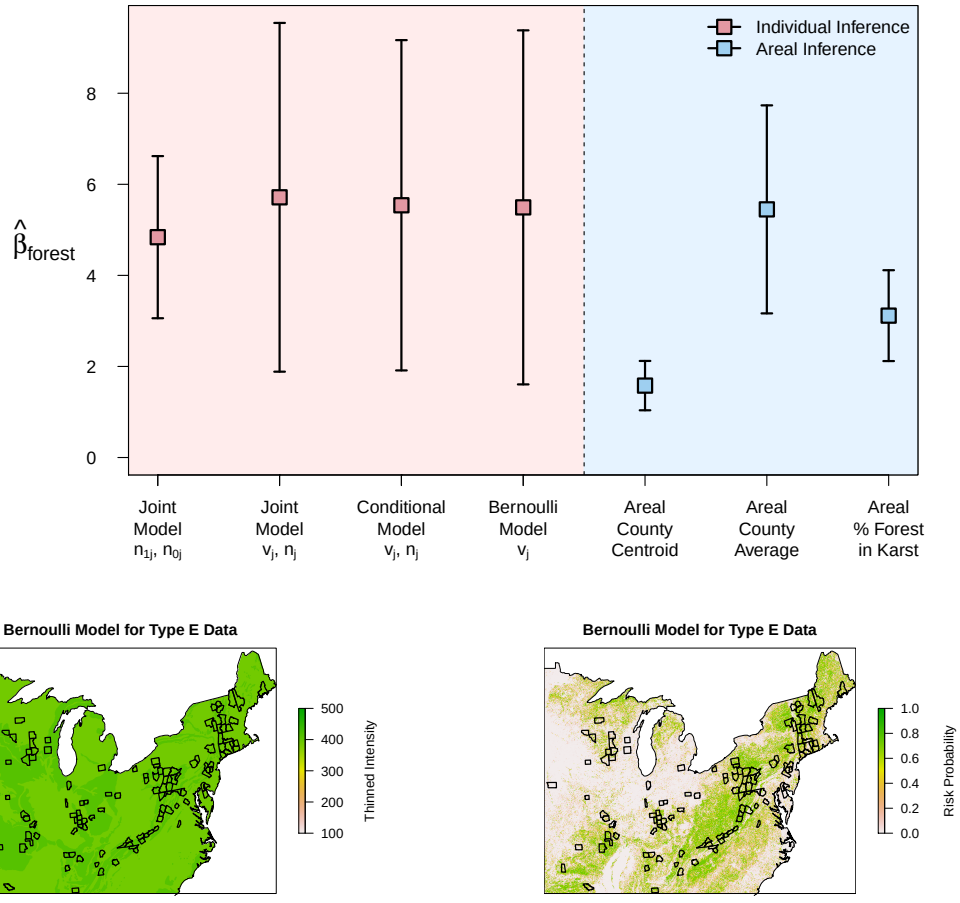

Figure A.15: **Top:** The comparison of estimates and 95% CIs for β_{forest} across methods after refitting the Bernoulli model with a fixed value for α_0. All other estimates and CIs are identical to **Figure 1.4** in Chapter 1. For each model, we give the coefficient estimate (box) followed by the 95% CI limits (whisker ends). **Bottom Left:** The estimated $\lambda(\mathbf{s})$ surface across the northeastern United States obtained from the Bernoulli model with a fixed value for α_0. **Bottom Right:** The estimated $p(\mathbf{s})$ surface across the northeastern United States the Bernoulli model with a fixed value for α_0.

A.5 Disclaimer

Any use of trade, firm, or product names is for descriptive purposes only and does not imply endorsement by the U.S. Government.

Appendix B

The Analytical Solution for the Homogenized PDE

We now derive the analytical solution to the homogenized ecological diffusion with exponential growth PDE in (3.7). This analytical solution provides an approximation to the PDE in (3.2) for any location and time in the study domain and results in substantial computational savings. We obtained the analytical solution through collaboration with an applied mathematician, Dr. Ian McGahan, at the University of Wisconsin. This analytical solution may be derived through several steps outlined in Logan (2014). To simplify the derivation, we define $s_{1,j} = (s_1 - \omega_{1,j})$, $s_{2,j} = (s_2 - \omega_{2,j})$, $\mathbf{s}_j \equiv (s_{1,j}, s_{2,j})'$, and $t_j = (t - t_{0_j})$. First, we transform the homogenized PDE from (3.4) into a Fourier parameter space as follows:

$$\hat{c}_{t,j}(\mathbf{k}_j, t_j) = \text{FT}[c_j(\mathbf{s}_j, t_j)] = \frac{1}{2\pi} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} c(\mathbf{s}_j, t_j) e^{i(k_{1,j} s_{1,j} + k_{2,j} s_{2,j})} d\mathbf{s}_j, \tag{B.1}$$

where $\mathbf{k}_j \equiv (k_{1,j}, k_{2,j})'$ is a vector of real-valued parameters in Fourier space. We also note that $\text{FT}[\frac{\partial}{\partial s_{1,j}} c_j(\mathbf{s}_j, t_j)] = ik_{1,j}\hat{c}(\mathbf{k}_j, t_j)$ and $\text{FT}[\frac{\partial}{\partial s_{2,j}} c_j(\mathbf{s}_j, t_j)] = ik_{2,j}\hat{c}(\mathbf{k}_j, t_j)$. Thus, (3.4) becomes a separable ordinary differential equation:

$$\hat{c}_{t,j}(\mathbf{k}_j, t_j) = -\bar{\mu}(k_{1,j}^2 + k_{2,j}^2)\hat{c}_j(\mathbf{k}_j, t_j) + \bar{\lambda}\hat{c}_j(\mathbf{k}_j, t_j), \tag{B.2}$$

$$\hat{c}_{t,j}(\mathbf{k}_j, t_j) = (-\bar{\mu}(k_{1,j}^2 + k_{2,j}^2) + \bar{\lambda})\hat{c}_j(\mathbf{k}_j, t_j), \tag{B.3}$$

$$\frac{\hat{c}_{t,j}(\mathbf{k}_j, t_j)}{\hat{c}_j(\mathbf{k}_j, t_j)} = -\bar{\mu}(k_{1,j}^2 + k_{2,j}^2) + \bar{\lambda}. \tag{B.4}$$

Integrating both sides of (B.4) with respect to t_j results in:

$$\log(\hat{c}_j(\mathbf{k}_j, t_j)) = (-\bar{\mu}(k_{1,j}^2 + k_{2,j}^2) + \bar{\lambda})t_j + A(\mathbf{k}_j), \tag{B.5}$$

$$\hat{c}_j(\mathbf{k}_j, t_j) = e^{A(\mathbf{k}_j)}e^{(-\bar{\mu}(k_{1,j}^2+k_{2,j}^2)+\bar{\lambda})t_j}, \tag{B.6}$$

where $A(\mathbf{k}_j)$ is an arbitrary function of $\mathbf{k}_j$, similar to a constant of integration.

Second, we transform the Gaussian kernel initial conditions from (3.3) into Fourier space:

$$\hat{c}_j(\mathbf{k}_j, t_{0_j}) = \mathrm{FT}[c_j(\mathbf{s}_j, t_{0_j})] = \theta e^{\frac{-\sigma_1^2 k_{1,j}^2}{2}} e^{\frac{-\sigma_2^2 k_{2,j}^2}{2}}. \tag{B.7}$$

Solving (B.6) when $t_j = 0$ (i.e., when $t = t_{0_j}$) reveals that $\hat{c}_j(\mathbf{k}_j, t_j = 0) = e^{A(\mathbf{k}_j)} = \hat{c}_j(\mathbf{k}_j, t_{0_j})$. Thus, we arrive at the unique solution to (3.4) given the specified initial condition in Fourier space, written:

$$\hat{c}_j(\mathbf{k}_j, t_j) = \hat{c}_j(\mathbf{k}_j, t_{0_j})e^{(-\bar{\mu}(k_{1,j}^2+k_{2,j}^2)+\bar{\lambda})t_j}. \tag{B.8}$$

Third, we invert the Fourier transform to convert (B.8) back to the physical space:

$$c_j(\mathbf{s}_j, t_j) = FT^{-1}[\hat{c}_j(\mathbf{k}_j, t_j)], \tag{B.9}$$

$$= \frac{1}{2\pi} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \hat{c}_j(\mathbf{k}_j, t_{0_j}) e^{(-\bar{\mu}(k_{1,j}^2 + k_{2,j}^2) + \bar{\lambda})t_j} e^{-i(k_{1,j}s_{1,j} + k_{2,j}s_{2,j})} d\mathbf{k}_j, \tag{B.10}$$

$$= \frac{1}{2\pi} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \theta e^{\frac{-\sigma_1^2 k_{1,j}^2}{2}} e^{\frac{-\sigma_2^2 k_{2,j}^2}{2}} e^{(-\bar{\mu}(k_{1,j}^2 + k_{2,j}^2) + \bar{\lambda})t_j} e^{-i(k_{1,j}s_{1,j} + k_{2,j}s_{2,j})} d\mathbf{k}_j, \tag{B.11}$$

$$= \frac{\theta}{2\pi} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{\bar{\lambda}t_j} e^{\frac{-\sigma_1^2 k_{1,j}^2}{2}} e^{-\bar{\mu}k_{1,j}^2} e^{-i(k_{1,j}s_{1,j})} e^{\frac{-\sigma_2^2 k_{2,j}^2}{2}} e^{-\bar{\mu}k_{2,j}^2} e^{-i(k_{2,j}s_{2,j})} dk_{1,j} dk_{2,j}, \tag{B.12}$$

$$= \frac{\theta}{2\pi} e^{\bar{\lambda}t_j} \int_{-\infty}^{\infty} e^{\frac{-\sigma_2^2 k_{2,j}^2}{2}} e^{-\bar{\mu}k_{2,j}^2} e^{-i(k_{2,j}s_{2,j})} \left(\int_{-\infty}^{\infty} e^{\frac{-\sigma_1^2 k_{1,j}^2}{2}} e^{-\bar{\mu}k_{1,j}^2} e^{-i(k_{1,j}s_{1,j})} dk_{1,j} \right) dk_{2,j}. \tag{B.13}$$

After completing the square of the integrands and evaluating the integrals in (B.13), we have:

$$c_j(\mathbf{s}_j, t_j) = \frac{\theta}{2\pi} \sqrt{\frac{1}{\sigma_1^2 + 2\bar{\mu}t_j}} \sqrt{\frac{1}{(\sigma_2^2 + 2\bar{\mu}t_j)}} e^{\bar{\lambda}t_j} e^{\frac{-(s_{1,j})^2}{(\sigma_1^2 + 2\bar{\mu}t_j)}} e^{\frac{-(s_{2,j})^2}{(\sigma_2^2 + 2\bar{\mu}t_j)}}. \tag{B.14}$$

Recall that $u_j(\mathbf{s}, t) = \frac{1}{\mu(\mathbf{s})} c_j(\mathbf{s}, t)$, $s_{1,j} = (s_1 - \omega_{1,j})$, $s_2 = (s_{2,j} - \omega_{2,j})$, $t_j = (t - t_{0_j})$, and $\sigma^2 = \sigma_1^2 = \sigma_2^2$ from assumed symmetry. Then, (B.14) may be simplified and rewritten as:

$$u_j(\mathbf{s}, t) = \frac{1}{\mu(\mathbf{s})} \frac{\theta}{2\pi(\sigma^2 + 2\bar{\mu}(t - t_{0j}))} e^{\bar{\lambda}(\mathbf{s})(t - t_{0j})} e^{\frac{-||\mathbf{s} - \boldsymbol{\omega}_j||^2}{2(\sigma^2 + 2\bar{\mu}(t - t_{0j}))}}, \tag{B.15}$$

which is the same as (3.7).

Appendix C

MCMC Algorithm to Fit BHMEDM

We use a Markov chain Monte Carlo (MCMC) method to sample from the posterior distribution of the parameters in the BHMEDM in (3.8-3.9,3.17, 3.11-3.13).

Algorithm 1: The Markov chain Monte Carlo algorithm used to sample from the posterior distribution of parameters in the BHMEDM specified in (3.8-3.9,3.17, 3.11-3.13). In this algorithm, k is the current iteration and $n.mcmc$ is the total number of iterations. We denote $\mathbf{y} \equiv (y_1, y_2, ..., y_n)'$ where the i^{th} position contains a binary indicator that the i^{th} individual has the pathogen. The vector $\mathbf{v}^k \equiv (v_1^k, v_2^k, ..., v_{J^*}^k)$ contains the indicators that determine if the j^{th} pseudo-introduction is included in the model at the k^{th} iteration. We define j' as the set of all $j = 1, 2, ..., J^*$ where $j' \neq j$. We define $\log(\theta)$ as the log-transformed initial concentration of pathogen that is common for all introductions. Whenever M-H is used, we refer to the adaptive Metropolis-Hastings algorithm (Roberts and Rosenthal, 2007; Roberts and Rosenthal, 2009).

Result: We obtain samples from $p(\mathbf{v}, \alpha_0, \boldsymbol{\alpha}, \gamma_0, \boldsymbol{\gamma}, \log(\theta), \boldsymbol{\tau}|\mathbf{y})$

Set initial values for $\mathbf{v}$, α_0, $\boldsymbol{\alpha}$, γ_0, $\boldsymbol{\gamma}$, $\log(\theta)$, and $\boldsymbol{\tau}$;

while $k < n.mcmc$ **do**

  for $j \leftarrow 1$ **to** J **do**

    Gibbs sample $(v_j^k|\mathbf{y}, \mathbf{v}_{j'>j}^{k-1}, \mathbf{v}_{j'<j}^k, \alpha_0^{k-1}, \boldsymbol{\alpha}^{k-1}, \gamma_0^{k-1}, \boldsymbol{\gamma}^{k-1}, \log(\theta)^{k-1}, \boldsymbol{\tau}^{k-1})$;

  end

  M-H sample $(\alpha_0^k, \boldsymbol{\alpha}^k|\mathbf{y}, \mathbf{v}^k, \gamma_0^{k-1}, \boldsymbol{\gamma}^{k-1}, \log(\theta)^{k-1}, \boldsymbol{\tau}^{k-1})$;

  M-H sample $(\gamma_0^k, \boldsymbol{\gamma}^k|\mathbf{y}, \mathbf{v}^k, \alpha_0^k, \boldsymbol{\alpha}^k, \log(\theta)^{k-1}, \boldsymbol{\tau}^{k-1})$;

  Gibbs sample $(\log(\theta)^k, \boldsymbol{\tau}^k|\mathbf{y}, \mathbf{v}^k, \alpha_0^k, \boldsymbol{\alpha}^k, \gamma_0^k, \boldsymbol{\gamma}^k)$;

end

For **Algorithm 1**, the full-conditional distribution of v_j at the k^{th} iteration is:

$$(v_j^k|\mathbf{y}, \mathbf{v}_{j'>j}^{k-1}, \mathbf{v}_{j'<j}^k, \alpha_0^{k-1}, \boldsymbol{\alpha}^{k-1}, \gamma_0^{k-1}, \boldsymbol{\gamma}^{k-1}, \log(\theta)^{k-1}, \boldsymbol{\tau}^{k-1}) \sim \text{Bern}(\tilde{q}_j), \tag{C.1}$$

where

$$\tilde{q}_j = \frac{q_j * p(\mathbf{y}|..., v_j^k = 1)}{q_j * p(\mathbf{y}|..., v_j^k = 1) + (1 - q_j) * p(\mathbf{y}|..., v_j^k = 0)}, \tag{C.2}$$

and

$$q_j = P(v_j = 1) \sim \text{Beta}(0.5, 0.5). \tag{C.3}$$

Here, $p(\mathbf{y}|..., v_j^k = 1) = p(\mathbf{y}|\mathbf{v}_{j'>j}^{k-1}, \mathbf{v}_{j'<j}^k, \alpha_0^{k-1}, \boldsymbol{\alpha}^{k-1}, \gamma_0^{k-1}, \boldsymbol{\gamma}^{k-1}, \log(\theta)^{k-1}, \boldsymbol{\tau}^{k-1}, v_j^k = 1)$ is the likelihood of the data from (3.8) given $v_j^k = 1$.

Likewise, $p(\mathbf{y}|..., v_j^k = 0) = p(\mathbf{y}|\mathbf{v}_{j'>j}^{k-1}, \mathbf{v}_{j'<j}^k, \alpha_0^{k-1}, \boldsymbol{\alpha}^{k-1}, \gamma_0^{k-1}, \boldsymbol{\gamma}^{k-1}, \log(\theta)^{k-1}, \boldsymbol{\tau}^{k-1}, v_j^k = 0)$ is the likelihood of the data from (3.8) given $v_j^k = 0$. We sample from the posterior distributions of α_0, $\boldsymbol{\alpha}$, γ_0, and $\boldsymbol{\gamma}$ by using the adaptive Metropolis-Hastings algorithm (Roberts and Rosenthal, 2007; Roberts and Rosenthal, 2009). We use random-walk proposal distributions that are dependent on the posterior draws from the previous step. Following Hefley et al. (2020), the full-conditional distribution of $(\log(\theta), \boldsymbol{\tau})'$ at the k^{th} iteration is:

$$p(\log(\theta), \boldsymbol{\tau}|\cdot) \sim \text{MVN}(\mathbf{Ad}, \mathbf{A}), \tag{C.4}$$

where $\mathbf{A} = ((1, \mathbf{b}_i)'(1, \mathbf{b}_i) + \boldsymbol{\Sigma}_\tau^{-1})^{-1}$, $\mathbf{d} = ((1, \mathbf{b}_i)'(\mathbf{h} - \mathbf{u}) + \boldsymbol{\Sigma}_\tau^{-1}\boldsymbol{\mu}_\tau)$, and the prior on $(\log(\theta), \boldsymbol{\tau})'$ is $\text{MVN}(\boldsymbol{\mu}_\tau, \boldsymbol{\Sigma}_\tau)$. Further, $\mathbf{h} \equiv (h_1, h_2, ..., h_n)'$ and $\mathbf{u} \equiv (\log(\sum_{j=1}^J v_j * \frac{u_j(\mathbf{s}_1, t_1)}{\theta}), \log(\sum_{j=1}^J v_j * \frac{u_j(\mathbf{s}_2, t_2)}{\theta}), ..., \log(\sum_{j=1}^J v_j * \frac{u_j(\mathbf{s}_n, t_n)}{\theta}))'$. We obtain h_i as follows:

$$p(h_i|\cdot) \propto \begin{cases} \text{TN}((1, \mathbf{b}_i)'(\log(\theta), \boldsymbol{\tau}), 1)_0^\infty & , y_i = 1 \\ \text{TN}((1, \mathbf{b}_i)'(\log(\theta), \boldsymbol{\tau}), 1)_{-\infty}^0 & , y_i = 0 \end{cases}, \tag{C.5}$$

where $\text{TN}(\cdot)$ is the truncated normal distribution. Additionally, because $\log(u(\mathbf{s}_i, t_i)e^{\mathbf{b}_i'\boldsymbol{\tau}}) = \log(u(\mathbf{s}_i, t_i)) + \mathbf{b}_i'\boldsymbol{\tau}$, and $\log(u(\mathbf{s}_i, t_i)) = \log(\theta) + \log(\sum_{j=1}^J v_j * \frac{u_j(\mathbf{s}_i, t_i)}{\theta})$, then $\log(u(\mathbf{s}_i, t_i)e^{\mathbf{b}_i'\boldsymbol{\tau}}) = \log(\sum_{j=1}^J v_j * \frac{u_j(\mathbf{s}_i, t_i)}{\theta})e^{(1, \mathbf{b}_i)'(\log(\theta), \boldsymbol{\tau})})$.