Utilizing Network Structure to Flexibly Model Areal Data

by

Michael Fredrick Christensen

Department of Statistical Science
Duke University

Defense Date: June 7, 2024

Approved:

Peter D. Hoff, Supervisor

James S. Clark

David B. Dunson

Alan E. Gelfand

Dissertation submitted in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in the Department of Statistical Science
in the Graduate School of Duke University
2024

ABSTRACT

Utilizing Network Structure to Flexibly Model Areal Data

by

Michael Fredrick Christensen

Department of Statistical Science
Duke University

Defense Date: June 7, 2024

Approved:

Peter D. Hoff, Supervisor

James S. Clark

David B. Dunson

Alan E. Gelfand

An abstract of a dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Statistical Science in the Graduate School of Duke University
2024

Abstract

The majority of statistical methods for the analysis of spatially indexed data have been developed for settings in which data are observed at a collection of point-indexed locations. Region-indexed data, also known as areal data, are common in many disciplines including economics, sociology, public health, and ecology. While statistical methods methods for the analysis of spatially correlated, point-indexed data typically characterize covariance as a function of the distances between observed locations, methods for areal data tend to characterize between-region spatial dependence using a graphical representation of the spatial domain. In most cases this has been done by defining covariance as a function of the unweighted adjacency matrix of the data, resulting in the fairly rigid assumption that all pairs of neighboring regions interact in essentially the same manner. This dissertation presents multiple methods by which the intrinsic network structure of areal data may be utilized to define more flexible models for graph and areal data than those commonly used at present. In Chapter 2 we introduce the graph deformation (GDEF) framework which implicitly defines an embedding of a graph into high-dimensional Euclidean space, wherein between-node covariance is defined using the Matérn covariance function. The model is parameterized by an unknown edge weights matrix and defines a class of covariance matrices that are both highly flexible and interpretable. We compare the GDEF model to alternatives, illustrating the advantages of the approach, before concluding with an analysis of bird abundance data for several species in North Carolina. Chapter 3 extends the work presented in Chapter 2, proposing a method by which unknown edge weights matrices may be more efficiently estimated using a basis function representation. We show how this framework improves the GDEF model as well as other existing models for areal data.

We provide several illustrations of the properties of the GDEF and other models when integrated with the proposed method, concluding with a number of simulation studies and a data analysis that demonstrate the utility of the GDEF framework. In Chapter 4 we explore how crowdsourced observations of migratory birds may be utilized to better understand avian migratory trends. We propose a novel hidden Markov model that draws from circuit theory and electrical physics to characterize its transition structure, and conclude by fitting our model to observations of the Baltimore oriole and yellow-rumped warbler within the eastern United States, with data spanning a five year period. We conclude the dissertation with a brief discussion of the work presented.

Dedication

To Josie, for being so incredibly
supportive, patient, and loving these past five years,
and to Julia, for being so incredibly cute.
I love you both.

Contents

List of Tables

List of Figures

Acknowledgements

I'm incredibly grateful for the mentorship and guidance of my advisor, Dr. Peter Hoff. He encouraged and enabled me to pursue my own research interests, was patient when I needed it, and was insightful and generous with his time throughout it all. Asking Peter to be my advisor was the best decision I have made in my academic career; no comment on where Peter's decision to accept me as a student ranks in his.

Thanks as well to Jo Eidsvik, whose proactive interest in some of my early research gave me confidence and inspired me to pursue certain ideas further than I might have otherwise. His collaboration led to significant portions of the work contained within this dissertation.

Thanks to my committee members, Alan Gelfand, David Dunson and Jim Clark, each of whom sat down with me at various points to discuss and provide insight into my research. I'd like to express specific gratitude to Alan, for encouraging my interest in spatial statistics and for mentoring my first research project as a PhD student, to David, for always indulging my questions during his lectures and whose instruction stands out among the highlights of my experience as a student at Duke University, and to Jim, for helping me to think more clearly about the ecological questions and motivations underpinning much of this work.

Others who have helped me grow as a researcher include my master's thesis advisors Shane Reese and Matt Heaton, as well as Devin Francom and Abby Nachtsheim, who mentored me during my internship at Los Alamos National Laboratory. I'd also like to thank Phil White, who not only collaborated with Alan and me on that first research project at Duke, but who was also the first person to teach me about Gaussian processes during the summer of 2015, when he was a graduating master's student and I had just begun work as an undergraduate research assistant at Brigham Young University.

I'm grateful to all of the professors from whom I've received instruction and advice over the past five years, and to the staff members of the Duke Department of Statistical Science who keep everything running. I particularly appreciate Lori Rauch, who does an incredible job making sure that no one slips through the cracks, and who always provided me with prompt and helpful guidance on the steps I needed to take while navigating my PhD studies. I'm thankful for the wonderful graduate students and postdoctoral fellows who surrounded me during my time at Duke. Special thanks goes to my office mates and close friends Betsy Bersson, Joseph Lawson and David Buch, who where there for me from qualifying exams to the present.

Thanks to Duke University, The Graduate School, and the Department of Statistical Science for funding my tuition, stipend, insurance, and travel these past five years in addition to providing the infrastructure that has allowed me and so many others to study at Duke. It is a unique and amazing privilege to be paid to be a student.

Lastly, I wish to express my love and gratitude for my family. To my parents, siblings and in-laws, whose Sunday phone calls and occasional visits were a blessing while I was on the other side of the country. To my ten-month-old daughter Julia, whose arrival didn't necessarily make completing my dissertation easier, but who has brought to my life a more consistent source of joy and fulfillment than I thought possible. Most importantly, I'm grateful to my wife Josie, who keeps me grounded, and whose love, support, companionship, and selfless willingness to at times carry an inequitable load made this possible.

1. Introduction

1.1 Why areal data?

The discipline of spatial statistics is structured around the assumption that spatially indexed random variables observed close to one another in space tend to exhibit higher correlation than those which are further apart (Cressie, 1993). The majority of statistical methods for the analysis of spatially indexed data have been developed for settings in which data are observed at a collection of point-indexed locations $\{\mathbf{s}_1,...,\mathbf{s}_n\} \in \mathcal{D}$, with $\mathcal{D}$ typically defined as some subset of $\mathbb{R}^2$. In such settings, the modeled spatial process $\{Z(\mathbf{s}) : \mathbf{s} \in \mathcal{D}\}$ is defined continuously over $\mathcal{D}$, with covariance $\text{Cov}(Z(\mathbf{s}), Z(\mathbf{s}'))$ often characterized as a decreasing function of the distances between locations $\|\mathbf{s} - \mathbf{s}'\|$.

Settings also exist in which $\mathcal{D}$ is partitioned into a collection of disjoint regions, with data observed only at the regional level. Such region-indexed data, referred to as areal data, are common in many disciplines. Political, economic, sociological and public health data are frequently available only at the county, state, or national level due to the nature of how such data are collected and reported (e.g. Arbia, 2012; Garner & Raudenbush, 1991; Jin et al., 2005, etc.). Data which were originally point-indexed or observed continuously in space may also be aggregated within a gridded partitioning of $\mathcal{D}$ and reported as areal data, as is commonly the case for climate products which are produced using a combination of on-site observations, satellite data and physical models (Christensen et al., 2019).

A partitioned spatial domain may be represented using a graph, with nodes corresponding to each discrete region of the domain and edges corresponding to each pair of adjacent or neighboring regions. Characterizing spatial dependence for areal data as a function of

distances between locations is difficult to do in a coherent manner, as illustrated by the lack of simple answer to the question "What is the distance between North and South Carolina?" Instead, models for areal data are typically defined using the graphical representation of the spatial domain's partition, with proximity characterized by the existence of edges between nodes. The simultaneous autoregressive (SAR) model of Whittle (1954) and the conditional autoregressive (CAR) model of Besag (1974) are among the earliest approaches that formally defined the distribution of areal data in terms of network structure; subsequent extensions and variants of these two models represent the majority of the statistical literature regarding spatially dependent areal data (Ver Hoef, Hanks, & Hooten, 2018), with the CAR model being particularly ubiquitous. Researchers using such autoregressive models have typically defined them using an unweighted adjacency matrix to represent network structure, leading to the implicit assumption that all pairs of adjacent regions interact in a similar manner, which may not be appropriate in many applied settings.

Considerable advances have been made in the development of models for spatially correlated point-indexed data, allowing for fast computation on large datasets (e.g. Datta et al., 2016; Gramacy & Apley, 2015; Katzfuss & Guinness, 2021), the representation of nonstationary covariance structures (e.g. Lindgren et al., 2011; Paciorek & Schervish, 2006; Sampson & Guttorp, 1992) and irregular and non-Euclidean spatial domains (e.g. Yang & Dunson, 2016). Gelfand et al. (2001) proposes representing areal random variables as the integral of a continuous spatial process over discrete spatial regions, presenting a framework that allows areal data to be represented using models for point-indexed data, and which mitigates issues arising from model sensitivity to the spatial partition considered. Alternatively, one can simply treat areal data as point-indexed data observed at the centroids of each spatial region, though this is considerably less appropriate when regions are of irregular shapes and sizes.

Given the limitations of existing models for areal data (which are discussed further in later chapters) and the advances mentioned above, it is reasonable to ask whether it makes more sense to simply default to using methods developed for point-indexed data in

all settings, making whatever assumptions and adjustments are necessary to adapt to areal observations. Why should we care about models constructed specifically for areal data? Discussion of this question has existed in some form for over 50 years (Avadhani, 1969), and an exhaustive treatment of it would necessarily wind up being both highly theoretical and philosophical; such discussion ultimately resides outside of the immediate purview of the work presented in this dissertation. We do however wish to highlight a few potential answers and starting points to the question of why there is value in exploring the space of models for areal data.

A trivial but significant argument for why areal data models are necessary is that areal data exists! Independent of the situational validity of the assumptions necessary to treat areal data as discretized realizations of an underlying continuous process, it is useful from a practical perspective to have tools designed to model data that cannot straightforwardly be associated with specific geographic coordinates (such as county-level reports of COVID-19 infections, or statewide GDP levels). If such models are to be used, we would hope that modeling options exist that are computationally efficient and flexible enough to represent a range of spatial dependence patterns that may arise in real world settings. By extension, the fact that many contemporary methods used for point-indexed data are presently more advanced than methods used to model areal data simply means that there is currently more "low-hanging fruit" and potential for meaningful improvements within the research space of methods for areal and graphical data.

Another important consideration is whether the spatial partition over which areal data are observed is intrinsically meaningful. There is likely little practical or philosophical reason why the areal nature of data recorded over gridded regions should be considered sacrosanct. The discrete structure of such data may make use of a CAR model convenient, but it is reasonable to model such data as though it arises from a continuous spatial process. On the other hand, the distribution of some random variables may be deeply affected by political boundaries due to differences in governance, economy and infrastructure between neighboring regions. Consider for example that many social and economic outcomes for

residents of El Paso, Texas are more highly correlated with residents of other Texas cities than with residents of the bordering city of Ciudad Juárez, Mexico (McDonald & Grineski, 2012). While some differences between cities, counties or states may be accounted for through use of covariates, the total impact of governmental and economic policy on certain outcomes may be extremely complex. Even if one could perfectly account for qualitative differences between regions using some covariate-based model, it is important to note that the borders between adjacent regions may themselves have intrinsic meaning. Borders between states and countries often follow rivers or mountain ranges, and these features may affect the extent to which observations on either side of that border are correlated. Interactions between animal populations on either side the US-Mexico border for instance are impacted by the natural barrier of the Rio Grande river as well as the artificial barrier of the US-Mexico border wall (Schlyer, 2012). While methods for point-indexed data in continuous space may be constructed to account for the existence of hard or soft boundaries within $\mathcal{D}$ (e.g. Dunson et al., 2022), the graphical nature of areal data allows for explicit representation of these shared boundaries in the form of edges. In the chapters that follow, we find considerable utility in using variable edge weights within our models which allows for differing levels of connectivity between adjacent regions.

A final consideration deals with data observed on more general classes of graphs and not only simple partitions of $\mathcal{D}$. Areal data typically has spatial structure that can be represented using a planar graph, which is to say that it has a natural embedding in $\mathbb{R}^2$ (Jungnickel, 2012). Given spatially correlated data observed at the nodes of a non-planar graph, there is no obvious or straightforward way to characterize between-node covariance using methods developed for point-indexed data. This is relevant for settings in which the spatial relationships between geographic locations can be represented using a multiplex graph, as was the case in Strano et al. (2015) which considered how street and subway networks interact to characterize the movement of people in New York and London.

1.2 Chapter overview

The body of this dissertation is composed of three articles, each of which consider ways in which the graphical representation of a partitioned spatial domain may be utilized to define models for areal data that exhibit greater flexibility than many commonly used alternatives. The methods presented were developed with the importance of interpretive value in mind, and may be used to gain insight into the underlying data generating processes.

Chapter 2 is based on the work presented in Christensen and Hoff (2023). As previously stated, many models for areal data are constructed such that all pairs of adjacent regions are implicitly assumed to have near-identical spatial autocorrelation, but in practice, data can exhibit dependence structures more complicated than this assumption allows. Within the chapter, we develop a new model for spatially correlated data observed on graphs, which can flexibly represented many types of spatial dependence patterns while retaining aspects of the original graph geometry. The method presented implies an embedding of the graph into Euclidean space wherein covariance can be modeled using traditional covariance functions, such as those from the Matérn family (Matérn, 1960). The model is parameterized using a class of graph metrics compatible with such covariance functions, and which characterize distance in terms of network flow, a property useful for understanding proximity in many ecological settings (Thiele et al., 2018). By estimating the parameters underlying these metrics, we recover the "intrinsic distances" between graph nodes, which assist in the interpretation of the estimated covariance and allow us to better understand the relationship between the observed process and spatial domain. We compare our model to existing methods for spatially dependent graph data, primarily conditional autoregressive models and their variants, and illustrate advantages of our method over traditional approaches. We fit our model to bird abundance data for several species in North Carolina, and show how it provides insight into the interactions between species-specific spatial distributions and geography.

Chapter 3 is based on work from Christensen and Eidsvik (2024). We demonstrate

in Chapter 2 that complex spatial structures are better represented when the relationship between graph nodes is characterized by a variable edge weights matrix instead of the traditionally utilized unweighted adjacency matrix. We note however, that a potential issue with the method in Chapter 2 is that each edge weight is treated as an individual parameter, resulting in increasingly challenging parameter estimation as graph size increases. In Chapter 3 we propose a framework for estimating edge weight matrices that reduces their effective dimension via a basis function representation. We show that this method may be used to enhance the performance and flexibility of covariance models parameterized by such matrices in a series of illustrations, simulations and data examples.

Chapter 4 presents work from Christensen and Hoff (2024) and contains an applied data analysis for county-level observations of migratory birds. While the general patterns governing avian migration are well documented, there are significant questions pertaining to the phenomenon that remain poorly understood. Studies of animal movement are often conducted using telemetry data, but the collection of such data is time- and resource-intensive, which can result in small sample sizes (Hebblewhite & Haydon, 2010). Ecological surveys of animal populations are also indicative of species distribution trends, but may be constrained to a limited spatial domain (Caughlan & Oakley, 2001). Within Chapter 4, we propose utilizing crowdsourced observations from the eBird database (Sullivan et al., 2009) to model the abundance of migratory bird species in space and time. We introduce a hidden Markov model for observed bird counts utilizing a novel transition structure developed using principles from circuit theory. After illustrating model properties we fit it to observations of Baltimore orioles and yellow-rumped warblers within the eastern United States and discuss insight it provides into the migratory patterns for these species. We conclude the dissertation with Chapter 5, which contains a brief discussion of findings from previous chapters along with avenues for future work.

2. A Flexible and Interpretable Spatial Covariance Model for Data on Graphs

2.1 Introduction

Models for spatially indexed data traditionally assume that observations close to one another in space are more highly correlated than observations which are far apart. While models for point-indexed data within continuous space frequently characterize spatial correlation as a function of geographic distance between observations, models for area-indexed data, also known as areal data, tend to define spatial dependence as a function of the adjacency structure of the graphical representation of the spatial domain's partition (Ver Hoef, Peterson, et al., 2018). Such models often assume that the correlation between observations made at adjacent regions is the same throughout the network, an assumption that is often inappropriate for real world data (Guttorp et al., 1994).

As an example of this we consider a simple exploratory data analysis made using data which was downloaded and processed from the eBird database. This database is a project managed by the Cornell Lab of Ornithology that crowd-sources bird observation data by allowing any user to record and submit their bird watching observations to the website. Observations are recorded in the form of user-submitted, location-indexed checklists containing the counts of each species observed by the bird watcher along with the time spent bird watching at each location (Sullivan et al., 2009). The data used here consist of the county-level number of combined observations of laughing gulls, a species common along the Atlantic coast, made in eastern North Carolina during each month between January 2018 and December 2020. During this time period, there were nearly 350,000 individual

checklists submitted within the state of North Carolina, representing over 400,000 combined hours of bird watching.

To evaluate whether the degree of autocorrelation is consistent across our spatial domain we evaluate the month-to-month correlation in the data between pairs of adjacent counties. To account for the potential impact of environmental covariates on this simple analysis we restrict our attention the twenty coastal counties (as designated by the Coastal Area Management Act (1974)) of North Carolina. Figure 2.1 depicts the monthly log rate of observation for three pairs of adjacent counties, with each point corresponding to the frequency of observation at a pair of counties within a given month. If the assumption of constant spatial autocorrelation between adjacent counties were correct, then the correlations between adjacent counties should be approximately equal along the entirety of the coastline. As can be seen in the figure, the log rates of laughing gull observations in the adjacent Dare and Currituck counties are more correlated (0.84) than those of Onslow and Pender counties (0.24). To test this more more rigorously, we computed the between-county correlations for all 39 pairs of adjacent coastal counties. For each of those pairs, we obtained the statistic z_i^ρ given by Fisher transformation of the Pearson correlation coefficient (known to be approximately normal (Fisher, 1921)) using the monthly log rates of observations at each county. An F-test of $H_0 : \rho_1 = ... = \rho_{39}$ based on the statistics $\{z_i^\rho\}_{1:39}$ resulted in $p < 0.01$, indicating significant evidence of differences in the correlations between adjacent coastal counties, and implying that standard approaches to modeling areal data may fail to reflect this.

The majority of statistical methods designed to account for changing patterns of spatial dependence within a domain have been developed specifically for point referenced data. However, areal data, such as the county-level eBird data discussed in our example, are also common within many applications, either because some random variables are observed only at a regional level, or because raw point-referenced data might be aggregated and processed into areal or gridded regions out of convenience or due to scientific or policy interest in questions at the county, state, or national level. Such data are representable using graphs,

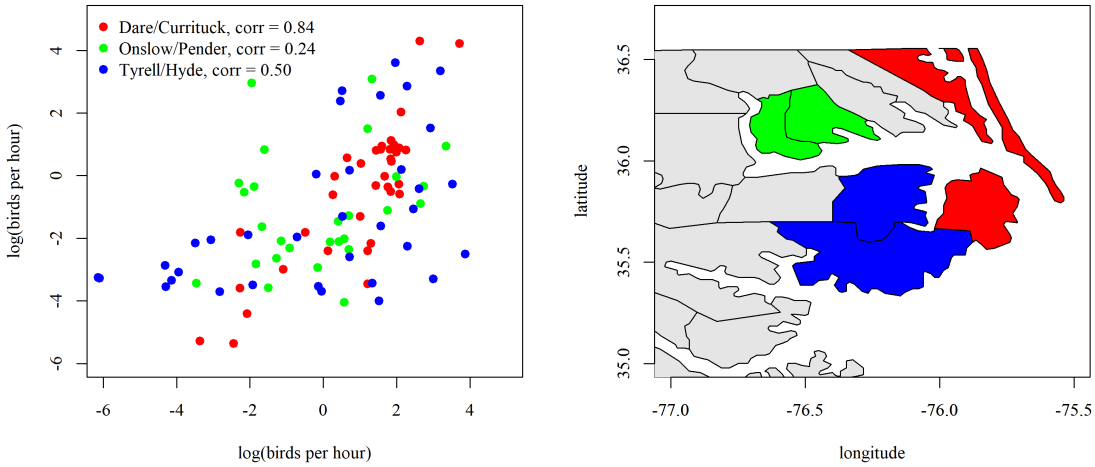


FIGURE 2.1: Three color-coded pairs of counties in North Carolina. Points correspond to the rate of laughing gull observation within a pair of counties during a month. Dare and Currituck counties (red) exhibit stronger correlation than Onslow and Pender counties (green) or Tyrell and Hyde counties (blue).

with a node representing each region and edges connecting each pair of adjacent regions.

Conditional autoregressive (CAR) models, first introduced by Besag (1974), are commonly used for modeling spatially dependent data observed on a graph, but are potentially limited when faced with more complex spatial dependence patterns. Given a random vector $\mathbf{y}$ with $y_1,...,y_p$ being observations at each of the p nodes (or vertices) of some graph $G = (V,E)$, where V contains the set of nodes, and E is set of edges between those nodes, a CAR model specifies a Markov random field in which each element of $\mathbf{y}$ has a conditional distribution determined by the values at neighboring nodes, for example

$$y_i|\mathbf{y}_{-i} \sim N\left(\sum_{j\sim i}\kappa w_{ij}y_j, \sigma^2\right) \tag{2.1}$$

where $j \sim i$ indicates that node j is adjacent to node i, $\mathbf{W}$ is a $p \times p$ adjacency or weights matrix (where $w_{ij} > 0$ if $j \sim i$ else $w_{ij} = 0$), κ controls the degree of spatial correlation (with $|\kappa| \leq 1$) and $\sigma^2 > 0$ is a variance parameter. This results in a marginal distribution for $\mathbf{y}$ of

$$\mathbf{y} \sim N(\mathbf{0}, \sigma^2(\mathbf{I}_n - \kappa\mathbf{W})^{-1}). \tag{2.2}$$

The CAR model may be thought of as the spatial generalization of the first order autoregressive (AR1) model often used in time series modeling. There are many varieties of CAR model that have been developed for use in generalized mixed models (Besag et al., 1991), with multivariate data (Gelfand & Vounatsou, 2003), and which can utilize weighting schemes incorporating geographic and covariate information (Ejigu & Wencheko, 2020; Hanks & Hooten, 2013). Most of the literature regarding spatially dependent graph data is limited to various forms of autoregressive models (Ver Hoef, Hanks, & Hooten, 2018). The majority of CAR models and their variants tend to to treat all first-order neighbors within a graph as equally proximate due to the use of graph adjacency structure alone in their definition of spatial dependence. This will rarely result in effective representation of more complex spatial processes. In instances where more complex weighting schemes have been utilized, such as in Hanks and Hooten (2013), and Ejigu and Wencheko (2020), edge weights are defined as linear functions of a small set of environmental covariates. While such approaches allow for more flexibility than those using adjacency alone, the space of possible weighting matrices under those models is restricted by the column space defined by the collection of environmental covariates being used. There are also instances in which covariates associated with graph edges may not be available to researchers, in which case a model must be defined using only the structure of the graph itself. One computational advantage of the CAR modeling framework is that the induced precision matrix will generally be sparse due to the assumption of independence for each observation conditional on the values of its neighbors. This assumption can come with trade-offs in terms of predictive efficiency; Gramacy and Apley (2015), Katzfuss and Guinness (2021), Guinness (2018) all demonstrate the shortcomings of models which condition only on local or nearest neighbor structure (as is the case in CAR models) in comparison to those which also condition on more distant observations. For more discussion on CAR models see Ver Hoef, Peterson,

et al. (2018) and D. Lee (2011) for review papers which provide detail on the use, development, and interpretation of these models and their variants. To address the concerns regarding limitations of CAR models for use with more complex spatial processes, we propose a model for the spatial covariance of graph data using a different set of assumptions, which we will describe in the following sections. We will also illustrate that our proposed method has greater capacity to represent certain correlation structures that may be of interest to statisticians and ecologists, especially in contrast to the most commonly used CAR variants.

Another approach to modeling the spatial random effects that may be present in areal data include basis function-based methods such as that of Hughes and Haran (2013), though such approaches are generally structured more around the estimation of the spatial random effects themselves rather than the covariance structure that produced them. Additionally, such approaches often produce overly-smooth spatial random effects surfaces unless a very large number of basis functions are employed.

Our goal within this article is not only to present a new covariance model for areal data that is flexible enough to account for varying patterns of spatial dependence across a domain, but to ensure that said model has interpretive value as well. The relationship between distance and spatial autocorrelation is often explicitly defined in the modeling of data observed within continuous space, but this relationship is complicated in the areal data setting, where the definition of a metric between regions is less straightforward (Cressie, 1993). When considering our earlier example, it may be reasonable to imagine that from the perspective of a laughing gull within a coastal county, the changing environment as one moves inland from the coast may act as a barrier to movement, and that as a result, a neighboring coastal county is in some senses “closer" than a neighboring inland county. This proximity should be reflected in the data by a greater degree of spatial correlation between regions with smaller “intrinsic distances." We could explicitly represent this relationship by parameterizing a unique distance metric characterizing the intrinsic distances between all counties within our spatial domain, and defining the covariance between counties as a simple

function of this metric. This approach may be viewed as similar to the class of methods for modeling nonstationarity in continuous space in which spatial covariance is modeled as a stationary and isotropic function of a distance metric that arises from an estimated transformation of the geographic space in which the data were originally observed. See Sampson and Guttorp (1992) and Schmidt and O'Hagan (2003) and Bornn et al. (2012) for examples of this approach.

Within this article we present a covariance model for graph data built under the idea that complex patterns of spatial dependence can be represented using a simple covariance function applied to a latent matrix of intrinsic distances between graph nodes. Fitting the model involves the estimation of a distance metric over our graph which in turn corresponds to an embedding of our graph into a high-dimensional Euclidean space. Covariance is then assumed to be a function of the distances between nodes in the embedding. In addition to being more flexible than most CAR model-based approaches, we find that intrinsic distance framework makes our model simple and intuitive to interpret. In the next section of this article we establish some of the necessary background information regarding covariance functions and graph metrics. In particular, we highlight a class of graph metrics which are compatible with most traditional covariance functions such as those from the Matérn family (Matérn, 1960). In Section 3 we formally introduce our model, and provide some detail on its interpretation as well as its implementation. In Section 4 we provide a case study using the eBird data discussed in the Introduction. We fit our model to bird abundance data for several species common to the state of North Carolina, and demonstrate the manner in which our model can provide valuable insight into the interaction between a spatial process and the underlying geography via the posterior distance matrix of the graph on which a process is observed. We conclude this article with a discussion of our model's strengths and limitations in comparison to existing methods, as well as potential extensions for future work.

2.2 Background

Suppose we are given a graph $G = (V, E)$ and random variables $\{Y_v : v \in V\}$ observed at the nodes V of G. In this article we propose a model for $\Sigma_{vv'} = \mathrm{Cov}(Y_v, Y_{v'})$ which is described visually in Figure 2.2. The covariance matrix Σ is parameterized by $\mathbf{W}$, a matrix of unknown positive edge weights, and $\sigma^2 > 0$, a scale parameter. The parameter $\mathbf{W}$ is an element of $\mathcal{W}_G$, the space of all possible edge weight matrices given G, and $\mathbf{W}$ is mapped to a distance matrix $\mathbf{D}$ via a graph metric $\Delta(\cdot)$. $\mathbf{D}$ is in turn mapped to a symmetric positive definite matrix Σ using a covariance function $C(\cdot)$ which also takes the parameter $\sigma^2 > 0$ as input. In the following two subsections, we establish the background information necessary to justify our model before presenting it formally in Section 3.

$$\mathbf{W} \in \mathcal{W}_G \xrightarrow{\Delta} \mathbf{D} \xrightarrow{C} \Sigma \qquad \sigma^2 > 0$$

FIGURE 2.2: Visual representation of the model presented in this article. The edge weight parameter $\mathbf{W}$ is mapped to a distance matrix, which in conjunction with the scale parameter σ^2 is mapped to the covariance matrix Σ.

2.2.1 Covariance functions and distance

Let $\mathcal{D}_p$ be the space of metric distance matrices between p points. That is, if $\mathbf{D} \in \mathcal{D}_p$ then $\mathbf{D}$ is a $p \times p$ matrix where all diagonal elements of $\mathbf{D}$ are zero, the off-diagonal entries are positive, $\mathbf{D}$ is symmetric, and all elements of $\mathbf{D}$ satisfy the triangle inequality (i.e. for any $\{i, j, k\} \subset \{1 \ldots p\}$ $d_{ij} \leq d_{ik} + d_{kj}$). In our model we define the $p \times p$ covariance matrix Σ as $\Sigma = C(\mathbf{D})$, for $\mathbf{D} \in \mathcal{D}_p$, where $C(\cdot)$ is a covariance function which maps distance matrices to $\mathcal{S}_p^+$, the space of $p \times p$ symmetric positive definite matrices. To allow for flexibility in the covariance, we will treat the distance matrix of a graph as an unknown parameter to be estimated from the data. This in contrast to most spatial models, which take the distances between observations as known and fixed. Instead we will parameterize distance as a function of edge weights on the graph, which are estimated from the data. This

approach allows us define covariance structures for graph data that are considerably more complex than ones obtained using the neighborhood structure of the graph alone.

Covariance functions defined only by the distances between locations, and not the locations themselves, are said to be stationary and isotropic. These covariance functions are often defined such that they can be applied element-wise to a distance matrix. For $\mathbf{D} \in \mathcal{D}_p$ we can also define Σ by $\Sigma_{ij} = c(d_{ij})$ where $c(\cdot)$ is some function applied to a distance between two points. For Σ to be a valid covariance matrix it must be symmetric and positive definite, and thus $c(\cdot)$ must be chosen carefully with respect to $\mathbf{D}$. There are many well known covariance functions that can be used to produce positive definite matrices using distances as inputs. Functions from the Matérn covariance family, originally developed by Matérn (1960) and given by

$$c_\nu(d) = \sigma^2 \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\sqrt{2\nu} \frac{d}{\tau} \right)^\nu K_\nu \left(\sqrt{2\nu} \frac{d}{\tau} \right) \tag{2.3}$$

are among the most commonly utilized within spatial statistics (Banerjee et al., 2003; Cressie, 1993). Here the input d is a distance between two points, σ^2 is a scale parameter, τ is a range parameter, ν is a smoothness parameter, and K_ν is a modified Bessel function of the second kind of order ν. The Matérn family includes the commonly used exponential and Gaussian (also known as the radial basis kernel or squared exponential) covariance functions (Banerjee et al., 2003). If $c(\cdot)$ is a member of the Matérn family of covariance functions it is necessary for the distances contained in $\mathbf{D}$ to be Euclidean in order to ensure that Σ is a positive definite matrix (Ver Hoef, 2018). Here we define the space of Euclidean distance matrices $\mathcal{D}_p^E \subset \mathcal{D}_p$ such that if $\mathbf{D} \in \mathcal{D}_p^E$, there exists $\{\mathbf{x}_1, ..., \mathbf{x}_p\} \subset \mathbb{R}^k$ for some k such that $d_{ij} = \|x_i - x_j\|$. In our model, for a given graph $G = (V, E)$ with p nodes we will require that $\mathbf{D}$, the matrix containing pairwise distances between the nodes of G, satisfies $\mathbf{D} \in \mathcal{D}_p^E$ for some p. This also implies that there exists an embedding of G into a Euclidean space such that the graph distances are equal to the distances between nodes in the embedding.

The fact that many of the most common covariance functions (including those from the Matérn family) require Euclidean distances as inputs to guarantee positive definiteness

is reasonably well known, and can be confirmed when one considers Bochner's theorem, which provides a necessary and sufficient condition for functions on Euclidean vectors to be positive definite (see Gihman and Skorohod (1974)). However, this requirement has at times been disregarded by researchers who were interested in utilizing common spatial modeling tools with non-Euclidean distances, which can result in non-positive definite covariance matrices and negative prediction variances in applications. Such scenarios arise especially often in analyses of graph and network data, where the most commonly used metrics, such as the shortest path distance, often result in non-Euclidean distances. Ver Hoef (2018) discusses this issue and provides several examples of published articles which failed to consider the non-compatibility of their chosen covariance functions and distance metrics.

2.2.2 Graph metrics

To parameterize the covariance of random variables observed on a graph by applying the Matérn covariance function to a matrix of estimated distances, we must be careful that $\mathbf{D}$ is parameterized in a manner that ensures that it is Euclidean over the entirety of the parameter space. To do this, we must consider what graph metrics produce Euclidean distances and how their properties will influence model interpretation.

Given an undirected, connected, simple graph $G = (V, E)$ with p nodes, we define a class of distance matrices parameterized by the symmetric edge weights matrix $\mathbf{W}$ of G. Many graph metrics are functions of $\mathbf{W}$ where $w_{ij} > 0$ if node i is adjacent to node j and $w_{ij} = 0$ otherwise (Jungnickel, 2012). We define $\mathcal{W}_G$ to be the space of all symmetric edge weights matrices possible given G, and note that all information regarding the edges of G is contained in each $\mathbf{W} \in \mathcal{W}_G$ based on whether or not an element of $\mathbf{W}$ is equal to zero. The distance matrix $\mathbf{D}$ for a graph G is given by $\mathbf{D} = \Delta(\mathbf{W})$ where $\Delta(\cdot)$ is the function for a particular metric which maps an edge weight matrix to a distance matrix.

For a given graph G and metric $\Delta(\cdot)$ we define $\mathcal{D}_G^{\Delta}$ as the space of distance matrices which are possible to obtain under different edge weight configurations for G, that is $\mathcal{D}_G^{\Delta} = \{\mathbf{D} = \Delta(\mathbf{W}) : \mathbf{W} \in \mathcal{W}_G\}$. We have established that for $\Sigma = C(\mathbf{D})$ to be a valid covariance

matrix with $C(\cdot)$ a Matérn family covariance function, $\mathbf{D}$ must be a Euclidean distance matrix. Thus, in choosing a graph metric $\Delta(\cdot)$, we must ensure that $\mathcal{D}_G^{\Delta} \subset \mathcal{D}_p^E$. In doing so, we can specify a parameterization of $\boldsymbol{\Sigma}$ in terms of an unobserved edge weights matrix $\mathbf{W} \in \mathcal{W}_G$ with the property that $\boldsymbol{\Sigma} = C(\Delta(\mathbf{W})) \in \mathcal{S}_p^+$ for all $\mathbf{W} \in \mathcal{W}_G$. This also implies that because $\mathbf{D} = \Delta(\mathbf{W}) \subset \mathcal{D}_p^E$, $\mathbf{W}$ characterizes an embedding of G into Euclidean space. In Section 3 we establish that for an appropriate choice of metric, our covariance model is identifiable, which implies that the embedding of G implied by $\mathbf{W}$ is unique up to isometry.

The most commonly used graph metric is the shortest path distance (Chebotarev, 2011) (denoted here as $\Delta^{SP}(\cdot)$) which defines the distance between nodes i and j as

$$d_{ij} = \Delta^{SP}(\mathbf{W})_{ij} = \min_{p_{ij} \in P_{ij}} \left(\sum_{e_{lk} \in p_{ij}} w_{lk} \right) \tag{2.4}$$

where P_{ij} is the set of all possible paths between nodes i and j, but can intuitively be described as the sum of the weights along the "shortest" path connecting those nodes. However, this metric generally produces non-Euclidean distance matrices (Klein & Zhu, 1998) and as such $\boldsymbol{\Sigma} = C(\mathbf{D})$ is not guaranteed to be positive definite if $C(\cdot)$ is Matérn.

A commonly used alternative graph metric is the resistance distance (Chebotarev, 2011), denoted $\Delta^{\Omega}(\cdot)$, where the distance between nodes i and j is defined as

$$d_{ij} = \Delta^{\Omega}(\mathbf{W})_{ij} = (\mathbf{e}_i - \mathbf{e}_j)^{\top} \mathbf{L}^{+} (\mathbf{e}_i - \mathbf{e}_j) \tag{2.5}$$

where $\mathbf{L}^{+}$ is the Moore-Penrose generalized inverse of the graph Laplacian, defined as $\mathbf{L} = \text{diag}(\mathbf{W}\mathbf{1}_p) - \mathbf{W}$ where $\text{diag}(\cdot)$ is a function taking a p-length vector and returning a $p \times p$ diagonal matrix, $\mathbf{1}_p$ is a vector of ones, and $\{\mathbf{e}_i\}_{1:p}$ are the standard basis vectors. This metric was introduced to the mathematical literature by Klein and Randić (1993) with roots in electrical physics and is based on the formula for calculating the effective resistance between nodes in a network of resistors.

Attractive properties of resistance distance include the fact that the distance between

two nodes is reduced by the number of paths connecting said nodes, a property that shortest path distance does not possess (Klein & Randić, 1993), and that it is proportional to commute time distance, i.e. the expected hitting time of a random walk between two nodes (Chandra et al., 1996). Use of resistance distance is widespread within ecological settings as both of the above properties characterize a type of flow-based graph connectivity which corresponds nicely with existing models for the movement of individual organisms as well as animal populations (Peterson et al., 2019; Thiele et al., 2018).

Because $\mathbf{L}^+$ is always a positive semi-definite matrix, resistance distance may be viewed as a squared Euclidean (or Mahalanobis) distance. Klein and Zhu (1998) highlight that a related class of metrics of the form

$$d_{ij} = \Delta^{(m)}(\mathbf{W})_{ij} = \sqrt{(\mathbf{e}_i - \mathbf{e}_j)^\top \{\mathbf{L}^+\}^m (\mathbf{e}_i - \mathbf{e}_j)} \tag{2.6}$$

where $m > 0$ is an exponent (i.e $\{\mathbf{L}^+\}^2 = \mathbf{L}^+\mathbf{L}^+$), are guaranteed to produce Euclidean distance matrices for any $m > 0$ and all $\mathbf{W} \in \mathcal{W}_G$. Thus, all metrics of this form can be used in conjunction with any covariance function which requires Euclidean distances as inputs, such as those from the Matérn class. The square root of resistance distance is one of these metrics (when $m = 1$), but we are particularly interested in using the second metric of this form (when $m = 2$)—often referred to as the quasi-Euclidean metric—because it shares many of the same properties regarding graph connectivity with resistance distance and has been demonstrated to be well behaved in comparison to other metrics, especially when used with planar graphs which are typically seen in spatial and ecological settings (Ivanciuc et al., 2001; Zhu & Klein, 1996). The quasi-Euclidean metric also scales linearly with the inverse of the edge weights, that is, $c\mathbf{D} = \Delta^{(2)}(\mathbf{W}/c)$ for $c > 0$. This property is convenient for computation and enables greater parameter interpretability. Subsection 3.2 contains a numeric illustration of how the quasi-Euclidean metric operates within our model.

2.3 Method

2.3.1 A flexible graph covariance model

Given an undirected, connected, simple graph $G = (V, E)$ with p nodes, we make repeated observations of the collection of random variables $\{Y_v : v \in V\}$; each repetition is recorded as $\mathbf{y}_i \in \mathbb{R}^p$ for $i \in 1, ..., n$. These observations may be stored in the $n \times p$ data matrix $\mathbf{Y}$, where y_{ij} is the ith observation at the jth node of G. We define Σ, the between-node covariance of each $\mathbf{y}_i$, in terms of the intrinsic distances between the nodes of G which are in turn a function of the known structure of the graph (V, E) and unobserved edge weight parameter matrix $\mathbf{W}$, which is estimated from the data. We highlight that we consider the structure of the graph G to be fixed. This stands in contrast to methods such as those presented in White and Ghosh (2009) and Ma et al. (2010) which treat the adjacency structure itself as something unknown to be estimated. Our model instead estimates the unknown edge weights for a fixed graph, though in some senses our approach has the ability to omit edges from the model by estimating their weights to be near zero. This choice may be particularly justified in settings involving animal and plant populations, where interaction or movement between non-physically adjacent regions can only occur through intermediate regions. Our parameterization of Σ could be used in conjunction with any number of distributional assumptions for $\mathbf{Y}$, including temporal dependence or non-normality, but to make presentation concrete we initially define our model using the assumption that $\{\mathbf{y}_i\}_{1:n}$ are independent draws from a $N(\mathbf{0}_p, \Sigma)$ distribution. We will discuss possible extensions later in the article. We parameterize Σ as follows:

$$\begin{aligned} \Sigma_{jk} &= \sigma^2 \rho_\nu(d_{jk}) \\ d_{jk} &= \sqrt{(\mathbf{e}_j - \mathbf{e}_k)^\top \{\mathbf{L}^+\}^2 (\mathbf{e}_j - \mathbf{e}_k)} \\ \mathbf{L} &= \mathrm{diag}(\mathbf{W}\mathbf{1}_p) - \mathbf{W} \\ \sigma^2 &> 0,\ w_{jk} > 0 \text{ if } j \sim k, \text{ else } w_{jk} = 0, \text{ and } w_{jk} = w_{kj}. \end{aligned} \tag{2.7}$$

Under this model, the data $\mathbf{Y}$ comes from a mean-zero matrix-normal distribution with column covariance Σ and row covariance $\mathbf{I}_n$. The parameter $\mathbf{W}$ is a $p \times p$ symmetric edge weight matrix, σ^2 is a scale parameter, and $\rho_\nu(\cdot)$ is the Matérn correlation function given below:

$$\rho_\nu(d) = \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\sqrt{2\nu}d\right)^\nu K_\nu \left(\sqrt{2\nu}d\right). \tag{2.8}$$

Common practice is to take the smoothness parameter ν as fixed rather than as a parameter to be estimated. In this article we set $\nu = 3/2$, which implies a spatial process that is smoother than those given by an exponential covariance function, but less smooth than those defined with Gaussian covariance; however, any value of $\nu > 0$ would be valid in this model (Stein, 1999).

Note that the correlation function defined above does not include an explicit range or spatial-decay parameter. Equation 2.3, the standard form of the Matérn covariance function, contains the parameter τ whereas the correlation function we are using, given in Equation 2.8, does not. Despite the lack of a parameter τ, which directly scales the distances used as inputs in the correlation function, our parameterization of Σ implicitly controls for the rate of spatial decay thru the edge weight matrix $\mathbf{W}$: Let $\mathbf{D}$ be the $p \times p$ distance matrix produced by applying the quasi-Euclidean metric to the edge weight matrix $\mathbf{W}$. The third and fourth lines of Equation 2.7 show how $\mathbf{D}$ is obtained: let $\Delta^{(2)}(\cdot)$ be the function such that $\mathbf{D} = \Delta^{(2)}(\mathbf{W})$. Recall that $c\mathbf{D} = \Delta^{(2)}(\mathbf{W}/c)$ for any positive constant c. The distances in our model are directly scaled by the magnitude of the edge weights, which are themselves unrestricted above zero rendering an additional parameter τ where correlation is a function of d/τ unnecessary. To interpret our model's estimate of the spatial decay in the covariance of $\mathbf{Y}$, we recommend re-scaling the distance matrix $\mathbf{D}$ after fitting. If we define $\mathbf{D}^s = \mathbf{D}/\max(\mathbf{D})$ and $\tau^s = 1/\max(\mathbf{D})$, then the covariance defined by $\Sigma_{jk} = \rho_\nu(d^s_{jk}/\tau^s)$ is equal to the one defined $\Sigma_{jk} = \rho_\nu(d_{jk})$, and τ^s has equivalent interpretation to a range parameter in a covariance model for point-referenced data in which the geographic distances

between observations were scaled to have a maximum distance of one.

As parameterized, our model is identifiable:

Proposition 1 (Identifiability). *Using the parameterization for* Σ *given in Equation 2.7,* $(\sigma_1^2, \boldsymbol{W}_1) \neq (\sigma_2^2, \boldsymbol{W}_2)$ *implies* $\Sigma_1 \neq \Sigma_2$ *for all* $\sigma_1^2, \sigma_2^2 > 0$ *and all* $\boldsymbol{W}_1, \boldsymbol{W}_2 \in \mathcal{W}_G$.

A proof of the model's identifiability is provided in Appendix A.

2.3.2 Interpretation

Rather than characterizing spatial covariance via a complex covariance function or kernel, our model uses a relatively simple covariance function, and pushes most of the modeling complexity into the estimation of latent distances between graph nodes, which are themselves functions of a set of edge weight parameters; nodes that are close to one another are more correlated, while nodes that are far apart are less correlated. The metric we utilize to define distances over the graph was chosen to ensure that the resulting covariance matrix will be positive definite for any combination of edge weight parameters, but what of the edge weight parameters themselves? Generally speaking, larger edge weight parameters indicate greater correlation between nodes in the regions of the graph where that edge is found, while smaller edge weight parameters indicate lower correlation in those regions. When edge weights approach zero, the corresponding edges are effectively omitted from the graph, which may be thought of as a type of automatic model selection for the design of the graph itself.

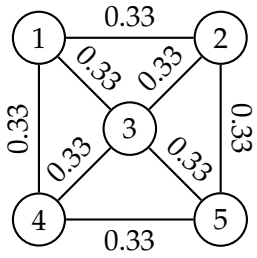
Figure 2.3 contains an illustration of how edge weights impact covariance in our model. It depicts four sets of edge weights for the same five-node graph and shows the resulting correlation matrix when a Matérn correlation function with $\nu = 3/2$ is applied to the distance matrix produced by the edge weights. Subfigures 2.3a and 2.3b highlight how the scale of edge weights matter, with larger edge weights producing greater between-node correlation. Subfigure 2.3c demonstrates the potential flexibility of our model when certain edge weights are larger than others, with nodes along the path from 4 to 1 to 2

exhibiting greater intercorrelation than the rest of the graph. Subfigure 2.3d shows that it is possible under our model to obtain covariance structures where non-adjacent nodes (1 and 5) are more correlated (in effect closer together) than any other pair of first-order neighbors, which is possible due to the properties of the quasi-Euclidean metric, for which the distance between non-adjacent nodes is decreased with an increase in the number of sufficiently weighted connecting paths.

2.3.3 Comparison to CAR models

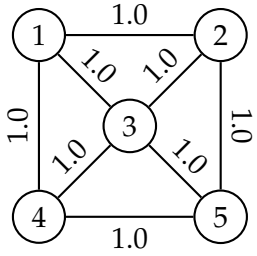
The covariance model specified here is quite flexible, especially in contrast to the commonly used graph-based methods which treat all first-order neighbors as equally proximate, such as most CAR models. We illustrate this with a simple example. The correlation component of our model for Σ is determined fully by $\mathbf{W} \in \mathcal{W}_G$, while σ^2 simply acts to scale the correlation matrix defined by $\mathbf{W}$. Applying the quasi-Euclidean metric to $\mathbf{W}$ in turn characterizes $\mathcal{D}_G^{\Delta}$, the space of distance matrices possible under the model, which in turn can be used to define $\mathcal{S}_G^{(C\circ\Delta)} \subset \mathcal{S}_p^+$, the subset of covariance matrices that can be obtained under our model for a given graph, metric, and covariance function. Fitting our model to data results in us finding the "best" $\hat{\Sigma}$ in $\mathcal{S}^{(C\circ\Delta)}$ to approximate Σ as defined by maximizing likelihood, minimizing Bayesian loss, or some other approach. Just as with our model, other specifications of Σ using fewer than $p(p-1)/2$ parameters (the number of free elements in a $p \times p$ covariance matrix) characterize a proper subset of $\mathcal{S}_p^+$ containing all covariance matrices that are possible to obtain under that model.

Figure 2.4 depicts a graph G with five nodes, and two correlation matrices, Σ_1 and Σ_2, which represent possible dependence structures for data observed on G. We note that neither Σ_1 or Σ_2 are elements of $\mathcal{S}_G^{(C\circ\Delta)}$. The matrix Σ_1 depicts a correlation structure in which one side of G (the path connecting nodes 4-1-2) exhibits greater correlation than the rest of the graph, while Σ_2 depicts a correlation structure in which non-adjacent nodes (1 and 5) are more correlated with one another than any pair of adjacent nodes in G. These two configurations of Σ were chosen to illustrate complex covariance structures which

$G_1 =$ 

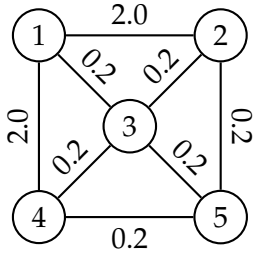
$\longrightarrow \Sigma_1 = \begin{bmatrix} 1.00 & 0.43 & 0.51 & 0.43 & 0.31 \\ 0.43 & 1.00 & 0.51 & 0.31 & 0.43 \\ 0.51 & 0.51 & 1.00 & 0.51 & 0.51 \\ 0.43 & 0.31 & 0.51 & 1.00 & 0.43 \\ 0.31 & 0.43 & 0.51 & 0.43 & 1.00 \end{bmatrix}$

(a)

$G_2 =$ 

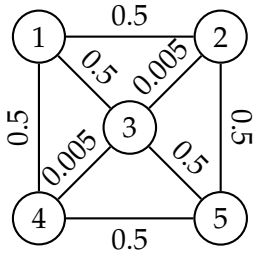
$\longrightarrow \Sigma_2 = \begin{bmatrix} 1.00 & 0.89 & 0.91 & 0.89 & 0.84 \\ 0.89 & 1.00 & 0.91 & 0.84 & 0.89 \\ 0.91 & 0.91 & 1.00 & 0.91 & 0.91 \\ 0.89 & 0.84 & 0.91 & 1.00 & 0.89 \\ 0.84 & 0.89 & 0.91 & 0.89 & 1.00 \end{bmatrix}$

(b)

$G_3 =$ 

$\longrightarrow \Sigma_3 = \begin{bmatrix} 1.00 & 0.90 & 0.41 & 0.90 & 0.24 \\ 0.90 & 1.00 & 0.41 & 0.78 & 0.25 \\ 0.41 & 0.41 & 1.00 & 0.41 & 0.24 \\ 0.90 & 0.78 & 0.41 & 1.00 & 0.25 \\ 0.24 & 0.25 & 0.24 & 0.25 & 1.00 \end{bmatrix}$

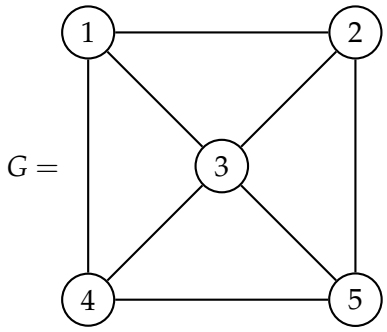
(c)

$G_4 =$ 

$\longrightarrow \Sigma_4 = \begin{bmatrix} 1.00 & 0.52 & 0.52 & 0.52 & 0.56 \\ 0.52 & 1.00 & 0.32 & 0.32 & 0.52 \\ 0.52 & 0.32 & 1.00 & 0.32 & 0.52 \\ 0.52 & 0.32 & 0.32 & 1.00 & 0.52 \\ 0.56 & 0.52 & 0.52 & 0.52 & 1.00 \end{bmatrix}$

(d)

FIGURE 2.3: The covariance matrices for a simple graph resulting from several different edge weight configurations when using a Matérn covariance function with $\nu = 3/2$.



$G =$

$$\Sigma_1 = \begin{bmatrix} 1.00 & 0.75 & 0.125 & 0.75 & 0.125 \\ 0.75 & 1.00 & 0.25 & 0.50 & 0.25 \\ 0.125 & 0.25 & 1.00 & 0.125 & 0.25 \\ 0.75 & 0.50 & 0.125 & 1.00 & 0.25 \\ 0.125 & 0.25 & 0.25 & 0.25 & 1.00 \end{bmatrix}$$

$$\Sigma_2 = \begin{bmatrix} 1.00 & 0.35 & 0.35 & 0.35 & 0.50 \\ 0.35 & 1.00 & 0.125 & 0.125 & 0.35 \\ 0.35 & 0.125 & 1.00 & 0.125 & 0.35 \\ 0.35 & 0.125 & 0.125 & 1.00 & 0.35 \\ 0.50 & 0.35 & 0.35 & 0.35 & 1.00 \end{bmatrix}$$

FIGURE 2.4: A graph G and two correlation matrices used for model fit comparisons. We identify the closest approximations of Σ_1 and Σ_2 from the subsets of correlation matrices defined by multiple models.

could be present within real world data, but are not meant to represent the full range of possible cases. In this example, we assume a multivariate normal, mean-zero distribution and find that best approximations of Σ_1 and Σ_2 from $\mathcal{S}_G^{(C\circ\Delta)}$ minimize Kullback-Leibler (KL) divergence to a greater degree than the equivalent best approximations from the CAR models we considered. Specifically, $\hat{\Sigma}_1$ and $\hat{\Sigma}_2$ are obtained by identifying the optimal set of edge weights such that

$$\hat{\Sigma}_i = \underset{\Sigma}{\arg\min}\, \mathrm{KL}(N(\mathbf{0},\Sigma), N(\mathbf{0},\Sigma_i)),\ \Sigma \in \mathcal{S}_G^{(C\circ\Delta)},\ i = 1,2 \tag{2.9}$$

The most widely used CAR variant is the first-order CAR model, which we will refer to as the CAR1 model; it incorporates only the basic adjacency structure of G into its design and has a marginal distribution which defines

$$\Sigma = \sigma^2(\mathbf{I}_n - \kappa\mathbf{A})^{-1}, \quad \sigma^2 > 0,\ \kappa \in (-1,1) \tag{2.10}$$

where $\mathbf{A}$ is the fixed adjacency matrix of G. Let $\mathcal{S}_G^{\mathrm{CAR1}}$ be the space of covariance matrices possible to obtain under the CAR1 model. Given that the CAR1 model has only two parameters, it may be unfair to compare it to a more complex model such as the one proposed in this article, which has number of parameters equal to the number of edges

in G (8 in the case of G from Figure 2.4) plus one. We thus propose a second more complex CAR model, which we refer to as the weighted CAR or CARw model, which represents the maximally flexible correlation structure possible under the CAR framework, (as characterized by this article) in which we treat all non-zero elements of the weights matrix as parameters, defining

$$\Sigma = \sigma^2(\text{diag}(\mathbf{W1}) - \kappa\mathbf{W})^{-1}, \quad \sigma^2 > 0, \kappa \in (-1,1), \mathbf{W} \in \mathcal{W}_G. \tag{2.11}$$

Let $\mathcal{S}_G^{\text{CARw}}$ be the space of covariance matrices possible to obtain under this CAR model, which is specified using a weights matrix $\mathbf{W}$ for which edge weights are estimated individually, just as with the model described in Section 3. This type of autoregressive model is uncommon within the literature, though a similar specification appears in Smith et al. (2015). In general, the limited number of existing CAR approaches utilizing a flexible weights matrix tend to define the elements of $\mathbf{W}$ as some function of environmental covariates, geographic distances, or the sizes of areal units rather than estimating individual edge weights directly (Hanks & Hooten, 2013; Ver Hoef, Peterson, et al., 2018; Wang et al., 2016).

For each of the three models discussed above and the two correlation matrices from Figure 2.4, we approximated the correlation matrix from each model which minimizes the KL-divergence between the distributions characterized by the true Σ and the model estimate $\hat{\Sigma}$. This was done by generating 100,000 samples from $N_5(\mathbf{0}_5, \Sigma)$ and obtaining a Bayes estimator for Σ using the simulated data (the details for obtaining this estimator are described in the following subsection). The Bayes estimator has been shown to concentrate about the parameter value which minimizes KL-divergence between a model and the true distribution as sample size increases (Gelman et al., 2014; Zellner, 1988). For this exercise we scaled each draw of Σ within our sampler to ensure that it was a correlation matrix; this was done to improve the fairness of the comparison between our model and the CAR models which will generally have heterogeneous diagonal variance if specified as in equations 2.10 and 2.11.

Table 2.1: $\mathrm{KL}(N(\mathbf{0}, \hat{\Sigma}_i), N(\mathbf{0}, \Sigma_i))$ for each model and correlation matrix. The KL-divergence of the best performing model for each "true" correlation matrix is in bold.

	Σ_1	Σ_2
$\hat{\Sigma}^{(C \circ \Delta)}$	**0.093**	**0.028**
$\hat{\Sigma}^{\mathrm{CAR1}}$	1.135	0.249
$\hat{\Sigma}^{\mathrm{CARw}}$	0.150	0.111

Table 2.1 contains the KL-divergences between each model and the truth for Σ_1 and Σ_2. $\hat{\Sigma}$ is the approximate optimal parameter under each model in terms of KL-divergence. Lower KL-divergence indicates that the subset of covariance matrices defined by a particular model gets closer to the true data-generating distribution. As can be seen, our model is able to more closely approximate both Σ_1 and Σ_2 than the CAR variants we considered. The CAR1 model performs notably worse than our method, which is unsurprising given that it uses far fewer parameters and the KL-divergence does not account for model complexity. The CARw model has one more parameter than our model, yet still performs worse for both Σ_1 and Σ_2 with the gap in performance being more apparent for Σ_2. This is due in part to the fact that our model has greater flexibility than most autoregressive models to produce covariance matrices for which non-adjacent nodes are more correlated with one another than to any of the intermediate nodes due how the quasi-Euclidean metric defines distance in the presence of multiple viable paths between locations.

2.3.4 Priors and model fitting

While one could obtain parameter estimates using maximum likelihood estimatation, the complexity of the relationship between edge weights and the covariance matrix makes maximizing the likelihood challenging in comparison to Bayesian estimation using a Markov chain Monte Carlo (MCMC) approach. We suggest priors in accordance with the parameter space of our model which requires that σ^2 and all edge weight parameters $\{w_{jk} : j \sim k\}$ must be positive real numbers. If the nodes of the graph correspond to a known orientation in physical space (as is frequently the case when dealing with areal data) one can obtain the

geographic distance matrix $\mathbf{D}^{Geo}$ by calculating the physical distances between centroids of areal units. It is then possible to encode this geographic information into the prior for the edge weights, which may be helpful for model fitting, especially when dealing with a limited number of observations of the process over the graph. In such cases, we propose using the following priors:

$$\begin{aligned} \sigma^2 &\sim \text{Inverse-gamma}(a_{\sigma^2}, b_{\sigma^2}) \\ w_{jk} &\sim \text{Gamma}(a_w / d_{jk}^{Geo}, b_w) \text{ if } j \sim k, \text{ else } w_{jk} = 0. \end{aligned} \tag{2.12}$$

Here d_{jk}^{Geo} is the geographic distance between nodes j and k, and the prior expectation of the associated edge weight is $a_w/(d_{jk}^{Geo} b_w)$ which is smaller for adjacent nodes which are far apart than for adjacent nodes which are closer in physical space. In instances where no geographic information is available about the graph on which data are observed, it is appropriate to simply use i.i.d. gamma priors for the edge weights. The choice of inverse-gamma prior for σ^2 is convenient for normally distributed $\mathbf{Y}$ due to built-in conjugacy. The choice to use independent gamma priors for all edge weights is somewhat simplistic, as it may be reasonable to assume some level of dependence between the weights of incident edges. Details for the MCMC sampler used to obtain posterior samples for σ^2 and edge weight matrix $\mathbf{W}$ when fitting this model are provided in Appendix A.

2.3.5 Simulation study

To assess the practical capacity of our model to reasonably perform parameter estimation we conducted the following simulation study. We defined 5-by-5, 10-by-10, and 15-by-15 lattices each having 40, 180, and 420 individual edges respectively. An edge weights matrix $\mathbf{W}$ was randomly generated with each nonzero edge having a Gamma(3,3) distribution, and $n = 10$, 25, or 50 samples $\{\mathbf{y}_i\}_{1:n}$ were independently drawn from a normal distribution with mean zero and covariance $\Sigma(\sigma^2, \mathbf{W}) + \psi \mathbf{I}_p$, with ψ acting as a nugget parameter, and with the setting $(\sigma^2 = 0.8, \psi = 0.2)$ representing a scenario in which the spatial signal to

noise ratio is high and the setting ($\sigma^2 = 0.2, \psi = 0.8$) a scenario in which the ratio is high.

All possible combinations of graph-size, samples generated, and signal-to-noise ratio represent 18 total scenarios, for each of which we performed 10 simulations and model fittings. When fitting each model we utilized a Gamma(3,3) prior for the edge weights and Inverse-gamma(2,2) priors for σ^2 and τ. Using the posterior samples we computed the average 90% coverage rate for the posterior distribution of all edge weight parameters, along with RMSE and bias. The results of our simulation study can be found in Table 2.2.

Table 2.2: Average coverage rates, RMSE, and bias for all 18 settings in simulation study.

Setting	Coverage	Bias	RMSE	Setting	Coverage	Bias	RMSE
$p = 5^2, n = 10, \sigma^2 = 0.2$	0.807	-0.314	0.568	$p = 5^2, n = 10, \sigma^2 = 0.8$	0.835	-0.334	0.625
$p = 5^2, n = 25, \sigma^2 = 0.2$	0.835	-0.209	0.483	$p = 5^2, n = 25, \sigma^2 = 0.8$	0.868	-0.297	0.603
$p = 5^2, n = 50, \sigma^2 = 0.2$	0.818	-0.177	0.435	$p = 5^2, n = 50, \sigma^2 = 0.8$	0.860	-0.257	0.584
$p = 10^2, n = 10, \sigma^2 = 0.2$	0.824	-0.267	0.551	$p = 10^2, n = 10, \sigma^2 = 0.8$	0.857	-0.314	0.610
$p = 10^2, n = 25, \sigma^2 = 0.2$	0.834	-0.222	0.495	$p = 10^2, n = 25, \sigma^2 = 0.8$	0.846	-0.300	0.614
$p = 10^2, n = 50, \sigma^2 = 0.2$	0.816	-0.190	0.452	$p = 10^2, n = 50, \sigma^2 = 0.8$	0.857	-0.277	0.593
$p = 15^2, n = 10, \sigma^2 = 0.2$	0.819	-0.284	0.563	$p = 15^2, n = 10, \sigma^2 = 0.8$	0.849	-0.303	0.610
$p = 15^2, n = 25, \sigma^2 = 0.2$	0.832	-0.224	0.504	$p = 15^2, n = 25, \sigma^2 = 0.8$	0.843	-0.310	0.616
$p = 15^2, n = 50, \sigma^2 = 0.2$	0.841	-0.176	0.435	$p = 15^2, n = 50, \sigma^2 = 0.8$	0.850	-0.276	0.602

From our simulation, we generally see improvement in average coverage, bias and RMSE as the number of samples increases, which is to be expected. Interestingly, the patterns relating to changes in lattice size or signal to noise ratio are less clear from our results. The consistent negative bias across simulation settings is potentially concerning, but is likely a result of the non-linear relationship between edge weights and covariance via the Matérn function. The fact that this bias decreases with sample size, along with our demonstration of model identifiability gives us confidence that this model functions in practice.

2.4 Numerical results

2.4.1 An application using NC eBird data

Returning to the data discussed in the introduction and which motivated the development of this method, we now fit our model to bird abundance data within the state of North Carolina. As alluded to in our introduction, there are varying patterns of between-site dependence among species within this spatial domain, and we are interested in obtaining the

species-specific intrinsic distances which characterize that spatial dependence. The data were retrieved from the eBird database, a project of the Cornell Lab of Ornithology that allows bird watchers across the globe to record and submit their bird watching observations in the form of checklists that contain the specimen count for each observed species as well as the observer's location and "effort," or time spent bird watching (Sullivan et al., 2009). The model was fit to the data for several common species within the state of North Carolina during the time period from January 2018 to December 2020. The species considered were the northern cardinal, the Carolina wren, the mourning dove, the turkey vulture, the Canada goose, the laughing gull, and the mallard. In addition to being abundant throughout the state, these species were chosen as they represent a considerable range of genetic and habitat diversity.

We treat each of the 100 counties of North Carolina as the nodes of a graph, with 256 edges connecting each pair of adjacent counties. For each species in the data set, observed counts were aggregated temporally by month and spatially by county. The data were stored in a 36×100 matrix $\mathbf{Y}$, in which y_{ij} represents the combined number of individual birds from that species observed during month $i = 1,...,36$ within the borders of county $j = 1,...,100$ by all birdwatchers contributing to the database. In addition to the species-specific counts matrices, there is a 36×100 effort matrix $\mathbf{T}$ common to all species, where t_{ij} contains the total time spent bird watching during month i within county j, as well as a geographic distance matrix $\mathbf{D}^{Geo}$, obtained by calculating the physical distances between the geographical centroid of all bird watching locations within each county. We also have the 100×3 matrix of normalized county-level environmental covariates $\mathbf{X}$ which contains the average county elevation, the percentage of county lands designated as urban vs. rural by the 2020 US Census, and percentage of county lands covered by water. Each of these covariates can influence the baseline abundance of various bird species throughout the region due to species' individual habitat preferences (Swick, 2016).

For each species, we model $\mathbf{Y}$ using the following generalized linear model with spatial random effects:

$$
\begin{aligned}
y_{ij} &\sim \text{Poisson}(\exp(\theta_{ij})t_{ij}) \\
\theta_{ij} &= \beta_{0i} + x_j^\top \beta + u_{ij} + e_{ij} \\
\mathbf{U} &\sim N_{n\times p}(\mathbf{0}_{n\times p}, \Sigma(\sigma^2, \mathbf{W}) \otimes \mathbf{I}_n) \\
\mathbf{E} &\sim N_{n\times p}(\mathbf{0}_{n\times p}, \psi\mathbf{I}_{np}) \\
\beta_0 &\in \mathbb{R}^n,\ \beta \in \mathbb{R}^3,\ \sigma^2, \psi > 0,\ \mathbf{W} \in \mathcal{W}_G.
\end{aligned}
\tag{2.13}
$$

Here θ_{ij} is the log of the expected rate at which birds are observed during month i within county j. The parameter β_{0i} is a temporally-varying intercept term and β the coefficient vector relating x_j, the vector of environmental covariates at county j to the mean of θ_{ij}. The spatial random effects matrix $\mathbf{U}$ has independent rows and between-county covariance given by Σ as defined as in equation 2.7; $\mathbf{E}$ is a matrix of iid Gaussian noise. Of primary interest to us within this article is Σ and the underlying distance matrix $\mathbf{D}$, which can be viewed as containing the species-specific intrinsic distances between counties in North Carolina, with other model parameters primarily serving to enable an appropriate fit for $\mathbf{Y}$. We note that this model is structurally quite similar to the one proposed in Besag et al. (1991) with the primary distinction being that a CAR model is used in their work to characterize the covariance of $\mathbf{U}$.

We use the priors given in equation 2.12 for σ^2 and $\mathbf{W}$ and use the following priors for the remaining model parameters:

$$
\begin{aligned}
\beta_0 &\sim N(\mathbf{0}_n, \sigma_0^2\mathbf{I}_n) \\
\beta &\sim N(\mathbf{0}_3, \sigma_\beta^2\mathbf{I}_3) \\
\psi &\sim \text{Inverse-gamma}(a_\psi, b_\psi) \\
\sigma_0^2 &\sim \text{Inverse-gamma}(a_0, b_0) \\
\sigma_\beta^2 &\sim \text{Inverse-gamma}(a_\beta, b_\beta).
\end{aligned}
\tag{2.14}
$$

In our implementation of this model we set hyperprior parameters $a_{\sigma^2} = a_\psi = a_0 = a_\beta = 2$

and $b_{\sigma^2} = b_w = b_\psi = b_0 = b_\beta = 1$, choices made to be relatively uninformative. We note that we originally considered using an autoregressive prior for the vector of temporally-varying intercepts, but found that in practice it had very little impact on parameter estimates, but a fairly large negative on computational efficiency because doing so forces us to update θ by inverting an $np \times np$ matrix rather than a $p \times p$ matrix n times. Additionally, because our interest is primarily in the structure of Σ and the corresponding distance matrix, the degree of temporal correlation in β_0 is not important here. If our goal was to use this model for the purposes of forecasting, then more nuance would need to be applied to handling the spatiotemporal structure and interactions.

2.4.2 Interpretation of results

The model is complex and the sampler used is somewhat computationally expensive, but we have found that it mixes well and within reasonable time periods for most small and medium sized graphs ($p < 500$ nodes). The data for the application described in the following section had $n = 36$ observations and $p = 100$ nodes. It required approximately eight hours to produce 10000 iterations of the sampler using code written in R and run single threaded on the lead author's personal laptop equipped with an Intel Core i7-9750H processor. The first 2500 iterations were treated as burn in and the 7500 we retained had an average effective sample size of approximately 3000, indicating reasonably good mixing.

To interpret the output of our model, which we subsequently refer to as the graph deformation (GDEF) model, it is useful to have a cursory understanding of the geography of North Carolina. The majority of physical geographic variability within the state of North Carolina tends to run on an east to west axis. Figure 2.5 depicts the four main geographic regions of the state, with the easternmost tidewater region, followed by the inner coastal plain, the Piedmont plateau in the center of the state, with the Blue Ridge mountains running through the westernmost part of North Carolina (NCpedia, 2015; Swick, 2016).

We are interested in how these geographic regions may manifest themselves within the intrinsic distances across the state for each species. Are there species for which distances

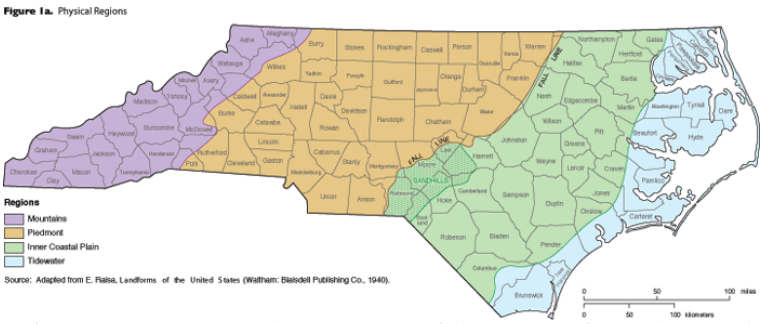
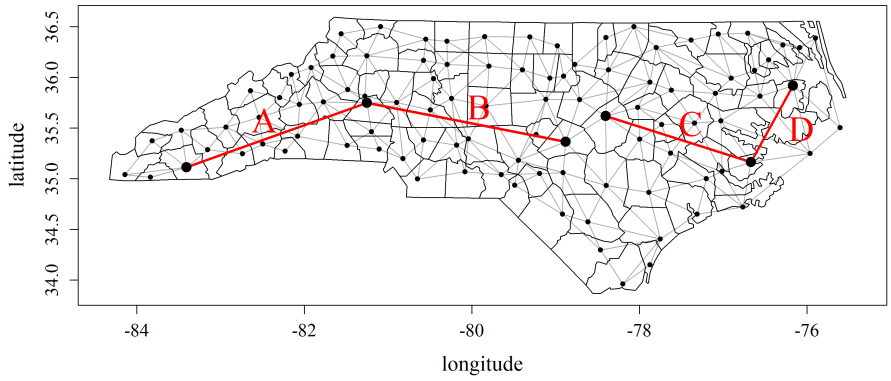
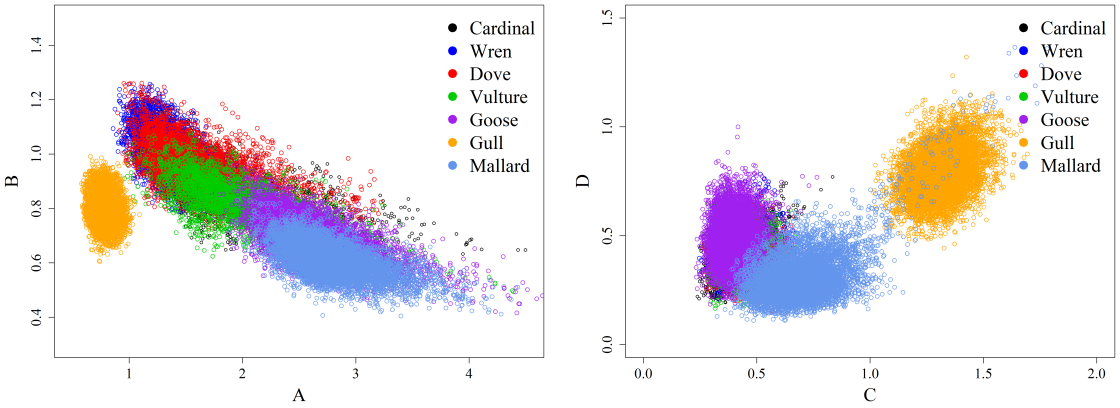


FIGURE 2.5: The four main geographic regions of the state of North Carolina. Image retrieved from NCpedia (ncpedia.org) and accessed on 20 February 2022.

along the eastern coast of the state are greater or smaller relative to the intrinsic distances from the coast moving inland? Are there species for which the mountain ranges in the west of the state significantly impact their spatial distribution while other species are unaffected? Figure 2.6 depicts the approximate posterior distribution for selected elements of the scaled posterior distance matrix for each species. Subfigure 2.6a depicts the state of North Carolina and four line segments labeled “A", “B", “C", and “D", with these segments chosen to reflect various physical regions within the state, or the borders between them. For each of the seven bird species, there is a different distance associated with each of the four line segments; these distances are shorter, and the counties at their endpoints are “closer" to one another if there is high correlation in the bird counts between locations. Subfigures 2.6b and 2.6c show the posterior distributions of the scaled intrinsic distances associated with two pairings of those segments. Both of these pairings corresponds to a geographic contrast that may be useful for understanding how these species spatial distributions interact with the physical properties of the environments they inhabit. We see that the intrinsic distances associated with each of the segments differ from species to species, with the posterior for the laughing gull being more divergent from the other six species in both subfigures. This divergence makes sense when one considers that the laughing gull it is the only seabird of the seven species, having a natural range of habitats with the least overlap among the species considered (Swick, 2016). We see this in subfigure 2.6c, which highlights the contrast between intrinsic distances running along the coast (segment “D") versus ones moving inland (segment “C").



(a)



(b) (c)

FIGURE 2.6: Visualization of the uncertainty in posterior distances associated with four line segments for each bird species

Subfigure 2.6b highlights the contrast between the distances associated with segment “A," which spans across a mountain range, and “B" which runs along the relatively flat Piedmont. The spatial distributions of birds with posteriors concentrated at higher values along the x-axis of this plot, such as the Canada goose and Mallard are more impacted by the changes in elevation and habitat along segment “A" than the others.

We would also like to visualize the entire mean posterior distance matrix for each species. Because this matrix corresponds to a high dimensional Euclidean embedding of the graph, it is challenging to represent with a single plot and without significant distortion. To this

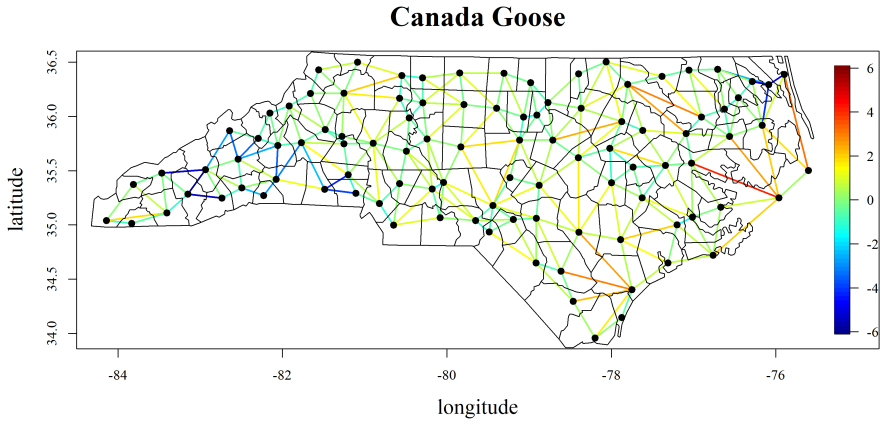
end, we present a visualization which highlights the contrast between the posterior mean intrinsic distance and the physical distance between the centroids of each pair of adjacent counties, which in the minds of the authors is more interpretable than the raw edge weights themselves; this visualization provides us not only a sense of the relative distances between counties, but more significantly of the deformation of the graph's geographic orientation necessary to produce the intrinsic distances. We have done this by calculating z^d_{jk}, a score designed to represent whether the intrinsic distance between adjacent nodes j and k is smaller or greater than suggested by the physical distance between those nodes.

$$z^d_{jk} = Z(\mathbf{D}^{Geo}(E)) - Z(\overline{(\mathbf{D}(E)|\mathbf{Y})})_{jk} \tag{2.15}$$

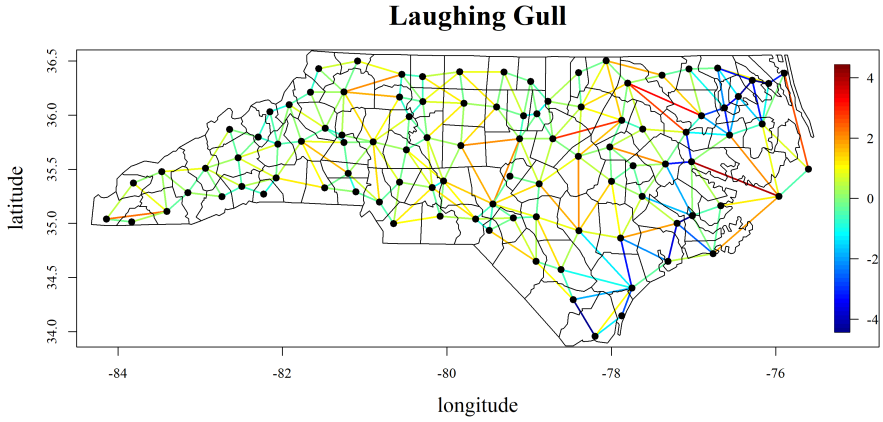
Here $\mathbf{D}(E)$ indicates the set of distances between nodes connected by edges of our graph, and $\overline{(\mathbf{D}(E)|\mathbf{Y})}$ is the set of posterior means for those same distances. The function $Z(\cdot)$ scales a set to have mean zero and standard deviation one. High values of z^d_{jk} indicate that nodes j and k are intrinsically "closer" together relative to their physical separation, while low values indicate that they are more distant.

Figure 2.7 contains two plots depicting z^d_{jk} for all pairs of adjacent counties. Subfigure 2.7a contains a map of North Carolina for the Canada goose, a migratory species which can be observed year round within the state, but is considerably more common during the winter (Swick, 2016). We note the bands of dark blue line segments running approximately along longitude $-82°E$, separating the westernmost portion of the state from its central regions, which indicates that observations of Canada geese on opposite sides of this border are less correlated than ones made on the same side. Significantly, that border corresponds well to the mountain ranges making up the westernmost parts of North Carolina. Ornithologists have established that mountain ranges present significant barriers to the flight paths of migratory birds (Aschwanden et al., 2020) which leads to a physical interpretation for the posterior correlations identified by the GDEF model.

Subfigure 2.7b depicts a comparable map for the laughing gull, which is found most abundantly along the coast of the state, and becomes progressively less common as one



(a)



(b)

FIGURE 2.7: Maps of $\mathbf{z}^d$ for Canada goose and laughing gull. Warm colors indicate areas which are "closer" than suggested by their geographic distance and cool colors more distant.

moves inland (Swick, 2016). While many of the edges between coastal counties are red and orange, indicating greater connectivity in the regions, the majority of the edges connecting the coastal counties bordering the Albermale sound (at $-76°E, 36°N$) are blue and green, indicating that not all coastal locations possess the same level of interdependence. We also note the existence of a blue band of edges running parallel to the coast at approximately $-77°E$, along the border between the coast and coastal plains regions of the state. This could be viewed as a boundary between coastal and inland regions over which laughing gull populations tend not to cross, and which is revealed within the posterior distance matrix

Table 2.3: Posterior mean and 95% credible intervals of ratio $\sigma^2/(\sigma^2+\psi)$ for all seven analyzed species

	Mean	95% CI
Northern Cardinal	0.584	(0.429,0.712)
Carolina Wren	0.064	(0.014,0.298)
Mourning Dove	0.129	(0.058,0.313)
Turkey Vulture	0.371	(0.262,0.476)
Canada Goose	0.422	(0.312,0.693)
Laughing Gull	0.857	(0.789,0.897)
Mallard	0.471	(0.400,0.543)

of the GDEF model.

For both species, these plots reveal significant interactions between their spatial distributions and the geography of where they were observed, even after having accounted for how environmental covariates may impact their baseline abundance rates within each county. Taken together, the plots (along with the plots not included for the other five species analyzed) reveal that there is a unique covariance structure and underlying distance between locations associated with each species.

An additional quantity of interest in many spatial data applications is the ratio $\sigma^2/(\sigma^2+\psi)$ which may be thought of as the percentage of total variability attributable to the spatial process, in comparison to the nugget effect represented by ψ. Table 2.3 contains a posterior summary of this ratio for each of the seven bird species. We note that it is particularly low for the Carolina wren and mourning dove, indicating weak spatial correlation in the data for county-level populations of those particular species.

2.5 Discussion

We now briefly discuss a few potential applications, extensions, and questions related to the ideas presented within this article. One of the most significant pieces of output from the GDEF model is the posterior distance matrix. By itself, it can be analyzed to better understand the relationships between nodes of a graph in terms of their latent proximity to one another; but as was the case in our application with the eBird data, there are instances

in which one may fit this model separately for multiple datasets (unique bird species in the case of the eBird example) which are observed on the same graph. In such cases, an analysis of how the posterior distance matrices for each dataset compare to one another may reveal significant information about the relationships between observed variables. Abdi et al. (2005) presents a framework for analyzing multiple distance matrices computed on the same set of areal units. Such a framework could be used for example to perform a clustering analysis of common bird species in North Carolina based on their unique distance matrices.

When modeling the eBird data, we included environmental covariates in the mean of our spatial process. An alternative or complimentary approach could be to include them in the specification of the GDEF model's edge weights, which would enhance the interpretability of the model at the edge weights level, and not just in terms of the distance matrix. One could imagine based on the findings illustrated in Subfigure 2.7a that edge weights for the Canada goose may be negatively associated with elevation. An approach along these lines was explored using CAR models by Hanks and Hooten (2013) and Wang et al. (2016), and could be augmented within our model framework by treating edge weights as a linear combination of spatial covariates and an additional spatial random effect over the line graph of our network. Additionally, species count observations are often zero-inflated. Accounting for this as in Agarwal et al. (2002) could prove beneficial to model quality in applied settings. Other potential extensions include joint species modeling using our methodology. One approach could be to perform factor analysis (Harman, 1976; Lawley & Maxwell, 1962) on the model parameters characterizing species-specific covariance enabling one to better understand which species interact with the spatial domain similarly in terms of covariance and intrinsic distances, while providing a natural approach to modelling the intrinsic distances for each species hierarchically.

The method described in this article provides a flexible and novel framework for modeling the covariance of data observed on a graph structure. It's use of an estimated distance metric to which standard covariance functions may be applied allows for interpretation more similar to that of models in continuous space settings as compared to CAR models.

We have demonstrated that the GDEF model is flexible enough to represent covariance structures in which patterns of spatial autocorrelation vary throughout the domain, and that there are contexts in which it performs better than advanced versions of autoregressive models which are the traditional tools for modeling spatially dependent graph data. We also find the specification of between-node covariance in terms of intrinsic distances a useful framing for understanding and interpreting model results, as exemplified by our discussion of model output in section 4.2 in which we found that an analysis of the posterior distance matrix revealed connections between the covariance in the eBird dataset and the physical geography of the spatial domain on which it was observed.

3. A Dimension Reduction Approach to Edge Weight Estimation for Use in Spatial Models

3.1 Introduction

Studies involving areal, or region indexed data are common in many settings including ecology (e.g. Hanks & Hooten, 2013), economics (e.g. Arbia, 2012), public health (e.g. Jin et al., 2005) and sociology (e.g. Garner & Raudenbush, 1991). Models for spatially indexed data are typically defined such that observations close to one another in space exhibit higher correlation than observations which are far apart. In contrast to spatial models for point-indexed data, which tend to characterize between-location dependence as a function of the geographic distances between observations, models for areal data generally define spatial dependence in terms of neighborhood or adjacency structure (Cressie, 1993).

The spatial structure of areal data may be represented using a graph, with each spatial region represented by a graph node and with edges existing between each pair of nodes corresponding to physically adjacent regions. First introduced by Besag (1974), the most commonly used spatial model for areal data is the conditional autoregressive (CAR) model. Given a graph $G = (V, E)$ having p nodes, we observe the random vector $\mathbf{y}$, where $y_1, ..., y_p$ are observations at each node of G. A CAR model defines a Markov random field where the conditional distribution of each element of $\mathbf{y}$ is determined by the values of $\mathbf{y}$ at adjacent nodes:

$$y_i | \mathbf{y}_{-i} \sim N\left(\sum_{j \sim i} \alpha w_{ij} y_j, \sigma^2\right). \tag{3.1}$$

Equation 3.1 results in a marginal distribution for $\mathbf{y}$ of

$$\mathbf{y} \sim N_p(\mathbf{0}, \sigma^2(\mathbf{I}_p - \alpha\mathbf{W})^{-1}) \tag{3.2}$$

where $j \sim i$ indicates that node j is adjacent to node i, $\mathbf{W}$ is a $p \times p$ adjacency or weights matrix (where $w_{ij} > 0$ if $j \sim i$ else $w_{ij} = 0$), $\sigma^2 > 0$ is a variance parameter, and α controls the degree of spatial correlation. In order to produce a valid covariance matrix, α must be bounded between the largest and smallest eigenvalue of $\mathbf{W}$ (Ver Hoef, Peterson, et al., 2018). An alternative form of the CAR model may be written

$$\mathbf{y} \sim N_p(\mathbf{0}, \sigma^2(\text{diag}(\mathbf{W}\mathbf{1}_p) - \kappa\mathbf{W})^{-1}) \tag{3.3}$$

where $\text{diag}(\cdot)$ returns a diagonal matrix with entries equal to the input vector. This model is valid for all $|\kappa| < 1$. The intrinsic conditional autoregressive (ICAR) model is a special case of Equation 3.3 when $\kappa = 1$. This results in an improper distribution, but it may be reconfigured as a $p - 1$ dimensional distribution using the constraint $\sum y_i = 0$ (Besag & Kooperberg, 1995).

While many variants of the CAR and other similarly defined spatial autoregressive models have been developed and utilized, Ver Hoef, Peterson, et al. (2018) notes that in practice almost all such models have been implemented using the unweighted adjacency matrix of G, with $w_{ij} = 1$ if $i \sim j$, and $w_{ij} = 0$ otherwise. The CAR model is frequently used to characterize spatial random effects or an "error process" within a larger generalized linear model (Besag et al., 1991); in many such contexts, interest in the covariance structure of the model has been secondary to inference on the fixed effects governing the model's mean structure. However, the lack of a more complex weighting scheme results in a pattern of spatial dependence where the partial correlations between all pairs of adjacent regions are identical regardless of location within the overall spatial domain. Such models may be thought of as discrete space equivalents to stationary and isotropic covariance models for point indexed data in continuous space, properties which are frequently inappropriate for real world data (Guttorp et al., 1994).

Christensen and Hoff (2023) introduced a covariance model (which we will refer to as the graph deformation model or GDEF model within this article) for areal data exhibiting greater flexibility than the traditional CAR model. Given a graph $G = (V, E)$ with p nodes and random vector $\mathbf{y}$ consisting of observations at each node of G, they define $\Sigma = \text{Cov}(\mathbf{y})$ as follows:

$$\begin{aligned} \Sigma_{ij} &= \sigma^2 \rho_\nu(d_{ij}) \\ d_{ij} &= \sqrt{(\mathbf{e}_i - \mathbf{e}_j)^\top \{\mathbf{L}^+\}^2 (\mathbf{e}_i - \mathbf{e}_j)} \\ \mathbf{L} &= \text{diag}(\mathbf{W}\mathbf{1}_p) - \mathbf{W} \\ \sigma^2 &> 0,\ w_{ij} > 0 \text{ if } i \sim j, \text{ else } w_{ij} = 0, \text{ and } w_{ij} = w_{ji}. \end{aligned} \tag{3.4}$$

Here σ^2 is a scale parameter and $\rho_\nu(\cdot)$ is the Matérn correlation function (Matérn, 1960) with smoothness parameter ν. The term d_{ij} may be thought of as the latent distance between nodes i and j and is defined as a function of unknown edge weights matrix $\mathbf{W}$. The matrix $\{\mathbf{L}^+\}^2 = \mathbf{L}^+\mathbf{L}^+$ is the square of the generalized inverse of the graph Laplacian $\mathbf{L}$, and $\{\mathbf{e}_i\}_{1:p}$ are the standard basis vectors. The edge weights $\{w_{ij}\}_{i\sim j}$ are estimated from the observed data.

The GDEF model described in Equation 3.4 results in a valid covariance matrix for any combination of positive edge weights and the matrix $\mathbf{W}$ is an identifiable parameter. While Christensen and Hoff (2023) demonstrate the flexibility and interpretive value of the GDEF approach, the authors also acknowledge that the model has limited scalability when applied to large graphs given the number of model parameters (each non-zero edge weight must be estimated individually) and computational intensity of the Markov chain Monte Carlo (MCMC) algorithm they used to obtain parameter estimates.

Both CAR and GDEF models are constructed using the edge weights matrix $\mathbf{W}$. While the GDEF model could be improved by coming up with a simpler and more efficient approach to estimation of $\mathbf{W}$, CAR models may be improved in practice by using a more complex approach. As noted, CAR models have traditionally been defined with $\mathbf{W}$ equal

to the unweighted adjacency matrix of the graphical representation of the spatial domain, a choice that is both simple and convenient but may be restrictive depending on the data and scientific question of interest. Despite likely under-utilization in practice, more complex weighting schemes for CAR models have been shown to be valuable, as in Hanks and Hooten (2013) and Ejigu and Wencheko (2020) which define edge weights using linear combinations of environmental covariates. Weighting schemes with even greater flexibility are also possible and potentially enable deeper understanding of the interaction between discretely observed spatial processes and the the spatial domain over which they are observed.

Within this article we introduce a new approach to estimating edge weights matrices that may be used within the GREF framework of Christensen and Hoff (2023) and which allows for the specification of CAR models that are considerably more flexible than existing variants. In the following section we provide a brief overview of necessary background information related to spatial basis functions, graphs, and spectral theory. In Section 3 we present our method and discuss some of its properties. Section 4 illustrates and evaluates the performance of our approach through a series of simulation studies and a real data example. We conclude the article with a discussion of potential applications and extensions.

3.2 Background

Given a simple, undirected graph $G = (V, E)$, random variables $\{y_v : v \in V\}$ observed at the nodes V of G, and defining $\Sigma_{vv'} = \text{Cov}(y_v, y_{v'})$, we parameterize the covariance matrix as a function of $\mathbf{W}$, a matrix of unknown positive edge weights. We define $\mathcal{W}_G$ to be the space of all edge weight matrices $\mathbf{W}$ possible under G such that $w_{ij} > 0$ if an edge exists between nodes i and j and $w_{ij} = 0$ otherwise. As the number of nodes in G increases, so too does the size of $\mathcal{W}_G$. As such, setting a prior distribution for $\mathbf{W}$ over the complete space $\mathcal{W}_G$ (as in Christensen and Hoff (2023)) may not be feasible for graphs containing more than a few hundred nodes. To address this, we restrict the parameter space to $\mathbf{W} \in \mathcal{W}_G^B \subset \mathcal{W}_G$, where $\mathcal{W}_G^B$ contains all possible edge weights matrices that can be obtained from a basis function representation (to be described in detail later) of the non-zero edge weights.

3.2.1 Basis functions and spatial models

If $y(\mathbf{s})$ is an observation of a random process at location $\mathbf{s} \in \mathcal{D}$ within some spatial domain ($\mathcal{D}$ is typically a subset of $\mathbb{R}^2$, but could also be the set of spatial regions under the areal data setting), with spatially indexed predictors $\mathbf{x}(\mathbf{s})$, a typical model for $y(\cdot)$ might be written as follows:

$$y(\mathbf{s}) = \mathbf{x}^\top(\mathbf{s})\beta + u(\mathbf{s}) + \epsilon(\mathbf{s}) \tag{3.5}$$

where β is a vector of coefficients, $u(\mathbf{s})$ is a spatial random effect and $\epsilon(\mathbf{s}) \overset{iid}{\sim} N(0, \sigma^2)$ for all $\mathbf{s} \in \mathcal{D}$. Let $\mathbf{u} = (u(\mathbf{s}_1), ..., u(\mathbf{s}_p))$ denote the vector of spatial random effects at some collection of p locations.

Instead of modeling $\mathbf{u} \sim N_p(\mathbf{0}, \Sigma)$ it is common to define $\mathbf{u} = \mathbf{M}\eta$, where $\mathbf{M}$ is an $p \times k$ matrix of $k << p$ basis functions, and $\eta \sim N_k(\mathbf{0}, \Phi)$. Reduced-rank statistical models representing spatial random effects using basis functions are widespread. Examples of this include, but are not limited to the kernel convolution approach (Higdon, 1998), predictive processes (Banerjee et al., 2008), spatial splines (Sangalli et al., 2013) the stochasistic partial differential equation approach (Lindgren et al., 2011), and various eigenvector and principle component based methods (Higdon et al., 2008; Hughes & Haran, 2013; B. S. Lee & Haran, 2022). Note that the majority of these methods were developed for point indexed data, though the framework of Hughes and Haran (2013) is designed specifically for areal data. Applying methods designed for one data type to another is possible, but requires assumptions that may not be appropriate in all settings (Gelfand et al., 2001).

Basis function methods tend to be very computationally efficient relative to full-rank Gaussian process models. Notably, some approaches to basis functions specification are also capable of addressing issues such as nonstationarity and anisotropy (Schmidt & Guttorp, 2020). It has often been observed that methods based on low-rank representations of spatial random effects may result in oversmoothing as well as bias in the estimation of variance components (Banerjee et al., 2010; Datta et al., 2016; Stein, 2014). Additionally,

when pre-specifying the design matrix $\mathbf{M}$ using some set of basis functions, the covariance of $\mathbf{u}$ is restricted to matrices of the form $\mathbf{M}\Phi\mathbf{M}^\top$: in many instances, Φ is assumed to be a diagonal matrix or multiple of the identity matrix, resulting in strong and potentially inappropriate assumptions about model covariance structure. This may be especially true if analysis goals go beyond spatial prediction and smoothing and include inference regarding model covariance. Rather than modelling spatial random effects for areal data as a linear combination of basis functions, we propose using basis functions in a novel parameterization of edge weights matrix $\mathbf{W}$ which is then incorporated into a CAR or GDEF model, resulting in a class of full-rank covariance matrices that are more flexible than traditional autoregressive models and less computationally demanding than the implementation of the GDEF used by Christensen and Hoff (2023).

3.2.2 Line graphs and an eigenvector basis

We now describe the basis used in this article. Given a graph $G = (V, E)$ having p nodes and q edges, one may define its line graph $L(G)$ as a graph with q nodes corresponding to each of the edges of G, and edges between each pair of nodes corresponding to coincident edges in the original graph G. (Edges are considered coincident if they share a common node). Figure 3.1 illustrates the relationship between a graph G and its line graph $L(G)$ for a graph containing five nodes.

Let $\mathbf{A}^L$ be the $q \times q$ adjacency matrix of $L(G)$, where $a^L_{ij} = 1$ if nodes i and j of $L(G)$ are adjacent and $a^L_{ij} = 0$ otherwise. Let $\mathbf{L}^L = \text{diag}(\mathbf{A}^L\mathbf{1}_q) - \mathbf{A}^L$ be the Laplacian matrix of $L(G)$. If $\mathbf{V}\Lambda\mathbf{V}^\top$ is the eigendecomposition of $\mathbf{L}^L$, the columns vectors of $\mathbf{V}$ form an orthonormal basis for $\mathbb{R}^q$. The eigenvalues of Laplacian matrices are non-negative and generally ordered from smallest to largest, with smallest non-zero eigenvalues corresponding to the most "important" eigenvectors. Laplacian matrices for connected graphs are positive-semidefinite and have exactly one eigenvalue (denoted λ_1) equal to 0 with corresponding eigenvector $\mathbf{v}_1 \propto \mathbf{1}$. Despite some referring to $\mathbf{v}_1$ as the "trivial" eigenvector, it has interpretive value as

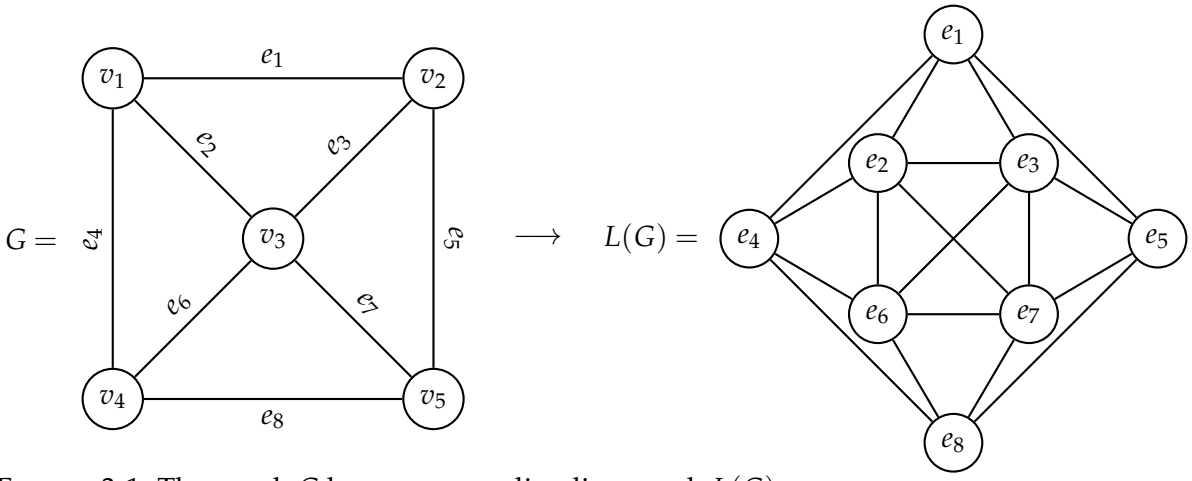


FIGURE 3.1: The graph G has corresponding line graph $L(G)$

an intercept column within an eigenvector basis. Considerable work has been done in the field of spectral graph theory investigating the properties of the eigenvalues and eigenvectors of Laplacian matrices. See Spielman (2012) for greater detail on this topic.

To visualize the basis characterized by the eigendecomposition of the line graph Laplacian (LGL) Figure 3.2 depicts a selection of LGL eigenvectors for a 30×30 lattice graph, while Figure 3.3 depicts the LGL eigenvectors for the graph of the 100 counties of North Carolina (the spatial domain analyzed in Christensen and Hoff (2023)). Note that each vector element corresponds to a specific edge of the original graph. We also highlight the harmonic behavior of these eigenvectors, with later eigenvectors exhibiting higher frequency oscillations between positively and negatively valued regions of the graph. This makes selection of the first k eigenvectors (corresponding to the k smallest eigenvalues) a natural choice of basis, with the "most important" initial eigenvectors capturing larger-scale trends in edge weight behavior and additional eigenvectors adding finer detail.

3.3 Method

Equations 3.3 and 3.4 describe two models for areal data, each parameterized by unknown edge weights matrix $\mathbf{W}$. Given a spatial domain representable by graph $G = (V, E)$ having p nodes and q edges, let $\mathbf{w}$ be the q-length vector containing the unknown weights

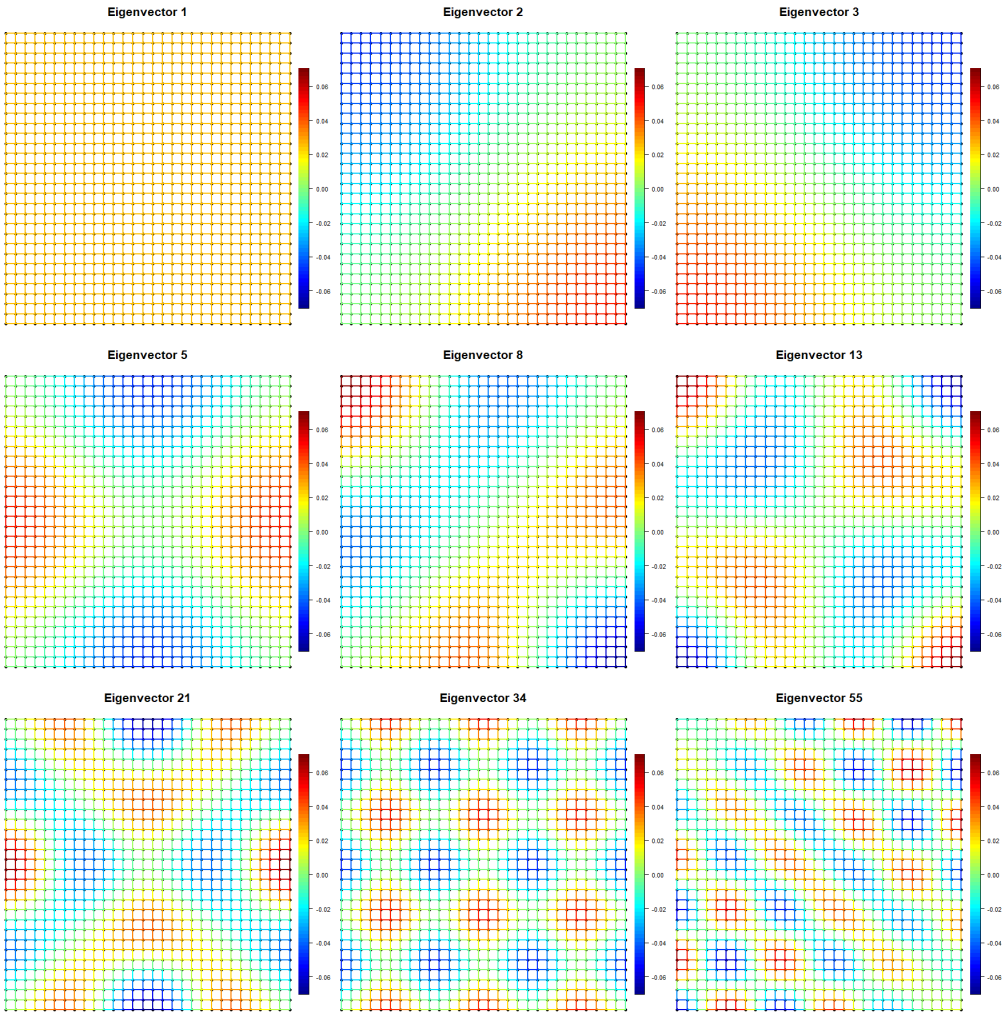


FIGURE 3.2: A selection of LGL eigenvectors for 30×30 lattice grid

associated with the edges of G. The $p \times p$ edge weights matrix $\mathbf{W}$ may be constructed from $\mathbf{w}$ and vice versa. We construct the line graph $L(G)$ and obtain the line graph Laplacian $\mathbf{L}^L$, a $q \times q$ matrix. If $\mathbf{V}\Lambda\mathbf{V}^\top = \mathbf{L}^L$ is the eigendecomposition of the LGL, let $\mathbf{V}^k$ be the $q \times k$ matrix containing the k eigenvectors of $\mathbf{L}^L$ associated with its k smallest eigenvalues. We can now represent $\mathbf{w}$ as follows:

$$\mathbf{w} = \exp(\mathbf{V}^k \eta),\ \eta \in \mathbb{R}^k \tag{3.6}$$

Here, η is a k-length vector of basis coefficients that fully characterize the $p \times p$ edge weights matrix $\mathbf{W}(\eta)$. For all $\eta \in \mathbb{R}^k$, (and $\sigma^2 > 0$ and $|\kappa| < 1$) the CAR covariance matrix $\mathbf{\Sigma} =$

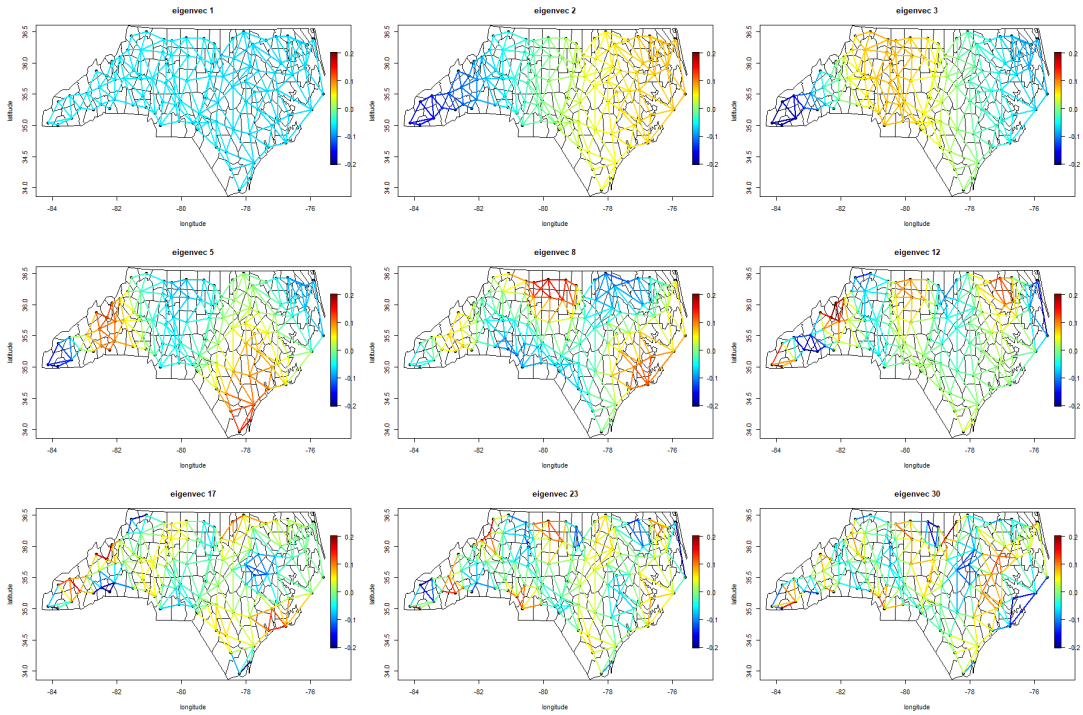


FIGURE 3.3: A selection of LGL eigenvectors for graph of the 100 counties of North Carolina

$\sigma^2(\text{diag}(\mathbf{W}(\eta)\mathbf{1}_p) - \kappa\mathbf{W}(\eta))^{-1}$ and the GDEF covariance matrix defined in Equation 3.4 are positive definite matrices. We now discuss some modeling considerations and illustrate properties of this method.

3.3.1 Contrasting the CAR and GDEF Approaches

We have stated that our framework for parameterizing the edge weights matrix of a given graph may be used when defining both CAR and GDEF models, but provided little discussion so far regarding the implications of using one covariance model or another. CAR models have sparse precision matrices, allowing for computationally efficient likelihood calculations. In contrast, the GDEF model specifies a dense covariance matrix; even when using the proposed basis function parameterization of $\mathbf{W}$, it is virtually impossible to use the GDEF covariance for graphs with over 10,000 nodes due to the impracticality of inverting matrices of that size. The computational advantages of the CAR approach do however

come with trade-offs.

Several authors have noted that models which condition only on local or nearest-neighbor structure perform more poorly than those incorporating longer range dependencies (Gramacy & Apley, 2015; Guinness, 2018; Katzfuss & Guinness, 2021). Realizations of spatial random fields produced by CAR models (including ICAR models, which induce perfect correlation between neighboring regions) are considerably less smooth than those generated by the GDEF model, a contrast that is more noticeable on graphs with a large number of nodes. Besag and Mondal (2005) shows that the standard ICAR model on a regular two-dimensional lattice grid converges asymptotically to the distribution of a two-dimensional Brownian motion as grid resolution increases, while Lindgren et al. (2011) shows that the ICAR model may be thought of as a discrete approximation to the limit of a Matérn random field whose range parameter goes to infinity and smoothness goes to zero. Paciorek (2013) demonstrates that CAR models using more than first order adjacency (extending graph structure to second order neighbors and beyond) do not result in meaningfully smoother fields. While computationally efficient, straightforward to implement and widely used, CAR models carry implicit assumptions that may not be desirable in practice.

In contrast to the CAR model, the GDEF model accommodates a wider range of spatial dependence patterns for areal data. Equation 3.4 defines the GDEF covariance, where $\rho_\nu(\cdot)$ is the Matérn correlation function given below:

$$\rho_\nu(d) = \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\sqrt{2\nu}d\right)^\nu K_\nu \left(\sqrt{2\nu}d\right). \tag{3.7}$$

Here d is the distance between two observations, $\nu > 0$ is a smoothness parameter, often fixed in practice, and K_ν is a modified Bessel function of the second kind of order ν. The exponential covariance function $\sigma^2\exp(-d)$ is a special case of the Matérn covariance when $\nu = 1/2$; $\nu = 3/2$ and $\nu = 5/2$ are popular choices for spatial models, and the squared exponential covariance $\sigma^2\exp(-d^2/2)$ is the limit of Equation 3.7 as ν goes to infinity. Those familiar with the Matérn family of covariance functions may note that Equation 3.7

omits an explicit range parameter. As discussed in Christensen and Hoff (2023), the range parameter of the GDEF model is implicitly defined by the scale of the edge weights matrix, e.g. multiplying $\mathbf{W}$ by a factor of two is equivalent to doubling the range parameter, a property we now illustrate.

Figure 3.4 highlights some distinctions between the CAR and GDEF approaches. Depicted samples were generated over a 30×30 grid from a mean zero normal distribution. Each row corresponds to a different covariance model: the CAR model from Equation 3.3 with $\kappa = 0.95$, the ICAR model with a sum-to-zero constraint (the "smoothest" possible CAR model) and the GDEF model with four different values of ν. In the first column, $\mathbf{W}$ was constructed such that $w_{ij} = 4$ for all $i \sim j$, with different constants used in subsequent columns. Across all settings the scale parameter σ^2 was set equal to 1. As expected, samples produced by the CAR models and the GDEF model for $\nu = 1/2$ are visibly less smooth than those from GDEF models with larger values of ν. The impact that the scale of $\mathbf{W}$ has on each model is also notable. As previous stated, the scale of $\mathbf{W}$ may thought of as the implicit range parameter within the GDEF; samples with edge weights equal to 4 exhibit slower decay in spatial correlation than those with smaller edge weights. For CAR models, the scale of $\mathbf{W}$ doesn't influence range, but rather the scale of marginal variances. (From Equation 3.3, it is clear multiplying $\mathbf{W}$ by some constant is equivalent to dividing σ^2 by that same constant, indicating that the CAR model as we have defined it is over parameterized, which we address later.)

The notion of range within CAR models is trickier to characterize than for GDEF models. While the correlation parameter κ ostensibly controls the smoothness of the Markov random field and the rate of spatial decay between areal regions, it is highly sensitive to the resolution of the spatial domain. This may be particularly relevant when the CAR model is viewed as the discretization of some continous spatial process. As areal resolution becomes finer, the space of region-averaged continuous spatial processes that could be represented using a CAR or ICAR model shrinks. Figure 3.5 highlights this, which depicts samples from an ICAR model at different spatial resolutions. If each grid represents the same subset of $\mathbb{R}^2$

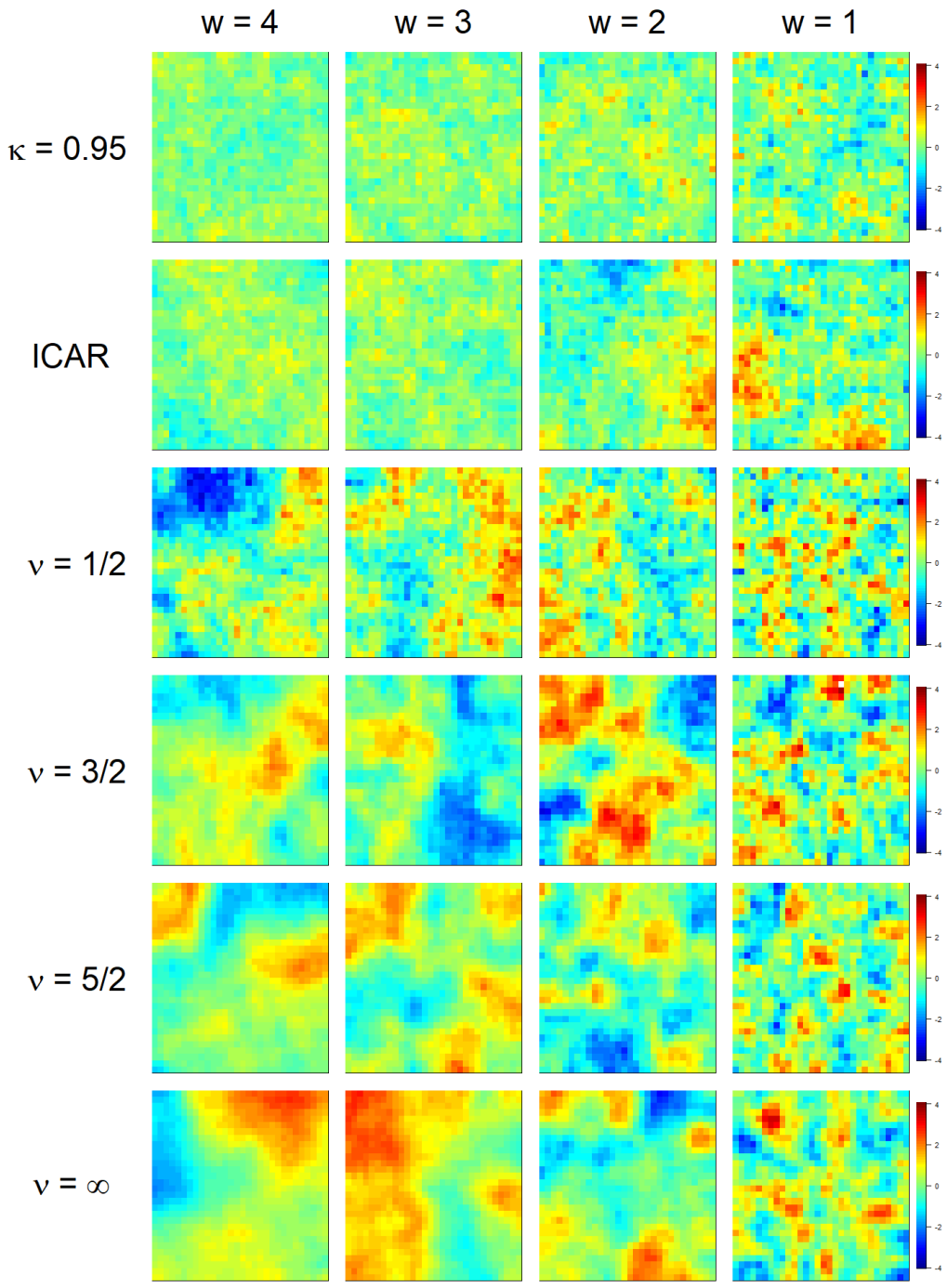


FIGURE 3.4: Samples from $N(\mathbf{0}, \Sigma)$ over a 30×30 grid with uniform edge weights. Columns correspond to the weight of all nonzero entries in $\mathbf{W}$, while each row corresponds to a different parameterization of Σ. The first row uses the CAR covariance with $\kappa = 0.95$, the second uses the ICAR covariance, and subsequent rows use the GDEF covariance for different values of ν.

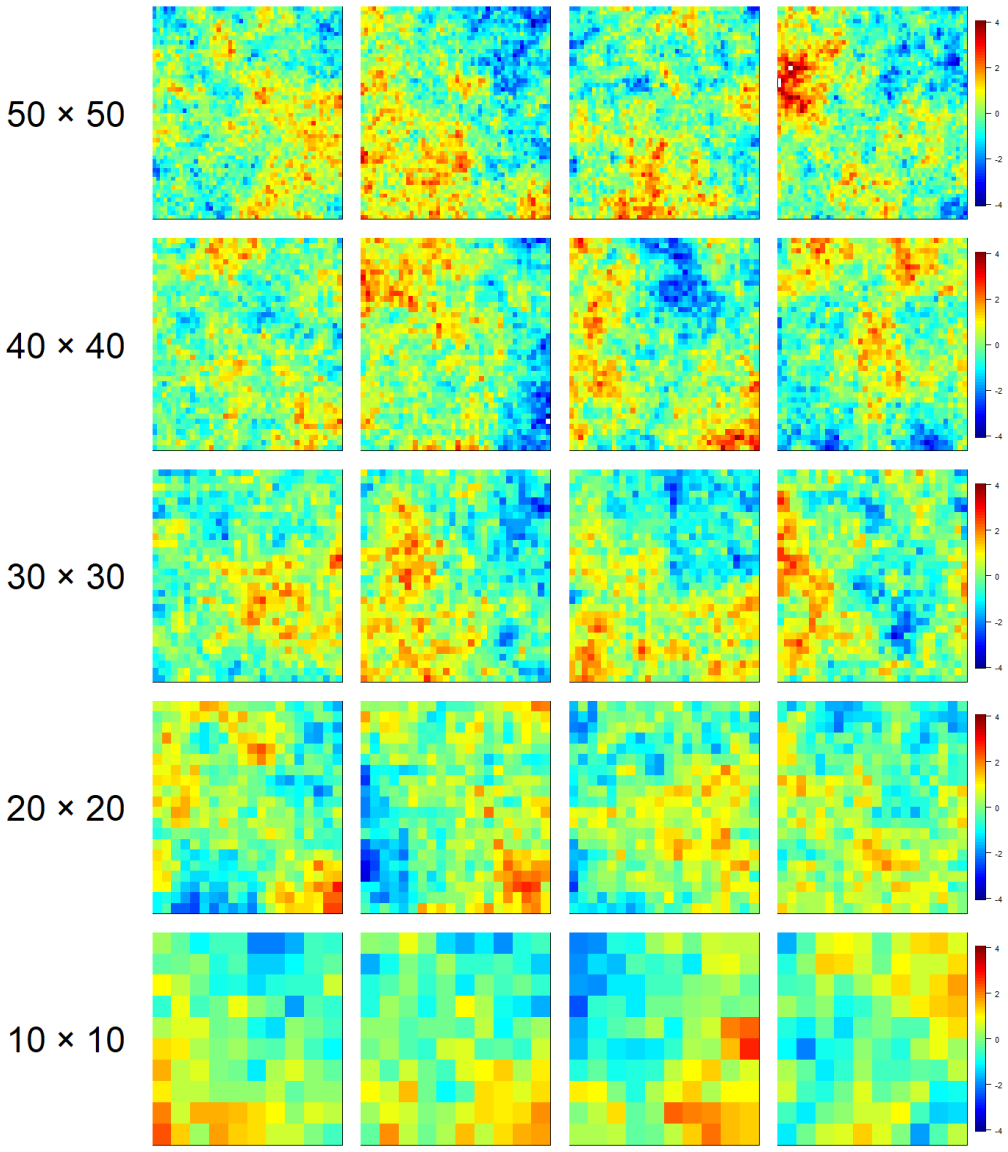


FIGURE 3.5: Samples from an ICAR model with uniform edge weights at different grid resolutions.

and samples are viewed as the region-averaged realizations of continuous spatial processes, samples over coarser grids could correspond to smoother underlying processes than those generated at higher resolutions. In general we find that the GDEF model can be used to represent a wider range of spatial dependence patterns, and is less sensitive to how areal units are defined, though there are practical settings in which the computational efficiency of the CAR model may trump these considerations.

3.3.2 Impact and interpretation of basis coefficients

The illustrations in Figures 3.4 and 3.5 depict samples from distributions defined by uniform edge weights, but a major aim of this article is to show how flexibly defining the edge weights matrix $\mathbf{W}$ can lead to more complex and potentially useful covariance structures. As discussed in Christensen and Hoff (2023) regarding the GDEF model, each possible edge weights matrix $\mathbf{W}$ corresponds to an embedding of the graph in high-dimensional Euclidean space which is unique up to isometry. The distances between nodes in this embedding may be thought of as the "intrinsic" distances between regions. The GDEF model may be viewed as the graph equivalent of the spatial deformation approach for modeling nonstationary data defined by Sampson and Guttorp (1992), where the deformation function is estimated by placing a prior on the edge weights. This interpretation leads us to say that nodes in regions containing high valued edges are "closer" to each other than regions with low valued edges. While the CAR model does not directly admit a distance based interpretation of $\mathbf{W}$, it can be said of both the CAR and GDEF approaches that edges with large weights correspond to greater connectivity between incident nodes.

Let $\mathbf{W}(\eta)$ be the $p \times p$ matrix of edge weights parameterized by $\eta \in \mathbb{R}^k$. Equation 3.6 states that the log of the q-length vector containing the unique edge weights in $\mathbf{W}(\eta)$ is a linear combination of a collection of k basis functions of length q, with $\mathbf{V}^k$ defined as the eigenvectors of the line graph Laplacian corresponding to the k smallest eigenvectors. We wish to provide some insight into the interpretation of the coefficient vector $\eta = (\eta_1, ..., \eta_k)'$ and illustrate some of the covariance patterns this construction can produce.

As previously noted, the first eigenvector of our basis, $\mathbf{v}_1^k$ is always constant. This means that $\eta_1 \mathbf{v}_1^k$ is constant and may be viewed as the intercept for the log edge weights. Because $\exp(\mathbf{V}^k \eta) = \exp(\eta_1 \mathbf{v}_1^k) \odot \exp(\mathbf{V}_{-1}^k \eta_{-1})$, the vector $\exp(\eta_1 \mathbf{v}_1^k)$ may be viewed as a scaling factor for the edge weights matrix $\mathbf{W}(\eta_{-1})$ which is defined using the other $k-1$ basis functions. In the GDEF model, this scaling factor is directly related to the range of spatial correlation, while in the CAR model, the scaling factor is redundant when σ^2 is

estimated. As such, we recommend fixing $\eta_1 = 0$ when using the basis function approach to characterizing the edge weights matrix for CAR models. Subsequent coefficients in η may be interpreted in context of the eigenvectors with which they are associated. As depicted in Figure 3.2, the second LGL eigenvector of a 30×30 grid characterizes a contrast between the upper left and lower right corners of the spatial domain. Figure 3.6 depicts samples from different covariance models defined using only the first two LGL eigenvectors. Across covariance models it can be seen how observations in one corner of the spatial domain are highly correlated, while the effective range of spatial correlation is much smaller in the opposite corner. Spatial covariance patterns such as this do appear in real world settings. Paciorek and Schervish (2006) presents a data set for precipitation in Colorado in which observations east of the Rocky mountains are highly correlated over long distances, whereas spatial correlations between observations in the mountainous western half of the state decay rapidly over short distances.

Different linear combinations of basis functions can produce a wide array of covariance structures. Figure 3.7 depicts a linear combination of four eigenvectors that results in covariance where observations at the borders of the spatial domain are highly correlated, while observations in the interior are not. In effect locations at opposite corners of the grid are "closer" together than locations that are both near the center. The more basis functions are included, the more flexible the covariance, though there are generally diminishing returns as the number of eigenvectors used increases due to the higher spatial frequencies exhibited by later eigenvectors. Omitting these high frequency eigenvectors essentially assumes that the edge weights themselves exhibit some degree of spatial smoothness. Intuition on how to select the number of basis functions is provided later in this article.

3.3.3 Incorporating covariates

We may be interested in the question of whether certain environmental features inhibit or facilitate connectivity between regions in our spatial domain. As such we may wish to model edge weights as a function of environmental covariates. Generally speaking, we will

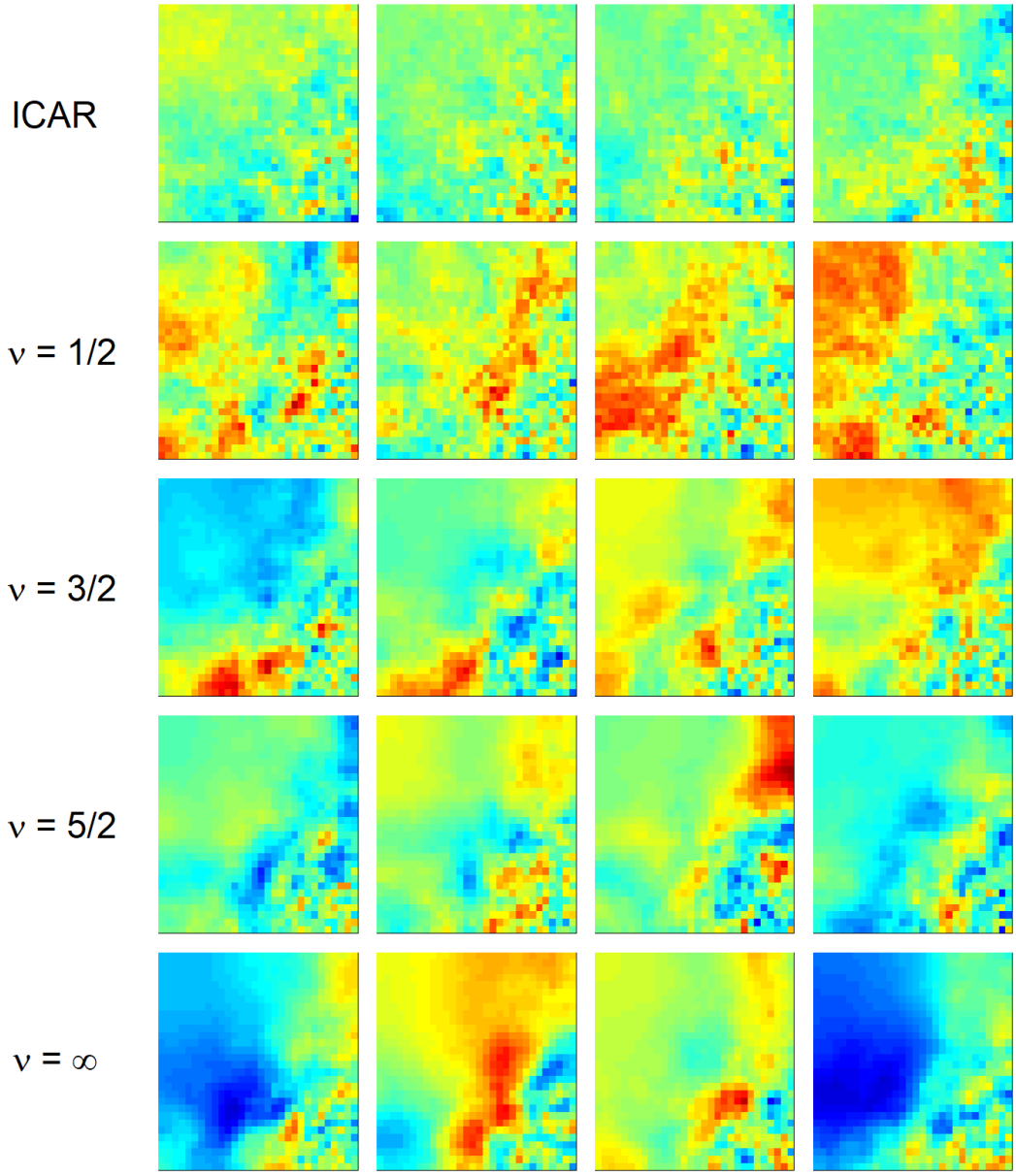


FIGURE 3.6: Samples from $N(\mathbf{0}, \Sigma)$ over a 30×30 grid using two basis functions with $\eta_1 = \eta_2 = 50$. Note how spatial correlation decays at different rates across the spatial domain.

not have covariates that are explictly associated with the edges of our graph. If $\mathbf{X}$ is a $p \times r$ matrix containing r features for each of the p nodes in a graph, Hanks and Hooten (2013) suggested defining $\mathbf{X}^E$, a $q \times r$ feature matrix for the graph's edges, by averaging the

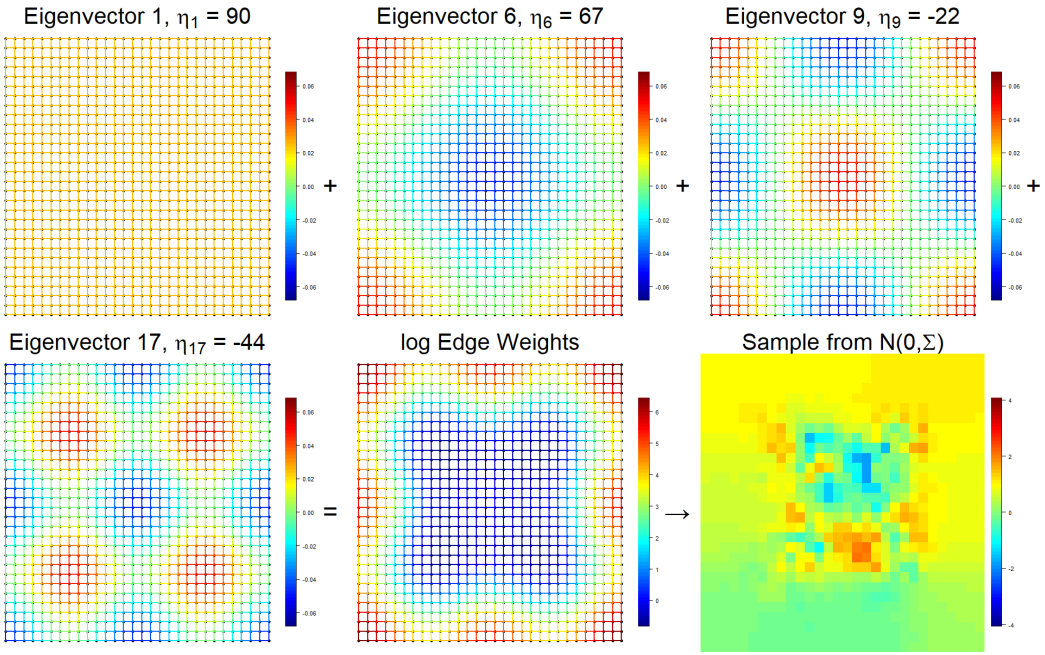


FIGURE 3.7: Illustration of how limited number of basis functions may combine to produce complex covariance structures. Σ defined using double exponential covariance.

features of adjacent nodes:

$$\mathbf{x}^{e_{ij}} = \frac{\mathbf{x}_i + \mathbf{x}_j}{2}. \tag{3.8}$$

Here $\mathbf{x}^{e_{ij}}$ is the row of $\mathbf{X}^E$ corresponding to the edge linking nodes i and j. The matrix $\mathbf{X}^E$ could also be constructed treating differences in the features of adjacent regions as a covariate itself. Others have suggested that the length of the shared border between two adjacent regions may be important to consider. However $\mathbf{X}^E$ is defined, we can augment the basis function model for edge weights in Equation 3.6 by setting

$$\mathbf{w} = \exp(\mathbf{X}^E\psi + \mathbf{V}^k\eta),\ \psi \in \mathbb{R}^r,\ \eta \in \mathbb{R}^k. \tag{3.9}$$

In order to avoid confounding between ψ and η due to correlation between $\mathbf{X}^E$ and $\mathbf{V}^k$, let $\mathbf{V}^\perp$ contain the eigenvectors of $\mathbf{X}^\perp\mathbf{L}^L\mathbf{X}^\perp$, where $\mathbf{X}^\perp = (\mathbf{I}_q - \mathbf{X}^E(\mathbf{X}^{E\top}\mathbf{X}^E)^{-1}\mathbf{X}^{E\top})$ is the projection onto the orthogonal compliment of the column space of $\mathbf{X}^E$. The $q \times k$ matrix $\mathbf{V}^{k\perp}$ contains the k eigenvectors corresponding to the smallest eigenvalues of $\mathbf{X}^\perp\mathbf{L}^L\mathbf{X}^\perp$. The matrix $\mathbf{X}^E$ is orthogonal to $\mathbf{V}^{k\perp}$, which may improve estimation and interpretation of ψ.

This approach to spatial deconfounding was introduced by Reich et al. (2006) and has seen widespread use since then, though there is debate as to whether such adjustments are appropriate (Zimmerman & Ver Hoef, 2022). While we have little to contribute to this discussion from a theoretical perspective, we find that using the orthogonalized $\mathbf{V}^{k\perp}$ results in a slightly more flexible covariance model when weights are defined using Equation 3.9.

3.3.4 Parameter estimation and model fitting

Given graph $G = (V, E)$ with p nodes and q edges, we make repeated observations of the collection of random variables $\{y_v : v \in V\}$; each repetition is recorded as $\mathbf{y}_i \in \mathbb{R}^p$ for $i \in 1, \ldots, n$. These observations may be stored in the $n \times p$ data matrix $\mathbf{Y}$, where y_{ij} is the ith observation at the jth node of G. For now, we assume that $\mathbf{y}_i \overset{iid}{\sim} N_p(\mathbf{0}, \Sigma)$, with Σ parameterized by $\theta = (\sigma^2, \eta)$ for the GDEF model and $\theta = (\sigma^2, \eta, \kappa)$ for the CAR model. Christensen and Hoff (2023) only considered settings with multiple observations of the spatial process. This was necessary due each edge weight being a unique parameter and the fact $q > p$ in almost all settings. The dimension reduction of the basis function approach to estimation of $\mathbf{W}$ allows us to model the covariance of spatial processes with only a single realization ($n = 1$).

The maximum likelihood estimator (MLE) for θ may be obtained using the Newton-Raphson algorithm which requires computation of the first and second derivatives of the log likelihood $\ell(\theta; \mathbf{Y})$ with respect to θ. Details on the form of $\nabla_\ell(\theta) = \frac{\partial \ell(\theta; \mathbf{Y})}{\partial \theta}$ and $\mathbf{H}_\ell(\theta) = \frac{\partial^2 \ell(\theta; \mathbf{Y})}{\partial \theta^2}$ for the models discussed in this article are provided in Appendix B. Given data $\mathbf{Y}$ and an initial set of parameter values θ^0 we can maximize the likelihood by iteratively calculating

$$\theta^l = \theta^{l-1} - \gamma \, \mathbb{E}\left[\mathbf{H}_\ell(\theta^{l-1})\right]^{-1} \left(\nabla_\ell(\theta^{l-1})\right) \tag{3.10}$$

until θ^l converges to the MLE. Here γ is a scaling parameter between 0 and 1. Smaller γ provides more robust optimization, while $\gamma = 1$ will generally be faster.

Uncertainty quantification may be conducted using the approximation

$$\hat{\theta} \sim N\left(\theta, -\left[\mathbf{H}_\ell(\hat{\theta})\right]^{-1}\right). \tag{3.11}$$

Note when employing the Newton-Raphson algorithm in Equation 3.10 that we use the expectation of the Hessian matrix $\mathbf{H}_\ell(\theta)$, which is less computationally intensive to obtain than the Hessian matrix itself. We only need to compute $\mathbf{H}_\ell(\theta)$ once after obtaining the MLE in order to use the approximation in Equation 3.11.

We have also used Bayesian estimation techniques to fit the model described in this article. Details on recommend prior choices and MCMC algorithms are also provided in Appendix B.

3.4 Illustrations

Within this section we evaluate the performance of our approach under a variety of settings and in comparison to alternative methods. We consider both simulated and real-world data examples.

3.4.1 Simulations

We begin with three simulation studies. In the first we evaluate approximate confidence interval coverage rates when the model is correctly specified. In the second, we consider how to select the number of basis functions when implementing our method. The third simulation assesses model performance under misspecification, evaluating the Kullback-Leibler divergence between the true data generating distribution and the estimated distribution.

Sim 1: Approximate confidence interval coverage

We defined 10×10 and 20×20 lattices, and obtained the LGL eigenvectors for each. We selected k eigenvectors and simulated $\eta \sim N_k(\mathbf{0}, \Phi)$, where Φ is a diagonal matrix with first entry (corresponding to the variance of the "intercept" eigenvector) equal to 0.5 and all other diagonal entries equal to 25. We then generated n samples from $N_p(\mathbf{0}, \Sigma(\eta))$ where

$\Sigma(\eta)$ is defined by Equations 3.4 and 3.6, using a Matérn covariance with $\nu = 3/2$ and fixing $\sigma^2 = 1$. This was done for each of $k \in \{10, 30\}$ and $n \in \{1, 10, 50\}$. We obtained the MLE $\hat{\eta}$ and calculated approximate 90% confidence intervals for η using Fisher's approximation. The simulation was repeated ten times, with coverage rates provided in Table 3.1.

Table 3.1: Coverage rates of approximate 90% confidence intervals. Monte Carlo standard error in parentheses.

	$n = 1$	$n = 10$	$n = 50$
$p = 100$, $k = 10$	0.880 ($\pm$ 0.039)	0.910 ($\pm$ 0.023)	0.870 ($\pm$ 0.040)
$p = 100$, $k = 30$	0.804 ($\pm$ 0.040)	0.867 ($\pm$ 0.027)	0.927 ($\pm$ 0.015)
$p = 400$, $k = 10$	0.920 ($\pm$ 0.025)	0.910 ($\pm$ 0.023)	0.910 ($\pm$ 0.031)
$p = 400$, $k = 30$	0.907 ($\pm$ 0.018)	0.900 ($\pm$ 0.021)	0.883 ($\pm$ 0.019)

In general coverage rates are appropriate, the exception being under the settings $k = 30$ and $n = 1$ over a 10×10 lattice. On top of having lower than expected coverage, Newton's method diverged in one of the ten simulations under these settings, indicating that one observation of a spatial process over a graph with $p = 100$ nodes may not be enough to inform the estimation of Σ using $k = 30$ basis functions. In general however, we feel that inference based on Fisher's approximation to the distribution of the MLE is appropriate. We investigate choice of k in the following simulation.

Sim 2: Number of basis functions

Once again, we defined 10×10 and 20×20 lattices obtaining LGL eigenvectors for each. Instead of simulating edge weights directly from Equation 3.6 we generate an edge weights matrix $\mathbf{W}$ from an ICAR model using the line graph adjacency matrix to define the associations between edges. Figure 3.8 depicts sampled edge weights over a 20×20 lattice. The spatial dependence in the simulated edge weights themselves means that $\mathbf{W}$ can be well approximated using our method. Larger numbers of basis functions will improve this approximation, but at the cost of model complexity and potential computational instability. The simulated $\mathbf{W}$ was used to define a covariance matrix according to Equation 3.4 with $\nu = 3/2$ and $\sigma^2 = 1$. Independent samples were generated from $N_p(\mathbf{0}, \Sigma)$ and stored in

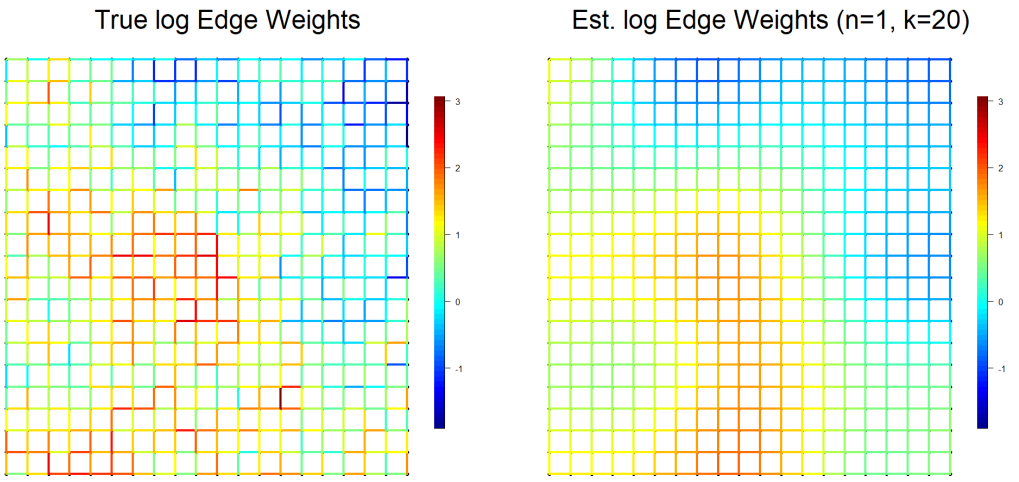


FIGURE 3.8: Simulated $\mathbf{W}$ on left, with $\hat{\mathbf{W}}$ on right (estimated from $n = 1$ observations and using $k = 20$ basis functions). Approximation "smooths" the true edge weights.

$n \times p$ matrix $\mathbf{Y}$ for $n \in \{1, 10, 50\}$. For each simulated $\mathbf{Y}$ we fit a GDEF model using Matérn covariance with $\nu = 3/2$ using $k \in \{10, 20, ..., 100\}$ basis functions and evaluated AIC and BIC for each of the 10 model fits, with lower values of

$$\begin{aligned} \text{AIC} &= -2\ell(\eta; \mathbf{Y}) + 2k \\ \text{BIC} &= -2\ell(\eta; \mathbf{Y}) + k \log(np) \end{aligned} \tag{3.12}$$

begin selected. The simulation was repeated 10 times. Table 3.2 contains information regarding the AIC- and BIC-preferred choices of k.

Table 3.2: Mode k indicates the most commonly selected number of basis functions, while Mean k indicates the average number of selected basis functions. Larger grids and larger sample sizes result in preference for higher k.

	Mode k (AIC)	Mean k (AIC)	Mode k (BIC)	Mean k (BIC)
$p = 100$, $n = 1$	10	12	10	10
$p = 100$, $n = 10$	40	58	10	11
$p = 100$, $n = 50$	100	100	40	42
$p = 400$, $n = 1$	20	37	10	10
$p = 400$, $n = 10$	100	99	30	39
$p = 400$, $n = 50$	100	100	100	99

Across simulation settings, we see that models with higher k perform better on the

larger grid and with bigger sample sizes. This is unsurprising as larger amounts of data should be expected to support more complicated models. AIC and BIC are considered not because they are objectively the best tools for model selection, but because they provide a likelihood-based starting point for the question of how to choose the number of eigenvectors when implementing our method. BIC is more conservative than AIC due to a larger model size penalty. The two criteria also differ in that AIC is calibrated for improving predictive accuracy, while BIC may be viewed as favoring the model that best approximates the marginal distribution of the data (Gelman et al., 2014). The fact that the two criteria frequently prefer different values of k provides some intuition for the range on model size that may be appropriate. Using the information in Table 3.2, a reasonable starting point is to select $k = \sqrt{np}$ when the model reasonably approximates the true data generating process, as was the case within this simulation. Optimal choice of k will also be contingent on the nature of the spatial process itself, as evidence of more complicated patterns of spatial dependence may indicate that larger numbers of basis functions should be used. Table 3.3 contains the rates at which the Newton-Raphson algorithm used to estimate $\hat{\eta}$ failed to converge across different simulation settings, with the majority of instances occurring for large k and $n = 1$. When computational instability is encountered, refitting with smaller k will frequently resolve the issue; the possibility of instability due to significant model misspecification should however be considered.

Table 3.3: Frequency with which Newton-Raphson algorithm failed to converge. Estimation was stable for all $k < 40$ and for all settings over the 20×20 lattice.

	$k = 40$	$k = 50$	$k = 60$	$k = 70$	$k = 80$	$k = 90$	$k = 100$
$p = 100$, $n = 1$	0.3	0.7	0.8	1.0	1.0	1.0	1.0
$p = 100$, $n = 10$	0.0	0.0	0.0	0.0	0.0	0.2	0.1
$p = 100$, $n = 50$	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Sim 3: Model performance under misspecification

The previous simulation evaluated the performance of the GDEF model for different values of k when the true data generating process was a GDEF model with an edge weights

matrix from outside of the parameter space characterized by Equation 3.6. We now consider the performance of the GDEF model when data comes from a distinct nonstationary covariance model. Given a nonstationary spatial process $Z(\mathbf{s})$ observed at $\mathbf{s} \in \mathbb{R}^2$ the spatial deformation approach of Sampson and Guttorp (1992) defines the process' covariance such that

$$\text{Cov}(Z(\mathbf{s},\mathbf{s}')) = \sigma^2\rho(||g(\mathbf{s}) - g(\mathbf{s}')||) \tag{3.13}$$

where $g(\cdot)$ is a transformation from the geographic location $\mathbf{s}$ to its location $g(\mathbf{s})$ in a "deformed" space wherein covariance is stationary and isotropic.

Inspired by this approach, we simulate data by first defining a 15×15 lattice. We then deform the grid by randomly selecting ten "deformation points" $d_{1:10} \in (1,15)^2$ before sequentially shrinking or expanding the grid towards or away from these locations, as illustrated by Figure 3.9. Let $\mathbf{D}^D$ be the 225×225 matrix of Euclidean distances between the nodes of the deformed lattice. Applying a Matérn correlation function to $\mathbf{D}^D$ results in a covariance matrix that is nonstationary with respect to $\mathbf{D}$, the matrix of Euclidean distances between the original grid point locations. We generate n observations from the following model

$$\begin{aligned} \mathbf{y}_i &= \beta_0\mathbf{1} + \mathbf{z}_i + \epsilon_i \\ \mathbf{z}_i &\sim N_p(\mathbf{0}, \sigma^2\Phi) \\ \epsilon_i &\sim N_p(\mathbf{0}, \tau^2\mathbf{I}_p) \end{aligned} \tag{3.14}$$

where $\mathbf{z}_i$ is a vector of spatial random effects with covariance matrix Φ defined using a Matérn correlation function with smoothness $\nu = 5/2$

$$\Phi = \left(1 + \frac{\sqrt{5}}{\rho}\mathbf{D}^D + \frac{5}{3\rho^2}\{\mathbf{D}^D\}^{\circ 2}\right) \odot \exp\left(\frac{\sqrt{5}}{\rho}\mathbf{D}^D\right) \tag{3.15}$$

We set $\beta_0 = 0$, $\sigma^2 = 0.9$, $\tau^2 = 0.1$, $\rho = 3$. We fit the GDEF and other models to the simulated $\mathbf{Y}$ as illustrated by Figure 3.10, which depicts a deformed lattice, samples of $\mathbf{y}_i$ produced by that deformation, and an estimated edge weights matrix.

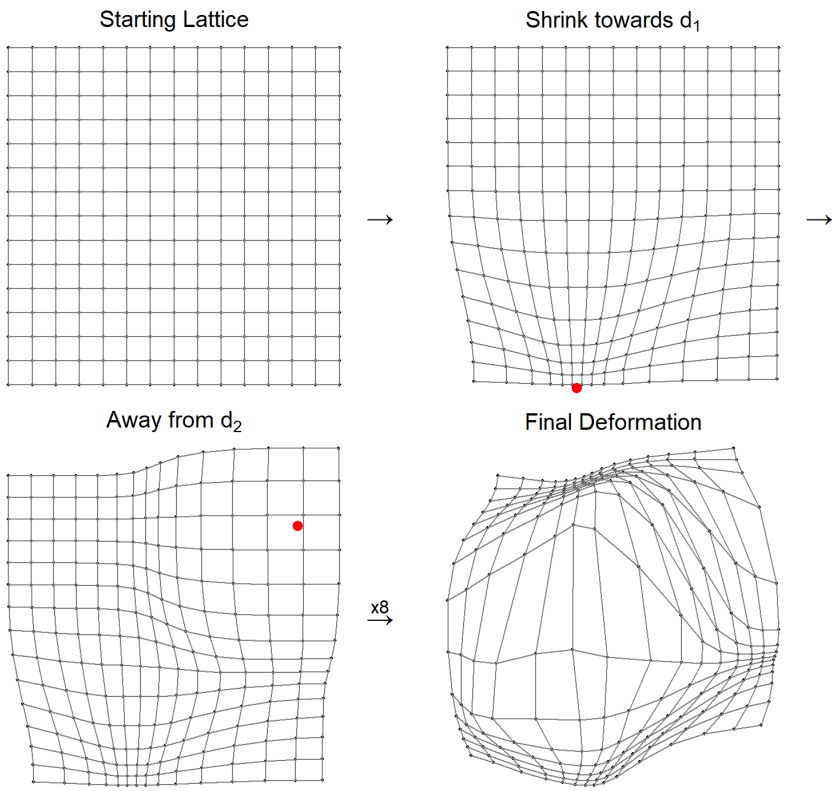


FIGURE 3.9: Deformed lattice obtained by repeatedly expanding and contracting grid around random locations. Each simulation replicate used a uniquely generated deformation.

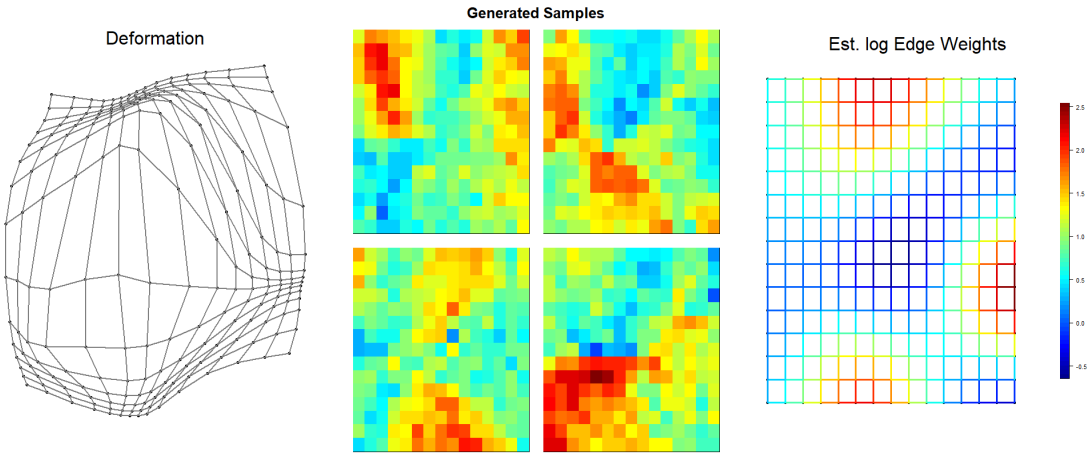


FIGURE 3.10: Distances between deformed grid locations define a nonstationary covariance matrix Σ. Data $\mathbf{Y}$ generated from $N_p(\mathbf{0}, \Sigma)$. GDEF model is fit to $\mathbf{Y}$, producing the depicted $\hat{\mathbf{W}}$ and corresponding $\hat{\Sigma}^{GDEF}$.

We randomly generated 125 unique grid deformations. For each deformation we sampled $\mathbf{Y}$ as described above for $n \in \{1,2,3,5,10,25,50\}$ and fit Equation 3.14 while defining the covariance of the random effects $\mathbf{z}_i$ using the following six submodels: a GDEF model with $\nu = 3/2$ and $k = 15$ basis functions, a GDEF model using the double exponential covariance function ($\nu = \infty$) and $k = 15$ basis functions, the spatial deformation approach of Sampson and Guttorp (1992) (S&G) as implemented by the `deform` R package (Youngman, 2023), a simple Matérn covariance model with $\nu = 5/2$ and using the Euclidean distances between non-deformed lattice nodes, the method presented in Hughes and Haran (2013) (H&H) for modeling random effects in areal data using 100 basis functions, and a standard CAR model using an unweighted adjacency matrix. The spatial deformation approach and the stationary Matérn covariance model treat the data as though it were point-indexed, while all other approaches utilize only the lattice's graphical structure; note however, that the true data generating process exists in continuous space.

Let $\Sigma = \sigma^2\Phi + \tau^2\mathbf{I}_p$ be the marginal covariance of $\mathbf{y}_i$ integrated over the random effects $\mathbf{z}_i$. For each $\mathbf{Y}$ and fitted model, we obtain the MLE $\hat{\Sigma}$ and compute $\mathrm{KL}(N_p(\mathbf{0},\Sigma),N_p(\mathbf{0},\hat{\Sigma}))$, the Kullback-Leibler (KL) divergence between the true data generating distribution and the distribution characterized by the estimated covariance. As both distributions are mean-zero and Gaussian, KL divergence may be computed using only the covariance matrices Σ and $\hat{\Sigma}$:

$$\mathrm{KL}(\Sigma,\hat{\Sigma}) = \frac{1}{2}\left(\log\frac{|\hat{\Sigma}|}{|\Sigma|} + \mathrm{tr}\left(\hat{\Sigma}^{-1}\Sigma)\right) - p\right) \tag{3.16}$$

Table 3.4 depicts the average KL divergence between the true distributions and the distributions estimated by the six models. Note that the true distribution is different for each of the 125 replicates which used a unique deformation. We also consider the frequency with which each model had the smallest KL divergence under the same settings. As can be seen, the GDEF model (for both choices of ν) performs well across sample sizes, with the simple Matérn model for $n = 1$ being the only alternative that was frequently competitive

in this simulation. While potentially surprising that the simple Matérn covariance model using the non-deformed distances outclasses many of the more complicated alternatives under some settings, it is clear that its relative performance is strongest for small n; the true data generating process also used a Matérn covariance function with $\nu = 5/2$, and a visual assessment of simulation output suggests that the simple Matérn performs best when the randomly generated deformation exhibits less dramatic patterns of warping.

Table 3.4: Average $\mathrm{KL}(\Sigma, \hat{\Sigma})$ for each model and number of samples. In parentheses is the frequency across simulations that a given model produced the smallest KL divergence for a sample of size n. The GDEF models perform best across all sample sizes.

	GDEF 3/2	GDEF ∞	S&G	Matérn 5/2	H&H	Std. CAR
$n = 1$	42.54 (0.48)	46.66 (0.192)	-	45.22 (0.328)	96.06	76.74
$n = 2$	28.2 (0.576)	29.61 (0.376)	73.6	42.35 (0.048)	144.81	79.92
$n = 3$	24.24 (0.6)	24.74 (0.384)	40.29	40.87 (0.016)	77.6	81.14
$n = 5$	21.93 (0.456)	21.94 (0.536)	32.33 (0.008)	39.75	55.57	83.24
$n = 10$	20.6 (0.304)	20.4 (0.688)	29.54 (0.008)	39.13	48.4	83.95
$n = 25$	19.83 (0.248)	19.56 (0.728)	28.53 (0.024)	38.65	46.31	84.45
$n = 50$	19.55 (0.24)	19.25 (0.744)	28.3 (0.016)	38.36	45.9	85.08

Note that the spatial deformation approach (S&G) cannot be fit for $n = 1$. This highlights one of the key advantages of our approach. Much of the literature regarding the spatial deformation framework requires that the spatial process be observed multiple times, and in some implementations requires that $n > p$. However, it is extremely common for spatial analyses to be conducted on datasets with little or no replication. Whether the deformation function is estimated using thin plate splines as in Sampson and Guttorp (1992) or a Gaussian process prior as in Schmidt and O'Hagan (2003), the inherent flexibility of the framework requires considerable amounts of data (or strong prior assumptions) to avoid over fitting. As seen in this and previous simulations, the GDEF model using the basis function representation of $\mathbf{W}$ as presented in this paper can effectively estimate complicated patterns of spatial dependence for areal data and performs well in a setting where the true spatial process exists in continuous rather than discrete space.

3.4.2 Example: Mercer and Hall wheat yield data

We conclude this section with an analysis of the classic wheat yield dataset from Mercer and Hall (1911) and which is available in the `spData` R package. The version of the data used in the package was taken from Cressie (1993). Mercer and Hall considered a one acre parcel of farmland that was divided into 500 rectangular plots and planted with wheat. Each plot received the same treatment and soil quality was perceived to be essentially uniform at the time of planting. At harvest, the yield of each plot in pounds of grain was recorded. Figure 3.11 depicts the yields for each plot, which were arranged in 20 rows each containing 25 plots. Mercer and Hall discuss the variability present in the data and explore the possibility of location-based effects on yield. Noting evidence of spatial correlation, the wheat yield data was analyzed in Whittle (1954), which first introduced the simultaneous autoregressive (SAR) model, and in Besag (1974), which first introduced the CAR model. The authors of both papers found the fit of their models to the wheat yield data unsatisfactory. Looking at Figure 3.11, one notes that certain columns (4,7 and 10) have larger average yields than others and that within-column correlation appears to be stronger than within-row correlation. By modeling the data using fixed row and column effects Cressie (1993) found that a stationary and isotropic kriging model with plot locations given by centroids resulted in satisfactory fit, though it is worth considering whether the inclusion of row and column effects in a model for this data provide interpretive value or merely serve to improve the fit of an otherwise poorly specified model.

We propose modeling this data according to Equation 3.14 with the covariance of the random effects defined by a GDEF model with smoothness $\nu = 3/2$ and $k = 20$ basis functions. To allow for a distinction between within-column and within-row dependency, we split the intercept vector $\mathbf{v}_1$ of our basis into two vectors, $\mathbf{v}_1^{row}$ and $\mathbf{v}_1^{col}$, where $v_{i1}^r = 1$ if edge i connects two nodes in the same row and is 0 otherwise, with $\mathbf{v}_1^{col}$ defined similarly for columns. We obtain maximum likelihood estimates $\hat{\eta}$, $\hat{\sigma^2}$, $\hat{\tau^2}$ and $\hat{\beta}_0$ which are reported in Table 3.5. Negative $\hat{\eta_1}^{row}$ and positive $\hat{\eta_1}^{col}$ indicates much stronger within-column than

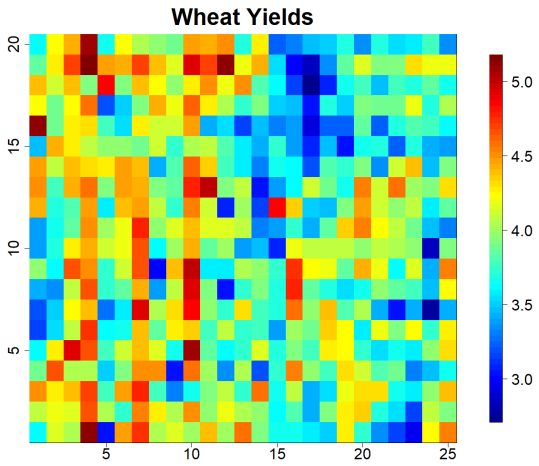


FIGURE 3.11: The Mercer Hall wheat yield data

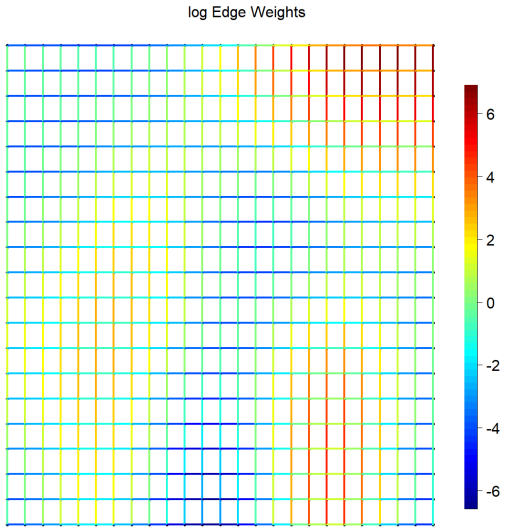


FIGURE 3.12: Estimated log edge weights

within-row correlations. Rather than reporting the other coefficients in $\hat{\eta}$ which are individually less interpretable, Figure 3.12 depicts $\log(\hat{\mathbf{w}})$ over the entire network; there are three regions within the spatial domain with larger edges indicating stronger spatial correlation between nodes in those locations.

We can also obtain an estimate of $\mathbf{z}$, which may be thought of as the smoothed or "de-

Table 3.5: MLEs and confidence intervals for model parameters.

	Est.	95% CI
η_1^{row}	-2.195	(-3.551, -0.838)
η_1^{col}	1.308	(0.286, 2.329)
σ^2	0.146	(0.101, 0.191)
τ^2	0.073	(0.027, 0.118)
β_0	3.949	(3.871, 4.026)

noised" version of the centered data. Because $\mathbf{y}$ and $\mathbf{z}$ are each normally distributed, the conditional distribution of $\mathbf{z}$ is straightforward to obtain:

$$\mathbf{z}|- \sim N_p\left(\left(\Phi^{-1}/\sigma^2 + \mathbf{I}_p/\tau^2\right)^{-1}(\mathbf{y} - \beta_0\mathbf{1})/\tau^2, \left(\Phi^{-1}/\sigma^2 + \mathbf{I}_p/\tau^2\right)^{-1}\right). \tag{3.17}$$

We define $\hat{\mathbf{z}} = \left(\hat{\Phi}^{-1}/\hat{\sigma^2} + \mathbf{I}_p/\hat{\tau^2}\right)^{-1}(\mathbf{y} - \hat{\beta_0}\mathbf{1})/\hat{\tau^2}$ and provide a plot of $\hat{\mathbf{z}}$ in Figure 3.13 alongside $\mathbf{y} - \hat{\beta_0}\mathbf{1}$. Moran's I (Moran, 1950) is a test statistic indicating the presence or absence of spatial autocorrelation. Moran's I for the residuals of our model $\epsilon = \mathbf{y} - \hat{\beta_0}\mathbf{1} - \hat{\mathbf{z}}$ is 0.022, with a p-value of 0.469, indicating a lack of spatial correlation in ϵ. The Shapiro-Wilk normality test (Shapiro & Wilk, 1965) on the residuals results in a p-value of 0.262 indicating that the normality assumption for ϵ is not violated. We find that the GDEF model fits the wheat yield data well, effectively representing the distinctive and irregular patterns of spatial dependence (seen especially in the columns of the data) in an uncontrived manner.

3.5 Discussion

Within this article we introduced a framework for estimating edge weight matrices using the eigenvectors of the line graph's Laplacian matrix. We show that this method may be used to enhance existing covariance models for areal data. We compared and contrasted the classically-used CAR model with the GDEF model introduced by Christensen and Hoff (2023), illustrating how both may be improved by this new framework. We evaluated properties of this method in a series of simulations and data examples. In general we find

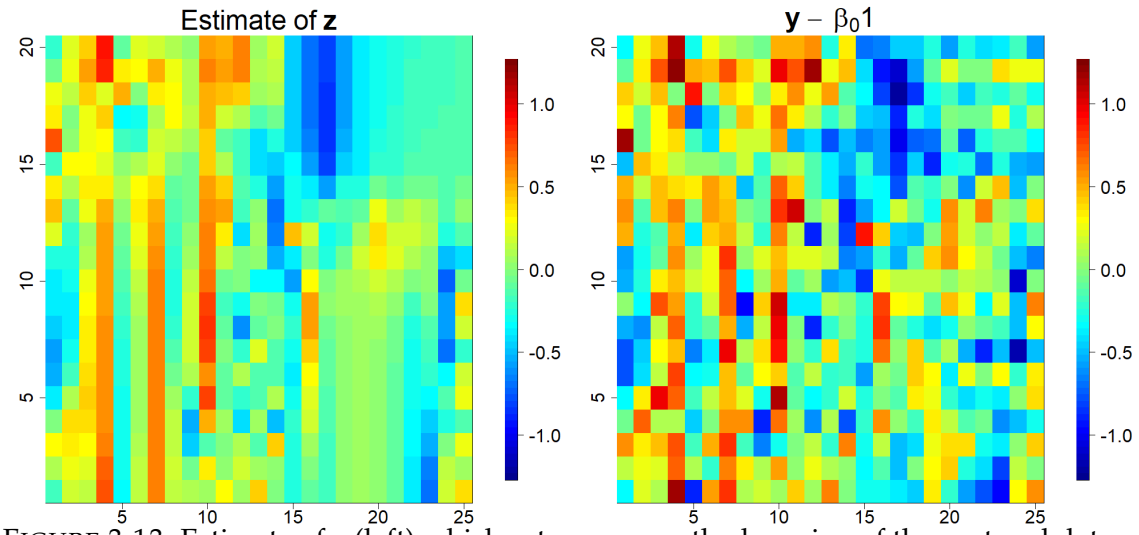


FIGURE 3.13: Estimate of **z** (left) which acts as a smoothed version of the centered data (right)

that the GDEF model is more flexible and capable of representing covariance structures with properties that are generally preferred by researchers, and that the introduced framework significantly improves the computational efficiency of the GDEF approach. We conclude this article with a brief discussion of potential avenues for future work and model extensions.

We are currently working on the development of an R package which may be used to easily implement our methodology. We also wish to evaluate how the INLA computing framework (Rue et al., 2009) may be used to complement or improve upon the model fitting approaches utilized within this article.

Part of our assessment that the GDEF model exhibits more desirable behavior than the CAR model stems from findings that the latter characterizes fundamentally non-smooth spatial processes. Rue and Held (2005) and Lindgren et al. (2011) demonstrate that it is possible to define Markov random fields (MRFs) that approximate smooth spatial processes by allowing some edge weights to be negative. The space of non-CAR MRFs is worth exploring due to their computational advantages. It may be possible to adapt our method for edge weight estimation to allow for the inclusion of negative weights in sparse precision matrix formulations, though additional work would be necessary to ensure that a valid

covariance matrix is always produced.

While we have discussed estimation of $\mathbf{W}$ within the context of explicitly defining a covariance matrix, there are other settings in which a coherent framework for flexibly defining an edge weights matrix may be valuable. For example, resistance distance is a graph metric constructed using a matrix of non-negative edge weights that has seen widespread use in many statistical and ecological applications (Dickson et al., 2019; Hanks & Hooten, 2013; Klein & Randić, 1993). Even if not utilized for the purpose of defining distances or covariances between locations, an edge weights matrix may be viewed as characterizing the connectivity of different regions or entities within a network.

An additional topic we would like to explore is how our method could be used to characterize the covariance of spatial processes on manifolds or restricted spatial domains. While most spatial models were been developed assuming that data are observed on some well-defined subset of $\mathbb{R}^2$, there is increased interest in the development of methods that can be applied to more complex domains in which Euclidean distances between observations do not accurately represent proximity with respect to the domain's underlying geometry (Yang & Dunson, 2016). Mathematical distinctions exist between settings in which data are observed in Euclidean space but concentrated near a lower-dimensional manifold such as the surface of a sphere and settings in which which an n-dimensional spatial process exists only on an irregular subset of $\mathbb{R}^n$ due to underlying physical constraints (Dunson et al., 2022). Sangalli et al. (2013) provides an example of the latter setting in which population density on the Island of Montréal is modeled using a spatial process which is undefined beyond island borders and within the boundaries of an airport and industrial park. Whether data is observed on a true manifold or a restricted domain, we could define a graph based on the set of observed locations, defining edges between locations based on neighbors within a small radius or through use of the Delauney triangulation (Hjelle & Dæhlen, 2006). The LGL eigenvectors of this graph could then be utilized as in this article to define a valid covariance model over the complex spatial domain; furthermore, the flexibility of our edge weight estimation approach would ultimately allow covariance to be nonstationary with

respect to the domain's underlying geometry. Some model extensions would be required in order to allow covariance to be defined continuously between observations, as our present framework is defined only at the nodes of a graph. Insight from Lindgren et al. (2011) and B. S. Lee and Haran (2022) provide potential frameworks for solving this issue.

4. Incorporating Circuit Theory Into a Dynamic Model for Crowd-sourced Observations of Migratory Birds

4.1 Introduction

The migratory patterns for many bird species found within the northern hemisphere can be broadly characterized by northward flight to breeding grounds in the spring, and southward flight to wintering grounds during the fall. While the overarching patterns observed during bird migration are well documented, there is scientific interest in understanding how these migratory patterns are evolving, or may be expected to evolve over time (Barton & Sandercock, 2018), in response to climate change (Visser et al., 2009) or loss of habitat (Sutherland, 1996). There is also interest in developing methods to better identify species- and population-specific flyways (Buhnerkempe et al., 1996), the routes used by migratory birds, and which tend to be characterized by environmental features conducive migration, such as abundance of food, water, and shelter and lack of mountain ranges or major shifts in elevation (Alerstam, 1993). In addition to these questions, many of the mechanisms behind bird migration are poorly understood at present. There is for example little consensus regarding what actually triggers the onset of migration season for any given species, whether a response to shifting day lengths, local climate conditions or innate biological factors (Helbig, 2003; The Cornell Lab, 2021).

Studies of avian migration generally require the collection of data on the location across time of individual birds or bird populations. Telemetry data, produced by capturing, tagging and monitoring the position of individual specimens in real time via satellite, GPS or radio (Dunn & Gipson, 1977; Perras & Nebel, 2012), is one of the most important data

types for studies of animal movement (Hooten et al., 2017). However, the collection of such data is time and resource intensive. and may often lead to small sample sizes and limited geographic domains (Hebblewhite & Haydon, 2010). It has also been noted that traditional telemetry data cannot be obtained for bird species which are too small to carry the tracking devices used to monitor location (Perras & Nebel, 2012). Professional monitoring surveys can also be a valuable tool for understanding species abundance within a studied region, (Lindenmayer & Likens, 2010) but once again, such data is both expensive and time consuming to obtain and provides only a snapshot of the species distribution during the time and within the vicinity that the survey was conducted (Caughlan & Oakley, 2001). Given that migratory birds may inhabit locations spanning thousands of kilometers during the course of a single year (Helbig, 2003), individual monitoring surveys are fundamentally limited in their ability to inform scientists of larger-scale migratory patterns. Given the limitations of these data types, we wish to consider the feasibility and potential value in using crowdsourced observations in the study of bird migration.

4.1.1 eBird data

Bird watching as a hobby has seen considerable growth in recent decades (Schwoerer & Dawson, 2022). The eBird database is a project managed by the Cornell Lab of Ornithology that crowdsources observational data by allowing any user to record and submit their bird watching observations to the website. Observations may be reported simply by indicating that a bird from a given species was observed at the user's location, or more rigorously in the form of time- and location-indexed "complete checklists" containing an exhaustive list of the counts and species of every bird observed by the submitting user, along with the length of time spent bird watching at that location (Sullivan et al., 2009). First launched in 2002 in collaboration with the Audubon society, the database has since recorded observations of over one billion birds and contains records from every country on earth (Team eBird, 2024). As the contributing user base and volume of freely available data has increased, so too has an awareness of the eBird database's value for scientific research. Numerous peer-reviewed

studies (e.g. Adde et al., 2021; Bianchini & Tozer, 2023; Hass et al., 2022; Hochachka et al., 2023; Tang et al., 2021) of bird population and species distribution trends utilizing the eBird database have been conducted and published. While the amateur collection process and crowdsourced nature of these data has raised questions about their reliability and bias (Hochachka et al., 2021; Zhang, 2020), studies have demonstrated that the population trends based on the complete checklists within the eBird database closely match those found in governmental monitoring surveys provided that a sufficient numbers of checklists were taken and reported within the same region (Horns et al., 2018; Stuber et al., 2022).

Within this article, we develop a model for data from the eBird database that is designed not only to reflect, but to help us better understand and identify the migratory trends present within this dataset. As part of this effort, we downloaded all completed checklists from the eBird database submitted during the years 2013 to 2017 within the seventeen easternmost U.S. states and Washington D.C. (a total of over five million checklists representing over five million combined hours of bird watching). This geographic region was chosen as it contains all U.S. states generally considered to be part of the Atlantic migratory flyway (Fritts, 2022). Data were aggregated temporally by week and spatially by county, for a total of 260 weeks and 805 counties (or county-equivalents). For each week-county pairing we obtain the vector containing total counts of each species observed, as well as the total combined hours of bird watching represented by all checklists submitted within each week-county pairing. The observation efforts that produced these data are highly variable in space and time, with some counties (often those containing major population centers or state/national parks) having hundreds of hours of combined observations during a given week, while many pairings have no observations whatsoever. Approximately 29% of the week and county combinations are represented by no checklists, indicating the potential ecological value in a model which can predict species trends for regions and time periods with little to no observational data.

Figure 4.1 contains a visualization for observations of the Baltimore oriole, a migratory bird common to the eastern United States, during the week beginning on February 20, 2017.

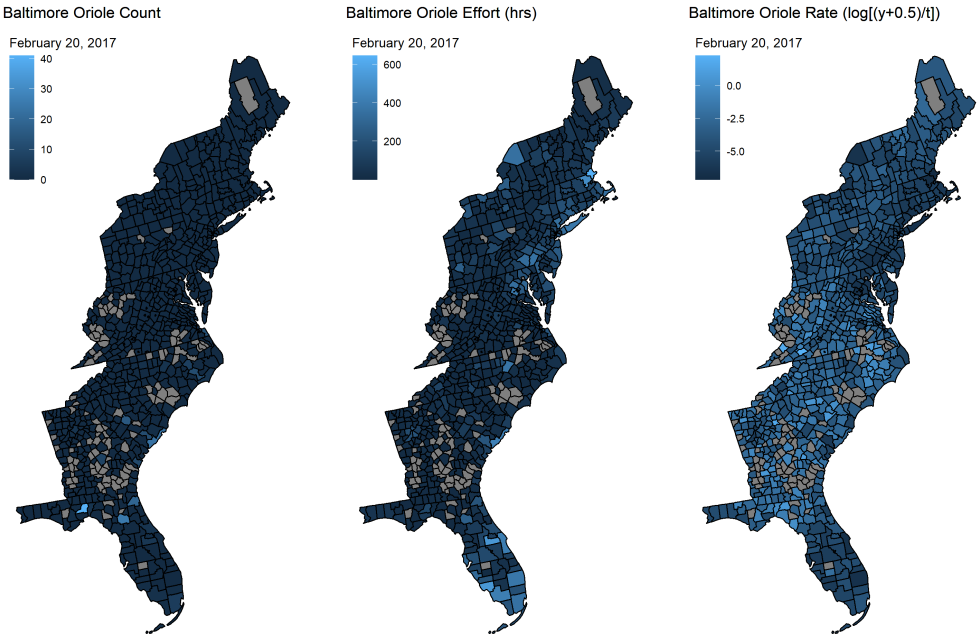


FIGURE 4.1: Counts, efforts, and log observation rates for Baltimore orioles from week beginning February 20, 2017. Grey regions indicate counties with zero submitted checklists.

It depicts the county-level counts of Baltimore orioles, observation efforts, and the log rates of observed birds per hour. Figure 4.2 provides a visualization of the "average latitude" within the eastern United States over time for the Baltimore oriole and the yellow-rumped warbler, another migratory species found in this region, computed using the weighted average latitude of all county centroids, with the count of birds observed per hour for each county as weights. Average latitude increases every spring and decreases every fall illustrating a strong cyclical pattern of migration for this species, and demonstrates that the empirical trends within this data are consistent with our expectations regarding migratory birds.

Within the following section we establish some of the notation, background, and conceptual motivation for our model for observations of migratory species found within the eBird dataset. This includes a discussion of precedent for the use of circuit theory in ecological applications. In section 3 we present the full model and discuss its construction,

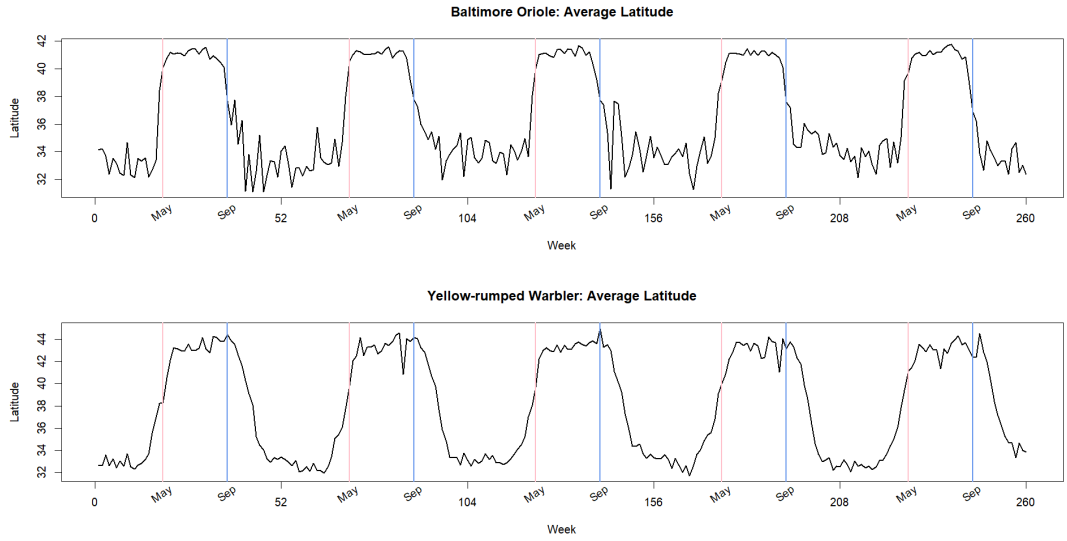


FIGURE 4.2: "Average latitude" of the Baltimore oriole and yellow-rumped warbler over time. County latitudes averaged, weights given by county-level counts per hour. The average location moves north by several degrees every May and south every September.

parameterization and interpretation. We also provide details regarding prior specification and computational implementation, along with it's performance using simulated datasets. In section 4 we provide results from our analysis of the eBird data before concluding with a discussion of ways the model may be used to better understand certain migratory trends, along with remarks on potential model extensions and future work we would like to undertake as part of this research.

4.2 Background

To model the observed species counts within the eBird dataset, we define $\mathcal{D}$ as the spatial domain consisting of the seventeen US states and the District of Columbia as depicted in Figure 4.1. The 805 counties contained within $\mathcal{D}$ form a partition that may be naturally represented using the graphical structure $G = (V, E)$ where V is a set of nodes corresponding to each county within $\mathcal{D}$ and E is the set of edges corresponding to pairs of adjacent counties. The structure of G may be mathematically represented by A, the 805×805 adjacency matrix constructed such that $a_{jj} = 0$, and $a_{jk} = 1$ if county j is adjacent to

county k (denoted $j \sim k$) and $a_{jk} = 0$ otherwise. As noted in the previous section, species counts were aggregated spatially by county and temporally by week. For a single species, e.g. the Baltimore oriole, let $\mathbf{y}_i$ be the length-805 vector containing the total observed specimen counts during week $i \in \{1,...,260\}$ for each county in $\mathcal{D}$. We define $\mathbf{t}_i$ to be the accompanying vector of observation efforts, containing the combined hours of bird watching within each county during week i. If no complete checklists were submitted for county j during week i then $t_{ij} = y_{ij} = 0$. We propose modeling $\{\mathbf{y}_1,..,\mathbf{y}_{260}\}$ using a hidden Markov model that incorporates a novel temporally evolving spatial transition structure designed to characterize migratory trends over time. This class of dynamic model is commonly used in the study of spatiotemporal and network data, and can be generalized to settings with non-Gaussian data, nonlinear dependence, and non-additive noise (Wikle & Hooten, 2010).

4.2.1 Hidden Markov models

Suppose we are given a collection of random vectors $\{\mathbf{y}_i\}_{1:n}$ which are observed at discrete times $i \in \{1,...,n\}$ within a spatial domain $\mathcal{D}$. Let $\mathbf{y}_i = \big(y(\mathbf{s}_1;i),...,y(\mathbf{s}_p;i)\big)'$, where $y(\mathbf{s}_j;i)$ is a noisy observation of some latent spatio-temporal process $\{z(\mathbf{s};i)\}_{\mathbf{s}\in\mathcal{D}}$ at time i and location $\mathbf{s}_j \in \mathcal{D}$. We wish to model $\mathbf{z}_i$, the latent process at time i realized at the same set of p locations $\{\mathbf{s}_1,...,\mathbf{s}_p\} \in \mathcal{D}$ where $\mathbf{y}_i$ was observed. A first-order hidden Markov model (HMM), utilizes the following conditional independence assumptions, where $\pi(\mathbf{y}_i|\mathbf{z}_i)$ denotes the distribution of $\mathbf{y}_i$ conditional on $\mathbf{z}_i$:

$$\begin{aligned} \pi(\mathbf{y}_i|\mathbf{z}_1,...,\mathbf{z}_n) &= \pi(\mathbf{y}_i|\mathbf{z}_i) \\ \pi(\mathbf{z}_i|\mathbf{z}_{i-1},...,\mathbf{z}_1,\mathbf{z}_0) &= \pi(\mathbf{z}_i|\mathbf{z}_{i-1}). \end{aligned} \tag{4.1}$$

This implies that the joint distribution of $(\mathbf{y}_1,...,\mathbf{y}_n,\mathbf{z}_1,...,\mathbf{z}_n)$ may be written as follows:

$$\pi(\mathbf{y}_1,...,\mathbf{y}_n,\mathbf{z}_1,...,\mathbf{z}_n) = \Pi_{i=1}^{n}[\pi(\mathbf{y}_i|\mathbf{z}_i)\pi(\mathbf{z}_i|\mathbf{z}_{i-1})]\pi(\mathbf{z}_0) \tag{4.2}$$

HMMs have seen widespread usage in many environmental and ecological applications (Glennie et al., 2023; Holsclaw et al., 2016; McClintock et al., 2020) especially those in-

volving models for animal movement (Hooten et al., 2017; Patterson et al., 2008). As shown in Thygesen et al., 2009, which models the location of fish populations using a HMM, this model may be thought of as the discrete (in space and time) approximation of the advection-diffusion equation which describes the transfer of particles or energy within a physical system and which has been used to characterize models for animal movement (Prima et al., 2018; Turchin, 1998).

Hidden Markov models are generally constructed using two components, the first being a model for observed data $\mathbf{y}_i$ given the latent process $\mathbf{z}_i$, while the second component characterizes the evolution of the latent process in time. The most commonly used model for $\mathbf{z}_i|\mathbf{z}_{i-1}$ is to specify that $\mathbb{E}[\mathbf{z}_i] = \mathbf{M}\mathbf{z}_{i-1}$, where $\mathbf{M}$ is a $p \times p$ matrix of transition probabilities or weights characterizing between-location dependence (McClintock et al., 2020; Wikle & Hooten, 2010).

While there are many ways to define the transition matrix $\mathbf{M}$, a common choice when dealing with spatial or graphical data is to define transition probabilities as a one-step random walk, such that $m_{jk} > 0$ if $j \sim k$ or $j = k$ and $m_{jk} = 0$ otherwise (McClintock et al., 2012; Thygesen et al., 2009). While random walks are commonly used to model the movement of individual animals (Ahmed et al., 2023; Fagan & Calabrese, 2014; Turchin, 1998), there are two reasons such a parameterization for $\mathbf{M}$ is inappropriate for this application. Firstly, our interest is in modeling the underlying distribution of a migratory species over time rather than the movement of individual birds. Indeed the crowdsourced nature of the eBird data precludes the possibility of linking any two observations to the same specimen, a contrast to telemetry data where all positions of a tagged individual over time are known. (See Turchin (1991) for early work linking models for the movement of individuals to models for the evolution of a population's distribution over time.) Secondly, and of greater practical importance for our application, consider the significance of timescale when evaluating the possible behavior of a discrete time random walk. If $\mathbf{M}$ is constructed such that $m_{jk} = 0$ for all $j \nsim k$, then it will be impossible for the overall spatial distribution of $\mathbf{z}_i$ to shift more than one step in any direction from its state at time $i-1$. Given that our data has a weekly

temporal resolution, and the fact that some species of bird may cover distances of hundreds of kilometers per day (Hedenstrom & Alerstam, 1998), it will be necessary to define a transition structure that allows for communication between non-adjacent nodes within a single time step. This can also be seen in Figure 4.2, which illustrates that significant shifts in the empirical spatial distribution of the data can occur within a single time step, especially during periods of high migration.

4.2.2 Circuit theory and ecology

The hidden Markov model and parameterization for $\mathbf{M}$ presented in this article were developed in part using ideas from the fields of electrical physics and circuit theory. The introduction of electrical circuit theory to ecological models for animal movement and landscape connectivity dates to work published by Brad McRae and collaborators (McRae, 2006; McRae & Beier, 2007; McRae et al., 2008). Such methods, which use the physical and mathematical laws governing voltage, current, and resistance to represent animal movement and other ecological processes have been demonstrated to perform well in many applications, and are now widespread in ecology and (increasingly) statistics (Dickson et al., 2019; Peterson et al., 2019). For instance, electrical current may be used as a conceptual proxy for the movement of animals through space (Grafius et al., 2017; Koen et al., 2014), and resistance distance—a metric intrinsically related to the notion of random walks on graphs (Chandra et al., 1996)—is useful for characterizing proximity between locations as a function of flow within an ecological network, especially in instances in which ecological barriers such as mountain ranges or rivers may inhibit animal movement (Hanks & Hooten, 2013; Thiele et al., 2018). Christensen and Hoff (2023) defined a graphical covariance model using a metric closely related to resistance distance and demonstrated how it could be used to understand interactions between species distribution and environment for bird populations in North Carolina.

To introduce the model used within this article, we consider the following analogy, illustrated by Figure 4.3. Imagine the counties of the United States as nodes within an

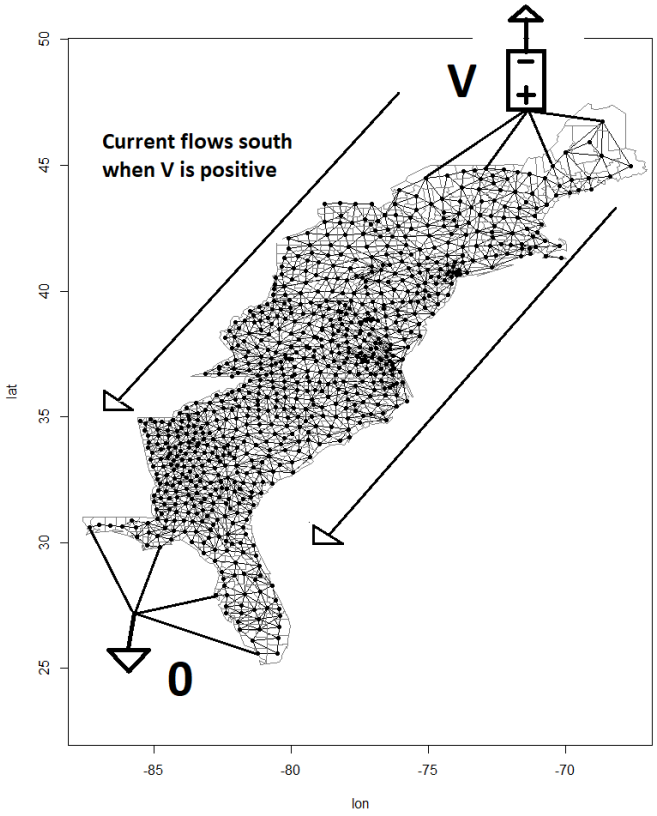


FIGURE 4.3: US counties as a circuit with power source attached to northern counties and grounded at southern counties.

electrical network, with resistors placed between each pair of adjacent counties. If a power source of voltage v and a ground (which by definition has voltage 0) were attached to two locations within the network, electrical current would be induced between all pairs of nodes within the network, the direction and magnitude of which would be determined by the voltage v and within network locations of the power source and ground. If v is positive, current will flow from the power source and through the network in the direction of the ground. If v is negative, current will flow in the opposite direction towards the power source. Furthermore, if v and the resistances between each pair of adjacent nodes are known, it is possible to use Kirchoff's and Ohm's circuit laws (see Hankin (2006)) to compute the effective current, resistance and voltage differential between every pair of nodes in the network. We demonstrate how this is done in the following subsection.

4.2.3 Obtaining effective resistances and currents

We begin with $\mathbf{A}$, the 805×805 unweighted adjacency matrix representing the neighborhood structure of all counties within our spatial domain. Under the previous analogy, $\mathbf{A}$ represents resistances of all edges within the electrical network. We augment our graph by defining two new nodes, one (hereafter the battery node) which is adjacent to 23 counties along the northern border of our spatial domain, and a second (the ground node) adjacent to 23 counties along the interior coast of Florida. Let $\mathbf{A}^*$ be the adjacency matrix for the augmented graph. We can construct $\mathbf{L} = \text{diag}(\mathbf{A}^*\mathbf{1}_p) - \mathbf{A}^*$, the Laplacian matrix of the graph. The resistance distance Ω_{jk} between nodes j and k is calculated as follows:

$$\Omega_{jk} = (\mathbf{e}_j - \mathbf{e}_k)'\mathbf{L}^+(\mathbf{e}_j - \mathbf{e}_k) \tag{4.3}$$

where $\mathbf{L}^+$ is the Moore-Penrose generalized inverse of $\mathbf{L}$ and $\{\mathbf{e}_1, ..., \mathbf{e}_{807}\}$ are the standard basis vectors. This metric characterizes the effective resistance between pairs of nodes in resistor networks and has seen widespread use within ecological applications, as previously noted (Dickson et al., 2019; Klein & Randić, 1993).

Suppose the voltage associated with the battery node is fixed to 1. Let $\mathbf{v}$ be the voltages associated with all nodes within our network. By definition, the voltage at the ground node is 0, and all other nodes will have voltages between 0 and 1. Let $\mathbf{i}$ be the vector of net current into each node. By Kirchoff's first law, this value will be 0 for all nodes except for the battery and ground. Ohm's law may be stated as

$$\mathbf{L}\mathbf{v} = \mathbf{i} \tag{4.4}$$

where $\mathbf{L}$ is the Laplacian matrix as before (Paul, 2001). The only unknowns within this system are the currents at the battery and ground nodes, and the voltages at all other nodes. Hankin (2006) shows that this results in the following partitioned matrix equation:

$$\begin{pmatrix} \mathbf{L}_{11} & \mathbf{L}_{12} \\ \mathbf{L}_{21} & \mathbf{L}_{22} \end{pmatrix} \begin{pmatrix} \mathbf{v}^k \\ \mathbf{v}^u \end{pmatrix} = \begin{pmatrix} \mathbf{i}^u \\ \mathbf{i}^k \end{pmatrix} \tag{4.5}$$

where the superscripts k and u correspond to the known and unknown components of $\mathbf{v}$ and $\mathbf{i}$, while $\mathbf{L}_{11}$ is the 2×2 portion of the Laplacian corresponding to the indices of the battery and ground nodes. The solution to this system is:

$$\begin{aligned} \mathbf{v}^u &= \mathbf{L}_{22}^+ \left(\mathbf{i}^k - \mathbf{L}_{21} \mathbf{v}^k \right) \\ \mathbf{i}^u &= \left(\mathbf{L}_{11} - \mathbf{L}_{12} \mathbf{L}_{22}^+ \mathbf{L}_{21} \right) \mathbf{v}^k + \mathbf{L}_{12} \mathbf{L}_{22}^+ \mathbf{i}^k. \end{aligned} \tag{4.6}$$

Once the full vector $\mathbf{v}$ has been obtained, the currents matrix $\mathbf{C}$ is defined as follows:

$$c_{jk} = (v_k - v_j)/\Omega_{jk} \text{ for } j \neq k, \text{ else } c_{jk} = 0. \tag{4.7}$$

Here c_{jk} represents the effective current from county k towards county j. From Equation 4.7 one can see that $c_{jk} = -c_{kj}$, and thus $\mathbf{C} = -\mathbf{C}^\top$. If $\mathbf{C}$ is computed using a battery node voltage of 1, it can be shown that the currents matrix when the battery node has voltage v^* is equal to $v^*\mathbf{C}$. This in conjunction with the skew-symmetry of $\mathbf{C}$ means that $\mathbf{C}^\top$ is the currents matrix when the battery node has voltage equal to -1.

Our general modelling approach is to define a hidden Markov model with temporally varying transition structure that can adapt to the shifting patterns of migratory flow throughout the year. The transition structure is conceptually based on how currents within the network shown in Figure 4.3 (with power source in the north and ground in the south) would adapt in response to voltage changing over time. For instance, the northward spring migration may correspond to a period of negative voltage, causing currents to flow towards the power source, while the southward fall migration would be produced by positive voltage. Within the following section we formalize our mathematical notation for the model and discuss specific choices made regarding its parameterization.

4.3 Method

4.3.1 Proposed model

We analyze five years worth of completed checklists submitted to the eBird database and reported within $\mathcal{D}$, the spatial domain consisting of 805 counties within the eastern United

States and depicted in Figure 4.1. Let $\mathbf{C}$ be the 805×805 matrix of effective currents between all counties as defined by Equation 4.7, and let $\mathbf{D}$ be the 805×805 matrix of Euclidean distances between county centroids. For a given bird species (e.g. the Baltimore oriole) let y_{ij} be the combined count of specimens observed by all bird watchers during week $i \in \{1,...,260\}$ within county $j \in \{1,...,805\}$, with $\mathbf{y}_i$ the vector of counts across all counties during week i. Let t_{ij} be the combined number of bird watching hours for that week-county pairing. We define the latent variable z_{ij} as the expected number of birds (of the given species) one would observe during a single hour of bird watching. A straightforward modelling choice for $\mathbf{y}_i|\mathbf{z}_i, \mathbf{t}_i$ is to use a Poisson distribution:

$$y_{ij} \sim \text{Poisson}(z_{ij}t_{ij}), \quad z_{ij} > 0. \tag{4.8}$$

In addition to the standard HMM conditional independence assumption that $\pi(\mathbf{y}_i|\mathbf{z}_1,...,\mathbf{z}_{260}) = \pi(\mathbf{y}_i|\mathbf{z}_i)$, we also assume that $\pi(y_{ij}|\mathbf{z}_i) = \pi(y_{ij}|z_{ij})$. Latent vectors $\{\mathbf{z}_i\}_{1:260}$ represent the rate at which birds are observed (or would have been observed for week-county combinations with no data), but can be interpreted as reflecting relative species abundance across our spatio-temporal domain. This interpretation requires us to make the assumption that the underlying rate at which birds are observed is purely a function of abundance and not of factors such as time of observation, remoteness of location, or expertise of the individual observer. While these assumptions are almost certainly violated at the level of individually submitted checklists, on aggregate much of this bias may be mitigated, as suggested by Horns et al. (2018). Variable observation effort is accounted for by including t_{ij} within the expectation of y_{ij}. While sampling efforts are uneven in both space and time (a greater number of checklists are reported during spring and summer months and close to major population centers) we assume that $\mathbf{t}_i \perp \mathbf{z}_i|\mathbf{z}_{i-1}, \mathbf{z}_{i+1}$, meaning that there are no systematic differences in underlying species abundance between high- and low-observation regions.

To characterize how the latent process evolves in time, we specify $\mathbf{z}_i|\mathbf{z}_{i-1}$ as follows:

$$\begin{aligned} \mathbf{z}_i &= \mathbf{M}(\theta_i)\mathbf{z}_{i-1} \odot \epsilon_i \\ \epsilon_{ij} &\overset{iid}{\sim} \text{Gamma}(\alpha, \alpha), \quad \alpha > 0 \end{aligned} \tag{4.9}$$

Here ϵ_i is a vector of multiplicative, Gamma-distributed noise, and $\mathbf{M}_i = \mathbf{M}(\theta_i)$ is a 805×805 transition matrix which is a function of time-indexed parameters θ_i (greater detail is provided later in this section), and α is a parameter controlling the concentration of z_{ij} about its mean. Our model for z_{ij} may equivalently be written as:

$$z_{ij} \sim \text{Gamma}(\alpha, \alpha / [\mathbf{m}'_{ij}\mathbf{z}_{i-1}]), \quad \alpha > 0 \tag{4.10}$$

where $\mathbf{m}_{ij}$ is the j^{th} row of $\mathbf{M}_i$, meaning that $\mathbb{E}[z_{ij}] = \mathbf{m}'_{ij}\mathbf{z}_{i-1}$ and $\text{Var}[z_{ij}] = \mathbf{m}'_{ij}\mathbf{z}_{i-1}/\alpha$.

4.3.2 Transition structure

To ensure that all elements of $\mathbf{M}_i$ and $\{\mathbf{z}_i\}_{1:260}$ are greater than 0, we will use only positive currents when defining transition probabilities. Let $\mathbf{C}^+ = \mathbf{C} \vee \mathbf{0}$, setting all negative elements of $\mathbf{C}$ to zero, be the currents matrix characterizing (southward) network flow when voltage is positive. Likewise, let $\mathbf{C}^- = \mathbf{C}^\top \vee \mathbf{0}$ be the currents matrix characterizing (northward) network flow when voltage is negative. Figure 4.4 depicts the effective currents from Albany County, New York (in grey) to all other counties based on the corresponding column of $\mathbf{C}^+$. One notes that the effective currents to all counties north of Albany County are equal to zero, and that currents to the southernmost counties in the network are largest. This latter feature indicates that raw currents alone can't be used to construct $\mathbf{M}_i$, as every transition would push the majority of the mass in $\mathbf{z}_{i-1}$ to one of the network poles. To account for the this, we incorporate the physical distances between locations (as contained in $\mathbf{D}$) in conjunction with the currents matrices $\mathbf{C}^+$ and $\mathbf{C}^-$. We parameterize $\mathbf{M}_i$ as follows:

$$\mathbf{M}(q_i, \rho_i, \nu_i, \delta_i) = \nu_i \text{ scaleCol}\left(\mathbf{C}^{+/0/-} \oslash \mathbf{D}^{\circ\rho_i}\right) + \delta_i \mathbf{I}, \quad q_i \in \{-1, 0, 1\},\ \rho_i, \nu_i, \delta_i > 0. \tag{4.11}$$

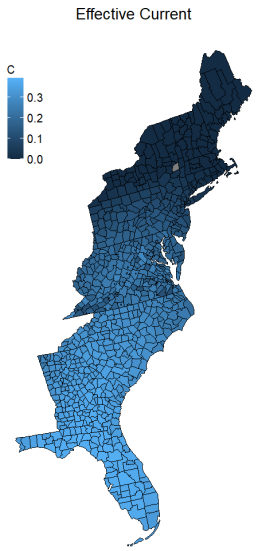


FIGURE 4.4: Effective current from Albany County, NY (in grey) towards all other counties when battery node voltage is 1.

Here $q_i \in \{-1,0,1\}$ is a parameter determining the direction of current in our network. If $q_i = 1$, then $\mathbf{C}^{+/0/-} = \mathbf{C}^{+}$. If $q_i = -1$, then $\mathbf{C}^{+/0/-} = \mathbf{C}^{-}$. If $q_i = 0$, then $\mathbf{C}^{+/0/-} = \mathbf{C}^{0} = \mathbf{1}\mathbf{1}^{\top}$ and the system is in a state of diffusion rather than flow in a particular direction. The function $\oslash$ indicates element-wise division while defining $0/0 = 0$, and $\mathbf{D}^{\circ\rho_i}$ is the matrix of between-county distances raised element wise to the ρ_i power. The purpose of ρ_i is to mitigate the effect seen in Figure 4.4 in which current is strongest towards the poles of the network. Constructing the transition matrix using only $\mathbf{C}$ would be akin to assuming that the majority of migrating birds always wind up in either Florida or Maine after one week, hence division by $\mathbf{D}^{\circ\rho_i}$. When ρ_i is large, shifts from $\mathbf{z}_{i-1}$ to $\mathbf{z}_i$ will cover smaller distances, while small ρ_i can lead to dramatic shifts in latent spatial distribution. Because changing ρ_i can significantly affect the scale of $\mathbf{C}^{+/0/-} \oslash \mathbf{D}^{\circ\rho_i}$ we incorporate the function scaleCol$(\cdot)$, which takes a matrix as input and rescales the columns such that they all sum to one. Lastly, ν_i and δ_i concern the rate of flow vs. self-transmission. If ν_i is high, the behavior dictated by $\mathbf{C}^{+/0/-} \oslash_0 \mathbf{D}^{\circ\rho_i}$ will dominate the transition pattern at time i. If δ_i is high, then $\mathbf{z}_i$ will tend to look more like $\mathbf{z}_{i-1}$. If overall species abundance within $\mathcal{D}$ is constant, $\nu_i + \delta_i$ should be equal to or close to 1. However, if there are a significant number

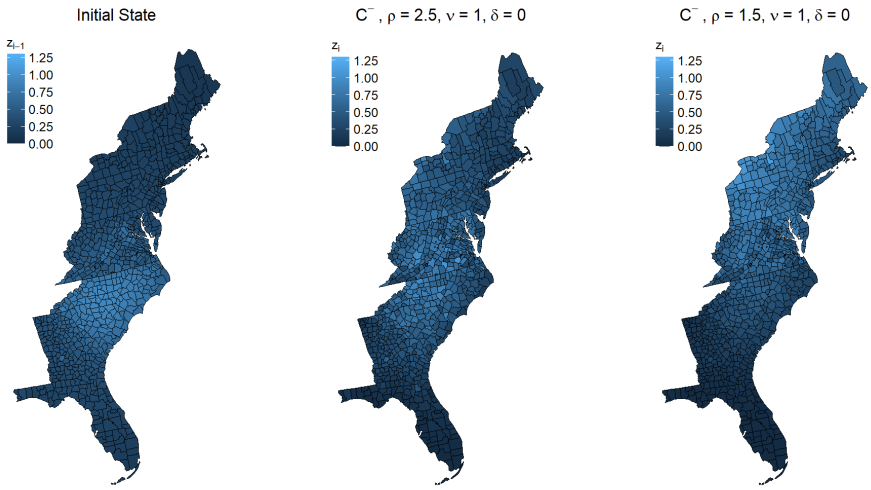


FIGURE 4.5: Left: Initial state of $\mathbf{z}_{i-1}$ for this figure, and Figures 4.6 and 4.7. Center: $\mathbf{z}_i$ with $q_i = -1$ (northward flow) and larger ρ_i. Right: $q_i = -1$, smaller ρ_i leading to greater northward flow.

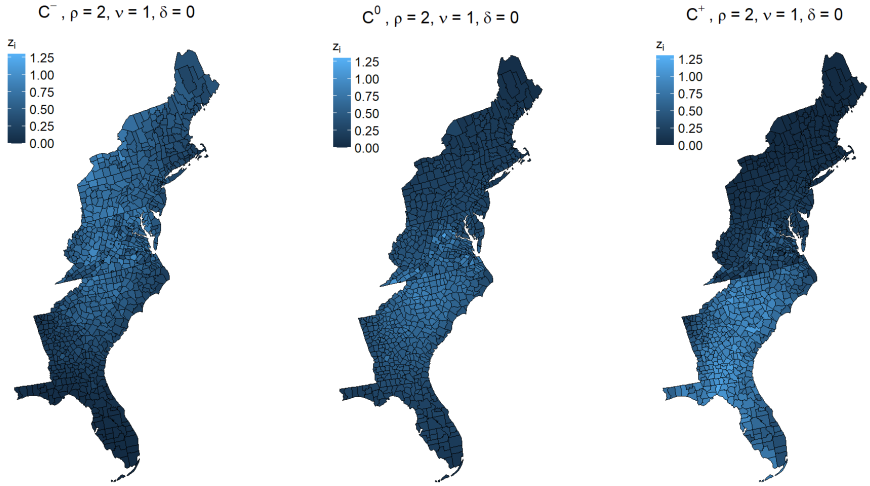


FIGURE 4.6: Left: $\mathbf{z}_i$ with $q_i = -1$ (northward flow). Center: $q_i = 0$ (diffusion). Right: $q_i = 1$ (southward flow).

of birds introduced to or removed from the system (via new births, deaths, or entry into or exit from Canada and Mexico) it is possible that this sum will deviate from that baseline.

To illustrate the behavior of our model, consider Figures 4.5, 4.6, 4.7 and 4.8, which depict the transitions between latent states under different parameter settings. The leftmost subfigure in Figure 4.5 represents the initial state of $\mathbf{z}_{i-1}$ while the other eight subfigures

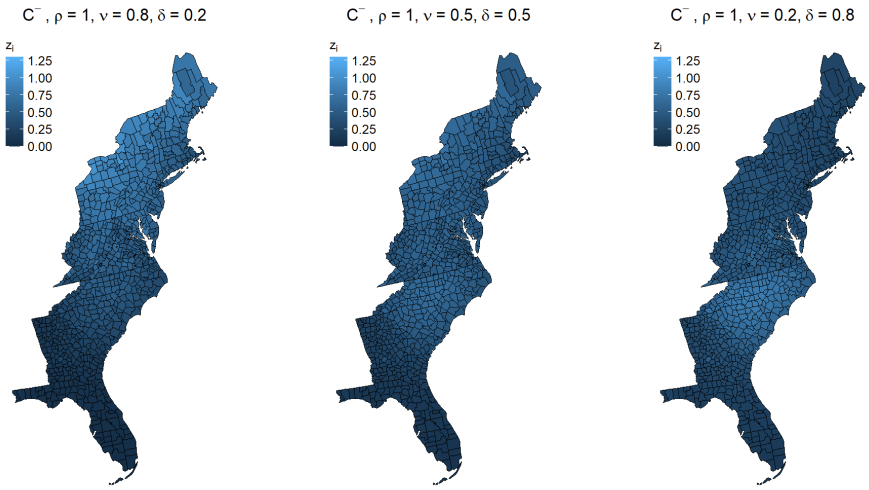


FIGURE 4.7: Left: $\mathbf{z}_i$ with large ν_i relative to δ_i. Center: Equal ν_i and δ_i. Right: Small ν_i relative to δ_i.

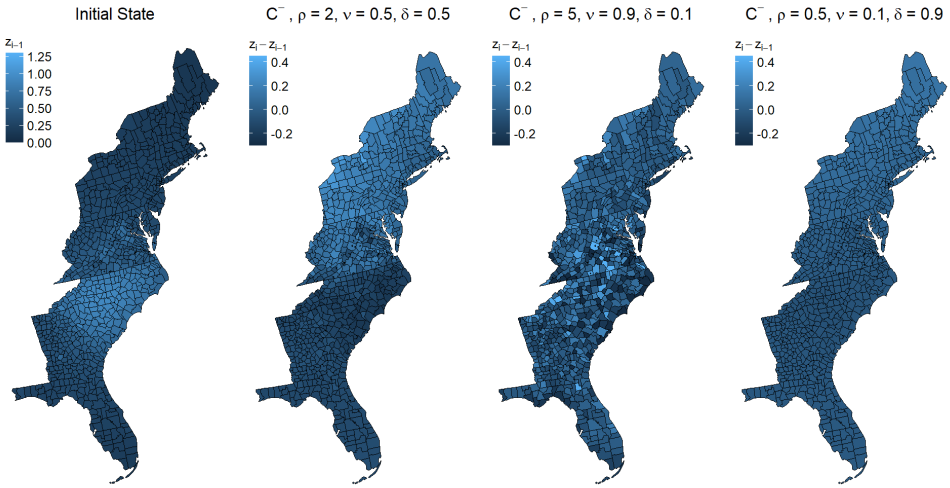


FIGURE 4.8: Change in latent rate $(\mathbf{z}_i - \mathbf{z}_{i-1})$ under different parameter settings.

from 4.5, 4.6 and 4.7 represent the state of $\mathbf{z}_i$ after transition according to the indicated parameters. As seen in Figure 4.5, the initial state is characterized by high concentration in the Carolinas and relatively low concentration in all other regions. Figure 4.5 depicts the impact of ρ_i on the transition from $\mathbf{z}_{i-1}$ to $\mathbf{z}_i$. The rightmost subfigure, with a smaller value of ρ_i depicts greater concentration of mass in the northern portions of the spatial domain, whereas higher ρ_i leads to a northward shift that is more muted. Figure 4.6 depicts the influence of q_i. The left subfigure depicts northward flow, the central subfigure diffusion,

and the right southward flow, with all other parameters constant. Figure 4.7 depicts the impact of ν_i and δ_i. While all three subfigures here depict northward flow under the same value of ρ_i, the rightmost figure differs notably less from the initial state due to higher δ_i. Lastly, consider the illustration in Figure 4.8, depicting the change from $\mathbf{z}_{i-1}$ to $\mathbf{z}_i$ under different parameter settings. Note that while all are characterized by northward flow, the transition structure characterized by moderate ρ_i and balanced ν_i and δ_i (second subfigure) produces a more substantive northward shift in latent rates than the transition matrix produced by high ρ_i and high $\nu_i/(\nu_i + \delta_i)$ (third subfigure), and the matrix produced by low ρ_i and low $\nu_i/(\nu_i + \delta_i)$ (fourth subfigure). This indicates that each transition matrix parameter can take on values that stifle flow, even if all other parameters are valued such that flow would be encouraged.

4.3.3 Priors and model fitting

We perform inference for the parameters of this model by approximating the posterior distribution under the following prior distribution:

$$\begin{aligned}
\alpha &\sim \text{Gamma}(a_\alpha, b_\alpha) \\
q_i &\overset{iid}{\sim} \begin{cases} P(q_i = 1) = p_+ \\ P(q_i = 0) = p_0 \\ P(q_i = -1) = p_- \end{cases} \\
\rho_i &\overset{iid}{\sim} \text{Gamma}(a_\rho, b_\rho) \\
\nu_i &\overset{iid}{\sim} \text{Gamma}(a_\nu, b_\nu) \\
\delta_i &\overset{iid}{\sim} \text{Gamma}(a_\delta, b_\delta).
\end{aligned} \tag{4.12}$$

Prior parameters and initial latent state vector $\mathbf{z}_0$ should be chosen in an application appropriate manner. The posterior distribution $\pi(\{\mathbf{z}_i\}_{1:260}, \{q_i\}_{1:260}, \{\rho_i\}_{1:260}, \{\nu_i\}_{1:260}, \{\delta_i\}_{1:260}, \alpha | \{\mathbf{y}_i\}_{1:260})$ is obtained using the Markov chain Monte Carlo (MCMC) sampler described in Algorithm 1. Greater detail regarding the sampler and proposal distributions used is provided in Appendix C.

Algorithm 1 MCMC Overview

Input: Observations $\{\mathbf{y}_i\}_{1:n}$, $\{\mathbf{t}_i\}_{1:n}$ and pre-computed currents matrix $\mathbf{C}$
Output: T posterior samples for parameters $\{\mathbf{z}_i\}_{1:n}$,$\{q_i\}_{1:n}$,$\{\rho_i\}_{1:n}$,$\{\nu_i\}_{1:n}$,$\{\delta_i\}_{1:n}$, and α
Initialize $\{\mathbf{z}_i^{(0)}\}_{1:n}$,$\{q_i^{(0)}\}_{1:n}$,$\{\rho_i^{(0)}\}_{1:n}$,$\{\nu_i^{(0)}\}_{1:n}$,$\{\delta_i^{(0)}\}_{1:n}$, and $\alpha^{(0)}$
for $t = 1$ to T **do**
  for $i = 1$ to n **do**
    MH update (MALA proposal) of $\mathbf{z}_i^{(t)}$ based on $\pi(\mathbf{z}_i|\mathbf{y}_i,\mathbf{z}_{i-1}^{(t)},\mathbf{z}_{i+1}^{(t-1)},\mathbf{M}_i^{(t-1)},\alpha^{(t-1)})$
    Direct update of $q_i^{(t)}$ based on $\pi(q_i|\mathbf{z}_i^{(t)},\mathbf{z}_{i-1}^{(t)},\rho_i^{(t)},\nu_i^{(t)},\delta_i^{(t)},\alpha^{(t-1)})$
    MH update (aMCMC proposal) of $\theta_i^{(t)} = (\rho_i^{(t)},\nu_i^{(t)},\delta_i^{(t)})'$ based on $\pi(\theta_i|\mathbf{z}_i^{(t)},\mathbf{z}_{i-1}^{(t)},q_i^{(t)},\alpha^{(t-1)})$
  MH update of $\alpha^{(t)}$ based on $\pi(\alpha|\{\mathbf{z}_i\}_{1:n},\{\mathbf{M}_i\}_{1:n})$

4.3.4 Simulation

Before fitting our model to the full data set, we assess performance on a smaller data set by simulating data from our model on the 12×5 lattice network depicted in Figure 4.9. Data $\{\mathbf{y}_i\}_{1:10}$ were produced by generating exposure efforts $t_{ij} \overset{iid}{\sim} \text{Exponential}(0.1)$ and simulating $\{\mathbf{z}_i\}_{1:10}$ by initializing $\mathbf{z}_0 = \mathbf{1}$ and generating transition parameters $q_i \overset{iid}{\sim} \text{Multinomial}(p_+ = 0.25, p_0 = 0.5, p_- = 0.25)$, $\rho_i \overset{iid}{\sim} \text{Gamma}(3, 1.5)$, $\nu_i \overset{iid}{\sim} \text{Gamma}(3, 6)$, $\delta_i \overset{iid}{\sim} \text{Gamma}(3, 6)$, and $\alpha \sim \text{Gamma}(10, 2)$ for $i \in \{1, ..., 10\}$ time periods. In order to assess the model's ability to recover the underlying rate z_{ij} in unobserved regions, 30 percent (approximately the percentage of the eBird data for which observation effort is 0) of total observations were censored at random such that $y_{ij} = t_{ij} = 0$. Samples from the posterior distribution were obtained using MCMC, from which coverage rates and other statistics were calculated. This simulation process was repeated twenty times. Tables 4.1 and 4.2 contain aggregated results from all model fittings.

Table 4.1: Average coverage, credible interval width, bias, and RMSE for z_{ij} when data are observed and censored.

	Coverage (90%)	Bias	RMSE
y_{ij}, t_{ij} observed	0.883	-0.003	0.349
$y_{ij} = t_{ij} = 0$	0.880	-0.002	0.589

We note that the model recovers the underlying rate parameters $\{\mathbf{z}_i\}_{1:10}$ effectively, even for nodes and time periods where no data are available, suggesting that this model

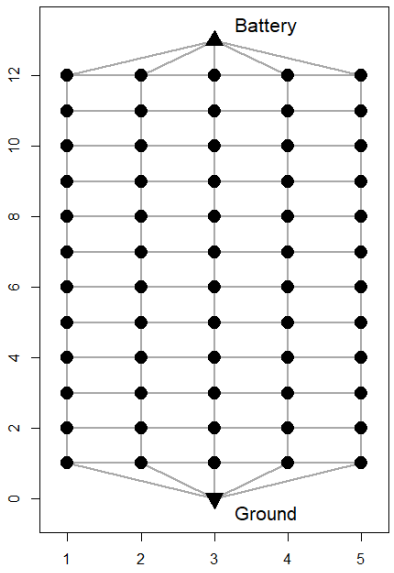


FIGURE 4.9: Network structure used for simulation.

Table 4.2: Average 90% coverage rates for all transition structure parameters.

	q_i	ρ_i	ν_i	δ_i	α
Coverage (90%)	0.895	0.610	0.905	0.855	0.900

could be valuable in estimating species abundance for regions with minimal observation effort. All other model parameters have appropriate coverage rates, with the exception of ρ_i. Investigation of simulation output indicated that increases in ρ_i have diminishing impact on likelihood, making higher true values of ρ_i difficult to correctly recover. In general, this issue does not seem to negatively impact inference on other parameters. We now turn to a discussion of model results when fit to observations from the eBird data set.

4.4 Analysis of eBird data for migratory species

To understand trends in the movement and relative abundance of migratory birds, we fit our model to five years worth of weekly county-level observations for the Baltimore oriole and yellow-rumped warbler. In order to reduce computational burden, the model was fit separately for each year, allowing some degree of parallelization at the cost of breaking

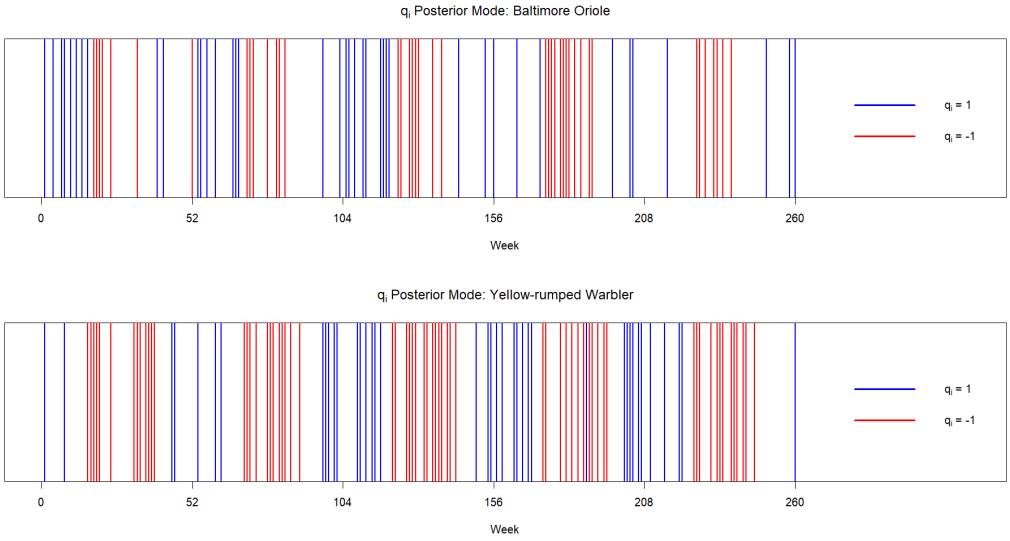


FIGURE 4.10: The posterior mode of $\mathbf{q}$: blue lines indicate weeks with high posterior probability that current is positive, resulting in southward flow, while red lines indicate weeks with high posterior probability that current is negative resulting in northward flow.

temporal dependence between fits. Using the priors given in Equation 4.12, the following hyperparameters were chosen: $a_\rho = a_\nu = a_\delta = 5$, $b_\rho = 2$, $b_\nu = b_\delta = 10$, and $(p_+, p_0, p_-) = (0.2, 0.6, 0.2)$. This results in ρ_i having a prior expectation of 2.5, a threshold we found to be consistent with modest but non-trivial levels of flow and diffusion when simulating current matrices, and balanced ν_i and δ_i. We fixed $\alpha = 2$ based on an ad hoc empirical Bayes procedure. This decision was made in order to keep α consistent between the separate model fits for each year, to ensure that the the multiplicative error term has a non-zero mode, and because we found in simulation that fixing α significantly improves mixing times while having limited impact on inference for other parameters. Computation was performed using the Duke University Compute Cluster. An average of approximately 15000 posterior samples were obtained, requiring two weeks of time and resulting in an average effective sample size of approximately 500 after burn-in.

Within this section, our primary focus to assess how our model informs our understanding of migratory patterns over time for each species. Recall from discussion of Figures 4.5-4.8 that q_i, ρ_i, ν_i and δ_i all work in concert to characterize the transition behavior at

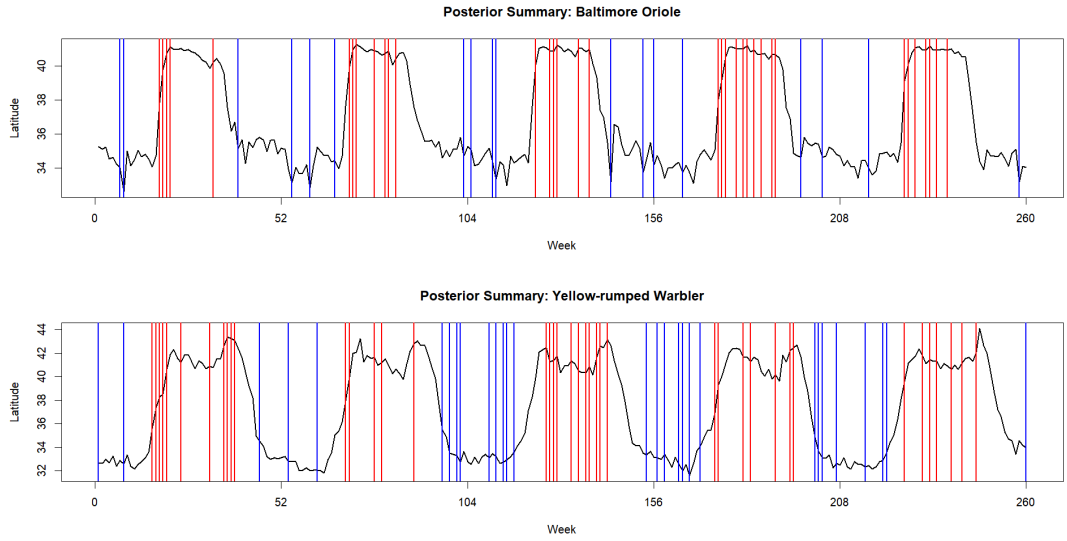


FIGURE 4.11: Black lines depict average latitude over time based on posterior distribution of $\{\mathbf{z}_i\}_{1:260}$. Vertical red lines indicate time periods of substantive northward flow and blue lines indicated periods of substantive southward flow, based on posterior of $\{q_i\}_{1:260}$, $\{\rho_i\}_{1:260}$, $\{\nu_i\}_{1:260}$, $\{\delta_i\}_{1:260}$

time i. While one could simply evaluate the posterior mode of $\{q_i\}_{1:260}$ over time, depicted in Figure 4.10, and infer that southward migration is occurring when the posterior distribution of q_i is concentrated at 1, and northward at -1, it may be useful to take other parameter estimates at time i into account. For example, we may wish to characterize a period of "substantive" flow as one for which the posterior probability of $q_i \in 1, -1$ exceeds 0.9, the posterior mean of $\rho_i > 3$ and the posterior mean of $\delta_i/(\nu_i + \delta_i) > 0.7$, thresholds corresponding approximately to the 75th percentile across i, thereby omitting any time periods with parameter values that significantly inhibit flow. Figure 4.11 contains plots of the average latitude over time (in black, and based on the posterior distribution of $\{\mathbf{z}_i\}_{1:260}$) for the population of the Baltimore-oriole and yellow-rumped warbler, with vertical red lines indicating weeks with substantive northward $(q_i = -1)$ migration according to the above heuristic, and blue lines for southward $(q_i = 1)$ migration.

As expected, our model identifies patterns of northward migration during spring and summer, with patterns of southward migration during fall and winter for each species. The consistency of certain patterns observable in Figure 4.11 is worth noting. In each of the

five years analyzed, the first substantive instance of northward migration for the Baltimore oriole occurs during the 17th or 18th week of the year (corresponding to the period between April 23 to May 6 in non-leap years), while first such period for the yellow-rumped warbler occurs each year (except in 2015) during the 15th to 17th weeks. Notably, the week of initial northward migration is correlated ($R = 0.72$) between these two species, suggesting potential common environmental factors driving the onset of spring migration for both species. The portion of the year during which our model is identifying northward shifts in distribution tends to last longer for the yellow-rumped warbler than for the Baltimore oriole. All of these findings are consistent with preexisting research that suggests that the yellow-rumped warbler has a more drawn out spring migratory season relative to the Baltimore oriole (Hunt & Flaspohler, 2020; Rising & Flood, 2020).

In contrast to the patterns observed with regards to northward migration, our model identifies fewer and more temporally varied periods of southward migration for each species. Ecologists have noted differences in patterns between spring and fall migration in terms of speed and duration of travel (Nilsson et al., 2013; Tryjanowski & Yosef, 2002) with spring migration occurring more rapidly over a shorter period, and seasonal differences in the physiology of migrating birds (Sharma et al., 2018). It has been theorized that sexual competition drives the intensity and speed of the spring migration (Kokko, 1999), whereas fall migration is driven more by local resource availability and climate conditions (Nilsson et al., 2013). Taken together, the patterns observed during spring migration tend to be more dramatic and consistent across populations, whereas fall migration is more protracted and exhibits more localized behavior, which may account for the weaker and less consistent signals of southward migration identified by our model.

4.5 Discussion

Within this article we utilized crowdsourced bird watching observations from the eBird database in order to model the latent abundance of migratory bird species over time and space. We constructed a hidden Markov model with novel transition matrix parameteriza-

tion utilizing principles from circuit theory in order to reflect migratory patterns present within the data. After illustrating model behavior and demonstrating its ability to recover the true underlying distribution in locations with missing data we fit our model to observations of the Baltimore oriole and yellow-rumped warbler, identifying migratory trends which are consistent with contemporary scientific understanding.

We conclude this article with a brief discussion of potential applications, model extensions, and follow up questions to the content of this article. As noted, the identification of migratory flyways (the routes birds use while travelling) is an important task in the study of avian migration (Buhnerkempe et al., 1996). The model as implemented within this article did not explicitly consider this question, essentially treating the entire spatial domain as a single flyway. As we consider potentially useful future extensions to this work, a major point of interest is to explore the use of variable edge weights in the construction of our model's adjacency structure. Under the analogy of the spatial domain as electrical network, the unweighted county adjacency matrix is akin to unit resistors being used exclusively in network construction. If edge weights were instead parameterized as functions of environmental covariates or spatial random effects (Christensen & Eidsvik, 2024; Hanks & Hooten, 2013), current flow could exhibit more flexible and scientifically noteworthy behaviors, especially within the context of flyway detection. As an illustration of this, see Figure 4.12, which depicts a 24×12 lattice network with battery node attached to the top of the graph and ground at the bottom. Cells marked with "X"s indicate high resistance regions; one could imagine that they correspond to mountains or other features inhibiting animal movement. Effective current is depicted relative to the battery node which has voltage equal to 1. As can be seen current flows around high-resistance regions and through low resistance regions. Defining a model such that edge resistances are estimated from the data would enable a detailed understanding of the paths birds are utilizing during migration; similar methods have already been used to characterize patterns of bird movement, albeit at more local scale and with resistances pre-specified according to expert belief rather than estimated (Grafius et al., 2017).

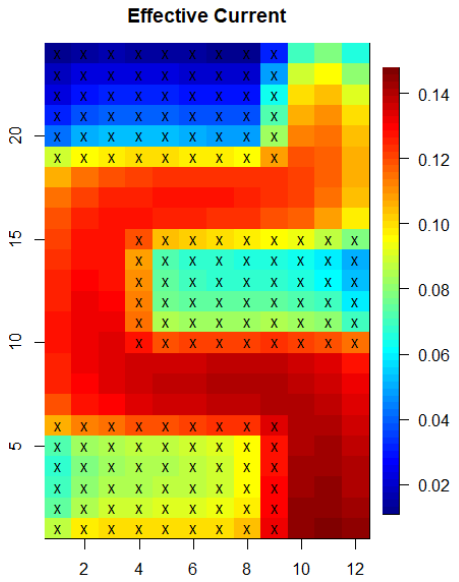


FIGURE 4.12: High and low resistance regions shape flow through a network, enabling us to learn migratory pathways.

Understanding the impact of climate change on bird populations and their migratory behaviors is of considerable ecological importance (Marra et al., 2005). Scientists have already observed shifting patterns in the timing and duration of migration season over the past decades (Covino et al., 2020). As noted in the previous section, our model identified that the onset of spring migration was correlated between the studied species. Incorporating historic climate information into our model, or comparing model output to such data could allow us to make more significant inferences about these relationships. Because climate conditions can vary dramatically across the migratory range of many species and local conditions can impact migratory timings (Tottrup et al., 2010), it could be valuable to adapt our model's transition structure such that flow patterns are defined locally rather than globally.

5. Conclusions

This dissertation explores the question of how we can more effectively model areal and graphical data. In Chapter 2 we described and introduced the graph deformation (GDEF) modeling framework, which defines a class of spatial covariance models for graph and areal data. The GDEF model improves upon commonly used alternatives through its capacity to flexibly approximate a wide range of nonstationary and anisotropic covariance patterns and by leveraging the general flexibility of the Matérn covariance class, and provides appealing interpretation in terms of "intrinsic distances" between spatial regions. In Chapter 3 we extended the work contained in Chapter 2 by introducing a basis function representation for unknown edge weights matrices using the eigenvectors of the line graph Laplacian (LGL). We went on to show how this method improves the CAR model, while significantly improving the practical usability of the GDEF framework, and also provided further detail and illustrations of the properties of the GDEF model. In Chapter 4 we utilized crowdsourced observations from the eBird database to model migratory trends for the Baltimore oriole and yellow-rumped warbler. We presented a hidden Markov model that utilized circuit theory and an electricity based interpretation of migratory flow to characterize transition patterns over time, and demonstrated how this model could be used to understand changes in the species distribution of migratory birds over time.

Within this dissertation we introduce several novel ideas to the space of statistical models for areal data, including the integration of the quasi-Euclidean graph metric within the Matérn covariance family and the introduction of the LGL eigenvector basis for edge weights. Important ideas highlighted within the dissertation that are presently underutilized include an emphasis on the use of unknown edge weight parameters to flexibly

characterize patterns of between node dependence and the use of electrical physics and circuit theory within dynamic space-time models.

The GDEF modeling framework discussed in Chapter 2 and Chapter 3 represents in our view a meaningful step forward in the space of models for areal data. A major objective of our future work in this area will be to develop an R package that allows others to easily implement GDEF models; to this end, an important focus will be the continued improvement of the method's computational efficiency. While the LGL eigenvector innovation of Chapter 3 significantly improved the GDEF's computational speed and practical usability, the code presently being used to implement this method has considerable space for further optimization. Potential improvement through integration of the GDEF model within the INLA computing framework (Rue et al., 2009) is a current research effort.

Appendix A. Supplemental Material for Chapter 2

1.1 Proof of identifiability

Proposition 1 (Identifiability). *Using the parameterization for* Σ *provided in Equation A.1,* $(\sigma_1^2, \boldsymbol{W}_1) \neq (\sigma_2^2, \boldsymbol{W}_2) \Rightarrow \Sigma_1 \neq \Sigma_2$ *for all* $\sigma_1^2, \sigma_2^2 > 0$ *and all* $\boldsymbol{W}_1, \boldsymbol{W}_2 \in \mathcal{W}_G$.

The equation relating Σ to model parameters $(\sigma^2, \mathbf{W})$ is provided again for reference:

$$\begin{aligned} \Sigma_{jk} &= \sigma^2 \rho_\nu(d_{ij}) \\ d_{jk} &= \sqrt{(\mathbf{e}_j - \mathbf{e}_k)^\top \{\mathbf{L}^+\}^2 (\mathbf{e}_j - \mathbf{e}_k)} \\ \mathbf{L} &= \mathrm{diag}(\mathbf{W}\mathbf{1}_p) - \mathbf{W} \end{aligned} \tag{A.1}$$

Because $\rho_\nu(\cdot)$ is a correlation function, and the distance from any node to itself is zero, the diagonal elements of Σ are equal to σ^2. Thus $\sigma_1^2 \neq \sigma_2^2 \Rightarrow \Sigma_1 \neq \Sigma_2$. For two distance matrices $\mathbf{D}_1 \neq \mathbf{D}_2$ there exists some pair of nodes (j,k) such that $d_{1jk} \neq d_{2jk}$. Because $\rho_\nu(\cdot)$ is a strictly decreasing function, $d_{1jk} \neq d_{2jk} \Rightarrow \Sigma_{1jk} \neq \Sigma_{2jk}$.

We now need only to prove that $\mathbf{W}_1 \neq \mathbf{W}_2 \Rightarrow \mathbf{D}_1 \neq \mathbf{D}_2$. The transformation from edge weights matrix to to distance matrix as defined by the quasi-Euclidean metric can be rewritten as follows:

$$\begin{aligned} \mathbf{D} &= \left(\mathbf{1}_p \mathbf{d}_{\{\mathbf{L}^+\}^2}^\top + \mathbf{d}_{\{\mathbf{L}^+\}^2} \mathbf{1}_p^\top - 2\{\mathbf{L}^+\}^2 \right)^{\circ \frac{1}{2}} \\ \mathbf{L} &= \mathrm{diag}(\mathbf{W}\mathbf{1}_p) - \mathbf{W} \end{aligned} \tag{A.2}$$

Here $\mathbf{d}_{\{\mathbf{L}^+\}^2}$ is defined to be the vector with elements equal to the diagonal entries of $\{\mathbf{L}^+\}^2$ and $(\cdot)^{\circ \frac{1}{2}}$ denotes the Hadamard (element-wise) square root of a matrix. The Hadamard root is an invertible transformation ($\mathbf{B}_1 \neq \mathbf{B}_2 \Leftrightarrow \mathbf{B}_1^{\circ \frac{1}{2}} \neq \mathbf{B}_2^{\circ \frac{1}{2}}$ for all matrices $\mathbf{B}_1, \mathbf{B}_2$) but the transformation $\mathbf{B} = \Delta(\mathbf{A})$ for any $p \times p$ matrix $\mathbf{A}$, given below, is not invertible.

$$\mathbf{B} = \Delta(\mathbf{A}) = \mathbf{1}\mathbf{d}_{\mathbf{A}}^\top + \mathbf{d}_{\mathbf{A}}\mathbf{1}^\top - 2\mathbf{A} \tag{A.3}$$

This transformation is fairly well known in other statistical applications involving distances, and appears within the formulation of the quasi-Euclidean metric. A linear transformation $f(\cdot)$ is injective iff its null space $\mathcal{N}(f) = \{\mathbf{0}\}$. The null space of Δ is $\mathcal{N}(\Delta) = \{\mathbf{A} : \mathbf{A} = \mathbf{1c}^\top + \mathbf{c1}^\top, \mathbf{c} \in \mathbb{R}^p\}$ as demonstrated below:

$$
\begin{aligned}
&\text{Suppose } \Delta(\mathbf{A}) = \mathbf{0}_{p\times p}: \\
&\Rightarrow \mathbf{A} = \mathbf{1d}_\mathbf{A}'/2 + \mathbf{d}_\mathbf{A}\mathbf{1}'/2 \\
&\forall \mathbf{c} \in \mathbb{R}^p, \text{ let } \mathbf{A} = \mathbf{1c}^\top + \mathbf{c1}^\top \\
&\Rightarrow \mathbf{d}_\mathbf{A} = 2\mathbf{c} \\
&\Rightarrow \mathbf{A} = \mathbf{1d}_\mathbf{A}'/2 + \mathbf{d}_\mathbf{A}\mathbf{1}'/2 \\
&\Rightarrow \mathcal{N}(\Delta) = \{\mathbf{A} : \mathbf{A} = \mathbf{1c}^\top + \mathbf{c1}^\top, \mathbf{c} \in \mathbb{R}^p\}
\end{aligned}
\tag{A.4}
$$

Let $\mathcal{A}_p$ be the space of $p \times p$ symmetric matrices such that $\forall \mathbf{A} \in \mathcal{A}_p$, $\mathbf{A1}_p = \mathbf{0}_p$. We note that for any graph with p nodes, both its Laplacian matrix $\mathbf{L}$ and $\{\mathbf{L}^+\}^2$ are elements of $\mathcal{A}_p$. It can be seen that $\mathcal{N}(\Delta) \cap \mathcal{A}_p = \{\mathbf{0}_{p\times p}\}$, which implies that $\forall \mathbf{A} \in \mathcal{A}_p$, $\Delta(\mathbf{A}) = \mathbf{0}_{p\times p} \Rightarrow \mathbf{A} = \mathbf{0}_{p\times p}$. We now show that this condition implies the injectivity of Δ over $\mathcal{A}_p$.

$$
\begin{aligned}
&\text{Suppose } (\forall \mathbf{A} \in \mathcal{A}_p, \Delta(\mathbf{A}) = \mathbf{0} \Rightarrow \mathbf{A} = \mathbf{0}): \\
&\text{If } \Delta(\mathbf{A}_1) = \Delta(\mathbf{A}_2) \text{ and } \mathbf{A}_1, \mathbf{A}_2 \in \mathcal{A} \\
&\Rightarrow \Delta(\mathbf{A}_1) - \Delta(\mathbf{A}_2) = \mathbf{0} \\
&\Rightarrow \Delta(\mathbf{A}_1 - \mathbf{A}_2) = \mathbf{0} \quad \text{(b.c. } \Delta \text{ is linear)} \\
&\Rightarrow \mathbf{A}_1 - \mathbf{A}_2 = \mathbf{0} \quad \text{(from supposition, note that } (\mathbf{A}_1 - \mathbf{A}_2) \in \mathcal{A}_p) \\
&\Rightarrow \mathbf{A}_1 = \mathbf{A}_2 \\
&\Rightarrow \Delta \text{ is injective over } \mathcal{A}_p
\end{aligned}
\tag{A.5}
$$

Because $\{\mathbf{L}^+\}^2 \in \mathcal{A}_p$, $\{\mathbf{L}^+\}_1^2 \neq \{\mathbf{L}^+\}_2^2 \Rightarrow \mathbf{D}_1 \neq \mathbf{D}_2$. The uniqueness of the Moore-Penrose inverse in conjunction with the fact that Laplacian matrices are always positive semi-definite means that the transformation from $\mathbf{L}$ to $\{\mathbf{L}^+\}^2$ is injective. From the definition $\mathbf{L} =$

$\text{diag}(\mathbf{W1}) - \mathbf{W}$, it is clear that $\mathbf{W}_1 \neq \mathbf{W}_2 \Rightarrow \mathbf{L}_1 \neq \mathbf{L}_2$, and thus that $(\sigma_1^2, \mathbf{W}_1) \neq (\sigma_2^2, \mathbf{W}_2) \Rightarrow \Sigma_1 \neq \Sigma_2$ for all $\sigma_1^2, \sigma_2^2 > 0$ and all $\mathbf{W}_1, \mathbf{W}_2 \in \mathcal{W}_G$.

1.2 MCMC sampler

In order to sample from the edge weight parameters of our model, we implemented the Metropolis-adjusted Langevin algorithm (MALA), which requires the analytic computation of the partial derivatives of the log posterior density with respect to all model parameters.

Below we provide details on how to obtain the partial derivative of the log-likelihood ℓ (assuming $\mathbf{Y} \sim N(\mathbf{0}, \Sigma \otimes \mathbf{I}_n)$) with respect to a single edge weight (w_{jk}). Special thanks to Jo Eidsvik, who assisted us with computation and derivative calculations.

One may obtain the derivative $\frac{\partial \ell}{\partial w_{jk}}$ through repeated application of the chain rule as follows:

$$
\begin{aligned}
\frac{\partial \ell}{\partial w_{jk}} &= -\frac{n}{2}\text{trace}\left(\Sigma^{-1}\frac{\partial \Sigma}{\partial w_{jk}}\right) + \frac{1}{2}\sum_{i=1}^{n}\mathbf{y}_i'\Sigma^{-1}\frac{\partial \Sigma}{\partial w_{jk}}\Sigma^{-1}\mathbf{y}_i \\
\frac{\partial \Sigma}{\partial w_{jk}} &= -\sigma^2\frac{\partial \mathbf{D}}{\partial w_{jk}} \odot \mathbf{D} \odot \exp(-\mathbf{D}) = -\frac{\sigma^2}{2}\frac{\partial \mathbf{D}_2}{\partial w_{jk}} \odot \exp(-\mathbf{D}), \\
\frac{\partial \mathbf{D}_2}{\partial w_{jk}} &= \text{diag}\left(\frac{\partial \{\mathbf{L}^+\}^2}{\partial w_{jk}}\right)\mathbf{1}_n' + \mathbf{1}_n\text{diag}\left(\frac{\partial \{\mathbf{L}^+\}^2}{\partial w_{jk}}\right)' - 2\frac{\partial \{\mathbf{L}^+\}^2}{\partial w_{jk}}, \qquad \text{(A.6)} \\
\frac{\partial \{\mathbf{L}^+\}^2}{\partial w_{jk}} &= \mathbf{L}^+\frac{\partial \mathbf{L}^+}{\partial w_{jk}} + \frac{\partial \mathbf{L}^+}{\partial w_{jk}}\mathbf{L}^+, \\
\frac{d\mathbf{L}^+}{\partial w_{jk}} &= -\mathbf{L}^+\frac{\partial \mathbf{L}}{\partial w_{jk}}\mathbf{L}^+, \\
\frac{\partial \mathbf{L}}{\partial w_{jk}} &= \text{diag}\left(\frac{d\mathbf{W}}{\partial w_{jk}}\mathbf{1}_n\right) - \frac{\partial \mathbf{W}}{\partial w_{jk}}.
\end{aligned}
$$

Here $\mathbf{D}_2 = \mathbf{D} \odot \mathbf{D}$ and $\frac{\partial \mathbf{W}}{\partial w_{jk}}$ is a $p \times p$ matrix with elements jk and kj equal to 1 and all others equal to 0. In practice it may be helpful to sample $\log(w_{jk})$ instead to avoid boundary issues in which case $\frac{\partial \mathbf{W}}{\partial \log(w_{jk})}$ is a matrix with elements jk and kj equal to w_{jk}.

Assuming a Gamma(a,b) prior for all edge weights, the derivative of the log-posterior with respect to w_{jk} is

$$\frac{\partial \log \pi(\mathbf{w}|\mathbf{Y})}{\partial w_{jk}} = \frac{\partial \ell}{\partial (w_{jk})} + \frac{a-1}{w_{jk}} - b. \tag{A.7}$$

Let $\nabla_{\mathbf{w}} \log \pi(\mathbf{w}|\mathbf{Y})$ be the vector of length $e = |E|$ containing all partial derivatives of the log likelihood with respect to edge weights. The posterior for $\mathbf{w}$ and σ^2 can be updated as detailed below in Algorithm 1:

Algorithm 2 MCMC Sampler for Covariance Model

Input: $\mathbf{Y}$ an $n \times p$ data matrix, G a graph with p nodes, and s a tuning parameter, which may be adaptively updated during burn-in to ensure an appropriate acceptance rate.
Output: T posterior samples for model parameters σ^2 and $\mathbf{w}$
Initialize $\sigma^{2(0)}$ and $\mathbf{w}^{(0)}$ and set $\Sigma^{(0)} = c_\nu(\sigma^{2(0)}, \mathbf{w}^{(0)})$
for $t = 1$ to T **do**
  Update $\sigma^{2(t)}$ by taking draw from full conditional $\pi(\sigma^{2(t)}|\mathbf{Y}, \Sigma^{(t-1)})$
  Set $\mathbf{w}^{cur} = \mathbf{w}^{(t-1)}$
  Set $\Sigma^{cur} = c_\nu(\sigma^{2(t)}, \mathbf{w}^{(t-1)})$
  Propose $\mathbf{w}^* \sim g(\mathbf{w}^*|\mathbf{w}^{cur}) = N_e(\mathbf{w}^{(t-1)} + \frac{s^2}{2}\nabla_{\mathbf{w}} \log \pi(\mathbf{w}^{cur}|\mathbf{Y}, \sigma^{2(t)}), s^2\mathbf{I}_e)$
  Set $\Sigma^* = c_\nu(\sigma^{2(t)}, \mathbf{w}^*)$
  Compute $\alpha = \dfrac{\pi(\mathbf{w}^{cur}|\mathbf{Y}) g(\mathbf{w}^{cur}|\mathbf{w}^*)}{\pi(\mathbf{w}^{cur}|\mathbf{Y}) g(\mathbf{w}^*|\mathbf{w}^{cur})}$
  Generate u from $U(0,1)$
  if $u < \min(1, \alpha)$ **then**
    $\mathbf{w}^{cur} \leftarrow \mathbf{w}^*$
    $\Sigma^{cur} \leftarrow \Sigma^*$
  $\mathbf{w}^{(t)} \leftarrow \mathbf{w}^{cur}$
  $\Sigma^{(t)} \leftarrow \Sigma^{cur}$

Appendix B. Supplemental Material for Chapter 3

2.1 Derivatives of model likelihood

In Chapter 3 we defined $\nabla_\ell(\theta)$ as the derivative of the log likelihood with respect to θ. Equation A.6 of Appendix A provides the first derivative of the GDEF log likelihood with respect to the individual edge weights. The same equation may be used to find $\nabla_\ell(\theta)$, with the only modification being that in place of $\frac{\partial \mathbf{W}}{\partial w_{jk}}$ we utilize $\frac{\partial \mathbf{W}}{\partial \eta_i}$ which is equal to $\mathbf{W}$ with each of its nonzero elements multiplied by the corresponding element of basis vector $\mathbf{v}_i$.

In order to implement Newton's method, one needs $\mathbf{H}_\ell(\theta)$, the second derivative of the log likelihood with respect to θ. The second derivative of the GDEF covariance with respect to η is given as:

$$
\begin{aligned}
\frac{d^2\Sigma}{d\eta_i d\eta_j} &= -\sigma^2 \exp(-\mathbf{D}) \odot \left[\frac{d^2\mathbf{D}}{d\eta_i d\eta_j} \odot \mathbf{D} + \frac{d\mathbf{D}}{d\eta_i} \odot \frac{d\mathbf{D}}{d\eta_j} - \frac{d\mathbf{D}}{d\eta_i} \odot \frac{d\mathbf{D}}{d\eta_j} \odot \mathbf{D} \right], \\
&= -\frac{\sigma^2}{2} \exp(-\mathbf{D}) \odot \left[\frac{d^2\mathbf{D}_2}{d\eta_i d\eta_j} - \frac{d\mathbf{D}_2}{d\eta_i} \odot \frac{d\mathbf{D}_2}{d\eta_j} \oslash 2\mathbf{D} \right] \text{ (Let } 0/0 = 0), \\
\frac{d\mathbf{D}_2}{d\eta_i} &= 2\frac{d\mathbf{D}}{d\eta_i} \odot \mathbf{D}, \quad i = 1, \ldots, k \\
\frac{d^2\mathbf{D}_2}{dw_i d\eta_j} &= \operatorname{diag}\left(\frac{d^2\mathbf{P}^2}{d\eta_i d\eta_j} \right) \mathbf{1}_n' + \mathbf{1}_n \operatorname{diag}\left(\frac{d^2\mathbf{P}^2}{d\eta_i d\eta_j} \right)' - 2\frac{d^2\mathbf{P}^2}{d\eta_i d\eta_j}, \qquad \text{(B.1)} \\
\frac{d^2\mathbf{P}^2}{d\eta_i d\eta_j} &= \mathbf{P}\frac{d^2\mathbf{P}}{d\eta_i d\eta_j} + \frac{d\mathbf{P}}{d\eta_j}\frac{d\mathbf{P}}{d\eta_i} + \frac{d^2\mathbf{P}}{d\eta_i d\eta_j}\mathbf{P} + \frac{d\mathbf{P}}{d\eta_i}\frac{d\mathbf{P}}{d\eta_j}, \\
\frac{d^2\mathbf{P}}{d\eta_i d\eta_j} &= [\mathbf{L}^{-1}\frac{d\mathbf{L}}{d\eta_j}\mathbf{L}^{-1}]\frac{d\mathbf{L}}{d\eta_i}\mathbf{L}^{-1} - \mathbf{L}^{-1}\frac{d^2\mathbf{L}}{d\eta_i d\eta_j}\mathbf{L}^{-1} + \mathbf{L}^{-1}\frac{d\mathbf{L}}{d\eta_i}[\mathbf{L}^{-1}\frac{d\mathbf{L}}{d\eta_j}\mathbf{L}^{-1}], \\
\frac{d^2\mathbf{L}}{d\eta_i d\eta_j} &= \operatorname{diag}\left(\frac{d^2\mathbf{W}}{d\eta_i d\eta_j}\mathbf{1}_n \right) - \frac{d^2\mathbf{W}}{d\eta_i d\eta_j}.
\end{aligned}
$$

Here, the second derivatives of matrix $\mathbf{W}$ with respect to η have entries equal to $\mathbf{w} \odot \mathbf{v}_i \odot \mathbf{v}_j$.

The second derivative with respect to the likelihood is given by

$$\begin{aligned}\frac{d^2l}{d\eta_i d\eta_j} &= -\frac{n}{2}\text{trace}\left(\Sigma^{-1}\frac{d^2\Sigma}{d\eta_i d\eta_j} - \Sigma^{-1}\frac{d\Sigma}{d\eta_i}\Sigma^{-1}\frac{d\Sigma}{d\eta_j}\right) \\ &- \sum_{r=1}^{n}\mathbf{y}_r'\Sigma^{-1}\frac{d\Sigma}{d\eta_i}\Sigma^{-1}\frac{d\Sigma}{d\eta_j}\Sigma^{-1}\mathbf{y}_r + \sum_{r=1}^{R}\frac{1}{2}\mathbf{y}_r'\Sigma^{-1}\frac{d^2\Sigma}{d\eta_i d\eta_j}\Sigma^{-1}\mathbf{y}_r. \end{aligned} \tag{B.2}$$

We have that $E(\mathbf{y}_r'\mathbf{A}\mathbf{y}_r) = \text{trace}(\mathbf{A}\Sigma)$. The expected value of this second derivative of the likelihood, with respect to the data $\mathbf{Y}$ is

$$E\left[\frac{d^2l}{d\eta_i d\eta_j}\right] = -\frac{n}{2}\text{trace}\left(\Sigma^{-1}\frac{d\Sigma}{d\eta_i}\Sigma^{-1}\frac{d\Sigma}{d\eta_j}\right). \tag{B.3}$$

The expected second derivative of the log likelihood can be computed considerably faster than $\mathbf{H}_\ell(\theta)$, which leads us to prefer it during the iterative portion of implementing Newton's method.

2.2 Bayesian model fitting

In Chapter 3 we used maximum likelihood estimation, obtaining uncertainty estimates based on the MLEs asymptotic distribution. We could also utilize a Bayesian approach to model fitting and suggest the following prior distribution for θ:

$$\begin{aligned}\eta &\sim N_k(\mathbf{0}, \phi\mathbf{I}_k) \\ \phi &\sim \text{Inverse-gamma}(a_\phi, b_\phi) \\ \sigma^2 &\sim \text{Inverse-gamma}(a_{\sigma^2}, b_{\sigma^2}) \\ \kappa &\sim \text{Beta}(a_\kappa, b_\kappa) \quad \text{(If using the CAR covariance).}\end{aligned} \tag{B.4}$$

Prior distribution parameters should be chosen with consideration for the application. Inverse-gamma priors on ϕ and σ^2 allow for posterior conjugacy. Particular attention should be given to the hyper-prior on ϕ. Because the length of an eigenvector is one, the relative scale of individual basis function elements will decrease as graph size increases; as such, the prior expectation of ϕ should generally be higher for larger graphs. We suggested a

Beta distribution prior for κ. Although the CAR model produces a valid covariance matrix for $|\kappa| < 1$, non-contrived settings where $\kappa < 0$ are unlikely to be encountered; any prior distribution over the interval (-1,1) may be used if desired.

Samples from the posterior distribution of θ may be obtained using any Markov chain Monte Carlo (MCMC) algorithm. We have found success using the standard form of the Metropolis adjusted Langevin algorithm (MALA), which utilizes the derivative of the log posterior $\frac{\partial \log \pi(\theta|\mathbf{Y})}{\partial \theta} = \nabla_\ell(\theta) + \frac{\partial \log(\pi(\theta))}{\partial \theta}$ to define a proposal distribution for θ. The derivative of the log prior distribution $\log \pi(\theta)$ as described in Equation B.4 is straightforward to obtain. A Langevin MH sampler for θ (or a block component of this vector) proposes new states according to $\phi_n(\theta'; \theta + h\frac{\partial \log \pi(\theta|\mathbf{Y})}{\partial \theta}, 2h^2\mathbf{I}_n)$, and the acceptance probability is

$$\alpha = \min\left\{1, \frac{p(\theta'|\mathbf{Y})\phi_n(\theta; \theta' + h\frac{\partial \log \pi(\theta'|\mathbf{Y})}{\partial \theta}, 2h^2\mathbf{I}_n)}{p(\theta|\mathbf{Y})\phi_n(\theta'; \theta + h\frac{\partial \log \pi(\theta|\mathbf{Y})}{\partial \theta}, 2h^2\mathbf{I}_n)}\right\}. \tag{B.5}$$

Here, h is a tuning parameter. The optimal acceptance rate of the Langevin MH sampler is 0.57 (Roberts & Rosenthal, 1998), and this is an easy way to tune h. More nuanced versions of the Metropolized Langevin sampler exist.

Alternatively, a Gaussian approximation of the posterior distribution may be used for reduced computational time:

$$\hat{\pi}_G(\theta|\mathbf{Y}) \equiv N\left(\hat{\theta}, -\left[\mathbf{H}_\ell(\hat{\theta}) + \frac{\partial^2 \log(\pi(\hat{\theta}))}{\partial \theta^2}\right]^{-1}\right). \tag{B.6}$$

Here, $\hat{\theta}$ is the posterior mode, obtained via Newton's method, with approximate covariance equal to the negative inverse of the log-posterior's Hessian matrix evaluated at $\hat{\theta}$.

Appendix C. Supplemental Material for Chapter 4

3.1 Details on computation

The MCMC sampler described in Algorithm 1 of Chapter 4 splits the model parameters into three blocks which are updated based the conditional posterior distribution for each block.

Updating $\mathbf{z}_i$

In the first step we update latent rate parameter $\mathbf{z}_i$, a vector of length equal to the number of graph nodes. Because of high dimensionality, a random walk Metropolis Hastings sampler for $\mathbf{z}_i$ will mix poorly. Furthermore, the individual elements of $\mathbf{z}_i$ are not conditionally independent of each other, which means that a multivariate updating scheme is likely necessary. Compounding these issues is the fact that highly variable observation efforts associated with $\mathbf{z}_i$ result in highly variable marginal posterior variances, which makes global tuning of our proposal scheme more challenging. To this end we used a variant of the manifold Metropolis adjusted Langevin algorithm (mMALA) introduced by Girolami and Calderhead (2011), which extended the traditional MALA algorithm referenced in Appendices A and B by scaling the gradient with the inverse of the expected Fisher Information evaluated at the current value of the parameter for which a proposal is being made. To reduce the computational burden at each iteration of our sampler, we implemented "diagonalized" MALA (dMALA) which scales the gradient using only the diagonal elements of the expected Fisher Information.

Implementing the algorithm requires computation of the gradient of the log posterior $\pi(\mathbf{z}_i|-)$, as well as the second derivatives of the log posterior. Because $z_{ij} > 0$ and the proposal distribution used in the dMALA algorithm is Gaussian, it makes sense to calculate derivatives with respect to $\eta_{ij} = \log(z_{ij})$:

$$\frac{\partial \log \pi(\mathbf{z}_i|-)}{\partial \eta_{ij}} = y_{ij} + \alpha - z_{ij}\left(\frac{\alpha}{\mathbf{m}'_{ij}\mathbf{z}_{i-1}} + t_{ij}\right) - \alpha z_{ij}\sum_k^p \left[\frac{m_{(i+1)kj}}{\mathbf{m}'_{(i+1)k}\mathbf{z}_i} - \frac{m_{(i+1)kj}z_{(i+1)k}}{\left(\mathbf{m}'_{(i+1)k}\mathbf{z}_i\right)^2}\right]$$

$$\frac{\partial^2 \log \pi(\mathbf{z}_i|-)}{\partial \eta_{ij}^2} = \left(\frac{\partial \log \pi(\mathbf{z}_i|-)}{\partial \eta_{ij}} - y_{ij} - \alpha\right) - \alpha z_{ij}^2 \sum_k^p \left[-\frac{m_{(i+1)kj}^2}{\left(\mathbf{m}'_{(i+1)k}\mathbf{z}_i\right)^2} + 2\frac{m_{(i+1)kj}^2 z_{(i+1)k}}{\left(\mathbf{m}'_{(i+1)k}\mathbf{z}_i\right)^3}\right] \tag{C.1}$$

The covariance of the proposal distribution is diagonal, allowing us to generate $\mathbf{z}'_i$ using

$$\eta'_{ij} \sim N\left(\eta_{ij} - \frac{h}{2}\frac{\partial \log \pi(\mathbf{z}_i|-)}{\partial \eta_{ij}} \Bigg/ \frac{\partial^2 \log \pi(\mathbf{z}_i|-)}{\partial \eta_{ij}^2}, -h\left[\frac{\partial^2 \log \pi(\mathbf{z}_i|-)}{\partial \eta_{ij}^2}\right]^{-1}\right) \tag{C.2}$$

where h is a tuning parameter. While the individual elements of $\mathbf{z}'_i$ are independent under this proposal distribution, the full proposal must be accepted or rejected in its entirety.

Updating transition matrix parameters

Because the prior distribution of q_i is discrete with only 3 possible values, the most straightforward way to sample from $\pi(q_i|-)$ is to compute the conditional posterior probability for each of $\pi(q_i = 1|-)$, $\pi(q_i = 0|-)$ and $\pi(q_i = -1|-)$, and to sample according to these probabilities scaled such that they sum to one.

We recommend updating the three remaining transition matrix parameters jointly. We found that a Metropolis-Hastings update using an adaptive MCMC proposal scheme such as those suggested in Andrieu and Thoms (2008) works well.

Bibliography

Abdi, H., O'Toole, A. J., Valentin, D., & Edelman, B. (2005). DISTATIS: The analysis of multiple distance matrices. *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)-Workshops*, 42–44.

Adde, A., Casabona i Amat, C., Mazerolle, M. J., Darveau, M., Cumming, S. G., & O'Hara, R. B. (2021). Integrated modeling of waterfowl distribution in western canada using aerial survey and citizen science (eBird) data. *Ecosphere*, *12*(10).

Agarwal, D. K., Gelfand, A. E., & Citron-Pousty, S. (2002). Zero-inflated models with application to spatial count data. *Environmental and Ecological Statistics*, *9*(4), 341–355.

Ahmed, D. A., Bailey, J. D., & Bonsall, M. B. (2023). On random walk models as a baseline for animal movement in three-dimensional space. *Ecological Modelling*, *475*.

Alerstam, T. (1993). *Bird migration*. Cambridge University Press.

Andrieu, C., & Thoms, J. (2008). A tutorial on adaptive mcmc. *Statistics and computing*, *18*, 343–373.

Arbia, G. (2012). *Spatial data configuration in statistical analysis of regional economic and related problems* (Vol. 14). Springer Science & Business Media.

Aschwanden, J., Schmidt, M., Wichmann, G., & et al. (2020). Barrier effects of mountain ranges for broad-front bird migration. *Journal of Ornithology*, *161*, 59–71.

Avadhani, T. (1969). An analysis of integrated observations. *Journal of the Indian Society of Agricultural Statistics*, *21*, 39.

Banerjee, S., Carlin, B. P., & Gelfand, A. E. (2003). *Hierarchical modeling and analysis for spatial data*. Chapman; Hall/CRC.

Banerjee, S., Finley, A. O., Waldmann, P., & Ericsson, T. (2010). Hierarchical spatial process models for multiple traits in large genetic trials. *Journal of the American Statistical Association*, *105*(490), 506–521.

Banerjee, S., Gelfand, A. E., Finley, A. O., & Sang, H. (2008). Gaussian predictive process models for large spatial data sets. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *70*(4), 825–848.

Barton, G. G., & Sandercock, B. K. (2018). Long-term changes in the seasonal timing of landbird migration on the pacific flyway. *The Condor: Ornithological Applications*, *120*(1), 30–46.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society: Series B (Methodological)*, *36*(2), 192–225.

Besag, J., & Kooperberg, C. (1995). On conditional and intrinsic autoregressions. *Biometrika*, *82*(4), 733–746.

Besag, J., & Mondal, D. (2005). First-order intrinsic autoregressions and the de wijs process. *Biometrika*, *92*(4), 909–920.

Besag, J., York, J., & Mollié, A. (1991). Bayesian image restoration with two applications in spatial statistics. *Annals of the institute of statistical mathematics*, *43*(1), 1–20.

Bianchini, K., & Tozer, D. C. (2023). Using breeding bird survey and eBird data to improve marsh bird monitoring abundance indices and trends. *Avian Conservation and Ecology*, *18*(1).

Bornn, L., Shaddick, G., & Zidek, J. V. (2012). Modeling nonstationary processes through dimension expansion. *Journal of the American Statistician*, *107*(497), 281–289.

Buhnerkempe, M. G., Webb, C. T., Merton, A. A., Buhnerkempe, J. E., Givens, G. H., Miller, R. S., & Hoeting, J. A. (1996). Identification of migratory bird flyways in north america using community detection on biological networks. *Ecological Applications*, *26*(3), 740–751.

Caughlan, L., & Oakley, K. L. (2001). Cost considerations for long-term ecological monitoring. *Ecological Indicators*, *1*(2), 123–124.

Chandra, A. K., Raghavan, P., Ruzzo, W. L., Smolensky, R., & Tiwari, P. (1996). The electrical resistance of a graph captures its commute and cover times. *Computational complexity*, *6*(4), 312–340.

Chebotarev, P. (2011). A class of graph-geodetic distances generalizing the shortest-path and the resistance distances. *Discrete Applied Mathematics*, *159*(5), 295–302.

Christensen, M. F., & Eidsvik, J. (2024). A dimension reduction approach to edge weight estimation for use in spatial models. *arXiv*.

Christensen, M. F., Heaton, M. J., Rupper, S., Reese, C. S., & Christensen, W. F. (2019). Bayesian multi-scale spatio-temporal modeling of precipitation in the indus watershed. *Frontiers in Earth Science*, *7*, 210.

Christensen, M. F., & Hoff, P. D. (2023). A flexible and interpretable spatial covariance model for data on graphs. *arXiv*.

Christensen, M. F., & Hoff, P. D. (2024). Incorporating circuit theory into a dynamic model for crowd-sourced observations of migratory birds. *arXiv*.

Coastal Area Management Act. (1974).

Covino, K. M., Horton, K. G., & Morris, S. R. (2020). Seasonally specific changes in migration phenology across 50 years in the black-throated blue warbler. *The Auk*, *137*(2).

Cressie, N. (1993). *Statistics for spatial data*. John Wiley & Songs.

Datta, A., Banerjee, S., Finley, A. O., & Gelfand, A. E. (2016). Hierarchical nearest-neighbor gaussian process models for large geostatistical datasets. *Journal of the American Statistical Association*, *111*(514), 800–812.

Dickson, B. G., Albano, C. M., Anantharaman, R., Beier, P., Fargione, J., Graves, T. A., Gray, M. E., Hall, K. R., Lawler, J. J., Leonard, P. B., et al. (2019). Circuit-theory applications to connectivity science and conservation. *Conservation biology*, *33*(2), 239–249.

Dunn, J. E., & Gipson, P. S. (1977). Analysis of radio telemetry data in studies of home range. *Biometrics*, 85–101.

Dunson, D. B., Wu, H.-T., & Wu, N. (2022). Graph based gaussian processes on restricted domains. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *84*(2), 414–439.

Ejigu, B. A., & Wencheko, E. (2020). Introducing covariate dependent weighting matrices in fitting autoregressive models and measuring spatio-environmental autocorrelation. *Spatial Statistics*, *38*, 100454.

Fagan, W. F., & Calabrese, J. M. (2014). The correlated rnadom walk and the rise of movement ecology. *The Bulletin of the Ecological Society of America*, *95*, 204–206.

Fisher, R. A. (1921). 014: On the “probable error" of a coefficient of correlation deduced from a small sample.

Fritts, R. (2022). Avian superhighways: The four flyways of north america [[Online; accessed 7-April-2023]]. https://abcbirds.org/blog/north-american-bird-flyways/

Garner, C. L., & Raudenbush, S. W. (1991). Neighborhood effects on educational attainment: A multilevel analysis. *Sociology of education*, 251–262.

Gelfand, A. E., & Vounatsou, P. (2003). Proper multivariate conditional autoregressive models for spatial data analysis. *Biostatistics*, *41*(1), 11–15.

Gelfand, A. E., Zhu, L., & Carlin, B. P. (2001). On the change of support problem for spatio-temporal data. *Biostatistics*, *2*(1), 31–45.

Gelman, A., Hwang, J., & Vehtari, A. (2014). Understanding predictive information criteria for bayesian models. *Statistics and computing*, *24*(6), 997–1016.

Gihman, I. I., & Skorohod, A. V. (1974). *The theory of stochastic processes*. Berlin: Springer-Verlag.

Girolami, M., & Calderhead, B. (2011). Riemann manifold langevin and hamiltonian monte carlo methods. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *73*(2), 123–214.

Glennie, R., Adam, T., Leos-Barajas, V., Michelot, T., Photopoulou, T., & McClintock, B. T. (2023). Hidden markov models: Pitfalls and opportunities in ecology. *Methods in Ecology and Evolution*, *14*(1), 43–56.

Grafius, D. R., Corstanje, R., Siriwardena, G. M., Plummer, K. E., & Harris, J. A. (2017). A bird's eye view: Using circuit theory to study urban landscape connectivity for birds. *Landscape Ecology*, *32*, 1771–1787.

Gramacy, R. B., & Apley, D. W. (2015). Local gaussian process approximation for large computer experiments. *Journal of Computational and Graphical Statistics*, *24*(2), 561–578.

Guinness, J. (2018). Permutation and grouping methods for sharpening gaussian process approximations. *Technometrics*, *60*(4), 415–429.

Guttorp, P., Meiring, W., & Sampson, P. D. (1994). A space-time analysis of ground-level ozone data. *Environmetrics*, *5*(3), 241–254.

Hankin, R. K. (2006). Resistor networks in r: Introducting the resistorarray package. *R News*, *6*(2), 52–54.

Hanks, E. M., & Hooten, M. B. (2013). Circuit theory and model-based inference for landscape connectivity. *Journal of the American Statistical Association*, *108*(501), 22–33.

Harman, H. H. (1976). *Modern factor analysis*. University of Chicago press.

Hass, E. K., La Sorte, F. A., McCaslin, H. M., Belotti, M. C., & Horton, K. G. (2022). The correlation between eBird community science and weather surveillance radar-based estimates of migration phenology. *Global Ecology and Biogeography*, *31*(11), 2219–2230.

Hebblewhite, M., & Haydon, D. T. (2010). Distinguishing technology from biology: A critical review of the use of gps telemetry data in ecology. *Philosophical Transactions of the Royal Society B*, *365*, 2303–2312.

Hedenstrom, A., & Alerstam, T. (1998). How fast can birds migrate? *Journal of Avian Biology*, *29*(4), 424–432.

Helbig, A. J. (2003). Evolution of bird migration: A phylogenetic and biogeographic perspective. *Avian Migration*.

Higdon, D., Gattiker, J., Williams, B., & Rightley, M. (2008). Computer model calibration using high-dimensional output. *Journal of the American Statistical Association*, *103*(482), 570–583.

Higdon, D. (1998). A process-convolution approach to modelling temperatures in the north atlantic ocean. *Environmental and Ecological Statistics*, *5*, 173–190.

Hjelle, Ø., & Dæhlen, M. (2006). *Triangulations and applications*. Springer Science & Business Media.

Hochachka, W. M., Alonso, H., Gutierrez-Exposito, C., Miller, E., & Johnston, A. (2021). Regional variation in the impacts of the COVID-19 pandemic on the quantity and quality of data collected by the project eBird. *Biological Conservation*, *254*.

Hochachka, W. M., Ruiz-Gutierrez, V., & Johnston, A. (2023). Considerations for fitting occupancy models to data from eBird and similar volunteer-collected data. *Ornithology*, *140*(4).

Holsclaw, T., Greene, A. M., Robertson, A. W., & Smyth, P. (2016). A bayesian hidden markov model of daily precipitation over south and east asia. *Hydrometeorology*, *17*(1), 3–25.

Hooten, M. B., Johnson, D. S., & McClintock, J. M. (2017). *Animal movement: Statistical models for telemetry data*. CRC Press.

Horns, J. J., Adler, F. R., & Sekercioglu, C. H. (2018). Using opportunistic citizen science data to estimate avian population trends. *Biological Conservation*, *221*, 151–159.

Hughes, J., & Haran, M. (2013). Dimension reduction and alleviation of confounding for spatial generalized linear mixed models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *75*(1), 139–159.

Hunt, P. D., & Flaspohler, D. J. (2020). Yellow-rumped warbler (steophaga coronata). In P. G. Rodewald (Ed.), *Birds of the world*. Cornell Lab of Ornithology.

Ivanciuc, O., Ivanciuc, T., & Klein, D. J. (2001). Intrinsic graph distances compared to euclidean distances for correspondent graph embedding. *MATCH Communications in Mathematical and in Computer Chemistry*, *44*, 251–278.

Jin, X., Carlin, B. P., & Banerjee, S. (2005). Generalized hierarchical multivariate car models for areal data. *Biometrics*, *61*(4), 950–961.

Jungnickel, D. (2012). *Graphs, networks and algorithms*. Berlin: Springer.

Katzfuss, M., & Guinness, J. (2021). A general framework for becchia approximations of gaussian processes. *Statistical Science*, *36*(1), 124–141.

Klein, D. J., & Randić, M. (1993). Resistance distance. *Journal of mathematical chemistry*, *12*(1), 81–95.

Klein, D. J., & Zhu, H.-Y. (1998). Distances and volumina for graphs. *Journal of mathematical chemistry*, *23*(1), 179–195.

Koen, E. L., Bowman, J., Sadowski, C., & Walpole, A. A. (2014). Landscape connectivity for wildlife: Development and validation of multispecies linkage maps. *Methods in Ecology and Evolution*, *5*(7), 626–633.

Kokko, H. (1999). Competition for early arrival in migratory birds. *Journal of Animal Ecology*, *68*(5), 940–950.

Lawley, D. N., & Maxwell, A. E. (1962). Factor analysis as a statistical method. *Journal of the Royal Statistical Society. Series D (The Statistician)*, *12*(3), 209–229.

Lee, B. S., & Haran, M. (2022). Picar: An efficient extendable approach for fitting hierarchical spatial models. *Technometrics*, *64*(2), 187–198.

Lee, D. (2011). A comparison of conditional autoregressive models used in bayesian disease mapping. *Spatial and Spatio-temporal Epidemiology*, *2*(2), 79–89.

Lindenmayer, D. B., & Likens, G. E. (2010). The science and application of ecologoical monitoring. *Biological Conservation*.

Lindgren, F., Rue, H., & Lindström, J. (2011). An explicit link between gaussian fields and gaussian markov random fields: The stochastic partial differential equation approach. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *73*(4), 423–498.

Ma, H., Carlin, B. P., & Banerjee, S. (2010). Hierarchical and joint site-edge methods for medicare hospice service region boundary analysis. *Biometrics*, *66*(2), 355–364.

Marra, P. P., Francis, C. M., Mulvihill, R. S., & Moore, F. R. (2005). The influence of climate on the timing and rate of spring bird migration. *Oecologia*, *142*, 307–315.

Matérn, B. (1960). Spatial variation - stochastic models and their applications to some problems in forest survey sampling investigations. *Report of the Forest Research Institute of Sweden*.

McClintock, B. T., King, R., Thomas, L., Matthiopoulos, J., McConnel, B. J., & Morales, J. M. (2012). A general discrete-time modeling framework for animal movement using multistate random walks. *Ecological Monographs*, *82*, 335–349.

McClintock, B. T., Langrock, R., Gimenez, O., Cam, E., Borchers, D. L., Glennie, R., & Patterson, T. A. (2020). Hidden markov models: Pitfalls and opportunities in ecology. *Ecology Letters*, *23*(12), 1878–1903.

McDonald, Y. J., & Grineski, S. E. (2012). Disparities in access to residential plumbing: A binational comparison of environmental injustice in el paso and ciudad juárez. *Population and environment*, *34*, 194–216.

McRae, B. H. (2006). Isolation by resistance. *Evolution*, *60*(8), 1551–1561.

McRae, B. H., & Beier, P. (2007). Circuit theory predicts gene flow in plant and animal populations. *Proceedings of the National Academy of Sciences*, *104*(50), 19885–19890.

McRae, B. H., Dickson, B. G., Keitt, T. H., & Shah, V. B. (2008). Using circuit theory to model connectivity in ecology, evolution, and conservation. *Ecology*, *89*(10), 2712–2724.

Mercer, W., & Hall, A. (1911). The experimental error of field trials. *The Journal of Agricultural Science*, *4*(2), 107–132.

Moran, P. A. (1950). Notes on continuous stochastic phenomena. *Biometrika*, *37*(1/2), 17–23.

NCpedia. (2015). Our state geography in a snap: Three regions overview — NCpedia [[Online; accessed 20-February-2022]]. http://www.ncpedia.org/our-state-geography-snap-three

Nilsson, C., Klaassen, R. H., & Alerstam, T. (2013). Differences in speed and duration of bird migration between spring and autumn. *The American Naturalist*, *181*(6), 837–845.

Paciorek, C. J. (2013). Spatial models for point and areal data using markov random fields on a fine grid. *Electron. J. Statist.*, *7*, 946–972.

Paciorek, C. J., & Schervish, M. J. (2006). Spatial modelling using a new class of nonstationary covariance functions. *Environmetrics: The official journal of the International Environmetrics Society*, *17*(5), 483–506.

Patterson, T. A., Thomas, L., Wilcox, C., Ovaskainen, O., & Matthiopoulos, J. (2008). State-space models of individual animal movement. *Trends in ecology & evolution*, *23*(2), 87–94.

Paul, C. R. (2001). *Fundamentals of electric circuit analysis*. John Wiley & Sons.

Perras, M., & Nebel, S. (2012). Satellite telemetry and its impact on the study of animal migration. *Nature Education Knowledge*, *3*(12), 4.

Peterson, E. E., Hanks, E. M., Ver Hoef, J. M., Hooten, M. B., & Fortin, M.-J. (2019). Intrinsic graph distances compared to euclidean distances for correspondent graph embedding. *Ecological Monographs*, *89*(2), e01355.

Prima, M.-C., Duchesne, T., Fortin, A., Rivest, L.-P., & Fortin, D. (2018). Combining network theory and reaction-advection-diffusion modelling for predicting animal distribution in dynamic environments. *Methods in Ecology and Evolution*, *9*(5), 1221–1231.

Reich, B. J., Hodges, J. S., & Zadnik, V. (2006). Effects of residual smoothing on the posterior of the fixed effects in disease-mapping models. *Biometrics*, *62*(4), 1197–1206.

Rising, J. D., & Flood, N. J. (2020). Baltimore oriole (icterus galbula). In P. G. Rodewald (Ed.), *Birds of the world*. Cornell Lab of Ornithology.

Roberts, G. O., & Rosenthal, J. S. (1998). Optimal scaling of discrete approximations to langevin diffusions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *60*(1), 255–268.

Rue, H., & Held, L. (2005). *Gaussian markov random fields: Theory and applications*. Chapman; Hall/CRC.

Rue, H., Martino, S., & Chopin, N. (2009). Approximate bayesian inference for latent gaussian models by using integrated nested laplace approximations. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *71*(2), 319–392.

Sampson, P. D., & Guttorp, P. (1992). Nonparametric estimation of nonstationary spatial covariance structure. *Journal of the American Statistical Association*, *87*(417), 108–119.

Sangalli, L. M., Ramsay, J. O., & Ramsay, T. O. (2013). Spatial spline regression models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *75*(4), 681–703.

Schlyer, K. (2012). *Continental divide: Wildlife, people, and the border wall*. Texas A&M University Press.

Schmidt, A. M., & Guttorp, P. (2020). Flexible spatial covariance functions. *Spatial Statistics*, *37*, 100416.

Schmidt, A. M., & O'Hagan, A. (2003). Bayesian inference for nonstationary spatial covariance structures via spatial deformations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *65*(3), 743–758.

Schwoerer, T., & Dawson, N. G. (2022). Evolution of bird migration: A phylogenetic and biogeographic perspective. *PLoS One*, *17*(7).

Shapiro, S. S., & Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, *52*(3-4), 591–611.

Sharma, A., Singh, D., Malik, S., Gupta, N. J., Rani, S., & Kumar, V. (2018). Difference in control between spring and autumn migration in birds: Insight from seasonal changes in hypothalamic gene expression in captive buntings. *Proceedings of the Royal Society B*, *285*(1885), 20181531.

Smith, T. R., Wakefield, J., & Dobra, A. (2015). Restricted covariance priors with applications in spatial statistics. *Bayesian analysis (Online)*, *10*(4), 965.

Spielman, D. (2012). Spectral graph theory. *Combinatorial scientific computing*, *18*, 18.

Stein, M. L. (1999). *Interpolation of spatial data: Some theory for kriging*. Springer Science & Business Media.

Stein, M. L. (2014). Limitations on low rank approximations for covariance matrices of spatial data. *Spatial Statistics*, *8*, 1–19.

Strano, E., Shai, S., Dobson, S., & Barthelemy, M. (2015). Multiplex networks in metropolitan areas: Generic features and local effects. *Journal of The Royal Society Interface*, *12*(111), 20150651.

Stuber, E. F., Robinson, O. J., Bjerre, E. R., Otto, M. C., Millsap, B. A., Zimmerman, G. S., Brasher, M. G., Ringelman, K. M., Fournier, A. M., Yetter, A., Isola, J. E., & Viviana, R.-G. (2022). The potential of semi-structured citizen science data as a supplement

for conservation decision-making: Validating the performance of ebird against targeted avian monitoring efforts. *Biological Conservation*, *270*.

Sullivan, B. L., Wood, C. L., Iliff, M. J., Bonney, R. E., Fink, D., & Kelling, S. (2009). Ebird: A citizen-based bird observation network in the biological sciences. *Biological conservation*, *142*(10), 2282–2292.

Sutherland, W. J. (1996). Predicting the consequences of habitat loss for migratory populations. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, *263*(1375), 1325–1327.

Swick, N. (2016). *American birding association field guide to birds of the carolinas*. Scott & Nix Inc.

Tang, B., Clark, J. S., & Gelfand, A. E. (2021). Modeling spatially biased citizen science effort through the eBird database. *Environmental and Ecological Statistics*, *28*(3), 609–630.

Team eBird. (2024). 2023 year in review: eBird, Merlin, Macaulay Library, and Birds of the World [[Online; accessed 16-April-2024]]. https://ebird.org/news/2023-year-in-review

The Cornell Lab. (2021). The basics of bird migration: How, why, and where [[Online; accessed 7-April-2023]]. https://www.allaboutbirds.org/news/the-basics-how-why-and-where-of-bird-migration/#

Thiele, J., Buchholz, S., & Schirmel, J. (2018). Using resistance distance from circuit theory to model dispersal through habitat corridors. *Journal of Plant Ecology*, *11*(3), 385–393.

Thygesen, U., Pedersen, M., & Madsen, H. (2009). Geolocating fish using hidden markov models and data storage tags.

Tottrup, A. P., Rainio, K., Coppack, T., Lehikoinen, E., Rahbek, C., & Thorup, K. (2010). Local temperature fine-tunes the timing of spring migration in birds. *Integrative and Comparative Biology*, *50*(3), 293–304.

Tryjanowski, P., & Yosef, R. (2002). Differences between the spring and autumn migration of the red-backed shrike lanius collurio: Record from the eilat stopover (israel). *Acta Ornithologica*, *37*(2), 85–90.

Turchin, P. (1991). Translating foraging movements in heterogeneous environments into the spatial distribution of foragers. *Ecology*, *72*(4), 1253–1266.

Turchin, P. (1998). *Quantitative analysis of movement: Measuring and modeling population redistribution in animals and plants*. Sinauer.

Ver Hoef, J. M. (2018). Kriging models for linear networks and non-euclidean distances: Cautions and solutions. *Methods in Ecology and Evolution*, *9*(6), 1600–1613.

Ver Hoef, J. M., Hanks, E. M., & Hooten, M. B. (2018). On the relationship between conditional (CAR) and simultaneous (SAR) autoregressive models. *Spatial statistics*, *25*, 68–85.

Ver Hoef, J. M., Peterson, E. E., Hooten, M. B., Hanks, E. M., & Fortin, M.-J. (2018). Spatial autoregressive models for statistical inference from ecological data. *Ecological Monographs*, *88*(1), 36–59.

Visser, M. E., Perdeck, A. C., Van Balen, J. H., & Both, C. (2009). Climate change leads to decreasing bird migration distances. *Global Change Biology*, *15*(8), 1859–1865.

Wang, X., Yang, J., Lee, C., Ji, Z., & You, S. (2016). Macro-level safety analysis of pedestrian crashes in shanghai, china. *Accident Analysis & Prevention*, *96*, 12–21.

White, G., & Ghosh, S. K. (2009). A stochastic neighborhood conditional autoregressive model for spatial data. *Computational Statistics and Data Analysis*, *53*(8), 3033–3046.

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

Wikle, C. K., & Hooten, M. B. (2010). A general science-based framework for dynamical spatio-temporal models. *Test*, *19*, 417–451.

Yang, Y., & Dunson, D. B. (2016). Bayesian manifold regression. *Ann. Statist.*, *40*(2).

Youngman, B. (2023). *Deform: Spatial deformation and dimension expansion gaussian processes* [R package version 1.0.0]. https://CRAN.R-project.org/package=deform

Zellner, A. (1988). Optimal information processing and bayes's theorem. with comments and a reply by the author. *American Statistician*, *42*(4), 278–284.

Zhang, G. (2020). Spatial and temporal patterns in volunteer data contribution activities: A case study of eBird. *ISPRS International Journal of Geo-Information*, *9*(10), 597.

Zhu, H.-Y., & Klein, D. J. (1996). Graph-geometric invariants for molecular structures. *Journal of chemical information and computer sciences*, *36*(6), 1067–1075.

Zimmerman, D. L., & Ver Hoef, J. M. (2022). On deconfounding spatial confounding in linear models. *The American Statistician*, *76*(2), 159–167.

Biography

Michael Christensen earned both his bachelor's and master's degree in statistics from Brigham Young University in April 2019. In August 2019 he began his PhD studies in the Department of Statistical Science at Duke University where he was advised by Dr. Peter Hoff. He has accepted a position as an assistant professor in the Department of Statistics at Brigham Young University, where he will begin working in July 2024.