DISSERTATION

STATISTICAL MODELS FOR ANIMAL MOVEMENT AND LANDSCAPE CONNECTIVITY

Submitted by
Ephraim M. Hanks
Department of Statistics

In partial fulfillment of the requirements
For the Degree of Doctor of Philosophy
Colorado State University
Fort Collins, Colorado
Summer 2013

Doctoral Committee:

Advisor: Mevin B. Hooten

Jennifer Hoeting
Haonan Wang
Mat Alldredge
David Theobald

ABSTRACT

STATISTICAL MODELS FOR ANIMAL MOVEMENT AND LANDSCAPE CONNECTIVITY

This dissertation considers statistical approaches to the study of animal movement behavior and landscape connectivity, with particular attention paid to modeling how movement and connectivity are influenced by landscape characteristics. For animal movement data, a novel continuous-time, discrete-space model of animal movement is proposed. This model yields increased computational efficiency relative to existing discrete-space models for animal movement, and a more flexible modeling framework than existing continuous-space models. In landscape genetic approaches to landscape connectivity, spatially-referenced genetic allele data are used to study landscape effects on gene flow. An explicit link is described between a common circuit-theoretic approach to landscape genetics and variogram fitting for Gaussian Markov random fields. A hierarchical model for landscape genetic data is also proposed, with a multinomial data model and latent spatial random effects to model spatial correlation.

ACKNOWLEDGEMENTS

Good science is rarely done by an individual working in a vaccuum. I am indebted to many statisticians and scientists who have been mentors, friends, and collaborators over the past five years. To reflect this, I have used the plural "we" in Chapters 2-4 of this dissertation. The writing is my own, and I alone am responsible for any errors, but these chapters are results of collaborative work with other statisticians and scientists, and I acknowledge their contributions.

Thanks to Devin Johnson, Jeremy Sterling, and others at the National Marine Mammal Laboratory for the northern fur seal data which were used in my first movement analysis. Thanks to Mat Alldredge at the Colorado Division of Parks and Wildlife for the lion data used in Chapter 2, and for many insights into movement behavior. Thanks to Karen Mock at Utah State University for introducing me to the field of landscape genetics and isolation by resistance. Thanks to Steve Knick, Sara Oyler-McCance, Todd Cross, and Jenny Fike for their work collecting, analyzing, and preparing the genetic data used in Chapter 4, as well as for their insights on greater sage-grouse. Thanks also to Jennifer Hoeting and Erin Schliep for a series of enlightening discussions on similar topics over the course of a year.

Most of all, thanks to Mevin Hooten. I couldn't have asked for a better mentor, advisor and friend.

TABLE OF CONTENTS

CHAPTER 1

STUDYING LANDSCAPE CONNECTIVITY THROUGH ANIMAL MOVEMENT AND GENETIC VARIATION

Landscape connectivity is “the degree to which the landscape facilitates or impedes movement among resource patches” (Taylor et al., 1993). Connectivity among subpopulations or habitat patches is an important factor in maintaining biodiversity, understanding the spread of infectious diseases, and allocating resources for conservation. Many management efforts devote considerable resources to maintaining or restoring landscape connectivity (e.g., Crooks and Sanjayan, 2006).

Understanding how the environment influences animal behavior and landscape connectivity is one of the most important questions in ecology and conservation (Dalziel et al., 2008; Gurarie et al., 2009; Cagnacci et al., 2010). Human disturbance and ongoing climate change alter the landscape, affecting connectivity and increasing uncertainty about the future of current ecosystems. In this environment, making data driven decisions is critical to effective management of species and ecosystems.

In this dissertation, my goal is to propose novel statistical approaches to studying connectivity by considering two common data types used in this field: animal movement data and spatially referenced genetic data. I begin by outlining current approaches in the modeling of animal movement data and spatially referenced genetic data.

1.1 Modeling Animal Movement

Recent advances in technology have made animal telemetry data easier to collect at finer temporal resolutions than previously possible (Tomkiewicz et al., 2010), but there are still significant challenges in linking telemetry data to animal movement and resource selection.

Some challenges arise in the use of telemetry data, which are typically irregular in time, have measurement error that is varying in severity, and may exhibit temporal autocorrelation (Kuhn et al., 2009; Tomkiewicz et al., 2010). Other challenges arise from modeling something as complex as animal movement, which typically exhibits changing behavior over time (Morales et al., 2004; Nathan et al., 2008; Getz and Saltz, 2008; Gurarie et al., 2009; Forester et al., 2009; Merrill et al., 2010; Polansky et al., 2010) and may be driven by a mix of external (environmental) and internal (biotic) factors. Still other challenges involve making population level inference based on telemetry data from multiple animals (Aarts et al., 2008; Morales et al., 2010; Hanks et al., 2011).

Movement models are primarily concerned with modeling the change in an animal's position, often in response to environmental factors. This sets them apart from occupancy or resource selection models, which are primarily concerned with modeling the animal's position itself. Thus, the "response" variable in a resource selection model is typically an animal's location, such as a telemetry location, while the "response" variable in a movement model is typically an animal's movement step between two observed animal locations. In a continuous space model for animal movement, the animal's movement step is a velocity vector pointing from the animal's location at a given time to the animal's next (in time) observed location. In a discrete space model for animal movement, the movement step is the change from one location to another (e.g., Turchin, 1998).

For continuous space models of animal movement, there are two decompositions of the movement step that are widely used in modeling. The first is a vector decomposition into orthogonal components (e.g., Easting and Northing components), while the second is a "polar" decomposition into step lengths and turning angles. Each of these decompositions are discussed in turn, highlighting strengths and weaknesses of each. In each case, existing approaches to modeling changing movement behavior over time are described. Without accounting for heterogeneity in animal behavior over time, important drivers of movement could appear to be insignificant, due to the temporally changing nature of the animal's response. Thus, particular attention is payed to existing approaches to modeling changing behavior over time.

1.1.1 Orthogonal Decomposition of the Movement Step

Let an animal's position in space at time $t \in \{t_0, t_1, \ldots, t_T\}$ be given by $\mathbf{s}(t) = [s_1(t), s_2(t)]$. For simplicity, consider only the case where the observations are regular in time (e.g., $t_k = t_0 + \Delta t \cdot k$) with temporal interval Δt, though this generalizes to irregular observations in a straightforward manner. A movement step $\mathbf{v}(t) = [v_1(t), v_2(t)]$ between two consecutive time steps is the first difference of consecutive (in time) locations:

$$\mathbf{v}(t) = \begin{bmatrix} v_1(t) \\ v_2(t) \end{bmatrix} = \mathbf{s}(t + \Delta t) - \mathbf{s}(t) = \begin{bmatrix} s_1(t + \Delta t) - s_1(t) \\ s_2(t + \delta t) - s_2(t) \end{bmatrix}.$$

The movement step $\mathbf{v}(t)$ is typically modeled as being a random vector with mean velocity vector $\nabla H(\mathbf{s}(t))$ equal to the gradient of a potential function $H(\mathbf{s}(t))$:

$$\mathbf{v}(t) = \nabla H(\mathbf{s}(t)) + \boldsymbol{\epsilon}(t). \tag{1.1}$$

The potential function $H(\mathbf{s})$ can be modeled nonparametrically using two dimensional basis splines (e.g., Brillinger et al., 2001; Preisler et al., 2004, 2013), or the gradient of the potential function can be modeled as a parametric function of gradients of continuous covariates (Hanks et al., 2011):

$$\nabla H(\mathbf{s}) = \sum_{k=1}^{K} \beta_k \cdot \nabla X_k(\mathbf{s}). \tag{1.2}$$

Under either formulation, $\nabla H(\mathbf{s})$ is the expected (mean) velocity vector of movement for an animal at location $\mathbf{s}$, and the vector $\boldsymbol{\epsilon}(t)$ is typically taken to be a bivariate Gaussian random variable with diagonal covariance matrix. This error specification corresponds to a stochastic differential equation (SDE) approach to modeling animal movement, where the movement can be modeled by two dimensional Brownian motion about the mean drift vector $\nabla H(\mathbf{s})$. The relationship between SDEs and their discrete approximations will be discussed in more depth in Chapter 5.

To allow for changing movement behavior over time, the potential function can be allowed to vary temporally. Preisler et al. (2013) partition each 24 hour period into 4 segments during which movement behavior is assumed to be homogeneous, and estimate a separate potential function for each time segment. Hanks et al. (2011) allow for changing response $\{\beta_k\}$ to covariate gradients $\{\nabla X_k\}$ over time through a change-point model on $\{\beta_k\}$ in (2):

$$\nabla H(\mathbf{s}) = \sum_{k=1}^{K} \beta_{tk} \cdot \nabla X_k(\mathbf{s}) \quad , \quad \beta_{tk} = \begin{cases} \beta_{1k} & , \; t \in [1, \tau_2 - 1] \\ \beta_{2k} & , \; t \in [\tau_2, \tau_3 - 1] \\ \vdots & \quad \vdots \\ \beta_{Pk} & , \; t \in [\tau_P, T - 1] \end{cases} ,$$

where the number P and temporal locations $\{\tau_p\}$ of change-points are allowed to be random. The resulting model has a variable parameter space, as the addition of new change-points adds new parameters to the model. Hanks et al. (2011) use a birth-death Monte-Carlo algorithm to allow for inference on parameters in this approach. This approach is computationally more expensive than partitioning the time *a priori* as is done by Preisler et al. (2013), but allows the data to flexibly partition the observation window into regions of homogeneous movement behavior, rather than relying on expert opinion.

The gradient-based nature of the orthogonal decomposition of a movement step makes it a natural approach for modeling animal movement toward (away) from attractive (repulsive) locations. In the parametric approach of Hanks et al. (2011), response to covariate gradients can be used to model migratory movement, movement around a central location, or attraction to conspecifics. These and other uses for gradient-based directional drivers of movement are discussed in Chapter 2.

1.1.2 Polar Decomposition of the Movement Step

The orthogonal decomposition of a movement step discussed above provides a natural framework for modeling an animal's response to environmental factors that can be expressed as gradients. However, it can be cumbersome to model responses that are not easily represented in gradient form. For example, it would be difficult to model an animal's absolute velocity (directionless speed) as a function of landcover type using the orthogonal decomposition of the movement step. A natural decomposition of a movement step to model an animal's speed is the polar decomposition of $\mathbf{v}(t)$ into a step length $r(t)$

$$r(t) = ||\mathbf{v}(t)|| = \sqrt{(s_1(t + \Delta t) - s_1(t))^2 + (s_2(t + \Delta t) - s_2(t))^2}$$

and a relative turning angle $\theta(t)$

$$\theta(t) = \arctan\left(\frac{s_2(t)/||\mathbf{s}(t)|| - s_2(t - \Delta t)/||\mathbf{s}(t + \Delta t)||}{s_1(t)/||\mathbf{s}(t)|| - s_1(t - \Delta t)/||\mathbf{s}(t + \Delta t)||}\right)$$

where $\theta(t)$ is the turning angle relative to the previous movement step $\mathbf{v}(t - \Delta t)$.

This decomposition provides a natural framework for modeling how an animal's behavior, as measured by speed (absolute velocity) and path tortuosity, differ over time. For example, one common approach to allow for changing behavior over time is to assume that the animal's movement can be classified into a number of distinct movement modes, and use state switching models to allow the animal to transition between movement regimes over time (Morales et al., 2004; Nathan et al., 2008; Getz and Saltz, 2008; Gurarie et al., 2009; Forester et al., 2009; Merrill et al., 2010; Polansky et al., 2010; McClintock et al., 2012). The state transition process can be parameterized by the landscape, allowing for varying propensity to transition to or from certain movement states in different environments (e.g., Morales et al., 2004). Gurarie et al. (2009) utilize a behavioral change-point model to identify structural changes in an animal's movement, and allow for the number and location of change-points to be inferred from the telemetry data, but their model does not incorporate environmental effects; the inference on behavioral changes is based solely on the telemetry data. Polansky et al. (2010) utilize Fourier and wavelet analysis to examine periodicity in movement behavior. Morales et al. (2004) propose a model based on a mixture of random walks, though their analysis requires specification of the number of movement states an animal can exhibit, which may not be known beforehand. McClintock et al. (2012) propose a method for incorporating a mixture of an unknown number of random walks.

The choice between an orthogonal decomposition or a polar decomposition of the movement step should be driven primarily by examining what effects one wants to model. If directional behavior is important, then the orthogonal decomposition and resulting gradient based modeling should be attractive. If, on the other hand, the modeler wishes to model changes in an animal's speed (absolute velocity) and path tortuosity, the polar decomposition is a natural choice. In many respects, the questions that can be addressed by one decomposition are difficult to address with the other. However, a discrete space approach to modeling movement offers some advantages in this respect. Both directional (gradient based) behavior and location based behavior such as changes in absolute velocity in different cover types (Hooten et al., 2010) can be modeled naturally in a discrete space framework. This allows for a flexible framework for modeling the relationship between movement behavior and multiple environmental and biotic factors of interest. Discrete space approaches are discussed in detail in Chapter 2.

1.2 Current Genetic Approaches to Landscape Connectivity

Animal movement data can be used to directly examine the effect that the landscape has on connectivity by estimating how movement behavior changes across a heterogeneous landscape. Genetic data, on the other hand, do not contain explicit information about how the landscape influences movement and connectivity; instead genetic data can tell us implicitly about connectivity. Genetic variation over space is the result of a spatio-temporal process over many generations of the species, in which genetic mutations occur, animals mate and pass on genetic information to their offspring, and animals disperse across the landscape. Genetic data consist of a snapshot in time of this spatio-temporal process occuring at a genetic time scale.

The study of how landscape features influence the flow of genetic information is known as landscape genetics (Manel et al., 2003). This relatively young field has developed quickly, and multiple approaches exist for making inference on the interface between landscape features and gene flow. The most common approaches can be grouped into two categories: spatial clustering and effective distance analysis.

1.2.1 Spatial Clustering Approaches in Landscape Genetics

One common approach to landscape genetics is to consider clustering the genetic observations spatially into sub-regions that are genetically similar. The result is a partitioning of the study region into sub-regions that represent the spatial domain of sub-populations that are genetically distinct from one another. These approaches typically employ Bayesian hierarchical models (BHMs) for inference, and often are known colloquially by the name of the software package used to implement the approach. Examples include Geneland (Guillot et al., 2005) and TESS (Chen et al., 2007; Durand et al., 2009).

These methods allow for inference about the location of genetic populations and boundaries based on observed genetic samples. However, boundaries resulting from such an analysis are typically linked to landscape features through comparing the locations of potential barriers (e.g., roads or rivers) to the mode of the posterior distribution through a geographic information system (GIS) overlay (Sahlsten et al., 2008; Wheeler et al., 2010), without formal statistical inference. Thus, while landscape genetics is the study of how landscape features influence gene flow (Manel et al., 2003), these spatial partitioning models do not allow the effects of landscape features to be modeled explicitly. Rather, a post-hoc analysis is required to link the posterior distribution of genetic subpopulation boundaries to landscape features which may influence gene flow.

1.2.2 Effective Distance Approaches in Landscape Genetics

An alternative approach to estimating the effects of landscape on gene flow involves computing a pairwise genetic distance measure between observed spatially referenced genetic samples, and comparing this pairwise distance with an effective distance that is a function of the spatial locations where the genetic samples were collected and the landscape features of the study region (e.g., McRae, 2006; Cushman et al., 2006; McRae and Beier, 2007; Cushman and Landguth, 2010; Cushman and Lewis, 2010). The three most common effective measures of spatial distance are Euclidean distance (e.g., Broquet et al., 2006), the resistance distance or effective resistance from electric circuit theory (e.g., McRae et al., 2008), and the least cost path (LCP) distance (e.g., Cushman and Lewis, 2010). If Euclidean distance is used, this effective distance approach is termed "isolation by distance" (IBD). Likewise, if the effective spatial distance is the resistance distance from circuit theory, the approach is termed "isolation by resistance"(IBR).

The IBR approach was first described by McRae (2006), who showed that there are theoretical links between the resistance distance of the circuit representing the heterogeneous landscape and the genetic distance between animal populations. The fixation index F_{st} is a measure of genetic dissimilarity between sub-populations based on comparing the diversity of randomly chosen alleles within each sub-population to the diversity found within the entire population (Holsinger and Weir, 2009). If it is assumed that each node in the graph (e.g., each raster cell G_i on a landscape) contains a sub-population, then pairwise F_{st} values can be computed for each pair of nodes. The major finding of McRae (2006) is that under a random walk model of animal movement and gene flow on an undirected graph, the resistance distance of an analogous circuit is proportional to linearized F_{st}. That is, if F_{st}^{ij} is the pairwise fixation index between the i^{th} and j^{th} node in the graph, then the resistance distance Γ_{ij}

between those two nodes is

$$\Gamma_{ij} \propto \frac{F_{st}^{ij}}{1 - F_{st}^{ij}} \tag{1.3}$$

(see McRae (2006) for details). Thus, resistance distance is proportional to a common formulation of genetic distance between sub-populations. This relationship between gene flow, random walks on a graph, and circuit theory forms the basis for IBR analysis (McRae, 2006; McRae and Beier, 2007; McRae et al., 2008; Cushman et al., 2006; Cushman and Landguth, 2010; Spear et al., 2010).

In a LCP analysis, the effective distance between two nodes is the smallest possible resistance of a single path through the graph that connects the two nodes. Thus, using LCP distance to model animal movement and gene flow is based on the assumption that animals know the landscape fully and are able to pick the "best" path between any two locations of their choosing. An analysis based on resistance distance, instead of LCP distance, assumes that the effective distance between two locations is based on a weighted average of all possible paths through the graph that connect the two locations. Using resistance distance to model animal movement and gene flow is based on the assumption that animals move through the landscape at random, with movement choices being based only on the local environment, as reflected in the connectivity of the graph. At small temporal scales, movement is affected by memory, interactions with other animals, and many other factors not directly related to the local environment. At large temporal scales (i.e., genetic time), the assumption that the environment drives population level movement and gene flow may be more plausible, lending credence to the use of resistance distance. Studies of observed genetic data have found that the correlation between circuit based resistance distance and observed genetic distance is generally higher than the correlation between LCP distance and genetic distance (e.g., McRae and Beier, 2007). However, the LCP distance can be computed much more efficiently than the resistance distance, allowing analysis of larger landscapes at finer resolutions.

Both IBR and LCP approaches rely on a discretized spatial support, where the landscape is viewed as a graph (typically a regular grid) with edge weights between connected nodes being a function of the landscape characteristics ($\mathbf{X}$) at and between the neighboring nodes, and parameters ($\boldsymbol{\theta}$) which determine how the edge weights are affected by the landscape characteristics in $\mathbf{X}$. In conventional approaches, the parameters $\boldsymbol{\theta}$ are often set *a priori* using expert opinion. When $\boldsymbol{\theta}$ is estimated, the estimation is typically achieved using a correlation analysis in which an observed genetic distance matrix is compared to a number of possible resistance distance matrices that result from various hypothesized resistance values $\boldsymbol{\theta}$ for landscape characteristics. The hypothesized resistance values that result in the highest correlation with the observed genetic distance are selected.

1.3 Overview

In this thesis, the study of landscape connectivity is advanced through the introduction of novel statistical approaches to analyzing animal movement and landscape genetic data. In Chapter 2, a continuous-time, discrete-space (CTDS) model for animal movement data is proposed. This CTDS model can be seen as a model based on the sufficient statistics of the discrete-time discrete-space movement model of Hooten et al. (2010), and a latent variable representation of the likelihood is presented that allows for inference to be made on movement parameters using standard generalized linear modeling software. A computationally efficient multiple imputation approach for estimating the posterior predictive distribution of movement parameters is also developed and compared to a full Bayesian analysis. The gains in computational efficiency due to this approximation allow for inference on changing movement behavior over time at finer temporal resolutions than have been possible using existing approaches, and for variable selection using a lasso penalty. The result is a framework for modeling animal movement that allows for flexible modeling of environmental and biotic covariates, changing behavior over time, and variable selection using computationally efficient software. This approach is illustrated by a study of telemetry data obtained from a pair of mountain lions in Colorado, USA, in which movement behavior in response to landscape features, conspecifics, and potential kill sites over time is examined.

In Chapter 3, the circuit theoretic IBR approach to modeling the landscape using effective distance analyses is considered from a statistical perspective. The generalized Wishart distribution for squared distance matrices (McCullagh, 2009) is proposed as a data model for observed pairwise genetic distance matrices, and the relationship between the resistance distance of an undirected landscape graph and the variogram of a zero mean intrinsic conditional auto-regressive Gaussian Markov random field is explored. This provides a spatially explicit model based framework for the analysis of squared genetic distance matrices under the IBR hypothesis that allows for estimation of uncertainty about point estimates for landscape resistances, something not typically available in current distance based genetic approaches. This approach is applied to a study of the effect of elevation on landscape connectivity for alpine chamois in the Bauges mountains of France.

In Chapter 4, the spatially explicit modeling approach for landscape genetics from Chapter 3 is expanded, and existing methodology in spatial statistics is be applied to landscape genetic approaches to studying connectivity. Spatial covariance functions that match the assumptions of the IBD, IBR, and LCP approaches to spatial genetic analysis are proposed, and a comparison is made between variogram fitting approaches in the analysis of geostatistical data and effective distance approaches in landscape genetics. Instead of modeling an observed pairwise genetic matrix, as in Chapter 3, a multinomial model for genetic microsatellite allele data is used in Chapter 4. By including a latent Gaussian random effect in this multinomial model, landscape connectivity can be modelled in a spatial generalized linear mixed modeling framework. This modeling approach is illustrated through a study of sage-grouse in the western U.S. in which recommendations for the optimal allocation of sampling effort in 2013 are provided, based on a retrospective analysis of sage grouse genetic data collected from 2009-2012.

Chapter 5 concludes with a summary and discussion of potential directions for future research, focusing on issues of scale in studying connectivity.

CHAPTER 2

A CONTINUOUS-TIME DISCRETE-SPACE MODEL FOR ANIMAL MOVEMENT

2.1 Introduction

[1] In Chapter 1, the use of animal movement data to study landscape connectivity and animal response to landscape features was introduced, with particular attention paid to the analysis of continuous-space movement steps. In this chapter, discrete-space representations of movement are considered, and a continuous-time discrete-space model for animal movement data is proposed.

There is no shortage of existing statistical models of animal movement; however, most of these models are computationally demanding. For example, consider the agent based model of animal movement of Hooten et al. (2010). The agent based framework is highly flexible, allowing for location-based and directional drivers of movement, but is computationally expensive. Analyzing the movement path of one animal with hourly locations over the course of a week using the approach of Hooten et al. (2010) can require computational time on the order of days using standard computing resources. The velocity-based framework for modeling animal movement of Hanks et al. (2011) allows for time-varying behavior through a changepoint model of response to drivers of movement and is more computationally efficient than the approach of Hooten et al. (2010), requiring computational time on the order of hours for a similar problem. Similarly, the mechanistic approach of McClintock et al. (2012) allows for time-varying behavior through a state-switching approach but is also computationally demanding. These three approaches use Bayesian statistical models, and both Hanks et al.

[1] The material in Chapter 2 is based on the following publication: Hanks, E.M., M.B. Hooten, and M.W. Alldredge. 2013. Continuous-Time, Discrete-Space Models of Animal Movement. *arXiv*:1211.1992.

(2011) and McClintock et al. (2012) allow for time-varying behavior by letting the model parameter space vary, either through a reversible-jump Markov chain Monte Carlo approach (Green, 1995) or the related birth-death Markov chain Monte Carlo approach (Stephens, 2000). These methods can be quite computationally demanding, require the user to tune the algorithm to ensure convergence, and can be inaccessible to many practitioners.

The agent based model of Hooten et al. (2010) assumes a representation of the animal's movement path that is discrete in both space (grid cells) and time (fixed time intervals). The velocity-based movement model of Hanks et al. (2011) assumes a representation of the movement path that is continuous in space and discrete in time. The state-switching model of McClintock et al. (2012) assumes a representation of the movement path that is discrete in time and continuous in space.

In this chapter, we present a continuous-time, discrete-space (CTDS) model for animal movement which allows for flexible modeling of an animal's response to drivers of movement in a computationally efficient framework. We consider a Bayesian approach to inference, as well as a multiple imputation approximation to the posterior distribution of parameters in the movement model. Instead of a state-switching or change-point model for changing behavior over time, we adopt a time-varying coefficient model. We also allow for variable selection using a lasso penalty. This CTDS approach is highly computationally efficient, requiring only minutes or seconds to analyze movement paths that would require hours using the approach of Hanks et al. (2011) or days using the approach of Hooten et al. (2010), allowing the analysis of longer movement paths and more complex behavior than has been previously possible.

In Section 2.2, *Modeling Animal Movement*, we describe the CTDS model for animal movement and present a latent variable representation of the model that allows for inference within a standard generalized linear model (GLM) framework. In Section 2.3, *Posterior Predictive Inference*, we present a Bayesian approach for inference and describe the use of multiple imputation (Rubin, 1987) to approximate the posterior predictive distribution of parameters in the CTDS model. In Section 2.4, *Time-Varying Behavior and Shrinkage Estimation*, we use a varying coefficient approach to model changing behavior over time, and use a lasso penalty for variable selection. In Section 2.5, *Drivers of Animal Movement*, we discuss modeling potential covariates in the CTDS framework. In Section 2.6, *Example: Mountain Lions in Colorado*, we illustrate our approach through an analysis of mountain lion (*Puma concolor*) movement in Colorado, USA. Finally, in Section 2.7, *Discussion*, we discuss possible extensions to the CTDS approach.

2.2 Modeling Animal Movement

Our goal is to specify a model of animal response to drivers of movement that is flexible and computationally efficient. In doing so, we will first consider an existing model for animal movement that is continuous in both time and space. We then consider a discrete (e.g., gridded) model for space (e.g., Hooten et al., 2010), and model animal movement as a continuous-time random walk through the discrete, gridded space.

2.2.1 Continuous-Time Continuous-Space Movement Model

Johnson et al. (2008) propose a continuous time correlated random walk (CTCRW) model of an animal's continuous path conditioned on observed telemetry data. Let $\mathbf{S} = \{\mathbf{s}(t), t = t_0, t_1, \ldots, t_T\}$ be a collection of time-referenced telemetry locations for an animal. If the animal's location and velocity at an arbitrary time t are $\mathbf{s}(t)$ and $\mathbf{v}(t)$, respectively, then the CTCRW model can be specified as follows, ignoring the multivariate notation for simplicity

$$s(t) = s(0) + \int_0^t v(u)du \ ,$$

$$v(t) = \psi_1 + \frac{\psi_2 e^{-\psi_3 t}}{\sqrt{2\psi_3}} \omega \left(e^{2\psi_3 t} \right) \ , \tag{2.1}$$

where $\boldsymbol{\psi} = [\psi_1, \psi_2, \psi_3]$ control the movement and $\omega(t)$ is standard Brownian motion. This model can be discretized and formulated as a state space model, which allows for efficient computation of discretized paths $\tilde{\mathbf{S}}$ at arbitrarily fine time intervals via the Kalman filter (Johnson et al., 2008). If a Bayesian framework is used for inference on $\boldsymbol{\psi}$, then Johnson et al. (2008) show how the posterior distribution $[\boldsymbol{\psi}|\mathbf{S}]$ can be obtained and how the posterior predictive distribution of the animal's continuous path $\tilde{\mathbf{S}}$ can be approximated using importance sampling. We will refer to the posterior predictive path distribution as $[\tilde{\mathbf{S}}|\mathbf{S}]$, where the bracket notation '$[\cdot]$' denotes a probability distribution.

2.2.2 Continuous-Time Discrete-Space Movement Model

We now consider a transformation of the animal's path $\tilde{\mathbf{S}}$ to a discrete (gridded) space. Let the study area be defined as a graph $(\mathbf{G}, \mathbf{A})$ of M locations $\mathbf{G} = (G_1, G_2, \ldots, G_M)$ connected by "edges" $\boldsymbol{\Lambda} = \{\lambda_{ij} : i \sim j, i = 1, \ldots, M\}$ where $i \sim j$ means that the nodes G_i and G_j are directly connected. For example, in a gridded space each grid cell is a node and the edges connect each grid cell to its first-order neighbors (e.g., cells that share an edge). In many studies, the spatial resolution of the grid cells in $\mathbf{G}$ will be determined by the resolution at which environmental covariates that may drive animal movement and selection are available.

A path realization $\tilde{\mathbf{S}}$ from the CTCRW model is continuous in time and space (Figure 1). If we consider a discrete, gridded space $\mathbf{G}$, then the continuous-time, continuous-space path $\tilde{\mathbf{S}}$ is represented by a continuous-time, discrete-space path $(\mathbf{g}, \boldsymbol{\tau})$ consisting of a sequence of grid cells $\mathbf{g} = (G_{i_1}, G_{i_2}, \ldots, G_{i_T})$ transversed by the animal's continuous-space path and the residence times $\boldsymbol{\tau} = (\tau_1, \tau_2, \ldots, \tau_T)$ in each grid cell.

2.2.3 Random Walk Model

The discrete-space representation $(\mathbf{g}, \boldsymbol{\tau})$ of the movement path allows us to use standard discrete-space random walk models to make inference about possible drivers of movement.

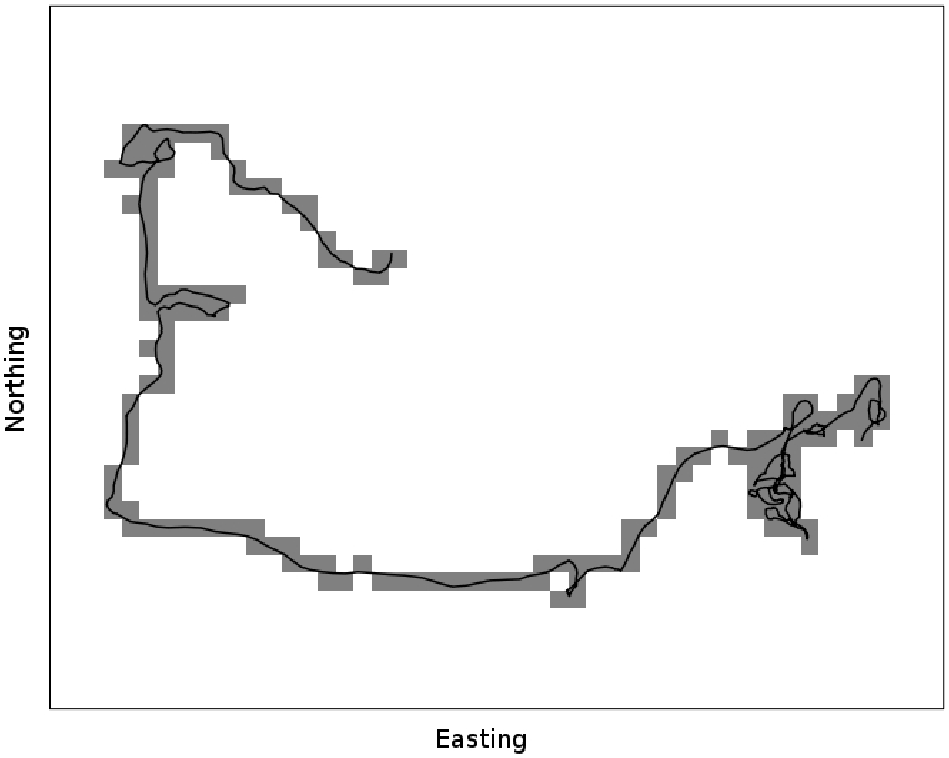

Figure 2.1: Continuous-time continuous-space and continuous-time discrete-space representations of an animal's movement path.

While we will relax this assumption later to account for temporal autocorrelation in movement behavior, we initially assume that the the t^{th} observation (G_{i_t}, τ_t) in the sequence is independent of all other observations in the sequence. Under this assumption, the likelihood of the sequence of transitions $\{(G_{i_t} \rightarrow G_{i_{t+1}}, \tau_t), t = 1, 2, \ldots, T\}$ is the product of the likelihoods of each individual observation. We will focus on modeling each transition $(G_{i_t} \rightarrow G_{i_{t+1}}, \tau_t)$.

If an animal is in cell G_{i_t} at time t, then define the Poisson rate of transition from cell G_{i_t} to a neighboring cell G_{j_t} at time t as

$$\lambda_{i_t j_t}(\boldsymbol{\beta}) = \exp\{\mathbf{x}'_{i_t j_t}\boldsymbol{\beta}\} \tag{2.2}$$

where $\mathbf{x}_{i_t j_t}$ is a vector containing covariates related to drivers of movement specific to cells G_{i_t} and G_{j_t}, and $\boldsymbol{\beta}$ is a vector of parameters that define how each of the covariates in $\mathbf{x}_{i_t j_t}$ are correlated with animal movement. The total transition rate λ_{i_t} from cell G_{i_t} is the sum of the transition rates to all neighboring cells: $\lambda_{i_t}(\boldsymbol{\beta}) = \sum_{j_t \sim i_t} \lambda_{i_t j_t}(\boldsymbol{\beta})$ and the time τ_t that the animal resides in cell G_{i_t} is exponentially distributed with rate parameter equal to the total transition rate $\lambda_{i_t}(\boldsymbol{\beta})$:

$$[\tau_t|\boldsymbol{\beta}] = \lambda_{i_t}(\boldsymbol{\beta}) \exp\left\{-\tau_t \lambda_{i_t}(\boldsymbol{\beta})\right\}. \tag{2.3}$$

When the animal transitions from cell G_{i_t} to one of its neighbors, the probability of transitioning to cell G_{k_t}, an event we denote as $G_{i_t} \to G_{k_t}$, follows a multinomial distribution with probability proportional to the transition rate $\lambda_{i_t k_t}$ to cell G_{k_t}:

$$[G_{i_t} \to G_{k_t}|\boldsymbol{\beta}] = \frac{\lambda_{i_t k_t}(\boldsymbol{\beta})}{\sum_{j_t \sim i_t} \lambda_{i_t j_t}(\boldsymbol{\beta})} = \frac{\lambda_{i_t k_t}(\boldsymbol{\beta})}{\lambda_{i_t}(\boldsymbol{\beta})}. \tag{2.4}$$

Under this formulation, the residence time and eventual destination are independent events, and the likelihood of the observation $(G_{i_t} \to G_{k_t}, \tau_t)$ is the product of the likelihoods of its parts:

$$\begin{aligned}
[G_{i_t} \to G_{k_t}, \tau_t|\boldsymbol{\beta}] &= \frac{\lambda_{i_t k_t}(\boldsymbol{\beta})}{\lambda_{i_t}(\boldsymbol{\beta})} \cdot \lambda_{i_t}(\boldsymbol{\beta}) \exp\left\{-\tau \lambda_{i_t}(\boldsymbol{\beta})\right\} \\
&= \lambda_{i_t k_t}(\boldsymbol{\beta}) \exp\left\{-\tau_t \lambda_{i_t}(\boldsymbol{\beta})\right\}. \qquad (2.5)
\end{aligned}$$

The transformation of the movement path from quasi-continuous space to discrete space results in a compression of the data to a temporal scale that is relevant to the resolution of the environmental covariates that may be driving movement and selection. For example, if an animal is moving slowly relative to the time it takes to traverse a grid cell in $\mathbf{G}$, then the quasi-continuous path $\tilde{\mathbf{S}}$ may contain a long sequence of locations within one grid cell. Under the discrete-space, discrete-time dynamic occupancy approach of Hooten et al. (2010), each discrete-time location is modeled as arising from a multinomial distribution reflecting transition probabilities from the animal's location at the previous time. If the animal is in cell $G_{i_{t-1}}$ at time $t-1$, then define the probability of transitioning to the j^{th} cell at the t^{th} time step as P_{ij_t} and the probability of remaining in cell i as P_{ii_t}. Hooten et al. (2010) recommend choosing a temporal discretization Δt of the quasi-continuous movement path fine enough to ensure that the animal remains in each cell for a number of time steps before transitioning to a neighboring cell. For sufficiently small Δt, discrete-time transition probabilities are approximated by $P_{ij_t} \approx \lambda_{i_t j_t} \cdot \Delta t$ and $P_{ii_t} \approx 1 - \lambda_{i_t} \cdot \Delta t$. Under this model, the probability of the animal remaining in cell G_i for time equal to τ_t is

$$\prod_{t=1}^{\tau_t/(\Delta t)} P_{ii_t} = P_{ii}^{\tau_t/\Delta t} = \left(1 - \lambda_{i_t} \cdot \Delta t\right)^{\tau_t/\Delta t}.$$

Letting $\Delta t \to 0$ results in

$$\lim_{\Delta t \to 0} \left(1 - \lambda_{i_t} \cdot \Delta t\right)^{\tau_t/\Delta t} = \exp\left\{-\tau_t \lambda_{i_t}\right\}. \tag{2.6}$$

Likewise, taking the limit as $\Delta t \to 0$ and using L'Hopital's rule, the probability of transitioning from cell G_i to G_k, given that the animal is transitioning to some neighboring cell, is

$$\lim_{\Delta t \to 0} \frac{P_{ik_t}}{\sum_j P_{ij_t}} = \lim_{\Delta t \to 0} \frac{\lambda_{i_t k_t} \cdot \Delta t}{\lambda_{i_t} \cdot \Delta t} = \frac{\lambda_{i_t k_t}}{\lambda_{i_t}} \tag{2.7}$$

and (2.5) is obtained by multiplying the right hand sides of (2.6) and (2.7). Thus the CTDS specification could be obtained by using the sufficient statistics $(\tau_t, \{\lambda_{i_t j_t}\})$ of the discrete-time, discrete-space approach of Hooten et al. (2010) in the limiting case as $\Delta t \to 0$. This data compression is especially relevant for telemetry data, in which observation windows can span years or even decades for some animals.

We note that the transition probabilities in (2.4) are similar in form to step selection functions (e.g., Boyce et al., 2002) in multinomial logit discrete choice models for movement data. The key distinction between the step selection function approach and the approach of Hooten et al. (2010) (and by extension the approach we present) is the imputation of the quasi-continuous path between telemetry locations. This requires modeling the continuous path and accepting the assumptions of the CTCRW movement model of Johnson et al. (2008). If these assumptions are justified, the continuous path distribution allows us to examine movement and resource selection between telemetry locations, providing a more complete picture of an animal's response to landscape features and other potential drivers of movement.

2.2.4 Latent Variable Representation

We now introduce a latent variable representation of the transition process that is equivalent to (2.5), but allows for inference within a standard generalized linear modeling framework. For each j_t such that $i_t \sim j_t$, define $z_{i_t j_t}$ as

$$z_{i_t j_t} = \begin{cases} 1 & , \ G_{i_t} \to G_{j_t} \\ 0 & , \ \text{o.w.} \end{cases}$$

and let

$$[z_{i_t j_t}, \tau_t | \boldsymbol{\beta}] \propto \lambda_{i_t j_t}^{z_{i_t j_t}} \exp\left\{-\tau_t \lambda_{i_t j_t}(\boldsymbol{\beta})\right\}. \tag{2.8}$$

Then the product of $[z_{i_t j_t}, \tau_t | \boldsymbol{\beta}]$ over all j_t such that $i_t \sim j_t$ is proportional to the likelihood (2.5) of the observed transition:

$$\begin{aligned}\prod_{j_t : i_t \sim j_t} [z_{i_t j_t}, \tau_t | \boldsymbol{\beta}] \propto \prod_{j_t : i_t \sim j_t} \lambda_{i_t j_t}^{z_{i_t j_t}} \exp\left\{-\tau_t \lambda_{i_t j_t}(\boldsymbol{\beta})\right\} \\ = \lambda_{i_t k_t}(\boldsymbol{\beta}) \exp\left\{-\tau_t \lambda_{i_t}(\boldsymbol{\beta})\right\} \ \text{, where } G_{i_t} \to G_{k_t} \\ = [G_{i_t} \to G_{k_t}, \tau_t | \boldsymbol{\beta}]\end{aligned}$$

The benefit of this latent variable representation is that the likelihood of $z_{i_t j_t}, \tau_t | \boldsymbol{\beta}$ in (2.8) is equivalent to the kernel of the likelihood in a Poisson regression with the canonical log link, where $z_{i_t j_t}$ are the observations and $\log(\tau_t)$ is an offset or exposure term. The likelihood of the entire continuous-time, discrete-space path $(\mathbf{g}, \boldsymbol{\tau})$ can be written as:

$$[\mathbf{g}, \boldsymbol{\tau} | \boldsymbol{\beta}] = [\mathbf{Z}, \boldsymbol{\tau} | \boldsymbol{\beta}] \propto \prod_{t=1}^{T} \prod_{i_t \sim j_t} \left[\lambda_{i_t j_t}^{z_{i_t j_t}}(\boldsymbol{\beta}) \exp\{-\tau_t \lambda_{i_t j_t}(\boldsymbol{\beta})\}\right] \tag{2.9}$$

where $\mathbf{Z} = (\mathbf{z}_1, \ldots, \mathbf{z}_T)'$ is a vector containing the latent variables $\mathbf{z}_i = (z_{i_1}, z_{i_2}, \ldots, z_{i_K})'$ for each grid cell in the discrete-space path.

2.3 Posterior Predictive Inference

Inference can be made on $\boldsymbol{\beta}$ in (2.9) using standard Poisson GLM approaches (e.g., maximum likelihood), if the continuous path $\tilde{\mathbf{S}}$ is known. In practice, only the posterior predictive distribution $[\tilde{\mathbf{S}} | \mathbf{S}]$ is known. In this chapter, we consider posterior predictive inference on $\boldsymbol{\beta}$.

Under a Bayesian framework, we could specify a Gaussian prior on $\boldsymbol{\beta}$ such that

$$\boldsymbol{\beta} \sim N(\mathbf{0}, \boldsymbol{\Sigma}_\beta) \tag{2.10}$$

and then the posterior predictive distribution of $\boldsymbol{\beta}$ conditioned only on the telemetry data $\mathbf{S}$ is given by

$$[\boldsymbol{\beta}|\mathbf{S}] = \int_{\tilde{\mathcal{S}}} [\boldsymbol{\beta}|\tilde{\mathbf{S}}][\tilde{\mathbf{S}}|\mathbf{S}]d\tilde{\mathbf{S}}. \tag{2.11}$$

We note that (2.11) implies a two-step procedure for inference in which the path distribution $[\tilde{\mathbf{S}}|\mathbf{S}]$ is first found and then used to make inference on $\boldsymbol{\beta}$ through (2.2). Hooten et al. (2010) and Hanks et al. (2011) use composition sampling to obtain samples from the posterior predictive distribution $[\boldsymbol{\beta}|\mathbf{S}]$ in (2.2) by sampling iteratively from $[\boldsymbol{\beta}|\tilde{\mathbf{S}}]$ and $[\tilde{\mathbf{S}}|\mathbf{S}]$. Sampling from $[\tilde{\mathbf{S}}|\mathbf{S}]$ can be accomplished using the “crawl” package (Johnson, 2011) in R (R Core Team, 2013), while sampling from $[\boldsymbol{\beta}|\tilde{\mathbf{S}}]$ can be accomplished using Metropolis-Hastings steps in the MCMC algorithm using the likelihood (2.4) and prior (2.5).

2.3.1 Multiple Imputation

As an alternative to fully Bayesian posterior predictive inference, we also consider approximating $[\boldsymbol{\beta}|\mathbf{S}]$ by treating the unobserved continuous path $\tilde{\mathbf{S}}$ as missing data. Inference on $\boldsymbol{\beta}$ can then be accomplished using multiple imputation (Rubin, 1987). We motivate multiple imputation as posterior predictive inference on the imputation distribution within a Bayesian framework. Our treatment is similar to that of Rubin (1987) and Rubin (1996).

In the multiple imputation literature, the posterior predictive path distribution $[\tilde{\mathbf{S}}|\mathbf{S}]$ is called the imputation distribution. The imputation distribution is typically specified by the modeler as a statistical model for the missing data $\tilde{\mathbf{S}}$ conditioned on the observed data $\mathbf{S}$. We will use the CTCRW model for animal movement to define an imputation distribution, the posterior predictive path distribution $[\tilde{\mathbf{S}}|\mathbf{S}]$. The CTCRW model has been successfully applied to studies of aquatic (Johnson et al., 2008) and terrestrial (Hooten et al., 2010) animals, and can represent a wide range of behavior.

Under the multiple imputation framework, the distribution $[\boldsymbol{\beta}|\mathbf{S}]$ is assumed to be asymptotically Gaussian. This assumption holds under the conditions that the joint posterior is unimodal (see e.g., Chapter 4 of Gelman et al., 2004, for details). This distribution can then be approximated using only the posterior predictive mean and variance, which can be obtained using conditional mean and variance formulae

$$
\begin{aligned}
E(\boldsymbol{\beta}|\mathbf{S}) &\approx \int_{\beta} \boldsymbol{\beta} \int_{\tilde{S}} [\boldsymbol{\beta}|\tilde{\mathbf{S}}][\tilde{\mathbf{S}}|\mathbf{S}] d\tilde{\mathbf{S}} d\boldsymbol{\beta} \\
&= \int_{\tilde{S}} \left(\int_{\beta} \boldsymbol{\beta}[\boldsymbol{\beta}|\tilde{\mathbf{S}}] d\boldsymbol{\beta} \right) [\tilde{\mathbf{S}}|\mathbf{S}] d\tilde{\mathbf{S}} \\
&= E_{\tilde{\mathbf{S}}|\mathbf{S}} \left(E(\boldsymbol{\beta}|\tilde{\mathbf{S}}) \right) \qquad (2.12)
\end{aligned}
$$

and

$$
\mathrm{Var}(\boldsymbol{\beta}|\mathbf{S}) \approx E_{\tilde{\mathbf{S}}|\mathbf{S}} \left(\mathrm{Var}(\boldsymbol{\beta}|\tilde{\mathbf{S}}) \right) + \mathrm{Var}_{\tilde{\mathbf{S}}|\mathbf{S}} \left(E(\boldsymbol{\beta}|\tilde{\mathbf{S}}) \right). \qquad (2.13)
$$

If we condition on $\tilde{\mathbf{S}}$, then the posterior distribution $[\boldsymbol{\beta}|\tilde{\mathbf{S}}]$ converges asymptotically to the sampling distribution of the maximum likelihood estimate (MLE) of $\boldsymbol{\beta}$ under the likelihood $[\tilde{\mathbf{S}}|\boldsymbol{\beta}]$, and we can approximate $[\boldsymbol{\beta}|\tilde{\mathbf{S}}]$ by obtaining the asymptotic sampling distribution of the MLE. This allows us to use standard maximum likelihood approaches for inference, which can be much more computationally efficient than their Bayesian counterparts for this class of models.

The multiple imputation estimate $\hat{\boldsymbol{\beta}}_{MI}$, and its sampling variance, are typically obtained by approximating the integrals in (2.12) and (2.13) using a finite sample from the imputation distribution. The procedure can be summarized as follows:

1. Draw K different realizations (imputations) $\tilde{\mathbf{S}}^{(k)} \sim [\tilde{\mathbf{S}}|\mathbf{S}]$ from the quasi-continuous path distribution (imputation distribution). This can be accomplished using the 'crawl' R-package (Johnson, 2011).

2. Transform each continuous path $\tilde{\mathbf{S}}^{(k)}$ into a CTDS path $(\mathbf{g}^{(k)}, \boldsymbol{\tau}^{(k)})$.

3. For each realization, find the MLE $\hat{\boldsymbol{\beta}}^{(k)}$ and asymptotic variance $\mathrm{Var}(\hat{\boldsymbol{\beta}}^{(k)})$ of the estimate under the likelihood $[(\mathbf{g}^{(k)}, \boldsymbol{\tau}^{(k)})|\boldsymbol{\beta}]$ in (2.9).

4. Combine results from different imputations using finite sample versions of the conditional expectation (2.12) and variance (2.13) results:

$$\hat{\boldsymbol{\beta}}_{MI} = \frac{1}{K}\sum_{k=1}^{K}\hat{\boldsymbol{\beta}}^{(k)} \tag{2.14}$$

$$\mathrm{Var}(\hat{\boldsymbol{\beta}}_{MI}) = \frac{1}{K}\sum_{k=1}^{K}\mathrm{Var}(\hat{\boldsymbol{\beta}}^{(k)}) + \frac{1}{K}\sum_{k=1}^{K}\left(\hat{\boldsymbol{\beta}}^{(k)} - \hat{\boldsymbol{\beta}}_{MI}\right)^2. \tag{2.15}$$

Equations (2.14) and (2.15) are the commonly used combining rules (Rubin, 1987) for multiple imputation estimators. This provides a computationally efficient framework for the statistical analysis of potential drivers of movement within the multiple imputation framework.

2.4 Time-Varying Behavior and Shrinkage Estimation

In this section we describe how covariate effects can be allowed to vary over time using a varying coefficient model, and how variable selection can be accomplished through regularization.

2.4.1 Changing Behavior Over Time

Animal behavior and response to drivers of movement can change significantly over time. These changes can be driven by external factors such as changing seasons (e.g., Grovenburg et al., 2009) or predator/prey interactions (e.g., Lima, 2002), or by internal factors such as internal energy levels (e.g., Nathan et al., 2008). The most common approach to modeling time-varying behavior in animal movement is through state-switching, typically within a Bayesian framework (e.g., Morales et al., 2004; Jonsen et al., 2005; Getz and Saltz, 2008; Nathan et al., 2008; Forester et al., 2009; Gurarie et al., 2009; Merrill et al., 2010). Often, the animal is assumed to exhibit a number of behavioral states, each characterized by a distinct pattern of movement or response to drivers of movement. The number of states can be either known and specified in advance (e.g., Morales et al., 2004; Jonsen et al., 2005) or allowed to be random (e.g., Hanks et al., 2011; McClintock et al., 2012).

State-switching models are an intuitive approach to modeling changing behavior over time, but there are limits to the complexity that can be modeled using this approach. Allowing the number of states to be unknown and random requires a Bayesian approach with a changing parameter space. This is typically implemented using reversible-jump MCMC methods (e.g., Green, 1995; McClintock et al., 2012; Hanks et al., 2011), which are computationally expensive and can be difficult to tune. Our approach is to use a computationally efficient GLM (2.9) to analyze parameters related to drivers of animal movement. Instead of using the common state-space approach, we employ varying coefficient models (e.g., Hastie and Tibshirani, 1993) to model time-varying behavior in animal movement.

For simplicity in notation, consider the case where there is only one covariate x in the model (2.2) and no intercept term. The model for the Poisson transition rate (2.2) will typically contain an intercept term and multiple covariates $\{x\}$, and the varying coefficient approach we present generalizes easily to this case. In a time-varying coefficient model, we allow the parameter $\beta(t)$ to vary over time in a functional (continuous) fashion. The transition rate (2.2) then becomes

$$\lambda_{i_t j_t}(\beta(t)) = \exp\left\{x_{i_t j_t}\beta(t)\right\},$$

where t is the time of the observation and x_{ij} is the value of the covariate related to the Poisson rate of moving from cell i to cell j. The functional regressor $\beta(t)$ is typically modeled as a linear combination of n_{spl} spline basis functions $\{\boldsymbol{\phi}_k(t), k = 1, \ldots, n_{spl}\}$

$$\beta(t) = \sum_{k=1}^{n_{spl}} \alpha_k \boldsymbol{\phi}_k(t).$$

B-spline basis functions are among the most widely used choices for $\{\phi_k(t)\}$, and are appropriate in most cases.

Under this varying coefficient specification, (2.2) can be rewritten as

$$\begin{aligned}\lambda_{i_t j_t} &= \exp\left\{x_{i_t j_t}\beta(t)\right\} \\ &= \exp\left\{x_{i_t j_t}\sum_{k=1}^{n_{spl}}\alpha_k \boldsymbol{\phi}_k(t)\right\} \\ &= \exp\left\{\boldsymbol{\psi}'_{i_t j_t}\boldsymbol{\alpha}\right\}, \qquad (2.16)\end{aligned}$$

where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_{n_{spl}})'$ and $\boldsymbol{\psi}_{i_t j_t} = x_{i_t j_t} \cdot (\phi_1(t), \ldots, \phi_{n_{spl}}(t))'$. The result is that the varying coefficient model can be represented by a GLM with a modified design matrix. This specification provides a flexible framework for allowing the effect of a driver of movement (x) to vary over time that is computationally efficient and simple to implement using standard GLM software.

2.4.2 Regularization

The model we have specified in (2.9) is likely to be overparameterized, especially if we utilize a varying coefficient model (2.16). Animal movement behavior is complex, and a typical study could entail a large number of potential drivers of movement, but an animal's response to each of those drivers of movement is likely to change over time, with only a few drivers being relevant at any one time. Under these assumptions, many of the parameters α_k in (2.16) are likely to be very small or zero. Multicollinearity is also a potential problem, as many potential drivers of movement could be correlated with each other.

We propose a shrinkage estimator of $\boldsymbol{\alpha}$ using a lasso penalty (Tibshirani, 1996). The typical maximum likelihood estimate of $\boldsymbol{\alpha}$ is obtained by maximizing the likelihood $[\mathbf{Z}, \boldsymbol{\tau}|\boldsymbol{\alpha}]$ from (2.9), or equivalently by maximizing the log-likelihood $\log[\mathbf{Z}, \boldsymbol{\tau}|\boldsymbol{\alpha}]$. The lasso estimate is obtained by maximizing the penalized log-likelihood, where the penalty is proportional to the sum of the absolute values of the regression parameters $\{\alpha_k\}$:

$$\hat{\boldsymbol{\alpha}}_{\text{lasso}} = \max_{\boldsymbol{\alpha}} \left\{ \log[\mathbf{Z}, \boldsymbol{\tau}|\boldsymbol{\alpha}] - \gamma \sum_{k=1}^{K} |\alpha_k| \right\}. \tag{2.17}$$

As the tuning parameter γ increases, the absolute values of the regression parameters $\{\alpha_k\}$ are "shrunk" to zero, with the parameters that best describe the variation in the data being shrunk more slowly than parameters that do not. Cross-validation is typically used to set the tuning parameter γ at a level that optimizes the model's predictive power.

Shrinkage approaches such as the lasso are well developed for GLMs, and computationally-efficient methods are available for fitting GLMs to data (e.g., Friedman et al., 2010). Recent work has also applied the lasso to multiple imputation estimators (e.g., Chen and Wang, 2011). The main challenge in applying the lasso to multiple imputation is that a parameter may be shrunk to zero in the analysis of one imputation but not in the analysis of another. The solution is to use a group lasso (Yuan and Lin, 2006), in which a group of parameters is constrained to either all equal zero or all be non-zero together. In the case of multiple imputation, we consider the joint analysis of all imputations, and constrain the set of $\{\alpha_p^{(k)}, k = 1, \ldots, K\}$, where p indexes the parameters in the model and k indexes the imputations, to either all equal zero or all be non-zero together. This group lasso sets the requirement that a parameter must be zero for all imputations, or non-zero for all imputations.

One approach to this group lasso is the likelihood-based stacked lasso estimate of Chen and Wang (2011). In this estimate, instead of computing the lasso estimate $\boldsymbol{\alpha}_{\text{lasso}}$ for each imputation individually, and then combining the results using (2.14) and (2.15), the imputed data from all estimates are "stacked" together and a lasso estimate is obtained for the combined data. This stacking together of imputations is very similar to the data cloning approach of Lele et al. (2011) in which the data are replicated (cloned) many times and analyzed as if each replicate were collected independently. The resulting estimates of parameter uncertainty are then adjusted to account for the replication. We note that this likelihood-based stacked lasso approach does not allow for a straightforward estimation of the variance of $\boldsymbol{\alpha}_{\text{lasso}}$, though we can accomplish this by conducting a full Bayesian analysis using the Bayesian lasso (Park and Casella, 2008).

Thus, we consider specifying a shrinkage prior distribution on $\boldsymbol{\alpha}$ such that the posterior mode of $\boldsymbol{\alpha}|\mathbf{S}$ is identical to the lasso estimate (2.17). Instead of the Gaussian prior in (2.10), we follow Park and Casella (2008) and consider a hierarchical prior specification:

$$\alpha_k|\sigma_k^2 \sim N(0, \sigma_k^2) \ , \ k = 1, \ldots, K \tag{2.18}$$

where the prior on σ_k^2 is a function of the shrinkage parameter γ:

$$[\sigma_k^2|\gamma^2] \propto \gamma^2 \exp\{-\gamma^2\sigma_k^2/2\} \ , \ k = 1, \ldots, K. \tag{2.19}$$

Then, marginalizing over the σ_k^2 gives a Laplace prior distribution (Park and Casella, 2008) on α conditioned only on γ:

$$\begin{aligned}
[\alpha_k|\gamma] &= \int_0^\infty [\alpha_k|\sigma_k^2][\sigma_k^2|\gamma]\mathrm{d}\sigma_k^2 \\
&\propto \int_0^\infty \frac{1}{\sqrt{2\pi\sigma_k^2}} \exp\{-\alpha_k^2/(2\sigma_k^2)\}\gamma^2 \exp\{-\gamma^2\sigma_k^2/2\}\mathrm{d}\sigma_k^2 \\
&= \frac{\gamma}{2} \exp\{-\gamma|\alpha_k|\}
\end{aligned}$$

where the last step uses the representation of the Laplace distribution as a scale mixture of Gaussian random variables with exponential mixing density (e.g., Park and Casella, 2008). Maximizing the resulting log-posterior predictive distribution for $\boldsymbol{\alpha}$ gives us the lasso estimate (2.17).

The hyperparameter γ controls the amount of shrinkage in the Bayesian lasso. While a prior distribution could be assigned to γ, we take an empirical approach and estimate γ

using cross-validation in the penalized likelihood approach (2.17) to the lasso. This estimate can then be used to set the value of the hyperparameter γ in the Bayesian lasso analysis.

2.5 Drivers of Animal Movement

We now provide some examples showing how a range of hypothesized drivers of movement could be modeled within the CTDS framework. Following Hooten et al. (2010), we consider two distinct categories for drivers of movement from cell G_i to cell G_j: location-based drivers ($\{p_{ki}, k = 1, 2, \ldots, K\}$) which are determined only by the characteristics of cell G_i, and directional drivers ($\{q_{lij}, l = 1, 2, \ldots, L\}$) which vary with direction of movement. Under a time-varying coefficient model for each driver, the transition rate (2.2) from cell G_i to cell G_j is

$$\lambda_{ij}\left(\boldsymbol{\beta}(t)\right) = \exp\left\{\beta_0(t) + \sum_{k=1}^{K} p_{ki}\beta_k(t) + \sum_{l=1}^{L} q_{lij}\beta_l(t)\right\} \tag{2.20}$$

where $\beta_0(t)$ is a time-varying intercept term, $\{\beta_k(t)\}$ are time-varying effects related to location-based drivers of movement, and $\{\beta_l(t)\}$ are time-varying effects related to directional drivers of movement. We consider both location-based and directional drivers in what follows.

2.5.1 Location-Based Drivers of Movement

Hooten et al. (2010) denote static, non-directional drivers of movement as location-based drivers of movement. Location-based drivers of movement can be used to examine differences in animal movement rates that can be explained by the environment an animal resides in. For example, if the animal is in a patch of highly desirable terrain, surrounded by less desirable terrain, a location-based driver of movement could be used to model the animal's propensity to stay in the desirable patch and move quickly through undesirable terrain. In the CTDS context, location-based drivers would be covariates dependent only on the characteristics of the cell where the animal is currently located. Large positive (negative) values of the corresponding $\beta_k(t)$ would indicate that the animal tends to transition quickly (slowly) from a cell containing the cover type in question.

2.5.2 Directional Drivers of Movement

In contrast to location-based drivers, which describe the effect that the local environment in which the animal resides has on movement rates, directional drivers of movement (Brillinger et al., 2001; Hooten et al., 2010; Hanks et al., 2011) capture directional selection by the individual.

A directional driver of movement is defined by a vector which points toward (or away) from something that is hypothesized to attract (or repel) the animal in question. Let $\mathbf{v}_l$ be the vector corresponding to the l^{th} directional driver of movement. In the CTDS model for animal movement, the animal can only transition from cell G_i to one of its neighbors $G_j : j \sim i$. Let $\mathbf{w}_{ij}$ be a unit vector pointing from the center of cell G_i in the direction of the center of cell G_j. Then the covariate q_{lij} relating the l^{th} directional driver of movement to the transition rate from cell G_i to cell G_j is the dot (or inner) product of $\mathbf{v}_l$ and $\mathbf{w}_{ij}$:

$$q_{lij} = \mathbf{v}_l' \mathbf{w}_{ij}.$$

Then p_{lij} will be positive when $\mathbf{v}_l$ points nearly in the direction of cell G_j, negative when $\mathbf{v}_l$ points directly away from cell G_j, and zero if $\mathbf{v}_l$ is perpendicular to the direction from cell G_i to cell G_j.

2.5.3 Examples

We now provide several examples of drivers of movement to illustrate the range of effects that can be modeled using this framework.

Overall Movement Rate

The intercept term $\beta_0(t)$ in (2.20) can be seen as a driver of movement in which $p_{0i} = 1$ for every cell $G_i \in \mathbf{G}$. This intercept term controls the animal's overall rate of transition from any cell, and thus models the animal's overall movement rate. Allowing the intercept parameter $\beta_0(t)$ to vary over time could reveal changes in activity levels over time. For example, we might expect $\beta_0(t)$ to be larger at night for nocturnal species and smaller during the day.

Movement Response to Land Cover Types

Indicator variables could be used to examine how animal movement differs between different landscape cover types (e.g., forest vs. plains) by setting $p_{ki} = 1$ for each cell G_i that is classified as containing the k^{th} cover type. As in the case of the intercept, allowing the parameter $\beta_k(t)$ related to the k^{th} cover type to vary over time can reveal variation in an animal's movement pattern through the cover type. For example, an animal may move quickly through open terrain during the day, but may move more slowly through the same terrain at night.

Environmental Gradients

Animals may use environmental gradients for navigation. For example, a mule deer might move predominantly in the direction of increasing elevation during a spring migration (e.g., Hooten et al., 2010), or a seal might follow gradients in sea surface temperature to navigate toward land (e.g., Hanks et al., 2011). Such effects can be modeled by including a directional driver of movement in (2.20) defined by a gradient vector $\mathbf{v}_l$ which points from the center of cell G_i in the direction of steepest increase in the covariate x_l. Positive values of β_l indicate that the animal moves generally towards cells with higher values of x_l, while negative values of β_l indicate that the animal moves generally towards cells with lower values of x_l.

Activity Centers

Many animals exhibit movement patterns that are centered on a location in space. This central location may be temporary, such as a kill site for a predator (e.g., Knopff et al., 2009), or more permanent, such as a den for a central place forager (e.g., Hanks et al., 2011; McClintock et al., 2012). The relatively new class of spatial capture-recapture models (e.g., Royle and Young, 2008) model detection probability as a decreasing function of distance from a central location (e.g., the "center" of an animal's home range). In the context of the CTDS model, movement around an activity center can be modeled by including a directional driver of movement in (2.20) defined by a vector $\mathbf{v}_l$ which points from the center of cell G_i to the location of the activity center. Then a positive value for β_l would indicate that the animal is generally drawn toward this activity center. If the activity center is considered to be temporary (such as a kill site for a predator), then a time-varying coefficient model should be used. The variable selection obtained through the lasso estimate can indicate the range of time in which the animal's movement is centered around the activity center. If the activity center is considered to be permanent through the duration of the study, a varying coefficient model may not be needed.

Under the likelihood-based specification of the CTDS model for animal movement, it is necessary to specify the locations of all hypothesized activity centers before proceeding with the analysis. In Section 2.6, we show an example of the specification of hypothesized activity centers (potential kill sites for mountain lions) using the original telemetry data. Under a Bayesian formulation of the CTDS model, the location of hypothesized activity centers could be random, and inference could be made on their locations jointly with inference on the movement parameters, similar to what is done in spatial capture-recapture models (e.g., Royle and Young, 2008).

Conspecific Interaction

An animal's movement patterns can be greatly affected by interaction with conspecifics. For example, one animal could follow the trail left by another animal, two animals could avoid one another by changing course when they become close enough to sense the other animal, or a pair of animals could maintain proximity as they move together across the landscape. While there are many possible approaches to modeling such dependence in behavior, we choose to model each of these interactions through the inclusion of directional effects in the CTDS modeling framework. For example, a directional driver could be included in the movement model for one animal that is defined by a vector pointing to the current location of another animal to examine whether the animal being modeled is attracted to ($\beta_l > 0$) or avoids ($\beta_l < 0$) the conspecific.

Directional Persistence

The CTCRW model of Johnson et al. (2008) is based on a correlated random walk model for velocity that allows for directional persistence in animal movement. So far, we have assumed that each discrete movement step in our CTDS model is independent, but this assumption is not met if the animal exhibits any directional persistence. To account for directional persistence in the CTDS approach, we use an autoregressive approach by including a directional driver of movement at each discrete movement step that is defined by a vector pointing in the direction of the previous move. For example, if the animal moved west in the previous discrete movement step, then the autoregressive vector for the next step points west as well. Positive values of the β related to this directional driver of movement indicate that the animal is likely to maintain its direction of movement over time.

2.5.4 Spatial and Temporal Scale

The choice of scale for a study can greatly influence results (e.g., Boyce, 2006). When speaking of the scale of a study, one could look at the grain, or resolution, at which the process is modeled, or the extent (coverage) over which the process is modeled. The spatial and temporal extent of a study of animal movement are determined by the telemetry data and the posterior predictive path distribution $[\tilde{\mathbf{S}}|\mathbf{S}]$. However, when implementing the CTDS approach, the researcher must make three choices pertaining to the grain or resolution: (1) the temporal scale at which the CTCRW movement path of the animal is sampled, (2) the spatial scale of the grid over which the discrete-space movement will be modeled, and (3) the temporal scale of the varying coefficient model, which is determined by the number and resolution of spline knots in the spline basis expansion.

As the CTCRW model of Johnson et al. (2008) is a continuous-time model, we recommend sampling from the movement path at as fine an interval as is feasible. In practice, this will be limited by computational resources and the size of the study. The temporal resolution needs to be fine enough that realizations from the posterior predictive path distribution $[\tilde{\mathbf{S}}|\mathbf{S}]$ are quasi-continuous and adequately capture the residence time τ in each grid cell in the CTDS representation of the movement path.

The choice of spatial resolution of the raster grid on which the CTDS process occurs implicitly specifies a time scale at which the movement process is modeled. Coarser spatial resolution (larger grid cells) will correspond to longer residence times $\boldsymbol{\tau}$ in the CTDS model. The spatial resolution should be chosen so that the time scale at which an animal transitions from one grid cell to another is a time scale at which the animal in question can make choices about movement and resource selection. The time scale implicit in the choice of spatial resolution can be examined by plotting a histogram of the residence times in the CTDS representation of the movement path.

If the lasso penalty is used, then it is common to choose a saturated spline basis expansion in the varying coefficient model, where one spline knot is specified at each data point in time. If, for example, movement behavior is hypothesized to vary diurnally, and telemetry locations are obtained at 3 hour intervals throughout the day, then a set of spline knots corresponding to the eight times throughout the day at which location fixes are obtained would represent a saturated basis expansion. The lasso penalization will shrink this expansion to a more parsimonious model if such a model better fits the data. While a finer temporal resolution could be used, the posterior predictive path distribution is unlikely to show changes in behavior at time scales smaller than the time scale of the original data. Using a coarser temporal resolution will force $\beta(t)$ to be smooth. This would imply that changes in animal behavior are gradual and occur at time scales larger than the time scale of the data.

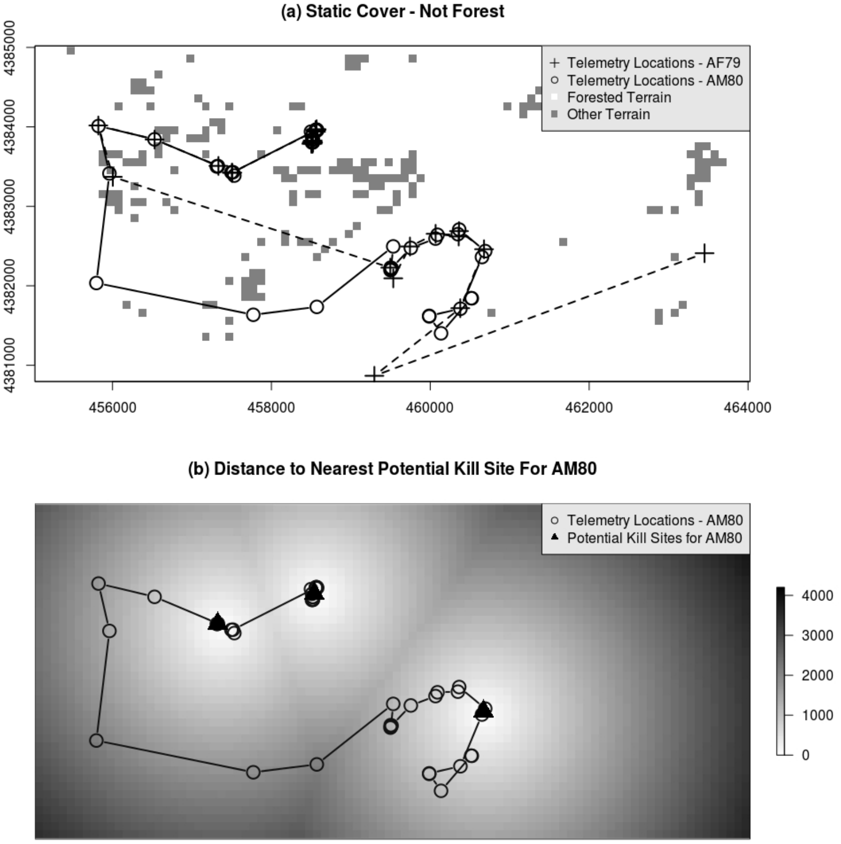
(a) Static Cover - Not Forest
Telemetry Locations - AF79
Telemetry Locations - AM80
Forested Terrain
Other Terrain
4381000
4382000
4383000
4384000
4385000
456000
458000
460000
462000
464000
(b) Distance to Nearest Potential Kill Site For AM80
Telemetry Locations - AM80
Potential Kill Sites for AM80
4000
3000
2000
1000
0

Figure 2.2: Telemetry data for a female mountain lion (AF79) and her male cub (AM80). A location-based covariate was defined by landcover that was not predominanty forested (a). Potential kill sites were identified, and a directional covariate defined by a vector pointing toward the closest kill site (b) was also used in the CTDS model.

2.6 Example: Mountain Lions in Colorado

We illustrate our CTDS random walk approach to modeling animal movement through a study of mountain lions (*Puma concolor*) in Colorado, USA. As part of a larger study, a female mountain lion, designated AF79, and her subadult cub, designated AM80, were fitted with global positioning system (GPS) collars set to transmit location data every 3 hours. We analyze the location data $\mathbf{S}$ from two weeks (14 days) of location information for these two animals (Figure 2.2).

We fit the CTCRW model of Johnson et al. (2008) to both animals' location data using the 'crawl' package (Johnson, 2011) in the R statistical computing environment (R Core Team, 2013). Ten imputations from the posterior distribution of the quasi-continuous path distribution $[\tilde{\mathbf{S}}|\mathbf{S}]$ were obtained at one minute intervals. The result is a quasi-continuous path at extremely fine temporal resolution for each imputation.

For covariate data, we used a landcover map of the state of Colorado created by the Colorado Vegetation Classification Project (http://ndis.nrel.colostate.edu/coveg/), which is a joint project of the Bureau of Land Management and the Colorado Division of Wildlife. The landcover map contained gridded landcover information at 100m square resolution. Figure 2.3 shows a histogram of the residence times $\boldsymbol{\tau}$ in each grid cell in the CTDS representation of the movement path of AM80. This gives some indication of the temporal scale of inference implied by our choice of spatial resolution. Increasing the grid cell size would result in inference at larger time scales (e.g., daily or monthly movements). The area traveled by the two animals in our study was predominantly forested. To assess how the animals' movement differed when in terrain other than forest, we created an indicator covariate where all forested grid cells were assigned a value of zero, and all cells containing other cover types, including developed land, bare ground, grassland, and shrubby terrain, were assigned a value of one (Figure 2.2a). This covariate was used as location-based covariate in the CTDS model.

For the subadult male AM80, we created a set of potential kill sites (PKS) by examining the original GPS location data (Figure 2.2). Knopff et al. (2009) classified a location as a PKS if two or more GPS locations were found within 200m of the site within a six day period. We added an additional constraint that at least one of the GPS locations be during nighttime hours (9 pm to 6 am) for the point to be classified a PKS. We then created a covariate raster layer containing the distance to the nearest PKS for each grid cell (Figure 2.2b). A directional covariate defined by a vector pointing towards the nearest PKS was included in the CTDS model.

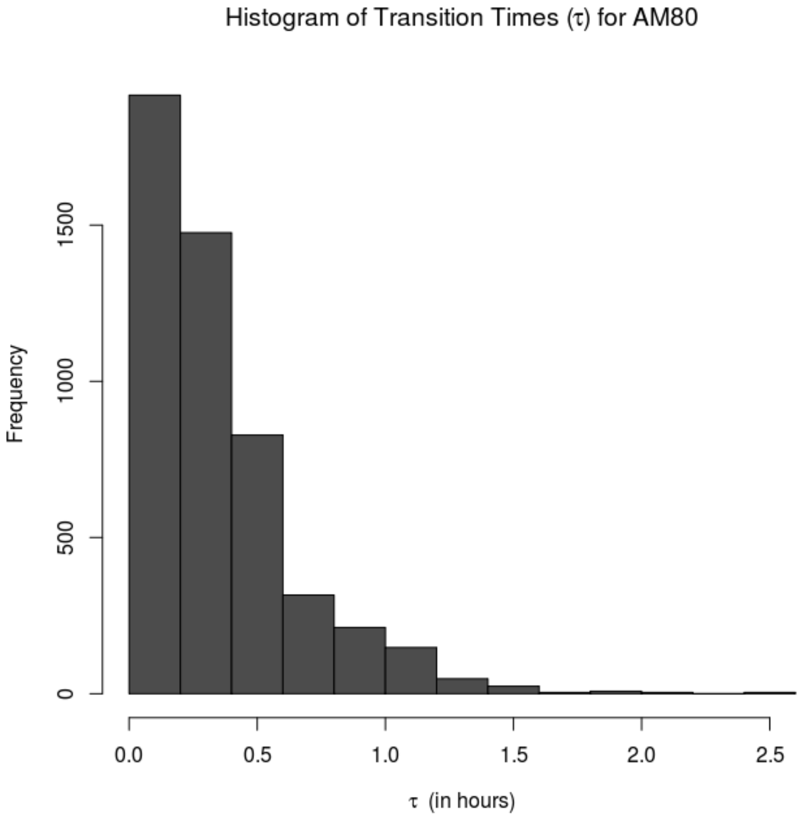

Figure 2.3: Histogram of residence times in each 100m square grid cell in the continuous-time discrete-space representation of the movement path of a male mountain lion (AM80).

To examine how the movement path of the mother AF79 affected the movement path of the cub AM80, we included a directional covariate in the CTDS model for AM80 defined by a vector pointing from the cub's location to the mother's location at each time point.

We also included a directional covariate pointing in the direction of the most recent movement at each time point. This covariate measures the strength of correlation between moves and thus the strength of the directional persistence shown by the animal's discrete-space movement path. As we are assuming an underlying correlated movement model (the CTCRW model of Johnson et al. (2008)), we expect the CTDS movement to be correlated in time as well.

We first compare a full Bayesian analysis of the path of AM80 using the CTDS model with the multiple imputation approximation to the posterior mean (2.12) and variance (2.13). For this analysis, we do not assume any time-varying behavior, but rather model the cub's mean response over time to the landscape, identified PKSs, and the movement path of AF79. We used a Markov chain Monte Carlo algorithm to draw 10,000 samples from the posterior predictive distribution of $\boldsymbol{\beta}|\mathbf{S}$ for AM80. We discarded the first 5,000 as burn-in and used the remaining samples to approximate the posterior predictive distribution. Histograms of the marginal distributions are shown in Figure 2.4. We then applied the multiple imputation approach to approximate the posterior distribution using the 10 quasi-continuous paths drawn from the path imputation distribution: $[\tilde{\mathbf{S}}|\mathbf{S}]$. The resulting mean and equal-tailed 95% credible interval bounds are shown in Figure 2.4. The multiple imputation results approximate the mean and variance of the posterior predictive distribution in this example with reasonable precision.

We next examine varying behavior over time, using a varying coefficient model for each covariate in the model, where behavior was allowed to vary with time of day. For all covariates we specified a B-spline basis expansion with regularly-spaced spline knots at 6 hour intervals over the course of a 24 hour period. Observations over multiple days (14 days in this study)

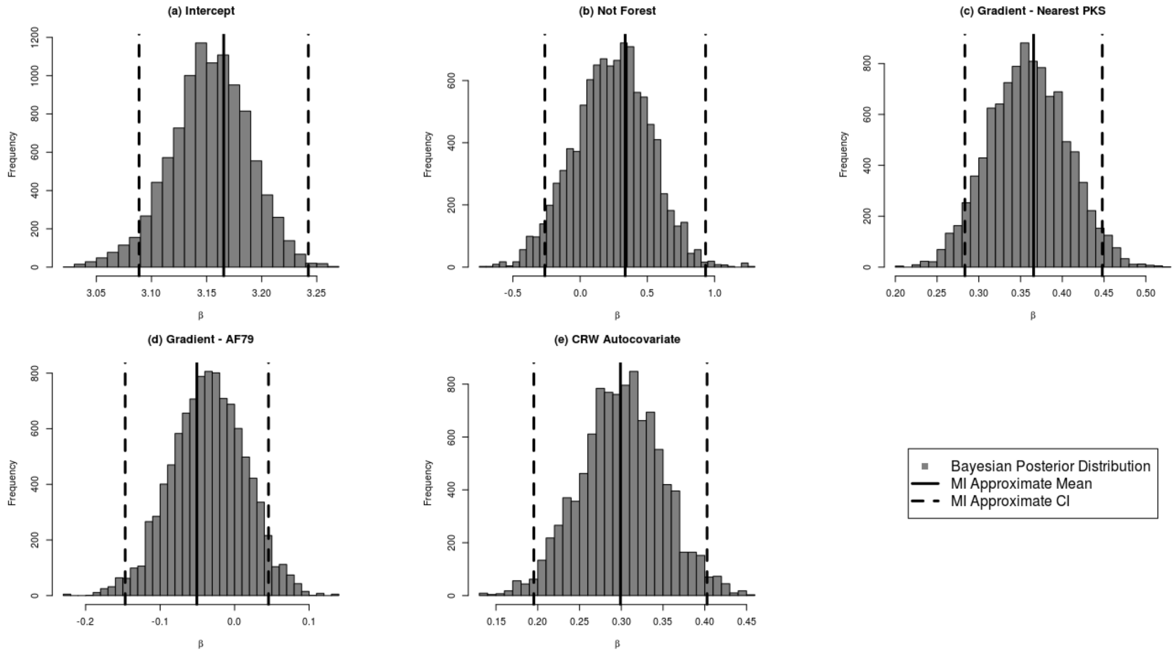

Figure 2.4: Histograms of posterior predictive distribution of $\boldsymbol{\beta}|\mathbf{S}$ for the male mountain lion (AM80), together with approximate means (solid lines) and 95% equal-tailed credible limits (dashed lines) obtained using the multiple imputation approximation.

are replications in this model and allow for inference about diurnal changes in movement behavior.

For this analysis, we fit the CTDS model for AM80 using the ‘glmnet’ R package (Friedman et al., 2010), using a lasso penalty, with tuning parameter chosen to minimize the average squared error of the fit in a 10-fold cross-validation. We used the resulting estimate of the lasso tuning parameter γ as a hyperparameter in the Bayesian model with lasso shrinkage prior. The resulting posterior predictive mean and equal-tailed 95% credible interval bounds for $\boldsymbol{\beta}(t)$ are shown in Figure 2.5.

2.6.1 Results

The results in Figure 2.4 show that the multiple imputation approximation adequately captures the location and spread of the posterior predictive distributions of movement parameters for AM80. The results show that much of the subadult male’s movement can be

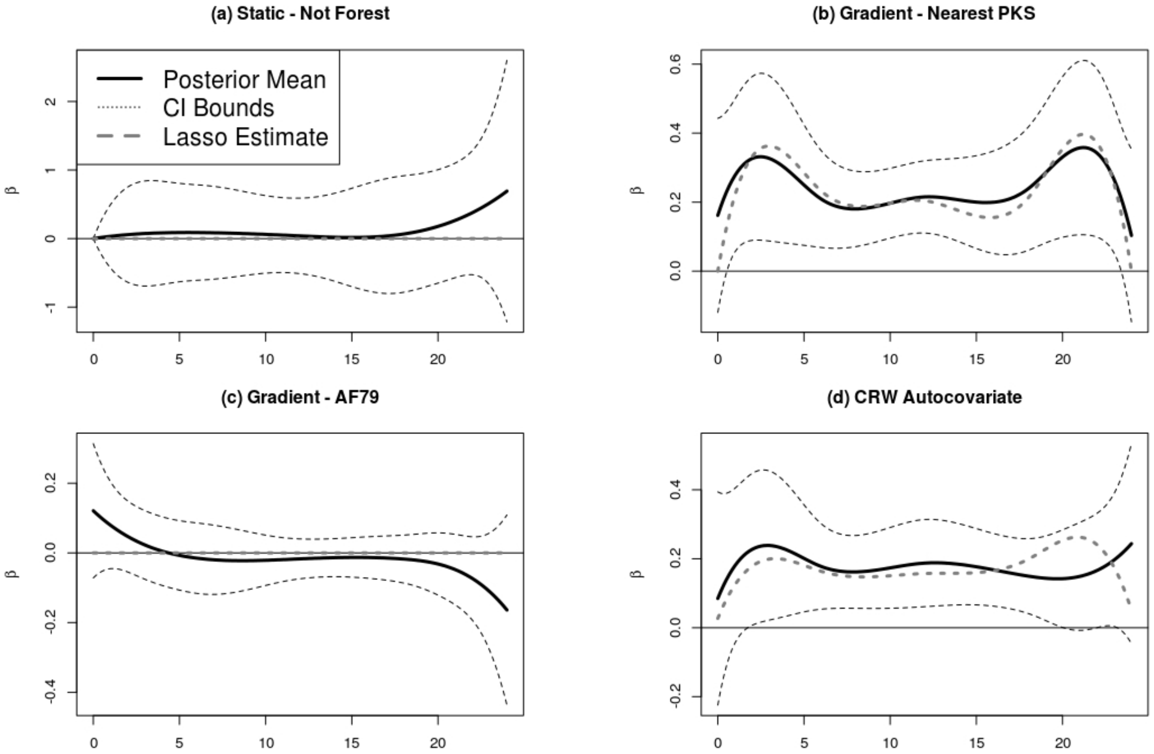

Figure 2.5: Time-varying results for the location-based and directional covariates in the continuous-time discrete-space model for a male mountain lion (AM80) obtained using a lasso shrinkage prior.

explained by a correlated random walk (Figure 2.4f) with attractive points at PKSs (Figure 2.4c). The results also show that the animal's movement behavior is fairly homogeneous when in forested and in non-forested terrain (Figure 2.4a).

The time-varying results for the location-based and directional drivers of movement for AM80 are shown in Figure 2.5. In Figure 2.5b, the nearly constant positive value or the $\beta(t)$ associated with movement towards the nearest PKS indicates the animal's behavior around PKS's is fairly consistent throughout the day, with some preference for returning to a PKS during the crepuscular periods. In Figure 2.5d, the positive value for the $\beta(t)$ associated with the autoregressive parameter indicates that the lion often engages in correlated movement, especially during the middle of the day. The confidence bands of the other parameters include zero throughout the day, indicating that the animal's behavior is is not consistent with the diurnal cycle.

2.7 Discussion

While we have couched our CTDS approach in terms of modeling animal movement, we can also view this approach in terms of resource selection (e.g., Manly et al., 2002). Johnson et al. (2008) describe a general framework for the analysis of resource selection from telemetry data using a weighted distribution approach where an observed distribution of resource use is seen as a re-weighted version of a distribution of available resources, and the resource selection function (RSF) defines the preferential use of resources by the animal. Warton and Shepherd (2010) and Aarts et al. (2012) describe a point process approach to resource selection that can be fit using a Poisson GLM, similar to the CTDS model we describe here. In the context of Warton and Shepherd (2010), the CTDS approach can be viewed as a resource selection analysis with the available resources at any time defined as the neighboring grid cells. The transition rate (2.7) of the CTDS process to each neighboring cell contains information about the availability of each cell, as well as the RSF that defines preferential use of the resources in each cell.

It is notable that the computationally-efficient multiple imputation approximation provided accurate estimates of the mean and variance of the posterior predictive distribution of $\boldsymbol{\beta}|\mathbf{S}$. The entire multiple imputation analysis required less than ten minutes using a computer with 4 GB of memory and a 1.67 GHz quad-core processor. In contrast, the fully Bayesian approach required nearly ten hours using the same computer. This increase in computational efficiency relative to the approaches of Johnson et al. (2008), Hooten et al. (2010), Hanks et al. (2011), and McClintock et al. (2012) allows for inference on complex behavior at finer temporal resolution than has been possible previously.

We have focused on building an individual model for animal movement that is intuitive and computationally-efficient. If population level inference on data from multiple animals is desired, there are multiple potential approaches. The first is to analyze movement paths from multiple animals jointly, with population level parameters in the GLM being shared by all animals and individual variation modeled using standard random effects approaches. However, this approach may not be straightforward to implement or interpret in the context of a varying coefficient model. Each individual animal is likely to encounter and be influenced by different drivers of movement (e.g., local environmental factors or nearby conspecifics) at differing times throughout the observation window. In this situation, it may make little sense to examine a population level response to a particular driver of movement at a particular time, as in a typical random effects analysis.

Instead, one approach to a population level analysis could be a post-hoc analysis of the time-varying response to covariates $\boldsymbol{\beta}(t)$. For example, Hanks et al. (2011) used a cluster analysis to examine different movement regimes shared across individuals in the population. Differences in the movement patterns exhibited by subgroups (e.g., male vs. female) were then examined by a comparison of the proportion of time spent by each subgroup in each of the movement regimes.

The use of directional drivers of movement has a long history. Brillinger et al. (2001) model animal movement as a continuous-time, continuous-space random walk where the drift term is the gradient of a “potential function” that defines an animal’s external drivers of movement. Tracey et al. (2005) use circular distributions to model how an animal moves in response to a vector pointing towards an object that may attract or repel the animal. Hanks et al. (2011) and McClintock et al. (2012) make extensive use of gradients to model directed movements, and movements about a central location. In our study of mountain lion movement data, we used directional drivers of movement to model conspecific interaction between a mother (AF79) and her cub (AM80). Interactions between predators and prey could also be modeled using directional covariates defined by vectors pointing between animals. Some movements based on memory could also be modeled using directional covariates. For example, a directional covariate defined by a vector pointing to the animal’s location one year prior could be used to model seasonal migratory behavior. The ability to model both location-based and directional drivers of movement make the CTDS framework a flexible and extensible framework for modeling complex behavior in animal movement.

CHAPTER 3

CIRCUIT THEORY AND MODEL-BASED INFERENCE FOR LANDSCAPE CONNECTIVITY

3.1 Introduction

[2] Having examined a class of models for animal movement data in Chapter 2, the analysis of genetic data for landscape connectivity will be the focus of Chapters 3 and 4. In this chapter, we focus on circuit theoretic approaches to studying landscape genetics and connectivity.

Circuit theory has been successfully used to study connectivity in a wide range of fields, including molecular chemistry (Zhu and Klein, 1996; Klein et al., 2004), collaborative recommendation (Fouss et al., 2007), communications network analysis (Tizghadam and Leon-Garcia, 2010, 2011), social network analysis (Kunegis et al., 2009), and random walks on a graph (Chandra et al., 1996; Volchenkov, 2011). Circuit theory has also seen extensive recent use in landscape ecology, where it has been theoretically linked to animal movement and gene flow in heterogeneous landscapes (McRae, 2006; Cushman et al., 2006; McRae and Beier, 2007; McRae et al., 2008; Urban et al., 2009; Cushman and Landguth, 2010; Dyer et al., 2010; Lookingbill et al., 2010; Owen-Smith et al., 2010; Rayfield et al., 2010; Saura and Rubio, 2010). In these latter cases, the landscape is specified as a raster grid with connectivity between grid cells determined by landscape characteristics and modeled based on circuit theory (Figure 3.1). Circuit theory provides a flexible framework for mod-

[2] The material in Chapter 3 is based on the following publication: Hanks, E.M., M.B. Hooten. 2013. Circuit Theory and Model-Based Inference for Landscape Connectivity. *Journal of the American Statistical Association* 108(501), pp.22-33.

eling nonstationary connectivity, and shows promise for predicting effects of landscape and environmental change on connectivity (e.g., Storfer et al., 2007; Spear et al., 2010).

A key challenge in modeling landscape connectivity using circuit theory is to estimate the relative resistance values of various landscape characteristics (e.g., Spear et al., 2010). In applications of circuit theory other than landscape ecology, resistance values are typically known, or all resistors in the circuit are assumed to have equal resistance. In these cases, the focus is typically on exploratory analysis of the connectivity implied by viewing the system as a circuit, rather than on estimating resistance values based on observations. In contrast, the goal in landscape ecological applications of circuit theory is to understand the impact that different landscape characteristics have on connectivity. Observations are typically second-order, and come in the form of an observed pairwise distance matrix representing the process under study (e.g., spatial gene flow in landscape genetics). The most common approach used to estimate resistance values for different landscape characteristics is to choose between a set of pre-specified candidate resistance values for each landscape covariate hypothesized to have an effect on connectivity (e.g., Cushman et al., 2006, 2009). Each set of candidate resistance values is used to create a hypothesized resistance distance matrix between the observed spatial locations in the study, where the resistance distance is computed based on circuit theory. The correlations between each of the hypothesized distance matrices and the observed distance matrix is computed, and the set of candidate resistance values that result in the highest correlation to the observed distance matrix is chosen (e.g., Cushman et al., 2006, 2009; Wang et al., 2009; Shirk et al., 2010), with significance assessed through Mantel permutation tests (e.g., Legendre and Fortin, 2010).

One major drawback of this approach is that there is no obvious way to assess the uncertainty in the parameter estimates for the resistance values of the landscape covariates. This is a critical point, as the results of spatial connectivity studies are being used to influence policy decisions (e.g., Theobald et al., 2011) and predict the results of landscape change over time (e.g., Spear et al., 2010).

Our goal is to put the estimation of resistance values from observed genetic distance matrices within a model-based framework. Recent work by McCullagh (2009) shows that observed squared-Euclidean distance matrices can be modeled using the generalized (or intrinsic) Wishart distribution with a spatial covariance matrix as a parameter. However, it is not immediately obvious how to parameterize a covariance matrix in a way that models connectivity based on circuit theory. As circuits are based on a graph or network of nodes, it seems natural to consider spatial statistical models with discrete spatial support, such as Gaussian Markov random field (GMRF) models (e.g., Rue and Held, 2005). In what follows, we make use of intrinsic conditional autoregressive (ICAR) models (Besag, 1974; Besag and Kooperberg, 1995) for modeling spatial connectivity based on circuit theory. We show that the GMRF with ICAR covariance structure can be seen as a spatial model for a Gaussian process with connectivity defined by circuit theory, and that the expected squared difference between observations from an ICAR model is exactly the resistance distance defined by circuit theory. The ICAR covariance matrix is also identifiable in the generalized Wishart model for observed distance matrices, making it an appealing choice for modeling spatial structure when the observations come as pairwise distances. This link between circuit theory and ICAR spatial models provides a computationally-efficient framework for modeling circuits with Gaussian error and making inference on resistance values in the circuit based on observed data. Circuits and GMRFs both have a natural graph structure, but to the best of our knowledge this link between GMRFs with ICAR structure and circuits has not been described in the literature.

Our paper is organized as follows. In Section 3.2: *Graphs, Circuits and Resistance* we review graph and circuit theory. We then review how circuit theory is used to model gene flow in Section 3.3: *Landscape Genetics.* In Section 3.4: *Distance and Covariance* we review the relationship between a covariance matrix and the mean squared distance between observations. We also review how a symmetric positive semi-definite (SPSD) matrix can be used to calculate the resistance distance of a graph, and discuss the generalized

Wishart model for observed squared distance matrices. In Section 3.5: *Gaussian Markov Random Fields for Circuits* we show that an ICAR covariance induces the circuit theory resistance distance on a graph. This provides a framework for modeling observations with connectivity defined by circuit theory. We also discuss modeling in the case of partial or repeated observation of the nodes in the circuit. In Section 3.6: *Application*, we illustrate our approach through a simulation example and then use our approach to examine gene flow in alpine chamois (*Rupicapra rupicapra*) in the Borges Mountains of France. Finally, in Section 3.7: *Discussion*, we discuss possible extensions to our approach.

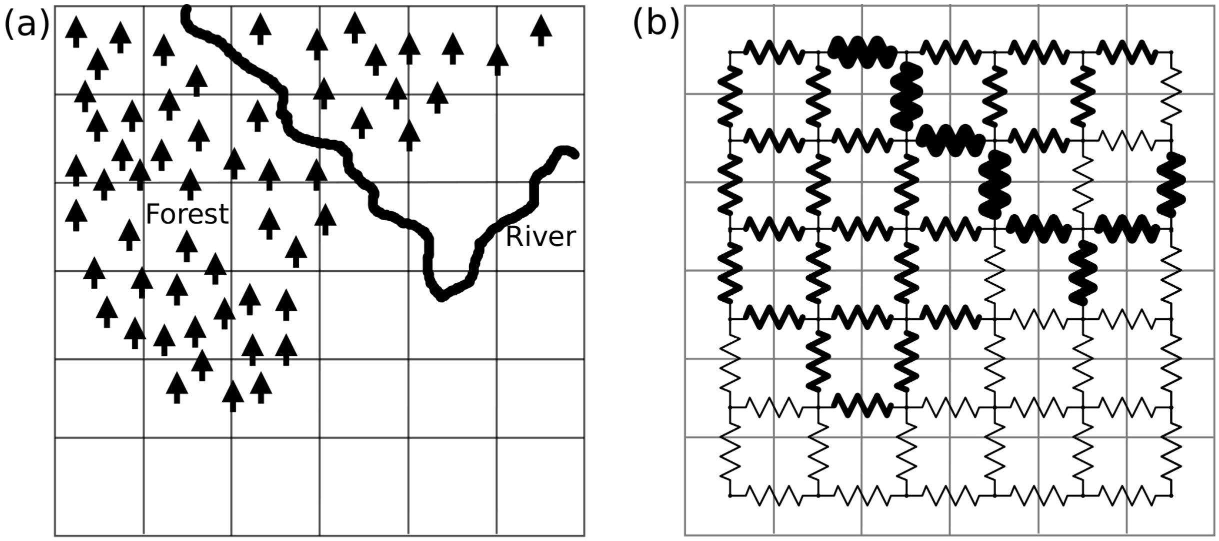

Figure 3.1: Illustration of the use of circuits to model spatial connectivity. The landscape is seen as a raster grid where connectivity between adjacent grid cells is a function of local landscape characteristics, and is computed using circuit theory. In the hypothetical landscape in (a), both the forest and the river impede connectivity. In the corresponding circuit (b), the raster cells are seen as being nodes connected by resistors, with thicker resistors graphically representing higher resistance and reduced conductance between nodes.

3.2 Graphs, Circuits, and Resistance

An undirected graph $(\mathbf{G}, \mathbf{A})$ is a collection of m "nodes" $\mathbf{G} = \{G_1, G_2, \ldots, G_m\}$ and the "edges" or "edge weights" $\mathbf{A} = \{\alpha_{ij}\}$ which are the connections between the nodes. If nodes i and j are first-order neighbors, denoted $i \sim j$, then $\alpha_{ij} > 0$ is a measure of their connectivity, with larger values leading to stronger connectivity between nodes. If nodes i and j are not first-order neighbors ($i \not\sim j$) then $\alpha_{ij} = 0$. The graph is said to be undirected if the edge weights are symmetric: $\alpha_{ij} = \alpha_{ji}$.

An electrical circuit can be represented by an undirected graph where the nodes $\mathbf{G}$ are connected by a set of resistors $\mathbf{A}$. Ohm's Law expresses the relationship between the resistance (R) of a resistor and the current (I) that flows through the resistor when a voltage (V) is applied as $V = I \cdot R$. The conductance α_{ij} of the resistor connecting neighboring nodes i and j is the current I that would flow through the resistor if it were removed from the circuit and a one-volt charge ($V = 1$) was applied across the resistor. The resistance R of the resistor is the inverse of the conductance: $1/\alpha_{ij}$.

A voltage applied across any two nodes i and j, which may or may not be directly connected (first-order neighbors) in the circuit, results in a current flow through the circuit. The effective resistance Γ_{ij} between nodes i and j is then defined as the resistance of a single resistor that would admit the same current flow if the voltage were applied across it (e.g., Dorf and Svoboda, 2004).

The effective resistance Γ_{ij} is a distance metric on $(\mathbf{G}, \mathbf{A})$ (Klein and Randić, 1993) that incorporates all possible pathways through the graph, where distance decreases with the addition of new pathways throughout the graph.

3.3 Landscape Genetics

Landscape genetics is the study of the effects of landscape on genetic diversity (Manel et al., 2003). Understanding how the landscape affects the genetic connectivity of a population of organisms can provide insight into the proper management of the population, the potential spread of disease through the population (e.g., Wheeler et al., 2010; Wheeler and Waller, 2010), and the effects of climate or landscape change on population connectivity (e.g. Spear et al., 2010). Linking genetic information (e.g., microsatellite data) to the landscape using a formal statistical model has been a continual challenge in the field of landscape genetics (e.g., Storfer et al., 2007; Spear et al., 2010). Spatial clustering methods (e.g., Guillot et al., 2005; Chen et al., 2007; Durand et al., 2009) are commonly used to identify genetic boundaries, but are typically based on the assumption of a homogeneous underlying landscape, with any link to the underlying landscape made *post hoc* (e.g., Wheeler et al., 2010). Approaches based on isolation by distance (Wright, 1943; Spear et al., 2005; Broquet et al., 2006) hypothesize that gene flow is based on a random walk in a homogeneous landscape, and thus that a genetic distance metric computed from observed genetic information is correlated with the straight-line distance between observed locations.

While the circuit theory model of gene flow (McRae, 2006) is formulated at a population level, with the assumption that each grid cell on a landscape contains a sub-population, it is often applied to individual-level genetic data (e.g., Cushman et al., 2006; Cushman and Landguth, 2010). In effect, the observed genetic information of an individual is viewed as a realization of genetic information in the theoretical sub-population in the grid cell. Pairwise genetic distances between individuals can be thought of as observations of the resistance distance between the grid cells where the individuals were observed, with noise.

Inference on resistance values of landscape characteristics is typically achieved using a correlation analysis. An observed genetic distance matrix is compared to a number of possible resistance distance matrices that result from various hypothesized resistance values for landscape characteristics. The hypothesized resistance values that result in the highest correlation with the observed genetic distance are selected.

We seek to improve the main features of this approach in two ways. First, little or nothing is learned through this approach about the uncertainty pertaining to the estimated resistance values. Second, the process of estimating the resistance values requires a grid search of the parameter space, and consequently a large number of candidate resistance values may need to be examined. For example, if the goal is to estimate the resistance to spatial gene flow due to p distinct landscape characteristics, a grid of hypothetical resistance values for each characteristic would need to be specified. If we consider a modest 10 hypothetical values for each resistance value, we would need to calculate the resistance distance and resulting correlation with observed data for each of the 10^p combinations of hypothesized values. The number of hypothetical values could grow much higher, depending on the degree of precision required.

Our goal is to put the estimation of resistance values from observed genetic distance matrices within a model-based framework. After developing a suitable model, Bayesian statistical methods will allow for estimation of the posterior distribution of the resistance values of landscape characteristics, providing point estimates through posterior means and estimates of uncertainty through posterior credible intervals.

3.4 Distance and Covariance

Covariance matrices and distance matrices both contain information about the connectivity between locations indexed by the matrix rows and columns. In this section, we review a link between covariance matrices and distance metrics, show how an improper (rank-deficient) covariance matrix can be used to induce the resistance distance $\mathbf{\Gamma}$ on the graph $(\mathbf{G}, \mathbf{A})$, and review recent developments in the modeling of observed distance matrices.

3.4.1 Induced Distance Matrices

We will refer to the space of symmetric positive semi-definite (SPSD) matrices of dimension $m \times m$ as $\mathcal{SPSD}(m)$. Let $\mathbf{G} = \{G_1, \dots, G_m\}$ be a set of m locations (e.g., nodes in a graph) and let $\mathbf{\Sigma} \in \mathcal{SPSD}(m)$ be a SPSD matrix with components $(\Sigma_{ij}),\ i, j \in \{1, 2, \dots, m\}$. It is helpful to think of Σ_{ij} as the covariance between the i^{th} and j^{th} locations in $\mathbf{G}$. Also, let $\mathbf{e}_j$ be the $m \times 1$ column vector with the j^{th} element equal to 1 and all other elements equal to zero. The metric defined by

$$\begin{aligned} D_{ij} = \Delta_{ij}(\mathbf{\Sigma}) &= (\mathbf{e}_i - \mathbf{e}_j)' \, \mathbf{\Sigma} \, (\mathbf{e}_i - \mathbf{e}_j) \\ &= \Sigma_{ii} + \Sigma_{jj} - 2\Sigma_{ij} \end{aligned} \tag{3.1}$$

is a critical transformation in both circuit theory and in the modeling of observed distance matrices. In Section 3.4.2 we will show how a particular matrix $\mathbf{\Sigma}$ results in D_{ij} being the resistance distance of the graph. In Section 3.4.3 we will follow McCullagh (2009) and use the transformation (3.1) to formalize the modeling of observed distance matrices. We first list some properties of (3.1) that will be important in our effort to model observed genetic distance matrices using covariance matrices based on circuit theory.

If we collect the $\{D_{ij}\}$ in an $m \times m$ matrix $\mathbf{D}$, the transformation (3.2) can be written in matrix form. Let $\mathbf{d}_\Sigma = \text{diag}(\mathbf{\Sigma})$ be a column vector with elements equal to the diagonal elements of $\mathbf{\Sigma}$, and $\mathbf{1}$ be a column vector with all elements equal to 1. Then

$$\mathbf{D} = \Delta(\mathbf{\Sigma}) = -2\mathbf{\Sigma} + \mathbf{d}_\Sigma \mathbf{1}' + \mathbf{1}\mathbf{d}_\Sigma' \tag{3.2}$$

is a linear transformation from $\mathcal{SPSD}(m)$ to the space of $m \times m$ matrices that are negative definite on contrasts. We will say that the covariance matrix $\mathbf{\Sigma}$ induces the distance matrix $\mathbf{D} = \Delta(\mathbf{\Sigma})$ on the set of locations $\mathbf{G}$.

This transformation Δ has an intuitive interpretation in the case where $\mathbf{\Sigma}$ is a covariance matrix. Let $\mathbf{y}$ be an m-dimensional random variable with common mean $\boldsymbol{\mu}$ (i.e., $\mu_i = \mu$ for $i = 1, 2, \ldots, m$) and covariance matrix $\mathbf{\Sigma}$. Then the induced distance $\Delta_{ij}(\mathbf{\Sigma})$ is a variogram: the expected squared distance between the observations at locations i and j:

$$
\begin{aligned}
E\left[(y_i - y_j)^2\right] &= \mathrm{Var}(y_i) + \mathrm{Var}(y_j) - 2\mathrm{Cov}(y_i, y_j) \\
&= \Sigma_{ii} + \Sigma_{jj} - 2\Sigma_{ij} \\
&= \Delta_{ij}(\boldsymbol{\Sigma}).
\end{aligned}
$$

While this interpretation only holds when $\boldsymbol{\Sigma}$ is a covariance matrix, the induced matrix $\Delta(\boldsymbol{\Sigma})$ is a distance for a class of rank deficient SPSD matrices. To see this, we first examine the null space or kernel of the operator Δ, which is the set of $\boldsymbol{\Sigma} \in \mathcal{SPSD}(m)$ such that $\Delta(\boldsymbol{\Sigma})$ is the zero matrix. McCullagh (2009) notes that the null space of Δ is the space of additive symmetric matrices $\mathcal{S} = \{\mathbf{S}_{m\times m} : S_{i,j} = v_i + v_j,\ \mathbf{v} \in \mathbb{R}^m\}$. We clarify this statement in the case where $\boldsymbol{\Sigma} \in \mathcal{SPSD}(m)$.

Proposition 1. *For $\boldsymbol{\Sigma} \in \mathcal{SPSD}(m)$, the transformation $\Delta(\boldsymbol{\Sigma})$ is a linear transformation with null space equal to $\{\boldsymbol{\Sigma} : \boldsymbol{\Sigma} = c\mathbf{1}\mathbf{1}'\ ,\ c \in \mathbb{R}\}$.*

This result is obtained by noting that $\boldsymbol{\Sigma} \in \mathcal{SPSD}(m)$ implies that there is a singular value decomposition $\boldsymbol{\Sigma} = \sum_{i=1}^m \lambda_i \mathbf{u}_i \mathbf{u}_i'$, where $\{\lambda_i \geq 0\}$ are the eigenvalues and $\{\mathbf{u}_i\}$ are orthonormal eigenvectors. Then we can rewrite (3.1) as

$$
\begin{aligned}
D_{ij} = \Delta_{ij}(\boldsymbol{\Sigma}) &= (\mathbf{e}_i - \mathbf{e}_j)' \boldsymbol{\Sigma} (\mathbf{e}_i - \mathbf{e}_j) \\
&= (\mathbf{e}_i - \mathbf{e}_j)' \left(\sum_{k=1}^m \lambda_k \mathbf{u}_k \mathbf{u}_k' \right) (\mathbf{e}_i - \mathbf{e}_j) \\
&= \sum_{k=1}^m \lambda_k \left[(\mathbf{e}_i - \mathbf{e}_j)' \mathbf{u}_k \right]^2
\end{aligned}
$$

As $\lambda_k \geq 0$, each term in the sum above is non-negative. From this, $\boldsymbol{\Sigma}$ is in the null space of the transformation Δ if and only if $\Delta_{ij}(\boldsymbol{\Sigma}) = 0$ for all $i = 1, \ldots, m$ and $j = 1, \ldots, m$, which in turn is true if and only if either $\lambda_k = 0$ or $\mathbf{u}_k = c\mathbf{1}\ ,\ c \in \mathbb{R}$ for each $k = 1, \ldots, m$. Since the $\mathbf{u}_k$ are orthogonal, we conclude that $\Delta_{ij}(\boldsymbol{\Sigma}) = 0$ for all i and j if and only if $\boldsymbol{\Sigma} = \lambda\mathbf{1}\mathbf{1}'$.

We will denote the null (column) space of a matrix $\mathbf{\Sigma}$ as $\mathcal{N}(\mathbf{\Sigma})$. If $\mathbf{\Sigma}$ is of full rank, or if $\mathcal{N}(\mathbf{\Sigma})$ is the space spanned by the $\mathbf{1}$ vector, then $\Delta(\mathbf{\Sigma})$ is a distance metric on $\mathbf{G} \times \mathbf{G}$. That is, $\mathbf{D} = \Delta(\mathbf{\Sigma})$ is symmetric ($D_{ij} = D_{ji}$), satisfies the triangle inequality ($D_{ij} + D_{jk} \geq D_{ik}$), and is zero only for the distance from a location to itself ($D_{ij} = 0$ iff $i = j$).

If $\mathcal{N}(\mathbf{\Sigma})$ is the space spanned by the $\mathbf{1}$ vector, then $\mathbf{\Sigma}$ is orthogonal to the null space of Δ and the transformation Δ is invertible. Note that this invertibility does not hold for any $\mathbf{\Sigma}$ of full rank. This will be important in Section 3.4.3 where we discuss the identifiability of covariance matrices in the generalized Wishart model for distance matrices induced by the transformation Δ in (3.2).

3.4.2 Inducing the Resistance Distance of a Graph

The effective resistance matrix $\mathbf{\Gamma} = (\Gamma_{ij})$ of a circuit represented by the undirected graph $(\mathbf{G}, \mathbf{A})$ can be computed by constructing a SPSD matrix that induces $\mathbf{\Gamma}$. Kirchoff's current law (e.g., Dorf and Svoboda, 2004) states that the electrical current flowing into a node must equal the current flowing out of the node. Applying Kirchoff's current law to the graph $(\mathbf{G}, \mathbf{A})$ leads to the formation of the Laplacian matrix (e.g., Babić et al., 2002) of the graph:

$$\mathbf{Q} = \begin{bmatrix} \sum_{j \neq 1} \alpha_{1j} & -\alpha_{12} & -\alpha_{13} & \cdots \\ -\alpha_{21} & \sum_{j \neq 2} \alpha_{2j} & -\alpha_{23} & \cdots \\ -\alpha_{31} & -\alpha_{32} & \sum_{j \neq 3} \alpha_{3j} & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix}. \tag{3.3}$$

The Laplacian matrix $\mathbf{Q}$ is of dimension $m \times m$, and each row sums to zero. The null space $\mathcal{N}(\mathbf{Q})$ is the space spanned by the $\mathbf{1}$ vector, and $\mathbf{Q}$ is singular with rank $m - 1$. The generalized inverse of the Laplacian $\mathbf{Q}$ is $\mathbf{Q}^{-}$, which is also of rank $m - 1$. As shown by Klein and Randić (1993), the effective resistance between the nodes of $(\mathbf{G}, \mathbf{A})$ is the distance induced by the SPSD matrix $\mathbf{Q}^{-}$:

$$\Gamma_{ij} = (\mathbf{e}_i - \mathbf{e}_j)' \, \mathbf{Q}^{-} \, (\mathbf{e}_i - \mathbf{e}_j) \, . \tag{3.4}$$

Generalized inverses are not unique, but the correspondence between the null space of the Laplacian matrix $\mathbf{Q}$ and the null space of the transformation (3.2) results in a useful invariance.

Proposition 2. *The resistance distance* $\Gamma_{ij} = (\mathbf{e}_i - \mathbf{e}_j)' \, \mathbf{Q}^{-} \, (\mathbf{e}_i - \mathbf{e}_j)$ *is invariant to the choice of generalized inverse* $\mathbf{Q}^{-}$ *of the Laplacian matrix* $\mathbf{Q}$.

Proposition 2 is a direct consequence of the lemma that $\mathbf{a}'\mathbf{Q}^{-}\mathbf{a}$ is invariant to the choice of generalized inverse $\mathbf{Q}^{-}$ if and only if $\mathbf{a}$ is orthogonal to $\mathcal{N}(\mathbf{Q})$ (see pp. 134-135 of Graybill (1983) and pp. 130-131 of Seber (2008)). The null space of the Laplacian $\mathbf{Q}$ is the space spanned by $\mathbf{1}$, and each contrast $\mathbf{e}_i - \mathbf{e}_j$ is orthogonal to $\mathbf{1}$.

This invariance was not noted by Klein and Randić (1993), though various generalized inverses have been used to compute the resistance distance (e.g., Babić et al., 2002; McRae, 2006).

3.4.3 Modeling Induced Distance Matrices

The observations in our landscape genetics example are genetic distance matrices, which we are viewing as distance matrices induced from an underlying spatial covariance motivated by circuit theory. The distributional properties of distance matrices induced by the transformation (3.2) have been studied by McCullagh (2009). We briefly review the results of McCullagh (2009) here, and then discuss identifiability in models for observed distance matrices.

Let $\mathbf{y}_i \sim N(\mu\mathbf{1}, \mathbf{\Sigma}), \; i = 1, 2, \ldots, \nu$ be ν independent realizations from a Gaussian process with common mean μ and covariance matrix $\mathbf{\Sigma}$. If we concatenate the observations as $\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_\nu]$, then $\mathbf{S} = \mathbf{Y}\mathbf{Y}'$ is Wishart distributed with ν degrees of freedom and covariance matrix $\mathbf{\Sigma}$:

$$\mathbf{S} = \mathbf{Y}\mathbf{Y}' \sim \mathcal{W}_\nu(\mathbf{\Sigma}).$$

McCullagh (2009) defines the generalized Wishart distribution as follows. $\mathbf{V}$ is said to come from a generalized Wishart distribution with ν degrees of freedom, covariance matrix $\mathbf{\Sigma}$, and kernel $\mathcal{K}$

$$\mathbf{V} \sim \mathcal{GW}_\nu(\mathcal{K}, \mathbf{\Sigma})$$

if for any linear transformation $\mathbf{L}$ such that the null space of $\mathbf{L}$ is $\mathcal{K}$ ($\mathcal{N}(\mathbf{L}) = \mathcal{K}$), $\mathbf{L}'\mathbf{V}\mathbf{L}$ is Wishart distributed:

$$\mathbf{L}'\mathbf{V}\mathbf{L} \sim \mathcal{W}_\nu(\mathbf{L}'\mathbf{\Sigma}\mathbf{L}).$$

McCullagh (2009) then shows that induced distance matrices (3.2) can be modeled using the generalized Wishart distribution. Let $\mathbf{S} \sim \mathcal{W}_\nu(\mathbf{\Sigma})$ and consider the induced distance matrix $\mathbf{D} = \Delta(\mathbf{S})$. Consider an $m \times (m-1)$ matrix $\mathbf{L}$ with full column rank and $\mathcal{N}(\mathbf{L})$ spanned by $\mathbf{1}$. For example, $\mathbf{L}$ could be a full rank matrix of contrast vectors:

$$\mathbf{L} = [\mathbf{e}_2 - \mathbf{e}_1, \mathbf{e}_3 - \mathbf{e}_1, \ldots, \mathbf{e}_m - \mathbf{e}_1].$$

Then $\mathbf{L}'\mathbf{1} = \mathbf{0}$ and it is easy to see from (3.2) that

$$\begin{aligned}\mathbf{L}'(-\mathbf{D})\mathbf{L} &= \mathbf{L}'\left(2\mathbf{S} - \mathbf{d}_{\mathbf{\Sigma}}\mathbf{1}' - \mathbf{1}\mathbf{d}_{\mathbf{\Sigma}}'\right)\mathbf{L} \\ &= \mathbf{L}'(2\mathbf{S})\mathbf{L}.\end{aligned}$$

which is Wishart distributed with covariance matrix $\mathbf{L}'(2\mathbf{\Sigma})\mathbf{L}$. Then, by definition,

$$-\mathbf{D} = -\Delta(\mathbf{S}) \sim \mathcal{GW}_\nu(\mathbf{1}, 2\mathbf{\Sigma}). \tag{3.5}$$

The likelihood of an observed distance matrix $\mathbf{D}$ under this model can be obtained by evaluating the likelihood $[\mathbf{L}'(-\mathbf{D})\mathbf{L} \mid \nu, \mathbf{L}'(2\mathbf{\Sigma})\mathbf{L}]$ in the Wishart distribution $\mathbf{L}'(-\mathbf{D})\mathbf{L} \sim$

$\mathcal{W}_\nu(\mathbf{L}'(2\mathbf{\Sigma})\mathbf{L})$, where the bracket '$[\cdot]'$ notation indicates a probability distribution. Alternately, McCullagh shows that this likelihood can be obtained equivalently by using a projection $\mathbf{A} = \mathbf{I} - \frac{\mathbf{1}\mathbf{1}'\mathbf{\Sigma}^{-1}}{\mathbf{1}'\mathbf{\Sigma}^{-1}\mathbf{1}}$ onto the space of $m \times m$ matrices orthogonal to the null space of the transformation Δ. The resulting expression for the likelihood does not rely on the choice of $\mathbf{L}$:

$$[-\mathbf{D}|2\mathbf{\Sigma}, \nu] \propto |\mathbf{\Sigma}^{-1}\mathbf{A}|^{\nu/2} \cdot \exp\left\{\frac{1}{4}\mathrm{tr}(\mathbf{\Sigma}^{-1}\mathbf{A}\mathbf{D})\right\}, \tag{3.6}$$

where the determinant $|\cdot|$ is the product of all non-zero eigenvalues.

As the transformation $\Delta(\mathbf{S})$ in (3.2) is invariant to the addition of a constant matrix $c\mathbf{1}\mathbf{1}'$ to $\mathbf{S}$, the likelihood (3.6) is invariant to the addition of a constant matrix $c\mathbf{1}\mathbf{1}'$ to $\mathbf{\Sigma}$, and $\mathbf{\Sigma}$ is not identifiable in general. While McCullagh (2009) is not concerned with identifiability (See pg. 635 of McCullagh, 2009), our goal is to parameterize and estimate a spatial covariance matrix $\mathbf{\Sigma}$ that describes gene flow under a circuit theory model for connectivity. Identifiability of $\mathbf{\Sigma}$ is important in this endeavor.

The fact that the generalized Wishart distribution is defined on distance matrices induced by applying the transformation (3.2) to observations from a Wishart distribution implies that the covariance matrix $\mathbf{\Sigma}$ in the generalized Wishart model (3.65) will be identifiable if the transformation (3.2) is invertible. The following proposition follows directly from Proposition 1, where we describe the null space of the transformation (3.2).

Proposition 3. *In the generalized Wishart model (3.5), $\mathbf{\Sigma}$ is identifiable if and only if the space spanned by $\mathbf{1}$ is contained in $\mathcal{N}(\mathbf{\Sigma})$.*

This implies that only a rank-deficient covariance matrix $\mathbf{\Sigma}$ will be identifiable. We also note that the deficiency implied here is identical to the null-space of the Laplacian matrix (3.3) of the graph under circuit theory. We now specify a GMRF with a covariance matrix that will be identifiable in the generalized Wishart model for distance matrices and that will induce the resistance distance on the graph, matching the second-order structure defined by circuit theory.

3.5 Gaussian Markov Random Fields for Circuits

Gaussian Markov random fields (GMRFs) are used extensively in statistical modeling, especially for spatial and temporal processes (Rue and Held, 2005; Lindström and Lindgren, 2008; Bolin et al., 2009). GMRFs on a graph $\mathbf{G}$ are characterized by the conditional independence of any two nodes that are not first-order neighbors when conditioned on all other nodes in $\mathbf{G}$. This conditional independence results in a sparse precision matrix, and sparse matrix methods can typically be applied to any statistical analysis utilizing GMRFs (Rue, 2001; Rue and Held, 2005; Rue et al., 2009). This allows for highly efficient computation, and can make inference possible at a finer spatial or temporal resolution than might be possible using a non-Markovian random field.

Our goal is to specify a GMRF on the graph $(\mathbf{G}, \mathbf{A})$ with a covariance matrix that induces the resistance distance matrix $\boldsymbol{\Gamma}$. The result is a statistical model for observations from a graph with Gaussian error and expected squared distance between observations equal to the resistance distance of the graph.

Consider a random variable $\mathbf{y} = [y_1, y_2, \ldots, y_m]'$ on the m nodes $\mathbf{G} = \{G_1, G_2, \ldots, G_m\}$ and define the conditional distribution of y_i given all other values in $\mathbf{y}$ as

$$y_i | \mathbf{y}_{-i} \sim N\Big(\sum_{j \sim i} w_{ij} y_j , \kappa_i \Big) \tag{3.7}$$

$$w_{ij} = \frac{\alpha_{ij}}{\sum_{l \sim i} \alpha_{il}} \quad , \quad \kappa_i = \frac{1}{\sum_{l \sim i} \alpha_{il}} .$$

Conditionally, each observation is normally distributed with mean equal to a weighted average of all nodes that are first order neighbors $j \sim i$, where the weights are proportional to the conductance α_{ij} of the edge connecting the nodes.

The conditional specification in (3.7) satisfies the Markov property and is a GMRF. In particular, $\mathbf{y}$ follows an intrinsic Gaussian conditional autoregressive (ICAR) model on the graph $(\mathbf{G}, \mathbf{A})$ (Besag and Kooperberg, 1995; Rue and Held, 2005). Using Brooks Lemma

(Besag, 1974; Besag and Kooperberg, 1995), the joint distribution of $\mathbf{y}$ is

$$\mathbf{y} \sim N(\mathbf{0}, \mathbf{Q}^{-}) \tag{3.8}$$

where the precision matrix $\mathbf{Q}$ is exactly the Laplacian matrix (3.3) of the graph $(\mathbf{G}, \mathbf{A})$. Then the generalized inverse $\mathbf{Q}^{-}$ is a SPSD matrix that induces the resistance distance $\mathbf{\Gamma}$ on the graph $(\mathbf{G}, \mathbf{A})$, and the mean squared distance between observations from the ICAR process (3.8) is the resistance distance of the graph.

The ICAR model (3.8) is proper under the constraint that $\mathbf{y}'\mathbf{1} = c$ for some constant $c \in \mathbf{R}$. If we observe ν independent realizations $\mathbf{y}_i \sim N(\mathbf{0}, \mathbf{Q}^{-})$, $i = 1, 2, \ldots, \nu$, with $\mathbf{y}_i'\mathbf{1} = 0$, and concatenate the observations as in Section 3.4.3: $\mathbf{Y} = [\mathbf{y}_i, \mathbf{y}_2, \ldots, \mathbf{y}_\nu]$, then $\mathbf{S} = \mathbf{Y}\mathbf{Y}' \sim \mathcal{W}_\nu(\mathbf{Q}^{-})$ is proper under the constraint that $\mathbf{S}\mathbf{1} = \mathbf{0}$. Under this constraint, $\mathbf{S}$ is orthogonal to the null space of the transformation Δ (3.2), and the transformation to $\mathbf{D} = \Delta(\mathbf{S})$ is invertible. The elements of $\mathbf{D}$ are the sum of the squared difference in observations $D_{jk} = \sum_{i=1}^{\nu}(y_{ji} - y_{ki})^2$ subject to the constraints that $\mathbf{y}_i'\mathbf{1} = 0$. Under this model, each element D_{jk} of the distance matrix $\mathbf{D}$ has mean $\nu\Gamma_{jk}$, the resistance distance Γ_{jk} between the j^{th} and k^{th} nodes in the graph $\mathbf{G}$ multiplied by the number of pairwise observations ν.

This provides a natural model for observed distance matrices with connectivity modeled by circuit theory:

$$-\mathbf{D} \sim \mathcal{GW}_\nu(\mathbf{1}, 2\mathbf{Q}^{-}) \tag{3.9}$$

where $\mathbf{Q}$ is the Laplacian of the graph $(\mathbf{G}, \mathbf{A})$, or, equivalently the ICAR precision matrix of the graph $(\mathbf{G}, \mathbf{A})$. From Proposition 3, $\mathbf{Q}$ is identifiable, and the observations in $\mathbf{D}$ can be thought of as observations of the resistance distance, with Gaussian noise.

3.5.1 Partial and Repeated Observation

We have assumed so far that the circuit in question is fully observed. That is, an observed effective distance D_{ij} is obtained for each pair of nodes in $\mathbf{G}$. In practice, however, it will be common to have observations from only a small subset of the nodes, and multiple observations may come from a single node. In a spatial analysis, for example, the study domain may include thousands of raster cells, each of which is considered a node in the resistance surface $(\mathbf{G}, \mathbf{A})$. If we obtain a set of spatially-referenced observations, multiple observations may lie within one raster cell, and, depending on the number of observations in the study, only a small fraction of the raster cells (perhaps dozens or hundreds) will contain observations. In the case where only a subset of the nodes are observed, we want to preserve the sparse nature of the precision matrix $\mathbf{Q}$ and the associated computational efficiency of the GMRF.

The precision matrix $\mathbf{Q}$ is sparse, but the generalized inverse $\mathbf{Q}^{-}$ will typically be dense. For large m, it is impossible to even store a dense $m \times m$ matrix in many standard computing environments, so we wish to avoid the express calculation of $\mathbf{Q}^{-}$ for the whole graph.

Consider the case where we have only n observations $y_i,\ i = 1, \ldots, n$ from the m nodes in $\mathbf{G}$, where $m >> n$ and let the observed nodes be indexed by the first n rows and columns of $\mathbf{Q}$. Let $\mathbf{Q}^*$ be the $(m-1) \times (m-1)$ upper-left submatrix of $\mathbf{Q}$. Then, since $\mathbf{Q}$ is of rank $m-1$, $\mathbf{Q}^*$ is of full rank and invertible. The inverse of $\mathbf{Q}^*$ then forms the $(m-1) \times (m-1)$ upper-left submatrix of the generalized inverse $\mathbf{Q}^{-1}$, with entries in the m^{th} row and m^{th} column of $\mathbf{Q}^{-1}$ being set to zero. In this formulation, the covariance matrix of the n observed nodes in $\mathbf{G}$ is just the $n \times n$ upper-left submatrix of $(\mathbf{Q}^*)^{-1}$. This submatrix can be found by partitioning $\mathbf{Q}^*$ in block form:

$$\mathbf{Q}^* = \left[\begin{array}{c|c} \mathbf{Q}^*_{11} & \mathbf{Q}^*_{12} \\ \hline \mathbf{Q}^*_{21} & \mathbf{Q}^*_{22} \end{array} \right] \tag{3.10}$$

where $\mathbf{Q}^*_{11}$ is the $n \times n$ block of $\mathbf{Q}$ corresponding to the n observed nodes of $\mathbf{G}$, $\mathbf{Q}^*_{22}$ is the $(m-n-1) \times (m-n-1)$ block of $\mathbf{Q}$ corresponding to the $m-1$ unobserved nodes, after the removal of the m^{th} row and column, and $\mathbf{Q}^*_{12} = (\mathbf{Q}^*_{21})^T$ is the $n \times (m-n-1)$ block of $\mathbf{Q}^*$ relating the observed nodes to the unobserved nodes. The precision matrix $\boldsymbol{\Phi}$ of the n observed locations is the Schur complement (Harville, 2008) of $\mathbf{Q}^*_{22}$: $\boldsymbol{\Phi} = \mathbf{Q}^*_{11} - \mathbf{Q}^*_{12}(\mathbf{Q}^*_{22})^{-1}\mathbf{Q}^*_{21}$. As $\mathbf{Q}^*_{22}$ is of dimension $(m-n-1) \times (m-n-1)$, we also wish to avoid the explicit calculation of the inverse $(\mathbf{Q}^*_{22})^{-1}$ when computing $\boldsymbol{\Phi}$. This can be accomplished by first computing the Cholesky decomposition: $\mathbf{Q}^*_{22} = \mathbf{U}^T\mathbf{U}$, which can be found efficiently using sparse matrix methods (Bates and Maechler, 2011). We then compute $\mathbf{V} = (\mathbf{U}^T)^{-1}\mathbf{Q}^*_{21}$ by solving the n linear equations $\mathbf{U}^T\mathbf{V} = \mathbf{Q}^*_{21}$. As $\mathbf{U}^T$ is lower-triangular, this can be solved efficiently using forward substitution. The precision matrix of the observed nodes is then obtained by

$$\boldsymbol{\Phi} = \mathbf{Q}^*_{11} - \mathbf{V}^T\mathbf{V}. \tag{3.11}$$

The advantage of this method is that we can use sparse matrix methods that are computationally efficient and do not require the explicit computation of large dense matrices (Rue, 2001).

In the case where multiple observations are obtained from a single node of $\mathbf{G}$, a nugget effect can be introduced to the covariance structure (Besag et al., 1991; Lindström and Lindgren, 2008; Cross et al., 2010). Let $\mathbf{y} = [y_1, y_2, \ldots y_{n_{\text{obs}}}]^T$ be the vector of n_{obs} observations from the graph $(\mathbf{G}, \mathbf{A})$, and let $s_i \in G$ be the node of the i^{th} observation, $i = 1, 2, \ldots, n_{\text{obs}}$. Since there may be repeated observations in some nodes, the total number of nodes of $\mathbf{G}$ for which we have observations is $n_{\text{nodes}} \leq n_{\text{obs}}$. Let $\boldsymbol{\Sigma}_{\text{nodes}} = \boldsymbol{\Phi}^{-1}$ be the $n_{\text{nodes}} \times n_{\text{nodes}}$

covariance matrix of the observed nodes, and let $\mathbf{K}$ be an $n_{\text{obs}} \times n_{\text{nodes}}$ matrix with $K_{ij} = 1$ if $s_i = j$, that is if the i^{th} observation y_i comes from the j^{th} node of $\mathbf{G}$, and $K_{ij} = 0$ otherwise. Then the covariance $\mathbf{\Psi}$ of the n_{obs} observations is

$$\mathbf{\Psi} = \mathbf{K}\mathbf{\Sigma}_{\text{nodes}}\mathbf{K}^T + \tau\mathbf{I}$$

where connectivity between the nodes of $\mathbf{G}$ is modeled based on circuit theory ($\mathbf{\Sigma}_{\text{nodes}}$) and τ is a spatial nugget parameter representing the variability in observations obtained from the same node in $\mathbf{G}$.

We can then model an observed distance matrix $\mathbf{D}$ as arising from a generalized Wishart distribution with covariance matrix of the observed locations $\mathbf{\Psi}$,

$$-\mathbf{D} \sim \mathcal{GW}(\mathbf{1}, 2\mathbf{\Psi}). \tag{3.12}$$

The covariance matrix $\mathbf{\Psi}$ is dependent on the edge weights $\{\alpha_{ij}\}$, and may contain a nugget effect τ to account for repeated observations.

3.5.2 Modeling Resistance

We now turn to the parameterization of the edge weights $\{\alpha_{ij}\}$ of the graph $(\mathbf{G}, \mathbf{A})$ in terms of landscape covariates. In doing so, it will be helpful to consider a few links between circuit theory and random walks on the graph. In a circuit, the edge weights are the conductances between neighboring nodes in the graph. If we consider the CTDS random walk model from Chapter 2, the edge weights $\{\alpha_{ij}\}$ of the graph are proportional to the rate at which a random walker transitions from node i to node j. As landscape covariates are typically available in gridded form, each node in $\mathbf{G}$ is a grid cell in the study area. The rate of transition from cell i to cell j could be driven by many factors, as discussed in Chapter 2, including the environment at the starting cell i, the environment at the neighboring cell j, and local environmental gradients. To maintain the links to circuit theory, however, the transition rates must be symmetric ($\alpha_{ij} = \alpha_{ji}$) and non-negative. This ensures that $(\mathbf{G}, \mathbf{A})$ is an undirected graph, and also ensures that the conditional specification (3.8) of the ICAR model results in a GMRF. As the edge weights must be symmetric, we cannot model drift due to directional (gradient-based) drivers of movement using circuit theory. Forcing the edge weights to be symmetric implies that we are modeling connectivity of a system that is stationary. While this assumption may be unreasonable for individual organisms at short time scales (e.g., Hanks et al., 2011), there is more support for it at a population-level and at long time scales (e.g., genetic time).

To specify symmetric edge weights, we consider an average of the landscape covariates at adjacent cells. This is motivated by considering the straight-line path between the centers of the two adjacent grid cells. Half of this path lies in each cell, and we can hypothesize a transition rate (or edge weight) that would be related to an average of the landscape covariates in the two cells. If $\mathbf{x}_i$ is a vector of p landscape characteristics of the i^{th} cell, and $\boldsymbol{\beta}$ is a p-vector of parameters related to the effect that each landscape characteristic has on gene flow, then the edge weights $\{\alpha_{ij}\}$ could be modeled as

$$\alpha_{ij} = \begin{cases} \exp\left[\frac{1}{d_{ij}}\left(\frac{\mathbf{x}_i' + \mathbf{x}_j'}{2}\right)\boldsymbol{\beta}\right] & , j \sim i \\ 0 & , j \nsim i \end{cases} \tag{3.13}$$

where d_{ij} is the straight-line distance between the centroids of cells i and j. In this model, connectivity is a function of the landscape characteristics of both cells $(\mathbf{x}_i, \mathbf{x}_j)$, the distance between the cells (allowing for differing distances between neighbors in, for example, a queen's neighborhood), and the conductance parameters $\boldsymbol{\beta}$. Using an exponential link function ensures that the $\{\alpha_{ij}\}$ are non-negative, and allows for convenient interpretation of the conductance parameters $\boldsymbol{\beta}$. If $\beta_h < 0$, then an increase in the h^{th} landscape characteristic results in a greater resistance to gene flow, while if $\beta_h > 0$, then an increase in the h^{th} landscape characteristic results in less resistance to gene flow.

Finally, assigning prior distributions to the coefficients $\boldsymbol{\beta}$ and nugget parameter τ

$$\boldsymbol{\beta} \sim N(\boldsymbol{\mu}_{\boldsymbol{\beta}}, \boldsymbol{\Sigma}_{\boldsymbol{\beta}}) \tag{3.14}$$

$$\tau \sim IG(a, b) \tag{3.15}$$

results in a statistical model (3.12)-(3.15) that can be fit using Bayesian techniques (e.g., Markov chain Monte Carlo, MCMC). In both the simulation example and data analysis that follow, we have specified a diffuse, mean-zero prior for $\boldsymbol{\beta}$: $\boldsymbol{\beta} \sim N(\mathbf{0}, 10^6\mathbf{I})$ and a relatively diffuse inverse-gamma prior for τ: $\tau \sim IG(10, 100)$. Samples from the posterior distribution of $\boldsymbol{\beta}$ and τ can be obtained using random walk Metropolis-Hastings updates in an MCMC sampler. After proposing a value for $\boldsymbol{\beta}$ or τ, the proposed value is used to compute the resulting precision matrix $\boldsymbol{\Psi}$ of the observations. The candidate $\boldsymbol{\beta}$ or τ value can then be accepted or rejected with the typical Metropolis-Hastings probability (e.g., Gelman et al., 2004).

One potentially useful extension to (3.13) is a varying coefficient model (Hastie and Tibshirani, 1993). For simplicity, consider one covariate x. The edge weights for neighboring cells could be modeled as

$$\alpha_{ij} = \exp\left[\frac{1}{d_{ij}}\left(\frac{\beta(x_i) + \beta(x_j)}{2}\right)\right]$$

where the functional regressor $\beta(x)$ is typically modeled as a linear combination of n_{spl} spline basis functions $\{\boldsymbol{\phi}_k, k = 1, \ldots, n_{spl}\}$: $\beta(x) = \sum_{k=1}^{n_{spl}} \gamma_k \boldsymbol{\phi}_k(x)$. This specification gives a flexible framework for examining nonlinear relationships between conductance and the covariate x.

3.6 Application: Landscape Genetic Analysis of Alpine Chamois

We now apply our approach, first in a study of simulated data, and then to observed genetic data.

3.6.1 Simulation Example

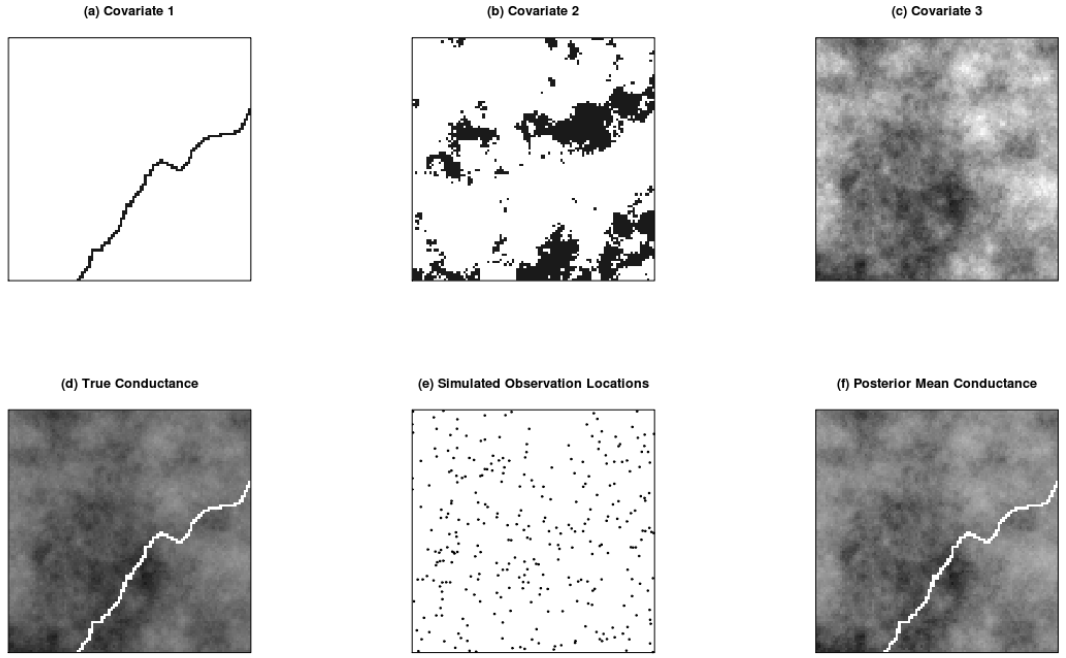

Figure 3.2: We simulated a resistance surface using two discrete cover types (a)-(b) and one continuous covariate (c). In the simulation study, Covariate 1 was set to impede connectivity (a), Covariate 2 was set to have no effect on connectivity (b), and Covariate 3 was set to facilitate connectivity (c). The resulting landscape conductance is shown in (d). Observations from the landscape were simulated (e), and inference on the contribution of each landscape covariate to connectivity was made (f).

Our approach for inference on gene flow across a heterogeneous landscape can be illustrated through a simulation example. A 100×100 landscape grid was created with a mix of continuous and discrete landscape characteristics (Figure 3.2). True values were assigned to the parameters $\boldsymbol{\beta}$ pertaining to the effect that each landscape characteristic has on gene flow, as well as for the nugget parameter τ. To simulate locations of organisms on the landscape, 300 completely spatial random locations were chosen and linked to grid cells. The covariance matrix $\mathbf{\Psi}_{\text{sim}}$ of these 300 locations was based on a rook's-neighborhood connectivity between grid cells (i.e., first order spatial neighbors) and the conductances $\{\alpha_{ij}\}$ specified by the landscape characteristics and true $\boldsymbol{\beta}$ values. We simulated ten realizations from a mean-zero Gaussian random field with covariance $\mathbf{\Psi}_{\text{sim}}$, and calculated the squared Euclidean distance between these realizations. The resulting pairwise distances between locations are proportional to the linearized pairwise F_{st} values that would be observed from the analysis of microsatellite alleles from ten loci of each organism.

The simulated pairwise distance matrix was used to fit the model (3.12)-(3.15). We used an MCMC algorithm to obtain 10,000 posterior samples of $\boldsymbol{\beta}$ and τ from four separate chains with distinct starting values. The posterior distributions for $\boldsymbol{\beta}$ and τ both indicate agreement with the true specified values (Table 3.1), providing evidence that the proposed model is able to recover resistance parameters in similar situations to that of our application data discussed in the following section.

Table 3.1: Simulation example results for connectivity parameters in the resistance surface.

Parameter	Truth	Posterior Mean	Lower 95% CI Bound	Upper 95% CI Bound
Intercept	2	1.364	0.744	2.015
Covariate 1	-5	-4.127	-6.340	-1.939
Covariate 2	0	0.040	-0.971	1.067
Covariate 3	1	1.386	0.815	1.901
Nugget	0.5	0.519	0.474	0.621

For comparison, we also employed a correlation analysis, the most common existing method in landscape genetics studies (e.g., Cushman et al., 2006, 2009; Wang and Xia, 2009; Shirk et al., 2010). This method requires specifying hypothesized resistance values, then computing the correlation between the observed genetic distance matrix and the resistance distance matrix that is the result of the hypothesized resistance values. We note that since this method relies on correlation, the intercept term in our model can be seen as a constant multiple of the distance matrices, and will have no effect on correlation. For the remaining three covariates, we employed a grid search of possible parameter values. Models with integer-valued parameters ranging from -6 to 3 were tried, resulting in 1000 candidate models for gene flow. The resulting 1000 hypothesized distance matrices were tested for correlation with the observed (simulated) genetic distance matrix, with the correlation being expressed as Mantel's z statistic as computed in the 'vegan' package (Oksanen et al., 2012) in the R statistical computing environment (R Core Team, 2013). The 5 models with the highest correlation are shown in Table 3.2. The model with the true parameter values used in the simulation was ranked 13 out of the 1000 hypothesized models, and is shown in bold in Table 3.2.

The correlation approach was able to identify the large negative effect of covariate 1, and correctly identifies that covariate 2 has no effect, but had some trouble with covariate 3. This highlights the need for a measure of uncertainty about the estimates of parameter values, as is obtained from the posterior credible intervals of resistance parameters $\boldsymbol{\beta}$ from the generalized Wishart model (Table 3.1).

Table 3.2: Simulation study results for the correlation method.

Model Rank	Mantel's z	Covariate 1	Covariate 2	Covariate 3
1	0.1450	-4	0	-1
2	0.1423	-4	0	-2
3	0.1421	-4	0	0
4	0.1416	-5	0	-1
5	0.1416	-5	0	0
⋮	⋮	⋮	⋮	⋮
13	**0.1378**	**-5**	**0**	**1**

3.6.2 Spatial Gene Flow in Alpine Chamois

The alpine chamois (*Rupicapra rupicapra*) is native to the mountains of Europe, and is a conserved species in France. Chamois live at moderately high altitudes and have adapted to living in steep, rocky terrain. Jombart (2012) have recently studied spatial gene flow among chamois in the Bauges mountains of France (Figure 3.3a). Their analysis, based on spatial principal component analysis (Jombart et al., 2008), concludes that gene flow in chamois may be impeded by lower elevation terrain (e.g., mountain valleys), possibly due to the increased risk of predation. The analysis of Jombart (2012) does not explicitly incorporate the landscape; rather, the results of the spatial principal component analysis are overlaid on the landscape, and *post hoc* inference is obtained for the effects of landscape characteristics (elevation in this case) on gene flow. To illustrate our approach to inference for gene flow based on circuit theory and the isolation by resistance hypothesis, we re-analyze the data of Jombart (2012) using our Bayesian model (3.12)-(3.15). This approach allows us to explicitly model and estimate the resistance that landscape characteristics (like elevation) have on gene flow in alpine chamois, something not possible using the approach of Jombart (2012).

Jombart (2012) studied microsatellite allele data from nine loci for 335 individual chamois. These data, together with spatial locations for each individual animal and elevation data on a 104×80 grid, where each grid cell measures 200m square, were obtained from the 'adegenet' package (Jombart, 2008) in the R statistical computing environment (R Core Team, 2013).

To examine the effect of elevation on gene flow in alpine chamois, we compared four competing models using DIC. The first model contained only an intercept term, hypothesizing no effect of elevation on gene flow. The second model contained an intercept term, as well as a linear effect for elevation in (3.13). The third model contained an intercept term, as well as linear and quadratic terms for elevation. The fourth model is a varying-coefficient model in elevation, in which we specified a B-spline basis expansion for the effect of elevation, using 7 equally spaced knots. This results in a flexible model of the effect that elevation has on gene flow. The varying-coefficient model could be difficult to implement using the correlation approach common in landscape genetics studies, as multiple hypothesized values for each of the parameters related to the spline basis functions would have to be specified.

We found the grid cell each animal was located in using the 'raster' package (Hijmans and van Etten, 2012) in R, and computed pairwise genetic distances between the chamois in the study using the mismatch genetic distance of Smouse and Peakall (1999). The resulting genetic distance matrix was used to fit each of the four models. For each model, we obtained 20,000 posterior samples using an MCMC algorithm from two separate chains with distinct starting values. Convergence was assessed using the potential scale reduction factor $\hat{R}$ (Gelman et al., 2004), which compares the within-chain variance to the between-chain variance. The resulting $\hat{R}$ values were less than 1.1 for all parameters in each of the four models, indicating convergence to the stationary posterior distribution.

To compare the four models, we used the deviance information criterion (DIC) of Spiegelhalter et al. (2002). The DIC of the varying-coefficient model for elevation was -2208, much lower than the DIC of the intercept-only model (DIC=8939), the linear model for elevation (DIC=7244), or the quadratic model for elevation (DIC=5963). This indicates strong support for the varying-coefficient model of the effect of elevation on gene flow. Posterior inference for the log-conductance of a grid cell as a function of elevation is shown in Figure 3.3(b), with upper and lower 95% credible limits. The posterior mean conductance of the study area is plotted in Figure 3.3(c). This plot reflects the nonlinear relationship between elevation and and gene flow. Higher elevation terrain (from about 1000m to 1750m) generally facilitates gene flow in alpine chamois, while lower elevations (valleys), and very high elevations (mountain peaks) impede gene flow. Jombart (2012) found that higher elevation terrain generally facilitates gene flow in alpine chamois using their spatial principal component analysis. Our model-based approach expands on their results by providing rigorous uncertainty estimates about the parameters related to gene flow, something not typically obtained in landscape genetics studies.

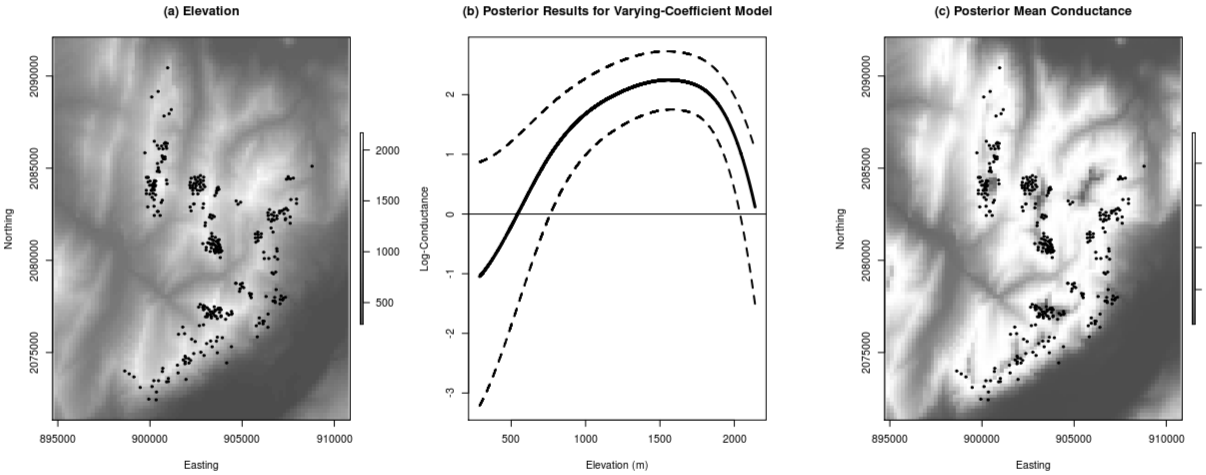

Figure 3.3: Spatial locations (a) of 335 alpine chamois (*Rupicapra rupicapra*) in the Bauges Mountains, France. A landscape genetic analysis of these chamois using a varying-coefficient model for the effect of elevation reveals that gene flow is impeded by low-elevation valleys and mountain peaks and facilitated by mid- to high-elevation terrain (b). The posterior mean conductance of the landscape to gene flow is shown in (c), with lower values corresponding to increased resistance to gene flow.

3.7 Discussion

Using circuit theory to model connectivity on a landscape requires viewing the continuous landscape as a graph $(\mathbf{G}, \mathbf{A})$. While this is a simplification, recent advances in GMRF theory have shown that GMRFs can be used to approximate solutions to stochastic partial differential equations on a continuous surface (Lindgren et al., 2011). The connectivity implied by circuit theory on a graph is analogous to a diffusion process on a continuous surface with heterogeneous diffusivity. Statistical models based on diffusion processes have been used extensively to model spatio-temporal processes on continuous surfaces (e.g., Hooten and Wikle, 2008; Wikle and Hooten, 2010), and could be used to model many of the systems that are currently modeled using circuit theory. The main advantage of viewing the landscape as a discrete graph is the ability to use computationally efficient GMRFs like the ICAR model (3.8) used in our analysis.

The link between circuit theory, GMRFs, and genetic distance allows for an efficient approach to simulate pairwise genetic distances on a heterogeneous landscape (e.g., Landguth and Cushman, 2010). Using (3.9), simulated resistance distances can be transformed to simulated linearized pairwise F_{st} values, which can provide significant information about population genetic structure across the landscape. This may be particularly useful in predicting the effects of landscape change on sub-population separation. For example, inference concerning parameter values $\boldsymbol{\beta}$ for a resistance surface related to gene flow could be made using the model (3.12)-(3.15). The effect of changes to the landscape, such as the addition of a road or the removal of existing open space, on sub-population separation could then be predicted using inferred values of $\boldsymbol{\beta}$ and a modified resistance surface reflecting the changes to the landscape.

While simulation is possible at extremely high resolutions, inference becomes increasingly computationally intensive as the landscape resolution increases. In practice, we have found that inference for landscape grids with $\mathcal{O}(10^5)$ grid cells is feasible on a machine with 4GB RAM and a 1.67 GHz quad-core processor, but inference for larger grids requires increased computing capabilities. This resolution is adequate for many studies, but larger landscapes may not be well-approximated by such a coarse discretization. Examining methods for increasing computational efficiency and thus allowing inference at higher resolutions is the subject of ongoing research.

The link between circuit theory and GMRFs has many potential applications. For example, agent-based models for a binary process on a graph (Hooten and Wikle, 2010) provide a highly flexible framework for modeling dynamic systems with discrete support. The evolution of agent-based systems is based on transition probabilities between neighboring nodes, and could be equivalently modeled as a resistance surface by envisioning a circuit with resistors connecting nodes. Circuit theory provides a flexible model for evaluating connectivity, and the link we have presented between circuit theory and GMRFs provides a natural framework that can provide inference for resistance values in systems modeled by circuits, and thus the methodology we propose here has wide applicability.

CHAPTER 4

LATENT SPATIAL MODELS AND SAMPLING DESIGN FOR LANDSCAPE GENETICS

4.1 Introduction

In Chapter 3, current approaches to studying landscape connectivity using spatially-referenced genetic data and connectivity models based on electric circuit theory were linked to Gaussian Markov random field spatial models. In this chapter, a general framework for the spatial modeling of microsatellite genetic data for landscape connectivity is proposed.

As was mentioned in Chapter 1 and Chapter 3, one common approach to estimating the effects of landscape on gene flow involves computing a pairwise genetic distance measure between observed spatially-referenced genetic samples, and comparing this pairwise distance with an effective distance that is a function of the spatial locations where the genetic samples were collected and the landscape features of the study region (e.g., McRae, 2006; Cushman et al., 2006; McRae and Beier, 2007; Cushman and Landguth, 2010; Cushman and Lewis, 2010). Under the isolation by distance (IBD) approach, the genetic distance is assumed to be correlated with Euclidean distance (e.g., Broquet et al., 2006). Under the isolation by resistance (IBR) and least cost path (LCP) approaches, the landscape is viewed as a graph with edge weights (resistances in the circuit interpretation) connecting neighboring cells being a function of the local landscape characteristics (e.g., McRae et al., 2008; Cushman and Lewis, 2010). The effective distance is either the effective resistance in the IBR approach, or the shortest single path distance in the LCP approach.

Two key challenges to estimating landscape effects on connectivity from observed genetic data are (1) the lack of spatially explicit statistical models for landscape genetic data in the literature and (2) the computational cost of obtaining the effective distances in the IBR and LCP approaches. We address each of these challenges by appealing to modern spatial statistical approaches (e.g., Cressie, 1993). In Section 4.2, we demonstrate how current techniques for parameter estimation in the IBD, IBR, and LCP approaches parallel the estimation of covariance parameters is spatial statistical models through variogram fitting. This expands on our work in Chapter 3, where we considered only the IBR case. In Section 4.3, we propose a spatial generalized linear mixed model (SGLMM) for microsatellite allele data, the most common form of genetic data used in current landscape genetics studies, and discuss how the IBD, IBR, and LCP approaches could be incorporated into this SGLMM. In Section 4.4, we illustrate our approach through a landscape genetic study of greater sage-grouse (*Centrocercus urophasiamus*) in the western United States, using genetic data collected during 2009-2012. In Section 4.5, we utilize optimal spatial sampling methodology to make recommendations for the allocation of sampling resources during 2013. In both Section 4.4 and 4.5, we rely on reduced rank approximations to the spatial process to improve computational efficiency. In Section 4.6, we discuss possible extensions to our approach.

4.2 Genetic Distance and Variogram Fitting

In this section, we briefly review variogram approaches to estimating covariance parameters in geostatistical models, and show how variogram estimation parallels current approaches for effective distance analysis in landscape genetics.

4.2.1 Variogram Approaches for Estimating Covariance Parameters

Consider the n observations $\mathbf{y} = (y_1, y_2, \ldots, y_n)'$ of a continuous (geostatistical) random field $\mathcal{Y}$ observed at the n (not necessarily distinct) spatial locations $\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_n$, where each $\mathbf{s}_i$ is contained in the study region $\mathcal{S}$. The variogram γ_{ij} between the spatial locations $\mathbf{s}_i$ and $\mathbf{s}_j$ is defined as the expected squared difference between the observations at different spatial locations:

$$\gamma_{ij} = E\left[(y_i - y_j)^2\right]. \tag{4.1}$$

We note that the variogram can be expressed as a function of the covariance. If $\mathbf{y} \sim N(\mathbf{0}, \mathbf{\Sigma})$, where the covariance matrix $\mathbf{\Sigma} = (\Sigma_{ij})$, then as discussed in Chapter 3,

$$\gamma_{ij} = \Sigma_{ii} + \Sigma_{jj} - 2\Sigma_{ij}. \tag{4.2}$$

If the covariance function is second-order stationary and isotropic, then the covariance Σ_{ij} between observations at $\mathbf{s}_i$ and $\mathbf{s}_j$ depends only on the Euclidean distance d_{ij} between the spatial locations (e.g., chapter 4.1 of Cressie and Wikle, 2011). In this case, it follows directly from (4.2) that the variogram γ_{ij} also depends only on the Euclidean distance between the spatial locations.

The empirical variogram $\hat{\gamma}_{ij}$ is calculated from observations $\mathbf{y}$ by approximating the expectation in (4.1) with a sum over repeated observations at the same locations, or by binning pairs of observations that are nearly equidistant in the stationary and isotropic case (e.g., Stein, 1999, page 81).

If the covariance is parameterized with $\boldsymbol{\theta}$ (e.g., $\mathbf{\Sigma} = \mathbf{\Sigma}(\boldsymbol{\theta})$), then estimating the parameters $\boldsymbol{\theta}$ in the Gaussian likelihood $[\mathbf{y}|\mathbf{\Sigma}(\boldsymbol{\theta})]$ requires inversion of the covariance matrix $\mathbf{\Sigma}(\boldsymbol{\theta})$. This inversion is numerically expensive and can be prone to numerical instability. A common alternative to estimating $\boldsymbol{\theta}$ is to fit the variogram model $\gamma_{ij}(\boldsymbol{\theta})$ to the observed empirical variogram $\hat{\gamma}_{ij}$ using ordinary least squares or weighted least squares. The procedure can be summarized in the following steps.

1. Calculate the empirical variogram $\hat{\gamma}_{ij}$ for all pairs of spatial locations $\mathbf{s}_i$ and $\mathbf{s}_j$.

2. Find $\hat{\boldsymbol{\theta}}$ that minimizes $\sum_{i,j}(\hat{\gamma}_{ij} - \gamma_{ij}(\boldsymbol{\theta}))^2$, or a weighted version of this loss function (See e.g., Cressie and Wikle, 2012, for details).

4.2.2 Variogram Approaches for Estimating Resistance Parameters

Estimation of the variogram parallels current approaches for estimating the parameters governing resistance between cells in the IBR and LCP approaches to landscape genetics, in which an observed pairwise genetic distance $\hat{g}_{ij}$ is related to an effective distance $d_{ij}(\boldsymbol{\theta})$. Contemporary approaches to estimating $\boldsymbol{\theta}$ can be summarized in the following steps.

1. Calculate an empirical measure of genetic dissimilarity $\hat{g}_{ij}$ for all pairs of spatial locations $\mathbf{s}_i$ and $\mathbf{s}_j$, for example using the fixation index (e.g., Holsinger and Weir, 2009) or the distance measure of Smouse and Peakall (1999).

2. Find $\hat{\boldsymbol{\theta}}$ that maximizes the correlation between $\hat{g}_{ij}$ and the effective distance $d_{ij}(\boldsymbol{\theta})$. In practice this is often accomplished by using a grid search and specifying "high", "medium" and "low" values for each element of $\boldsymbol{\theta}$ (e.g., Cushman et al., 2006).

For example, in the IBR case, $d_{ij}(\boldsymbol{\theta})$ is the resistance distance between the i^{th} and j^{th} locations, and in the LCP case, $d_{ij}(\boldsymbol{\theta})$ is the least-cost path distance between the i^{th} and j^{th} locations. Estimating the resistance parameters $\boldsymbol{\theta}$ parallels variogram approaches for estimating covariance parameters, where the effective distance $d_{ij}(\boldsymbol{\theta})$ is the variogram $\gamma_{ij}(\boldsymbol{\theta})$.

4.2.3 Spatial Covariance Models for Landscape Genetics

This relationship between variogram fitting in geostatistics and resistance parameter estimation in landscape genetics underscores the fact that the field of landscape genetics has developed in relative isolation from the existing field of spatial statistics, though the goal of landscape genetics (modeling correlation between genetic observations across space) lends itself naturally to a spatial statistical approach. We consider each of the IBD, IBR, and LCP approaches in turn, and describe spatial covariance functions that incorporate the assumptions about spatial variation in genetic diversity implied by each effective distance measure.

Covariance Models for Isolation by Resistance (IBR)

Under the IBR approach, the landscape is envisioned as an electric circuit, where each grid cell is a node and the resistors connecting neighboring cells have resistance which is a function of the local landscape characteristics (McRae, 2006). The resistance distance between two nodes in the circuit is the effective resistance of the entire circuit when a voltage is applied at those nodes. The resistance distance is a “multiple path” measure of distance, in that it allows for current to flow through all possible paths between nodes, and the resistance distance monotonically decreases with the addition of new pathways between nodes, regardless of how resistive the new pathways are. McRae (2006) note that under a random walk model for gene flow, the resistance distance is proportional to the linearized fixation index $F_{st} = \sigma_s^2/\sigma_t^2$, the ratio of the variance σ_s^2 in the frequency of alleles in different subpopulations and the variance σ_t^2 of allele frequencies in the total population. The IBR approach thus reflects a random walk model for gene flow across a heterogeneous landscape.

In Chapter 3, we described a link between the IBR approach and spatial statistics by showing that the resistance distance of the landscape circuit is exactly the variogram of an intrinsic conditional auto-regressive (ICAR) Gaussian Markov random field (GMRF) with edge weights equal to the inverse of the resistance between nodes in the circuit model. In that chapter, the genetic distance matrix $\mathbf{G}$ with entries $\hat{g}_{ij}$ was modeled using a generalized Wishart distribution (McCullagh, 2009). In Section 4.3 we will suggest modeling the observed allele data, instead of modeling the pairwise genetic distance matrix as we did in Chapter 3.

Covariance Models for the Isolation by Distance (IBD) Approach

Under the IBD approach, the key assumption is that individuals located close together in space (as measured by Euclidean distance) are likely to be more genetically similar than individuals located farther apart. Studies using the IBD approach typically examine the correlation between Euclidean distance and a genetic distance metric. This correlation analysis implies a linear relationship between genetic distance and Euclidean distance. However, simulations show that rather than increasing linearly with distance, simulated genetic distances first rise steeply with increasing Euclidean distance, then taper off to an asymptote (e.g., Graves et al., 2013). This tapering is common in many variogram models in geostatistics, and any stationary and isotropic covariance function has a variogram that is a function of Euclidean distance between observations. This suggests that the IBD approach could be modeled using a stationary covariance function. The Matern class of covariance functions (e.g., Stein, 1999, p. 31) contains the exponential, spherical, and squared-exponential (Gaussian) covariance functions, all of which have been used extensively to model stationary, isotropic spatial covariance in a wide variety of applications.

Covariance Models for the Least Cost Path (LCP) Approach

Under the LCP approach, the landscape is again envisioned as an electric circuit (or resistance surface), similar to in the IBR approach. Instead of utilizing the resistance distance between nodes (spatial locations) in the landscape graph, the LCP approach uses least-cost path distance, the cumulative resistance of the least resistive path between two nodes. While the IBR approach models gene flow by a random walk where only local landscape features are important, the LCP approach models gene flow under the assumption that it can always occur in the most efficient manner possible. The LCP approach is especially appealing for its computational efficiency which allows it to accomodate much larger landscape graphs than are possible in the IBR approach.

It is also worth noting that the IBR approach assumes that the correlation between genetic observations at two different spatial locations can be modelled by all possible pathways through the graph. The LCP approach, on the other hand, assumes that genetic correlation is only a function of the shortest path. This parallels the IBD approach and the associated stationary and isotropic covariance functions, though the resistance surface in the LCP approach is parameterized by landscape covariates, and is neither stationary nor isotropic.

One class of appropriate nonstationary covariance models for the LCP approach are spatial deformation approaches (e.g., Schmidt and O'Hagan, 2003), which assume that the process being modeled is stationary and isotropic in a transformed (deformed) geographic space. For the LCP approach, the spatial deformation could be parameterized by the landscape characteristics and the resulting resistance values of different grid cells in the landscape graph. We give additional discussion of spatial deformation approaches in Chapter 5.

4.3 Latent Spatial Models for Landscape Genetics

Having suggested Gaussian spatial covariance models for the IBD, IBR, and LCP approaches to landscape genetics, we now turn to the question of inference on parameters in these covariance models. As described in Section 4.2.1 and 4.2.2, common approaches for estimating resistance values for landscape characteristics in the IBR and LCP approaches parallel variogram fitting approaches to estimating covariance parameters in geostatistics, where the observations are Gaussian with appropriate covariance matrices. The pairwise genetic distance matrix can be thought of as an approximation to the empirical variogram matrix of a Gaussian random variable. In Chapter 3 we took this approach and modeled the genetic distance matrix using the generalized Wishart distribution of McCullagh (2009), which is an appropriate data model for the empirical variogram of a Gaussian random variable. However, it is unclear whether pairwise genetic distance measures (e.g., F_{st}) are good approximations to the variogram of a Gaussian random variable. A typical landscape genetic study involves microsatellite allele data, which come in the form of repeat lengths of short genome sequences. These data are not continuous, and the repeat lengths (alleles) are more naturally thought of as unordered categorical data, rather than ordered data. While continuous distance measures do exist for microsatellite allele data (e.g., Smouse and Peakall, 1999), it us unclear how well these continuous distance measures approximate variograms from Gaussian observations.

We instead propose a latent spatial model for spatially referenced allele data. In our specification, the categorical allele data are modeled using a multinomial probit model with latent Gaussian spatial random effects having a covariance appropriate for one of the IBD, IBR, or LCP approaches to landscape genetics.

Table 4.1: Subset of sage-grouse genetic data. Microsatellite allele data at 15 loci were obtained from 830 feather samples collected at sage-grouse leks across the western U.S. between 2009 and 2012. Data from five individual sage-grouse at two loci are shown below.

Sage-Grouse Sample	Locus 1a	Locus 1b	Locus 2a	Locus 2b
1510110	122	122	206	232
1510210	124	130	232	250
1510310	124	124	208	228
1510410	122	124	218	238
1510510	128	128	208	234

Consider microsatellite allele data observed at L distinct loci for each spatially referenced individual in the study. As an example, Table 4.1 shows microsatellite allele data at two loci for five individual sage-grouse.

At the ℓ^{th} locus, $\ell = 1, 2, \ldots, L$, denote the list of all distinct observed alleles from all individuals in the study as $\{a_{\ell 1}, a_{\ell 2}, \ldots, a_{\ell K_\ell}\}$. We consider the two observed alleles for each (diploid) individual to arise from a multinomial distribution with spatially varying allele probabilities $\mathbf{p}_{s\ell} = (p_{s\ell 1}\ p_{s\ell 2}\ \ldots\ p_{s\ell K_\ell})'$, where $s \in \{1, 2, \ldots, S\}$ indexes the spatial location.

Let $y_{sip\ell k} = 1$ if the p^{th} (indexing ploidy) observed allele at the ℓ^{th} locus is $a_{\ell k}$ for the i^{th} individual at the s^{th} spatial location, and $y_{sip\ell k} = 0$ otherwise. Then the multinomial probit model (e.g., Albert and Chib, 1993) for categorical data is often specified in terms of latent variables, $\mathbf{z}$, as follows. Let

$$y_{sip\ell k} = \begin{cases} 1 & , \ z_{sip\ell k} = \max\{z_{sip\ell a},\ a = 1, \ldots, K_\ell\} \\ 0 & , \ \text{o.w.} \end{cases} \tag{4.3}$$

where

$$z_{sip\ell k} \sim N(\mu_{\ell k} + \eta_{s\ell k}, 1). \tag{4.4}$$

Then the allele $a_{\ell k}$ makes up a fraction $p_{s\ell k}$ of the genetic makeup of the subpopulation at location s, where

$$p_{s\ell k} = \text{Prob}\left(z_{sip\ell k} = \max\{z_{sip\ell a},\ a = 1, \ldots, K_\ell\}\right)$$

The mean of the latent variable in (4.4) consists of the sum of two effects. The first is $\mu_{\ell k}$, an allele specific intercept which determines the relative frequency of the k^{th} allele at the ℓ^{th} locus across the entire population being studied. Large values of $\mu_{\ell k}$, relative to $\mu_{\ell k'}$ make it more likely that $z_{sip\ell k}$ will be larger than $z_{sip\ell k'}$, and so the k^{th} allele will be more prevalent than the $(k')^{\text{th}}$ allele. We note that the model (4.3)-(4.4) is invariant to a shift in all $\mu_{\ell k}$, as the likelihood is a function of the contrasts $z_{sip\ell k} - z_{sip\ell k'}$, and not the actual values of $z_{sip\ell k}$. Thus, if $\mu_{\ell k}$ were replaced by $\mu_{\ell k} + c$ for $k = 1, 2, \ldots, K_\ell$ and some constant c, the likelihood of the observed allele data would remain unchanged. To maintain model identifiability, we fix $\mu_{\ell 1} = 0$ for $\ell = 1, 2, \ldots, L$, as only the relative differences (contrasts) in $\mu_{\ell k}$ are identifiable.

The second term in the mean of (4.4) is $\eta_{s\ell k}$, which is a spatially varying random effect that allows the allele frequencies $\mathbf{p}_{s\ell}$ to vary over the spatial range of the species. Consider:

$$\boldsymbol{\eta}_{\ell k} = \begin{bmatrix} \eta_{1\ell k} \\ \eta_{2\ell k} \\ \vdots \\ \eta_{n\ell k} \end{bmatrix} \sim N(\mathbf{0}, \boldsymbol{\Sigma}(\boldsymbol{\theta})) \tag{4.5}$$

where $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ is the spatial covariance matrix of the spatially referenced locations for which we have observations, parameterized by $\boldsymbol{\theta}$. Then $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ can be specified using a covariance appropriate for the IBD, IBR, LCP or other landscape genetic approaches. For example, under the IBR approach, $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ could be an ICAR covariance matrix, with $\boldsymbol{\theta}$ being parameters related to the resistance of different landscape characteristics to gene flow.

As an illustration of our modeling approach, Figure 4.1 shows an example of categorical data simulated from the model (4.3-4.5), where the spatial support is discrete and defined by a graph of twenty nodes (locations) taken randomly from the unit square. Edges were allowed between any two nodes separated by less than a Euclidean distance of 0.2, and an ICAR spatial model was used to simulate three latent continuous random variables ($\mathbf{z}_1$, $\mathbf{z}_1$, and $\mathbf{z}_1$). Each of these latent variables represents a latent allelic process, and the observed categorical allele at a node is given by the index of the largest realization among the latent $\mathbf{z}$ variables at that node.

We adopt a Bayesian approach to modelling, and specify prior distributions for all parameters in (4.3)-(4.5):

$$\mu_{\ell k} \sim N(0, \sigma_\mu^2) \ , \ \ell = 1, 2, \ldots, L \ , \ k = 2, 3, \ldots, K_\ell \tag{4.6}$$

$$\boldsymbol{\theta} \sim [\boldsymbol{\theta}] \tag{4.7}$$

where the bracket notation $[\cdot]$ indicates a probability distribution. The prior on $\boldsymbol{\theta}$ will depend on the covariance model used, and we thus leave the prior specification (4.7) in a general form.

In treating our observations as arising from a multinomial distribution, we make the implied assumption that we have observed all possible alleles at each locus. If this is not the case, then the interpretation of $p_{s\ell k}$ changes to be the relative probability of observing the k^{th} allele at the ℓ^{th} locus of an individual at the s^{th} spatial location, *given* that one of $\{a_{\ell 1}, a_{\ell 2}, \ldots, a_{\ell K_\ell}\}$ are observed. Our interest is in estimating the spatial dependence in the observed alleles, as encoded in the latent covariance $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ of the allelic processes $\{\mathbf{z}_{ip\ell k}\}$. Under the assumption that $\mathbf{z}_{ip\ell k}$ is independent of $\mathbf{z}_{ip\ell k'}$, $k \neq k'$, ignoring the potential existence of unobserved alleles will not bias our results.

We have not included fixed covariate effects in our latent model (4.4), rather we have only modeled population level allele prevalence ($\mu_{\ell k}$) and latent spatial random variation $\eta_{s\ell k}$. If

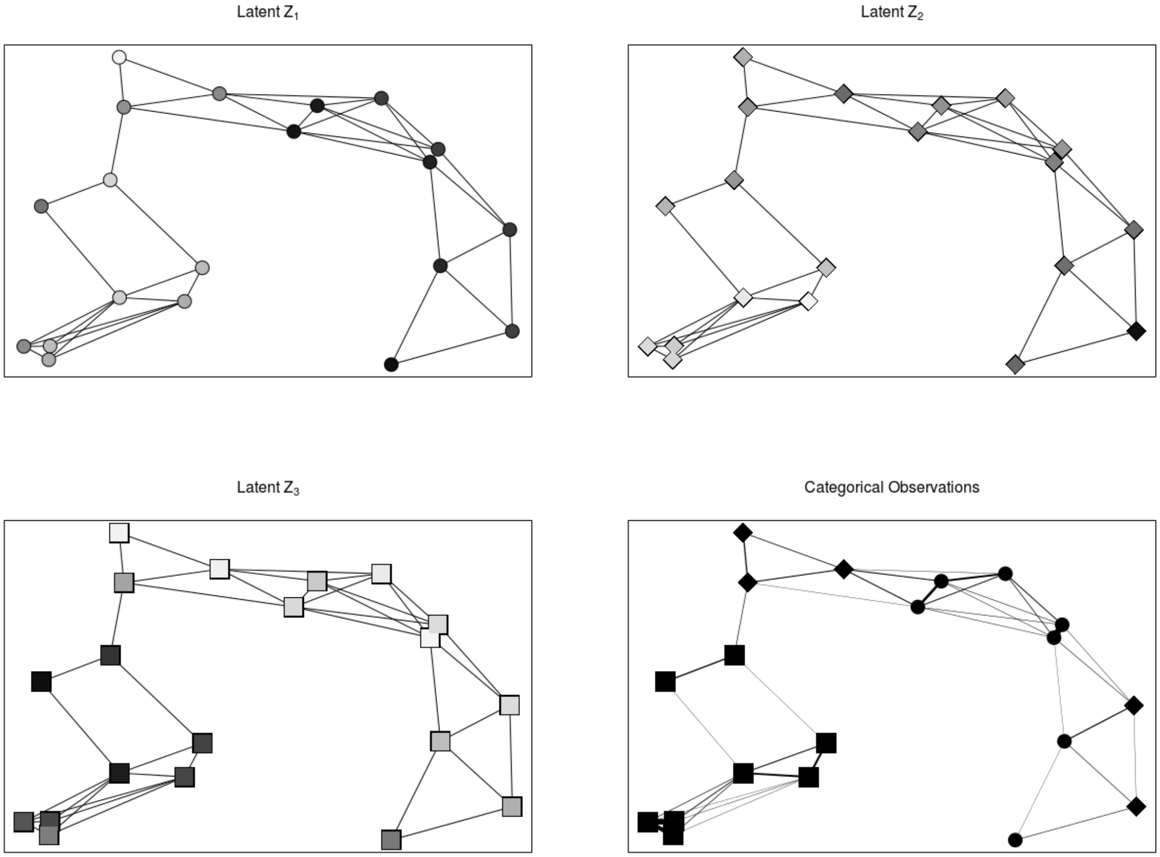

Figure 4.1: Simulation example of the Gaussian latent process model for categorical microsatellite allele data. Three realizations ($\mathbf{z}_1$, $\mathbf{z}_2$, and $\mathbf{z}_3$) were obtained from a specified ICAR GMRF on the graph shown. Each realization represents a latent process for a distinct allele, with larger values being shown as darker colors. The categorical allele observed at a node is given by the index of the latent allelic process which is the greatest at that node. The categorical observations are represented by the shape (circle=allele 1, diamond=allele 2, and square=allele 3) of the associated $\mathbf{z}_i$.

it is believed that the loci used in the analysis are from genes that might select for different landscape features, then a linear predictor $\mathbf{x}_s'\boldsymbol{\beta}_{\ell k}$ could be included in the mean of (4.4), where $\mathbf{x}_s$ is a vector containing landscape characteristics at the s^{th} spatial location, and $\boldsymbol{\beta}_{\ell k}$ are allele specific regression parameters which provide indications how different alleles might select for different landscape characteristics. Typical landscape genetic studies, however, seek to identify loci that are on non-coding and non-selecting areas of the genome. In this situation, there is no reason to assume that allele prevalence would be linked to the landscape features of the spatial locations where the individuals are observed.

We also assume in (4.5) that the latent allelic spatial effects $\boldsymbol{\eta}_{la}$ have the same spatial covariance matrix $\boldsymbol{\Sigma}$ for each locus and allele. This implies that the processes driving spatial variation in allele frequencies are the same for each locus. If the loci in question are from non-coding regions of the genome and have similar mutation rates, then this assumption is likely to be met, as the remaining processes driving gene flow (e.g., mating, survival, and movement) should be unaffected by the genotype at a non-coding locus on the genome. If mutation rates are highly variable between loci, then (4.5) could be generalized to allow for loci-specific covariance matrices (e.g., $\boldsymbol{\eta}_{\ell k} \sim N(\mathbf{0}, \boldsymbol{\Sigma}_\ell)$).

The model in (4.3)-(4.7) is a Bayesian hierarchical model (BHM) that takes into account the categorical nature of microsatellite allele data and allows for spatial variation in genetic diversity to be modeled using existing landscape genetic approaches. To our knowledge, this has not been described in the existing literature.

Modeling spatial genetic variation in terms of latent Gaussian random variables with spatial correlation (4.5) allows us to draw on the existing field of spatial statistics to approach landscape genetic studies. In Section 4.4, we make use of modern dimension reduction techniques in spatial statistics to increase the computational efficiency of the algorithms used to make inference on the model (4.3-4.7). In Section 4.5, we apply existing methodology for optimal spatial sampling of Gaussian random variables to suggest optimal sampling designs for landscape genetic studies.

4.4 Landscape Genetic Analysis for Sage-Grouse in the Western United States

In the previous section we specified a multinomial model for spatially referenced microsatellite allele data with latent spatial autocorrelation. We now illustrate the utility of this approach through a landscape genetic study of greater sage-grouse (*Centrocercus urophasiamus*) in the western United States. While their historical range covered 12 western states in the U.S., the greater sage-grouse now occupies only 56% of its historical range (Schroeder et al., 2004). This rapid decline in range has also been paralleled by a rapid decline in population (Connelly and Braun, 1997). The distribution of sage-grouse genetic variability across the range of the species could provide insight into regions of particular concern for conservation of the species.

From 2009-2012, feather samples were collected from sage-grouse leks (activity centers) in 11 western states (Figure 4.2). After the 2012 season, the feather samples were genotyped, and microsatellite allele data were collected at 15 disctinct loci. We analyze these data using the modeling framework described in Section 4.3. In Section 4.5 we will use the results of the analysis in this section to make recommendations for retrospective optimal sampling of sage-grouse leks during the 2013 data collection season.

4.4.1 Lek Network Connectivity Model

Sage-grouse leks are often found in natural openings in sagebrush communities, and are typically surrounded by potential nesting habitat (e.g., Connelly et al., 1991; Wakkinen et al., 1992; Connelly et al., 2004). During breeding seasons, male sage-grouse display at leks during crepuscular periods, and leks can be remarkably persistant, with some leks remaining active for over 28 years (Wiley, 1973).

Leks act as persistent activity centers for sage-grouse, and we will use this interpretation to motivate an IBR approach to landscape connectivity modeling. Consider a spatial network with nodes at known sage-grouse leks across the western U.S. and edge weights (resistances in the IBR framework) being a function of the Euclidean distance between leks.

For the s^{th} lek, consider the set of "neighboring" leks indexed by t: $\{t : t \in \mathcal{N}(s)\}$, where $\mathcal{N}(s)$ is the set of leks that are directly connected (neighbors) to the s^{th} lek. We consider lek s and lek t to be neighbors if it is judged to be scientifically reasonable for a sage-grouse

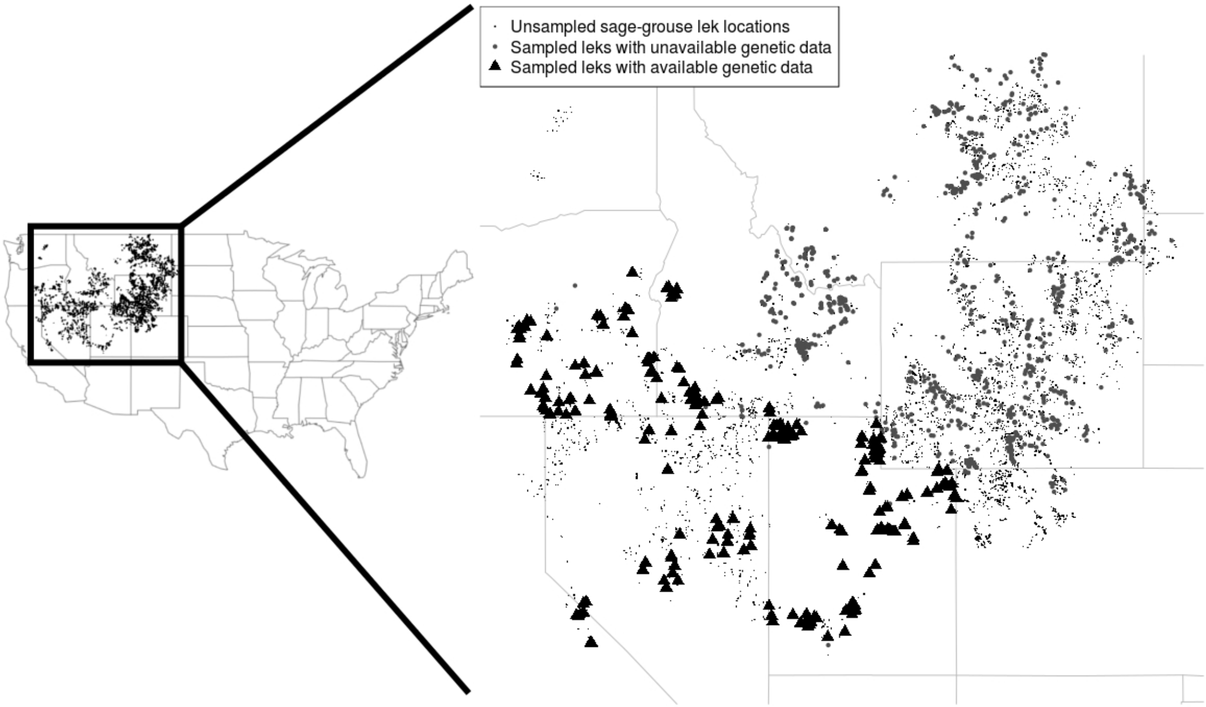
Unsampled sage-grouse lek locations
Sampled leks with unavailable genetic data
Sampled leks with available genetic data

Figure 4.2: Sage-grouse lek locations in the western United States. Feather samples were collected from over 1,000 sage-grouse leks between 2009 and 2012. Microsatellite allele data from 237 of these leks were available for analysis in March, 2013.

to migrate from lek s to lek t directly, without an intermediate stop at any other lek. This is based on the migration model for movement central to the IBR approach as described by McRae (2006).

Define the edge weight between the s^{th} and t^{th} leks as α_{st}/σ^2, where α_{st} is a decreasing function of the Euclidean distance between the leks and σ^2 is a scaling factor. Under the random walk model for migration in the IBR framework, α_{st}/σ^2 is proportional to the migration rate of sage grouse between the two leks. We consider edge weights that are decreasing functions $f(d_{st})$ of the Euclidean distance d_{st} between leks, and that are set to 0 for all leks that are more distant than a maximum distance d_{MAX}:

$$\alpha_{st} = \begin{cases} f(d_{st}) & , \ d_{st} \leq d_{\text{MAX}} \\ 0 & , \ d_{st} > d_{\text{MAX}} \end{cases} . \tag{4.8}$$

In our following analysis, we set $d_{\text{MAX}} = 25\text{km}$, where the distance between leks was computed using great circle distance, and consider multiple functional forms for $f(d)$, described in Section 4.4.4. Setting edge weights equal to zero for leks that are more distant than d_{MAX} indicates that migration (and the resulting flow of genetic information) is only possible between distant leks through the use of intermediate leks in the lek network. This is a Markov assumption, and our resulting random effects $\{\boldsymbol{\eta}_{\ell k}\}$ in (4.5) are Gaussian Markov random fields.

In Chapter 3, we have shown that the spatial connectivity implied by such a random walk migration model (or an equivalent electric circuit) can be modeled as an intrinsic conditional autoregressive (ICAR) Gaussian Markov random field (Besag, 1974; Besag and Kooperberg, 1995). The spatial precision matrix (inverse covariance matrix) for an ICAR on the entire lek network is given by

$$\mathbf{Q} = \frac{1}{\sigma^2} \begin{bmatrix} \sum_{j \neq 1} \alpha_{1j} & -\alpha_{12} & -\alpha_{13} & \cdots \\ -\alpha_{21} & \sum_{j \neq 2} \alpha_{2j} & -\alpha_{23} & \cdots \\ -\alpha_{31} & -\alpha_{32} & \sum_{j \neq 3} \alpha_{3j} & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix} . \tag{4.9}$$

We do not have observed genetic data for all known lek locations. If we divide the leks into observed (o) and unobserved (u) locations, the spatial random effects $\boldsymbol{\eta}$ (suppressing the locus ℓ and allele a subscripts) and precision matrix Q can be partitioned accordingly:

$$\boldsymbol{\eta} = \begin{bmatrix} \boldsymbol{\eta}_o \\ \boldsymbol{\eta}_u \end{bmatrix} \quad , \quad \mathbf{Q} = \frac{1}{\sigma^2} \left[\begin{array}{c|c} \mathbf{Q}_{oo} & \mathbf{Q}_{ou} \\ \hline \mathbf{Q}_{uo} & \mathbf{Q}_{uu} \end{array} \right]$$

and the covariance matrix $\boldsymbol{\Sigma}(\sigma^2)$ of the observed leks is given by the Schur complement:

$$\boldsymbol{\Sigma}(\sigma^2) = \sigma^2[\mathbf{Q}_{oo} - \mathbf{Q}_{ou}(\mathbf{Q}_{uu})^{-1}\mathbf{Q}_{uo}]^{-1}. \tag{4.10}$$

We assign a conjugate inverse-gamma prior distribution to σ^2:

$$\sigma^2 \sim IG(r, q) \tag{4.11}$$

with r and q chosen so that the prior has mean of 10 and variance of 100. Inference on the parameters ($\{\mu_{\ell k}\}$ and σ^2) in the full model, given by Equations (4.3)-(4.6) and (4.9)-(4.11), can then be made using a Markov chain Monte Carlo (MCMC) algorithm under a Bayesian statistical paradigm.

4.4.2 Reduced Rank Spatial Model

There are a large number ($\approx 6,000$) of known sage-grouse leks across the western U.S. This results in a large ($\approx$ 6000 × 6000) spatial precision matrix $\mathbf{Q}$ representing pairwise connections between leks. The magnitude of this spatial precision matrix will make model fitting and determination of optimal sampling effort very computationally intensive. We thus propose a reduced rank model for the spatial random effect $\boldsymbol{\eta}_{\ell k}$ in (4.5). Our reduced rank approach utilizes a spectral decomposition of the spatial precision matrix $\mathbf{Q}$ and is similar to that of Wikle and Cressie (1999); Berliner et al. (2000) and others.

For a fully observed lek network, $\boldsymbol{\eta}_{\ell k} \sim N(\mathbf{0}, \sigma^2\mathbf{Q}^-)$, where $\mathbf{Q}$ is an $n_s \times n_s$ precision matrix. Taking the spectral decomposition of $\mathbf{Q}$ gives

$$\mathbf{Q} = \mathbf{M}\mathbf{D}^-\mathbf{M}',$$

where the columns of $\mathbf{M}$ are the eigenvectors of $\mathbf{Q}$ and $\mathbf{D}$ is a diagonal matrix containing the respective eigenvalues $\{\lambda_1, \lambda_2, \ldots, \lambda_{n_s}\}$ of $\mathbf{Q}^-$. Then

$$\begin{aligned}\boldsymbol{\eta}_{\ell k} &= \mathbf{M}(\mathbf{M}'\boldsymbol{\eta}_{\ell k}) \\ &= \mathbf{M}\boldsymbol{\delta}_{\ell k}\end{aligned}$$

where

$$\boldsymbol{\delta}_{\ell k} \sim N(\mathbf{0}, \sigma^2\mathbf{D}).$$

Then a reduced rank version of the spatial random effect is given by taking only the first n_e eigenvectors of $\mathbf{Q}^-$ and setting $\lambda_m = 0$ for all $m > n_e$. This approximates the random effect $\boldsymbol{\eta}_{\ell k}$ with a random effect $\tilde{\boldsymbol{\eta}}_{\ell k}$ that captures a portion of the spatial structure in $\mathbf{Q}^-$. In practice, n_e can often be picked so that $\tilde{\boldsymbol{\eta}}_{\ell k} \approx \boldsymbol{\eta}_{\ell k}$ but n_e is still much smaller than n_s, the number of leks in the network. This leads to a computationally efficient representation of the spatial random effect, with (4.5) becoming:

$$\tilde{\eta}_{s\ell k} = \tilde{\mathbf{u}}_s'\tilde{\boldsymbol{\delta}}_{\ell k} \tag{4.12}$$

$$\tilde{\boldsymbol{\delta}}_{\ell k} \sim N(\mathbf{0}, \sigma^2\tilde{\mathbf{D}}) \tag{4.13}$$

where $\tilde{\mathbf{u}}_s'$ is the s^{th} row of $\tilde{\mathbf{M}}$, the $n_s \times n_e$ matrix of the first n_e eigenvectors of $\mathbf{Q}^-$, and $\tilde{\mathbf{D}}$ is the $n_e \times n_e$ diagonal matrix of the first n_e eigenvalues of $\mathbf{Q}^-$.

4.4.3 Full-Conditional Distributions

The complete hierarchical statistical model we have described for the multinomial allele model with latent reduced rank spatial random effects is

$$\begin{aligned}
y_{sip\ell k} &= \begin{cases} 1 & , \; z_{sip\ell k} = \max\{z_{sip\ell a}, \; a = 1, \ldots, K_\ell\} \\ 0 & , \; \text{o.w.} \end{cases} \\
z_{sip\ell k} &\sim N(\mu_{\ell k} + \tilde{\eta}_{s\ell k}, 1) \\
\tilde{\eta}_{s\ell k} &= \tilde{\mathbf{u}}_s' \tilde{\boldsymbol{\delta}}_{\ell k} \\
\tilde{\boldsymbol{\delta}}_{\ell k} &\sim N(\mathbf{0}, \sigma^2 \tilde{\mathbf{D}}) \\
\mu_{\ell k} &\sim N(0, \tau^2) \\
\sigma^2 &\sim IG(r, q)
\end{aligned}$$

Full-conditional distributions are available for all parameters in this hierarchical model. As noted by Albert and Chib (1993), the full conditional for $z_{sip\ell k}$ is truncated normal:

$$z_{sip\ell k} | \cdot \sim \text{TN}_{\mathcal{D}_{sip\ell k}}(\mu_{\ell k} + \tilde{\eta}_{s\ell k}, 1)$$

where $TN_{\mathcal{D}}(\cdot, \cdot)$ indicates a truncated normal distribution with support $\mathcal{D}$ and

$$\mathcal{D}_{sip\ell k} = \begin{cases} (0, \infty) & y_{sip\ell k} = 1 \\ (-\infty, 0) & y_{sip\ell k} = 0 \end{cases} .$$

The full-conditional distribution for $\tilde{\delta}_{\ell k}$ is normally distributed:

$$\tilde{\delta}_{\ell k} \sim N(\mathbf{A}^{-1} \mathbf{b}_{\ell k}, \mathbf{A}^{-1})$$

where

$$\mathbf{A} = 2\tilde{\mathbf{M}}'\mathbf{K}'\mathbf{K}\tilde{\mathbf{M}} + \frac{1}{\sigma^2}\tilde{\mathbf{D}}^{-1}$$

and

$$\mathbf{b}_{\ell k} = \tilde{\mathbf{M}}'\mathbf{K}'(\mathbf{z}_{1\ell k} + \mathbf{z}_{2\ell k} - 2\mu_{\ell k}).$$

The matrix $\mathbf{K}$ is a design matrix linking the individual sage-grouse to the lek. Then $\mathbf{K}'\mathbf{K}$ is a diagonal matrix with the s^{th} diagonal entry equal to the number of animals sampled at the s^{th} lek.

The full-conditional distribution for $\mu_{\ell k}$ is also normally distributed

$$\mu_{\ell k}|\cdot \sim N\left(\frac{h_{\ell k}}{c}, \frac{1}{c}\right)$$

with

$$\frac{1}{c} = \frac{1}{\tau^2} + \sum_s \sum_i \sum_p 1$$

and

$$h_{\ell k} = \sum_s \sum_i \sum_p \left(z_{sip\ell k} - \tilde{\eta}_{s\ell k}\right),$$

where the sums are taken over all possible values of s, i, and p.

The full-conditional distribution for σ^2 is inverse-gamma:

$$\sigma^2|\cdot \sim IG\left(\left(\frac{1}{r} + \frac{1}{2}\sum_\ell \sum_k \tilde{\boldsymbol{\delta}}'_{\ell k}\tilde{\mathbf{D}}^{-1}\tilde{\boldsymbol{\delta}}_{\ell k}\right)^{-1}, q + \frac{1}{2}\sum_\ell N_\ell\right).$$

The existence of full-conditional distributions for all model parameters allows for straightforward sampling from the posterior distribution of model parameters using the Gibbs sampler (e.g., Gelman et al., 2004).

4.4.4 Model Fitting

For our analysis of sage-grouse genetic data, two functional forms for $f(d)$ in (4.8) were specified, inverse distance ($f_1(d) = 1/d$), and inverse squared distance ($f_2(d) = 1/d^2$). The resulting spatial models were fit to the observed data and compared using the deviance information criterion (DIC) of Spiegelhalter et al. (2002). As mentioned previously, we set the maximal distance d_{MAX} equal to 25km.

The resulting models were each fit to the observed allele data from 830 feather samples collected at leks across the western U.S. Under a Bayesian statistical paradigm, an MCMC algorithm was used to obtain samples from the posterior distribution of parameters in each network model, conditioned on the observed sage-grouse genotype data. Multiple chains were simulated for each model, and convergence was assessed visually, with chains showing good mixing. The resulting samples were used to make inference on σ^2.

The two lek network models were compared using DIC. The best model (DIC=526,648) was the network model with edge weight between leks inversely proportional to the distance between leks ($f_1(d) = 1/d$). The network model with edge weight between leks inversely proportional to the square of the distance between leks ($f_2(d) = 1/d^2$) had a higher DIC (DIC=529,167). For the following analysis, we focus only on the model with the best (lowest) DIC.

The posterior distribution for σ^2 in the best model has mean of 0.169 and variance of 0.015. This quantity is not easily interpreted directly, but the latent representation of $z_{sip\ell k}$ lends an interpretation to the latent spatial covariance matrix $\tilde{\mathbf{\Sigma}}(\sigma^2)$. From (4.4),

$$z_{sip\ell k} = \mu_{\ell k} + \tilde{\eta}_{s\ell k} + \epsilon_{sip\ell k} \quad , \quad \epsilon_{sip\ell k} \sim N(0, 1)$$

where $\tilde{\eta}_{\ell k} \sim N(\mathbf{0}, \tilde{\mathbf{\Sigma}}(\sigma^2))$ and $\epsilon_{sip\ell k} \sim N(0, 1)$. The spatial covariance matrix $\tilde{\mathbf{\Sigma}}(\sigma^2)$ is nonstationary, and a summary of the posterior mean diagonal elements of $\tilde{\mathbf{\Sigma}}(\sigma^2)$ is given in Table 4.1. The diagonal elements of this spatial covariance matrix offer a comparison with the unit variance of the non-spatial noise (nugget) contained in $\epsilon_{sip\ell k}$ of (4.4). As the posterior mean diagonal elements of the spatial covariance matrix are greater than 1 by multiple orders of magnitude, we can conclude that spatial variability in the genetic data accounts for a much larger proportion of the variability than does non-spatial (within lek) variability. Summaries of posterior distributions for $\{\mu_{\ell k},\ \ell = 1, 2, \ldots, 15,\ k = 1, 2, \ldots, K_\ell\}$ are not shown, but correlate well with empirical allele frequencies taken over all 1,136 genetic samples.

Table 4.2: Summary of posterior mean diagonal elements of the spatial covariance matrix $\tilde{\boldsymbol{\Sigma}}$ for the selected network model.

1st Qu.	Median	Mean	3rd Qu.	Max.
980.80	1586.00	3947.00	4993.00	49020.00

4.5 Optimal Retrospective Sampling for Sage-Grouse in the Western United States

Having fit the model to microsatellite allele data from sage-grouse feather samples collected in 2009-2012, we now consider retrospective sampling design for the 2013 season. Our goal is to recommend optimal regions for sampling sage-grouse leks, given what is known from the spatially referenced genetic data already collected. This requires the definition of a criterion $\phi(\mathbf{d}_j)$ by which we can compare J potential sampling designs ($\mathbf{d}_j,\ j = 1, \ldots, J$).

4.5.1 Latent Gaussian Design

A common criterion to minimize for optimal sampling of Gaussian random variables is the mean squared prediction error (MSPE) at unobserved locations (e.g. Zimmerman, 2006). The observations $\{y_{sip\ell k}\}$ in our model (4.3)-(4.7) are not Gaussian; rather, they are categorical alleles which we model as arising from a multinomial distribution. However, each $z_{sip\ell k}$ is a latent representation of the true observation $y_{sip\ell k}$, and $z_{sip\ell k}$ is normally distributed:

$$z_{sip\ell k} = \mu_{\ell k} + \eta_{s\ell k} + \epsilon_{sip\ell k} \quad , \quad \epsilon_{sip\ell k} \sim N(0,1).$$

This leads us to consider the MSPE of the latent z variables as a design criterion. If we suppress all subscripts on z, except for the spatial index s, we can divide the spatial locations, and their corresponding z variables, into observed ($\mathbf{z}_o$) and unobserved ($\mathbf{z}_u$) categories. The joint distribution of $\mathbf{z}$ at all leks is Gaussian:

$$\mathbf{z} = \begin{pmatrix} \mathbf{z}_o \\ \mathbf{z}_u \end{pmatrix} \sim N\left(\mu\mathbf{1}, \begin{pmatrix} \mathbf{\Sigma}_{oo} & \mathbf{\Sigma}_{ou} \\ \mathbf{\Sigma}_{uo} & \mathbf{\Sigma}_{uu} \end{pmatrix}\right) \tag{4.14}$$

where the covariance matrix $\mathbf{\Sigma} = \sigma^2\mathbf{R}+\mathbf{I}$ in (4.14) has been partitioned according to observed and unobserved locations.

If $\mathbf{z}_o$ are known (observed), then the simple Kriging predictions at the unobserved locations are given by:

$$\hat{\mathbf{z}}_u = \mathbf{\Sigma}_{uo}\mathbf{\Sigma}_{oo}^{-1}\mathbf{z}_o \tag{4.15}$$

and the simple Kriging covariance matrix is given by:

$$\mathbf{\Psi} = E\left[(\mathbf{z}_u - \hat{\mathbf{z}}_u)(\mathbf{z}_u - \hat{\mathbf{z}}_u)'\right] = \mathbf{\Sigma}_{uu} - \mathbf{\Sigma}_{uo}\mathbf{\Sigma}_{oo}^{-1}\mathbf{\Sigma}_{ou}. \tag{4.16}$$

Each diagonal entry of $\mathbf{\Psi}$ contains the MSPE for z at an unobserved location, and we choose the sum of these entries (A-optimality in Harville (2008)) for our design criterion:

$$\phi(\mathbf{d}) = \mathrm{tr}(\mathbf{\Psi}) = \mathrm{sum}\left(\mathrm{diag}(\mathbf{\Psi})\right). \tag{4.17}$$

.

It is notable that neither the covariance matrix $\mathbf{\Psi}$ or the design criterion $\phi(\mathbf{d})$ depend on $\mathbf{z}_o$. Rather, they depend only on the covariance matrix $\mathbf{\Sigma}$ of the latent z variables. This allows for efficient comparison of the design criterion ϕ for different designs, as changing a design (e.g., adding a new lek to the set of previously sampled leks) only involves evaluating (4.16) and (4.17) for a different partition of locations into "observed" and "unobserved" in (4.14). This approach to optimal spatial sampling for Gaussian data is well known and has been used effectively in many systems (e.g., Wikle and Royle, 2005; Zimmerman, 2006; Hooten et al., 2009).

4.5.2 Design for Categorical Observations

Predicting the categorical allele response y at an unsampled lek is also possible, and an optimal sampling design could be obtained by minimizing the MSPE of the categorical response. However, the computational cost would be significantly greater than the computational cost of the latent approach of Section 4.5.1. For example, consider predicting the multinomial response $\mathbf{y}_{u\ell} = (y_{u\ell 1}, \ldots, y_{u\ell K})'$ for the ℓ^{th} locus at one unobserved location, conditioned on the actual observed responses $\mathbf{Y}_o = \{\mathbf{y}_{o\ell}\ ,\ o \text{ observed }\}$ at all observed locations. The MSPE is given by

$$E\left(\sum_{k=1}^{K}(y_{u\ell k} - \hat{y}_{u\ell k})^2\right) = P(\mathbf{y}_{u\ell} \neq \hat{\mathbf{y}}_{u\ell}|\mathbf{Y}_o) = P(k_{u\ell} = \hat{k}_{u\ell}) \tag{4.18}$$

where $k_{u\ell}$ and $\hat{k}_{u\ell}$ are the indices of the maximum $z_{u\ell a}$ and $\hat{z}_{u\ell a}$, respectively (e.g., $z_{u\ell k} = \max\{z_{u\ell a},\ a = 1, \ldots, K_\ell\}$ and $\hat{z}_{u\ell\hat{k}} = \max\{\hat{z}_{u\ell a},\ a = 1, \ldots, K_\ell\}$). The calculation of this design criterion can be accomplished by utilizing the latent Gaussian variables $\{z_{u\ell k}\}$. Each $z_{u\ell k}$ is marginally distributed as

$$z_{u\ell k} \sim N(\hat{z}_{u\ell k}, \psi_u^2)$$

where $\hat{z}_{u\ell k}$ is given by (4.15) and ψ_u^2 is the u^{th} diagonal of $\mathbf{\Psi}$ in (4.16). While the latent z random variables are correlated in space, they are uncorrelated between alleles. That is, $z_{u\ell k}$ is independent of $z_{u\ell k'}$ for $k \neq k'$, and $\mathbf{z}_{u\ell} = (z_{u\ell 1}, z_{u\ell 2}, \ldots, z_{u\ell K_\ell})'$ is distributed

$$\mathbf{z}_{u\ell} \sim N(\hat{\mathbf{z}}_{u\ell}, \psi_u^2 \mathbf{I}).$$

The MSPE in (4.18) is then the probability that the index $k_{u\ell}$ of the maximum $z_{u\ell a}$ in $\mathbf{z}_{u\ell}$ is not the index of the maximum $\hat{z}_{u\ell a}$, which corresponds to our categorical prediction.

The calculation of this probability (and associated design criterion) can be accomplished using multivariate Gaussian density functions. However, the computational cost can be much greater than that required to calculate the latent Gaussian design criterion (4.17).

Simulation studies (not shown) indicate that optimal designs based on the categorical design criterion tend to be closer to space filling designs than optimal designs based on the latent Gaussian design criterion. The optimal latent Gaussian designs tend to focus on remote regions of the spatial domain (i.e., the lek network). If a researcher considers spatial coverage to be highly important to a design, the categorical design criterion is recommended. This is especially relevant in a prospective design, or in a retrospective design based off of a small existing sample which does not adequately cover the spatial domain of the study. In our study of sage-grouse, the data collected from 2009-2012 provide a reasonable coverage of the spatial range of the species. We thus choose to use the latent Gaussian design criterion, which tends in simulation to put more weight on regions that are poorly connected to the rest of the lek network. This reflects a desire to preferentially sample in regions far removed from leks which have been previously sampled.

4.5.3 Optimal Sampling for Reduced Rank Spatial Models

Computing the design criterion (4.17) requires calculation of the inverse $\mathbf{\Sigma}_{oo}^{-1}$ which has computational cost of $O(n_d^3)$, where n_d is the number of locations in the proposed design. In the sage-grouse study, over $n_d = 1000$ leks were sampled during the years of 2009-2012. For our retrospective design, we consider adding leks to the sampling design, which will increase this number. Inverting a matrix of this size is computationally demanding, especially if we wish to compare a large number of potential sampling designs. However, we can take advantage of the reduced rank spatial model presented in Section 4.4.2 to significantly reduce the computational cost of computing the design criterion by replacing the Kriging covariance matrix $\mathbf{\Psi} = \mathbf{\Sigma}_{uu} - \mathbf{\Sigma}_{uo}\mathbf{\Sigma}_{oo}^{-1}\mathbf{\Sigma}_{ou}$ with a reduced-rank approximation $\tilde{\mathbf{\Psi}} = \tilde{\mathbf{\Sigma}}_{uu} - \tilde{\mathbf{\Sigma}}_{uo}\tilde{\mathbf{\Sigma}}_{oo}^{-1}\tilde{\mathbf{\Sigma}}_{ou}$.

Note that the reduced rank covariance matrix $\tilde{\mathbf{\Sigma}}$ of all observed and unobserved nodes can be decomposed and written in block form:

$$\begin{aligned}\tilde{\mathbf{\Sigma}} &= \sigma^2\tilde{\mathbf{M}}\tilde{\mathbf{D}}\tilde{\mathbf{M}}' + \mathbf{I} \\ &= \sigma^2 \cdot \begin{bmatrix} \tilde{\mathbf{M}}_o \\ \tilde{\mathbf{M}}_u \end{bmatrix} \tilde{\mathbf{D}} \; [\tilde{\mathbf{M}}_o' \;\; \tilde{\mathbf{M}}_u'] + \begin{bmatrix} \mathbf{I}_o & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_u \end{bmatrix}\end{aligned}$$

which leads to a reduced rank version of the Kriging covariance matrix:

$$\begin{aligned}\tilde{\mathbf{\Psi}} &= \tilde{\mathbf{\Sigma}}_{uu} - \tilde{\mathbf{\Sigma}}_{uo}\tilde{\mathbf{\Sigma}}_{oo}^{-1}\tilde{\mathbf{\Sigma}}_{ou} && (4.19)\\ &= \sigma^2\tilde{\mathbf{M}}_u\tilde{\mathbf{D}}\tilde{\mathbf{M}}_u' + \mathbf{I}_u - \left(\sigma^2\tilde{\mathbf{M}}_u\tilde{\mathbf{D}}\tilde{\mathbf{M}}_o'\right)\left(\sigma^2\tilde{\mathbf{M}}_o\tilde{\mathbf{D}}\tilde{\mathbf{M}}_o' + \mathbf{I}_o\right)^{-1}\left(\sigma^2\tilde{\mathbf{M}}_o\tilde{\mathbf{D}}\tilde{\mathbf{M}}_u'\right). && (4.20)\end{aligned}$$

This still requires the inverse of a $n_d \times n_d$ matrix, but the inverse in question can be expressed using the Sherman-Morrison-Woodbury identity (e.g., Gentle, 2007, p. 221)

$$\left(\tilde{\mathbf{M}}_o(\sigma^2\tilde{\mathbf{D}})\tilde{\mathbf{M}}_o' + \mathbf{I}_o\right)^{-1} = \mathbf{I}_o - \tilde{\mathbf{M}}_o\left(\frac{1}{\sigma^2}\tilde{\mathbf{D}}^{-1} + \tilde{\mathbf{M}}_o'\mathbf{I}_o\tilde{\mathbf{M}}_o\right)^{-1}\tilde{\mathbf{M}}_o'. \qquad (4.21)$$

The resulting expression only requires the inverse of a matrix of dimensionality $n_e \times n_e$, where n_e is equal to the number of eigenvectors retained in the reduced rank version of $\mathbf{Q}$ and typically $n_e << n_d$. This allows us to take advantage of the reduced rank approximation to the spatial covariance in the computation of the sampling criterion $\phi(\mathbf{d}) = \text{tr}(\tilde{\mathbf{\Psi}})$.

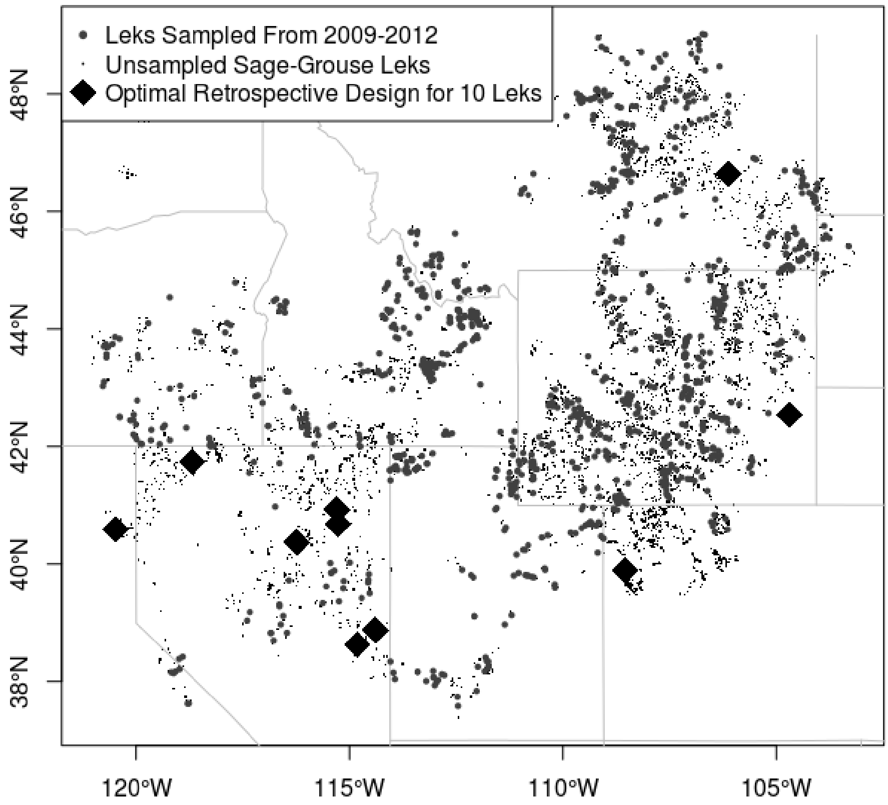
● Leks Sampled From 2009-2012
· Unsampled Sage-Grouse Leks
◆ Optimal Retrospective Design for 10 Leks
48°N
46°N
44°N
42°N
40°N
38°N
120°W
115°W
110°W
105°W

Figure 4.3: Sage-grouse lek sampling recommendations based on latent Gaussian MSPE. Leks shown as grey circles indicate leks that have been sampled in 2009-2012, or will be sampled in 2013 Leks shown as black diamonds are the optimal retrospective lek sampling locations under the constraint that no more than 10 additional leks are sampled in 2013.

4.5.4 Recommendations for 2013 Sage-Grouse Sampling Effort

To make recommendations for 2013 sampling effort of sage-grouse leks, we focus on identifying the optimal retrospective design under the latent Gaussian MSPE sampling criterion (4.17). Genetic data from 830 individuals at 243 distinct leks (Figure 4.2) were used to fit the statistical model in Section 4.4.4. Each of these leks were considered to be observed. We also included 934 additional leks where feather samples have been collected during 2009-2012, but had not been genotyped before the start of the 2013 field season, in the set of observed leks. The genetic information from these samples could thus not be used to estimate the spatial covariance parameters in our statistical model, but knowing that samples have been obtained from these additional sampling locations can aid in identifying regions of high sampling importance for the 2013 sampling effort.

Given this set of observed leks, we consider adding ten additional leks to the existing design. As a complete search of the space of retrospective designs would be computationally prohibitive, we constructed an iterative search algorithm to identify a pseudo-optimal retrospective design. We first considered 10,000 random retrospective designs of ten additional leks randomly chosen from all unsampled leks, and computed the latent Gaussian sampling criterion (4.17) for each. We also considered multiple space filling designs on the lek network. The best six designs were used as initial designs in an iterative exchange algorithm (Royle, 1998). At each iteration of the exchange algorithm, one lek in the retrospective design was replaced by a neighboring lek, with the neighbor selected from all unsampled leks within $d_{MAX} = 25$km. If the design criterion was improved by the switch in leks, then the change to the retrospective design is retained; otherwise, the switch is rejected and the design is unchanged. This algorithm was run until a local minimum was found for each initial design. The resulting six designs were then compared, and the design with the smallest design criterion is shown in Figure 4.3. This design is based on the estimated spatial covariance for a network model of lek connectivity and provides guidance on how states can direct their sage-grouse lek sampling effort in 2013.

4.6 Discussion

Landscape genetics is the study of how landscape features influence gene flow across space. While the process being studied is spatio-temporal in nature, the genetic time scale of the process is far greater than the time scale over which genetic observations are obtained. Spatially referenced genetic data thus represent a snapshot of the spatio-temporal process of gene flow, and are essentially spatial data, not spatio-temporal data. Our goal in this chapter has been to apply existing spatial statistical approaches to landscape genetic data. We have proposed a multinomial model with latent spatial random effects for microsatellite allele data and illustrated how optimal retrospective spatial designs can be obtained for landscape genetic studies.

While the multinomial model we proposed in Section 4.3 is appropriate for allele data that can be thought of as categorical, next generation genetic data such as single nucleotide polymorphisms (SNPs) may require modified data models. However, the general hierarchical framework of an appropriate data model linked to a latent spatial random effect should be broadly applicable to the field of landscape genetics as the tools and methods for analyzing genetic sequence data evolve.

We defined spatial covariance between leks based on a graphical lek network, with edge weights a function of Euclidean distance between leks. This model was developed to account for the lekking behavior of greater sage-grouse (e.g., Connelly et al., 2004), and will not be appropriate for most other species. It is far more common in landscape genetics to specify a landscape graph where the edge weights are defined by the local landscape features (e.g., McRae, 2006; McRae and Beier, 2007; McRae et al., 2008; Cushman et al., 2006; Cushman and Landguth, 2010; Spear et al., 2010), and connectivity is defined under the IBR or the LCP approach. This is also the approach we took in our study of alpine chamois in Chapter

3. As described in Section 4.2, covariance functions exist that match the assumptions of both the IBR and LCP approaches, as well as the stationary and isotropic IBD approach to landscape genetics. Hierarchical spatial modeling of landscape genetic data provides new opportunities to obtain inference for resistance parameters in these effective distance approaches within a formal statistical framework.

CHAPTER 5

ISSUES OF SCALE IN DISCRETE-SPACE MODELS FOR MOVEMENT AND CONNECTIVITY

In this dissertation, I have examined discrete-space approaches to modeling animal movement and landscape connectivity. In this concluding chapter, I briefly review and synthesize the work in the previous chapters, and discuss directions for future work in landscape connectivity.

In Chapter 2, a continuous time discrete-space (CTDS) model for animal movement was proposed that allows for natural modeling of a variety of potential drivers of animal movement behavior, including location-based drivers which model an animal's absolute speed through different environments, and gradient-based drivers which model an animal's mean response to directional covariates such as the direction to a conspecific. Considerable attention was paid to computational efficiency, first by proposing a latent-variable representation of the CTDS movement model that allowed for inference using standard Poisson GLM approaches, and second by suggesting a multiple imputation approach to approximating the posterior predictive distribution. The computational efficiency of our approach allowed for complex time-varying behavior to be modeled. These methods were applied to a study of a pair of mountain lions in Colorado, USA.

In Chapter 3, the circuit-theoretic approach to landscape genetic data was examined. The random walk model for gene flow that underlies the isolation by resistance (IBR) approach in this chapter is identical to the CTDS model for animal movement proposed in Chapter 2, as long as the drivers of movement chosen result in a reversible random walk, or equivalently an undirected graph with symmetric edge weights. The link between the IBR approach to landscape genetic analysis and GMRFs provides the first formal link between the effective distance approach to analysis of landscape genetic data and the field of spatial statistics. To compare and contrast spatially explicit models with the effective distance approach, observed pairwise genetic distance matrices were modeled using the generalized Wishart model of McCullagh (2009). This approach was illustrated with both a simulation study and a study of alpine chamois in the Bauges mountains of France.

The generalized Wishart model used in Chapter 3 is an appropriate statistical model for squared pairwise distances (variograms) of Gaussian-distributed random variables. However, it is not clear whether genetic distance matrices are distributionally similar to empirical variograms of Gaussian random variables. In Chapter 4, this discrepency was addressed by proposing a categorical data model for microsatellite allele data with spatial genetic correlation modeled using latent spatial random effects. This hierarchical approach to landscape genetic data shows great promise, as it allows for both modeling of the observations and modeling of spatial correlation under existing assumptions about gene flow, such as the assumptions present in isolation by resistance, isolation by distance, and least cost path approaches to effective distance analysis.

Two directions for future research are particularly appealing, both related to scale dependent inference for movement and connectivity. The first direction is a reconciliation of the mismatch of temporal scale between movement data and genetic data used to study connectivity. The second is the choice of spatial scale for discrete-space approaches to landscape connectivity. Each is considered in what follows.

5.1 Temporal Scale in Movement and Connectivity

The similarities between the CTDS movement model in Chapter 2 and the intrinsic conditional autoregressive (ICAR) GMRF model for spatial correlation in Chapter 3 are striking, and suggest that both movement data and genetic data might be used jointly to make inference about landscape effects of connectivity. Consider the transition rates (2.20) to neighboring cells in the CTDS model for animal movement from Chapter 2:

$$\lambda_{ij}\left(\beta(t)\right) = \exp\left\{\beta_0(t) + \sum_{k=1}^{K} p_{ki}\beta_k(t) + \sum_{l=1}^{L} q_{lij}\beta_l(t)\right\}$$

where $\beta_0(t)$ is a time varying intercept term, $\{\beta_k(t)\}$ are time varying effects related to location-based drivers of movement, and $\{\beta_l(t)\}$ are time varying effects related to directional drivers of movement. If we do not allow $\boldsymbol{\beta}$ to vary over time (e.g., $\boldsymbol{\beta}(t) = \boldsymbol{\beta}$), and do not include any location-based drivers p_{ki} in the model, then the transition rate between cells is only a function of the directional drivers of movement $\{q_{lij}\}$. If we define these drivers as

$$q_{lij} = \frac{1}{d_{ij}}\left(\frac{x_{li} + x_{lj}}{2}\right),$$

where d_{ij} is the distance between the centers of neighboring grid cells, and x_{li} is the value of the l^{th} landscape covariate at the i^{th} grid cell, then this transition rate is exactly the edge weight α_{ij} between neighboring grid cells from Chapter 3:

$$\alpha_{ij} = \begin{cases} \exp\left[\frac{1}{d_{ij}}\left(\frac{\mathbf{x}_i' + \mathbf{x}_j'}{2}\right)\boldsymbol{\beta}\right] & , j \in \mathcal{N}(i) \\ 0 & , j \notin \mathcal{N}(i) \end{cases}.$$

And, in fact, the interpretation of the parameters $\boldsymbol{\beta}$ is the same. This link between CTDS random walk models for animal movement and ICAR GMRF models for spatial correlation raises the possibility of making joint inference on both movement and genetic data. For example, consider observed telemetry data $\mathbf{S}$ and observed spatially-referenced microsatellite allele data $\mathbf{Y}$. One might consider a joint model of the data in which the movement data are considered to arise independent of the genetic data:

$$[\mathbf{S}, \mathbf{Y}|\boldsymbol{\beta}] = [\mathbf{S}|\boldsymbol{\beta}][\mathbf{Y}|\boldsymbol{\beta}]$$

$$\boldsymbol{\beta} \sim N(\mathbf{0}, \Sigma_\beta)$$

where $[\mathbf{S}|\boldsymbol{\beta}]$ is the CTDS model for animal movement from Chapter 2 and $[\mathbf{Y}|\boldsymbol{\beta}]$ is the multinomial categorical data model from Chapter 4 with a latent ICAR random effect. The assumption of independence between telemetry data and genetic data may seem counterintuitive, but is likely to hold due to the mismatch in temporal scales between the data generating processes. The genetic data are the result of a multi-generational process including dispersal, survival, and mating, while the telemetry data are collected at a much finer temporal resolution. This provides a formal approach for using both telemetry and genetic data to make inference about parameters related to landscape connectivity. Others have considered both telemetry data and genetic data from the same species (e.g., Cushman and Lewis, 2010), but each source of data was analyzed independently, and without formal correspondence between model parameters.

However, the mismatch in temporal scales between movement and genetic processes may complicate joint modeling of telemetry and microsatellite allele data. While the transition rate λ_{ij} in the random walk CTDS movement model models the movement of an animal, the transition rate (edge weight) α_{ij} from the ICAR spatial covariance model actually reflects rates of gene flow, a process which involves not only animal movement, but also survival and reproduction. Consider, for example, the effect of a patch of dangerous terrain, like a multilane highway. It is possible that animals are likely to move at a high rate of speed through this feature, but survival may be negatively impacted. In this scenario, results from a movement analysis could show that the highway facilitates movement at the short time scales present in telemetry data. In contrast, results from a genetic analysis could show that the highway impedes gene flow and genetic connectivity, as survival is negatively impacted by traveling through this feature.

This mismatch of temporal scales between movement and genetic data provides both challenges and opportunities. While joint modeling of these two sources of data requires careful consideration, it provides the opportunity to gain more scientific insight than could be obtained using either source of data individually. Analyzing movement and genetic data together may provide an approach to jointly model the effect that different terrain has on movement at short time scales, together with survival and reproduction rates at longer (genetic) time scales.

5.2 Spatial Scale in Landscape Connectivity

Animals inhabit a continuous, four-dimensional spatio-temporal world, and attempting to model animal response to the environment using a discretized version of that continuous space is not without its drawbacks. In particular, utilizing a discrete-space approach requires choosing a scale on which the discretized process will be modeled. In both the CTDS model for animal movement from Chapter 2 and the circuit approach to landscape genetics in Chapter 3, the spatial domain was discretized into a set of regular grid cells which partitioned the domain. This discretization is common in both IBR and LCP approaches to studying gene flow in heterogeneous landscapes, as well as in studies of resource selection (e.g., Hooten et al., 2013). However, the choice of spatial resolution (grid cell size) can greatly influence results, and there is little guidance on choosing an appropriate resolution (e.g., Borcard and Legendre, 2002; Cushman and Landguth, 2010). In Chapter 2, it was noted that choosing the grid cell size for the CTDS movement model specifies an implied time scale at which inference is made on animal movement decisions in response to environmental and biotic drivers of movement (Figure 2.3). For genetic analyses, the effects of spatial resolution are less clear.

One approach to considering multiple spatial resolutions for genetic analyses is to view a discrete-space analysis as an approximation to a continuous process. Lindgren et al. (2011) describe an explicit link between continuous spatial models and discrete-space spatial models by considering the following stochastic partial differential equation (SPDE):

$$\left(\kappa^2 - \Delta\right)^{(\nu+1)/2} \mathbf{y}(\mathbf{s}) = \mathcal{W}(\mathbf{s}), \quad \mathbf{s} \in \mathbb{R}^2, \quad \kappa > 0, \quad \nu > 0 \tag{5.1}$$

where the innovation process $\mathcal{W}(\mathbf{s})$ is spatial Gaussian white noise with unit variance, and Δ is the Laplace operator. Whittle (1954) showed that the solution to this SPDE is a Gaussian random field with Matern covariance function between the spatial locations

$$y(\mathbf{s}) \sim GP(\mathbf{0}, C(\cdot, \cdot)) \tag{5.2}$$

$$C(\mathbf{s}_i, \mathbf{s}_j) = \frac{\sigma^2}{2^{\nu-1}\Gamma(\nu)} \left(\kappa||\mathbf{s}_i - \mathbf{s}_j||\right)^{\nu} K_{\nu}\left(\kappa||\mathbf{s}_i - \mathbf{s}_j||\right) \tag{5.3}$$

where K_v is the modified Bessel function of the second kind and order $\nu > 0$. $\kappa > 0$ is a scaling parameter and σ^2 is the marginal variance

$$\sigma^2 = \frac{\Gamma(\nu)}{4\pi\kappa^{2\nu}\Gamma(\nu+1)}.$$

For example, setting $\nu = 1/2$ results in the exponential covariance function

$$C_{\exp}(\mathbf{s}_i, \mathbf{s}_j) = \sigma^2 \exp\{-\kappa||\mathbf{s}_i - \mathbf{s}_j||\}.$$

Thus, choosing κ and ν specifies a stationary, isotropic Gaussian random field of the Matern class that is a solution to the SPDE (5.1)

Lindgren et al. (2011) then show that a GMRF that approximates the Gaussian random field that is a solution to (5.1) can be obtained by taking a finite-element approach to solving the SPDE. The finite element method consists of choosing a grid, or mesh, of points across the spatial domain $\mathcal{S}$, and a set of basis functions, which are typically piecewise-linear functions between the mesh points. The solution to (5.1) is then approximated by a weighted sum of the basis functions.

To illustrate the relevance of this result, consider the multinomial data model with latent spatial random effects from Chapter 4. Under the LCP approach to connectivity, a discretized space is specified with local resistance to movement and gene flow being a function of the landscape characteristics of each grid cell. A spatial random effect with this discrete-space as its support is used to model connectivity and correlation in observed genetic data. Under the SPDE approach to specifying the spatially-correlated random effect, the spatial random effect is considered to have continuous spatial support, and the choice of spatial resolution (grid cell size) is seen as the choice of mesh points for the finite element method basis of the approximate solution to the SPDE. Thus, specifying a finer resolution of grid cells (mesh points) will result in a more accurate approximate solution to the SPDE (5.1), while a coarser resolution will result in a coarser approximation.

The SPDE (5.1) provides a natural approach for spatial deformation approaches to nonstationary covariance models (Schmidt and O'Hagan, 2003). The parameter κ in (5.1) controls the range of the spatial structure in the covariance function (5.3). Consider the case where κ is allowed to vary over space, with the range parameter being a function of the local landscape characteristics, e.g.,

$$\kappa(\mathbf{s}) = \exp\left\{-\mathbf{x}(\mathbf{s})'\boldsymbol{\beta}\right\}. \tag{5.4}$$

Then the resulting covariance will be nonstationary, with the nonstationarity defined by the landscape characteristics $\mathbf{x}(\mathbf{s})$ and the weighting parameters $\boldsymbol{\beta}$, where positive values of β_k indicate that the k^{th} landscape characteristic x_k facilitates gene flow and spatial connectivity, as positive values of β_k would result in greater correlation between locations separated by landscape high in x_k. This would provide a novel approach to modeling nonstationary correlation between genetic observations with assumptions similar to those being used in current LCP approaches.

5.3 Conclusion

Linking animal behavior to the landscape is one of the fundamental questions in ecology (Dalziel et al., 2008; Gurarie et al., 2009; Cagnacci et al., 2010). In this work, I have proposed novel statistical approaches to studying the effects of the landscape, as well as other potential biotic factors, on animal movement behavior and spatial gene flow. Throughout this treatment, particular attention has been paid to the following recurring themes:

1. Scientific relevance.
2. Statistical appropriateness.
3. Computational efficiency.

Each section in this dissertation was motivated by a scientific study, and the goal in each case was to advance the field of statistics to allow for the appropriate analysis of the data. Statistical models were chosen that are scientificially justified, and that allow for interesting scientific hypotheses to be investigated. It is important to recognize that the ability to collect data in the ecological sciences is growing rapidly (e.g., Tomkiewicz et al., 2010), and with it our need for sophisticated models that more accurately reflect the natural world. Together, these two trends make computational efficiency an essential consideration in the development of statistical approaches to studying connectivity.

As in any scientific endeavor, there is much work still to be done. I have outlined two avenues of future research in this concluding chapter, but anticipate many more opportunities in both the study of animal movement and spatial gene flow. The field of statistics is advancing at an exciting pace, together with our ability to collect new data of exceptionally high quality. These advances will spark new opportunities at the interface of statistics and the study of landscape ecology.

REFERENCES

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

Aarts, G., MacKenzie, M., McConnell, B., Fedak, M., and Matthiopoulos, J. (2008). Estimating space-use and habitat preference from wildlife telemetry data. *Ecography* **31,** 140–160.

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

Babić, D., Klein, D., Lukovits, I., Nikolić, S., and Trinajstić, N. (2002). Resistance-distance matrix: A computational algorithm and its application. *International Journal of Quantum Chemistry* **90,** 166–176.

Bates, D. and Maechler, M. (2011). *Matrix: Sparse and Dense Matrix Classes and Methods.* R package version 0.9996875-3.

Berliner, L. M., Wikle, C. K., and Cressie, N. (2000). Long-lead prediction of Pacific SSTs via Bayesian dynamic modeling. *Journal of Climate* **13,** 3953–3968.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society. Series B (Methodological)* **36,** 192–236.

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

Besag, J., York, J., and Mollié, A. (1991). Bayesian image restoration, with two applications in spatial statistics. *Annals of the Institute of Statistical Mathematics* **43,** 1–20.

Bolin, D., Lindström, J., Eklundh, L., and Lindgren, F. (2009). Fast estimation of spatially dependent temporal vegetation trends using Gaussian Markov random fields. *Computational Statistics & Data Analysis* **53,** 2885–2896.

Borcard, D. and Legendre, P. (2002). All-scale spatial analysis of ecological data by means of principal coordinates of neighbour matrices. *Ecological Modelling* **153,** 51–68.

Boyce, M. (2006). Scale for resource selection functions. *Diversity and Distributions* **12,** 269–276.

Boyce, M., Vernier, P., Nielsen, S., and Schmiegelow, F. (2002). Evaluating resource selection functions. *Ecological Modelling* **157,** 281–300.

Brillinger, D., Preisler, H., Ager, A., and Kie, J. (2001). The use of potential functions in modeling animal movement. In Salah, A. K., editor, *Data analysis from statistical foundations*, chapter 24, pages 369–386. Nova Publishers.

Broquet, T., Ray, N., Petit, E., Fryxell, J. M., and Burel, F. (2006). Genetic isolation by distance and landscape connectivity in the American marten (Martes americana). *Landscape Ecology* **21,** 877–889.

Cagnacci, F., Boitani, L., Powell, R., and Boyce, M. (2010). Animal ecology meets GPS-based radiotelemetry: a perfect storm of opportunities and challenges. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* **365,** 2157–2162.

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

Chen, C., Durand, E., Forbes, F., and Francois, O. (2007). Bayesian clustering algorithms ascertaining spatial population structure: a new computer program and a comparison study. *Molecular Ecology Notes* **7,** 747–756.

Chen, Q. and Wang, S. (2011). Variable selection for multiply-imputed data with application to dioxin exposure study. Technical Report 217, University of Wisconsin-Madison Department of Biostatistics and Medical Informatics.

Connelly, J., Knick, S., Schroeder, M., Stiver, S., et al. (2004). Conservation assessment of greater sage-grouse and sagebrush habitats. In *Unpublished Report.* Western Association of Fish and Wildlife Agencies Cheyenne, WY.

Connelly, J. W. and Braun, C. E. (1997). Long-term changes in sage-grouse (Centrocercus urophasianus) populations in western North America. *Wildlife Biology* **3,** 229–234.

Connelly, J. W., Wakkinen, W. L., Apa, A. D., and Reese, K. P. (1991). Sage-grouse use of nest sites in southeastern Idaho. *The Journal of Wildlife Management* **55,** 521–524.

Cressie, N. (1993). *Statistics for Spatial Data.* Wiley-Interscience.

Cressie, N. and Wikle, C. (2011). *Statistics for spatio-temporal data*, volume 465. Wiley.

Cressie, N. and Wikle, C. (2012). *Statistics for Spatio-Temporal Data*, chapter 4.1, pages 124–161. Wiley.

Crooks, K. R. and Sanjayan, M., editors (2006). *Connectivity conservation.* Cambridge University Press.

Cross, P., Heisey, D., Scurlock, B., Edwards, W., Ebinger, M., and Brennan, A. (2010). Mapping brucellosis increases relative to elk density using hierarchical Bayesian models. *PLoS One* **5,** e10322.

Cushman, S. and Landguth, E. (2010). Scale dependent inference in landscape genetics. *Landscape Ecology* **25,** 967–979.

Cushman, S., McKelvey, K., and Schwartz, M. (2009). Use of empirically derived source-destination models to map regional conservation corridors. *Conservation Biology* **23,** 368–376.

Cushman, S. A. and Lewis, J. S. (2010). Movement behavior explains genetic differentiation in American black bears. *Landscape Ecology* **25,** 1613–1625.

Cushman, S. A., McKelvey, K. S., Hayden, J., and Schwartz, M. K. (2006). Gene flow in complex landscapes: testing multiple hypotheses with causal modeling. *The American Naturalist* **168,** 486–499.

Dalziel, B. D., Morales, J. M., and Fryxell, J. M. (2008). Fitting probability distributions to animal movement trajectories: using artificial neural networks to link distance, resources, and memory. *The American Naturalist* **172,** 248–258.

Dorf, R. and Svoboda, J. (2004). *Introduction to electric circuits*. Wiley.

Durand, E., Jay, F., Gaggiotti, O., and François, O. (2009). Spatial inference of admixture proportions and secondary contact zones. *Molecular Biology and Evolution* **26,** 1963–1973.

Dyer, R., Nason, J., and Garrick, R. (2010). Landscape modelling of gene flow: improved power using conditional genetic distance derived from the topology of population networks. *Molecular Ecology* **19,** 3746–3759.

Forester, J. D., Im, H. K., and Rathouz, P. J. (2009). Accounting for animal movement in estimation of resource selection functions: sampling and data analysis. *Ecology* **90,** 3554–3565.

Fouss, F., Pirotte, A., Renders, J., and Saerens, M. (2007). Random-walk computation of similarities between nodes of a graph with application to collaborative recommendation. *IEEE Transactions on Knowledge and Data Engineering* **19,** 355–369.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software* **33,** 1–22.

Gelman, A., Carlin, B. P., Stern, H., and Rubin, D. B. (2004). *Bayesian Data Analysis*. Chapman and Hall/CRC, Princeton, New Jersey, USA, 2nd edition.

Gentle, J. E. (2007). *Matrix algebra: theory, computations, and applications in statistics*. Springer Verlag.

Getz, W. and Saltz, D. (2008). A framework for generating and analyzing movement paths on ecological landscapes. *Proceedings of the National Academy of Sciences* **105,** 19066–19071.

Graves, T., Beier, P., and Royle, J. (2013). Current approaches using genetic distances produce poor estimates of landscape resistance to inter-individual dispersal. *Molecular Ecology* **,** In Press.

Graybill, F. (1983). *Matrices with Applications in Statistics.* Wadsworth Inc.

Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* **82,** 711–732.

Grovenburg, T., Jenks, J., Klaver, R., Swanson, C., Jacques, C., and Todey, D. (2009). Seasonal movements and home ranges of white-tailed deer in north-central South Dakota. *Canadian Journal of Zoology* **87,** 876–885.

Guillot, G., Estoup, A., Mortier, F., and Cosson, J. F. (2005). A spatial statistical model for landscape genetics. *Genetics* **170,** 1261–1280.

Gurarie, E., Andrews, R. D., and Laidre, K. L. (2009). A novel method for identifying behavioural changes in animal movement data. *Ecology Letters* **12,** 395–408.

Hanks, E., Hooten, M., Johnson, D., and Sterling, J. (2011). Velocity-based movement modeling for individual and population level inference. *PLoS ONE* **6,** e22795.

Harville, D. (2008). *Matrix algebra from a statistician's perspective.* Springer Verlag.

Hastie, T. and Tibshirani, R. (1993). Varying-coefficient models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **55,** 757–796.

Hijmans, R. J. and van Etten, J. (2012). *raster: Geographic analysis and modeling with raster data.* R package version 1.9-92.

Holsinger, K. and Weir, B. (2009). Genetics in geographically structured populations: defining, estimating and interpreting Fst. *Nature Reviews Genetics* **10,** 639–650.

Hooten, M., Hanks, E., Johnson, D., and Alldredge, M. (2013). Temporal variation and scale in movement-based resource selection functions. *Statistical Methodology* **,** In Press.

Hooten, M. and Wikle, C. (2008). A hierarchical Bayesian non-linear spatio-temporal model for the spread of invasive species with application to the Eurasian Collared-Dove. *Environmental and Ecological Statistics* **15,** 59–70.

Hooten, M., Wikle, C., Sheriff, S., and Rushin, J. (2009). Optimal spatio-temporal hybrid sampling designs for ecological monitoring. *Journal of Vegetation Science* **20,** 639–649.

Hooten, M. B., Johnson, D. S., Hanks, E. M., and Lowry, J. H. (2010). Agent-based inference for animal movement and selection. *Journal of Agricultural, Biological, and Environmental Statistics* **15,** 523–538.

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

Johnson, D., London, J., Lea, M., and Durban, J. (2008). Continuous-time correlated random walk model for animal telemetry data. *Ecology* **89,** 1208–1215.

Johnson, D., Thomas, D., Ver Hoef, J., and Christ, A. (2008). A general framework for the analysis of animal resource selection from telemetry data. *Biometrics* **64,** 968–976.

Johnson, D. S. (2011). *crawl: Fit continuous-time correlated random walk models for animal movement data*. R package version 1.3-2.

Johnson, D. S., Thomas, D. L., Ver Hoef, J. M., and Christ, A. (2008). A general framework for the analysis of animal resource selection from telemetry data. *Biometrics* **64,** 968–976.

Jombart, T. (2008). adegenet: a R package for the multivariate analysis of genetic markers. *Bioinformatics* **24,** 1403–1405.

Jombart, T. (2012). A tutorial for the spatial analysis of principal components (spca) using adegenet 1.3-4. *Vignette for the R package 'adegenet'* .

Jombart, T., Devillard, S., Dufour, A.-B., and Pontier, D. (2008). Revealing cryptic spatial patterns in genetic variability by a new multivariate method. *Heredity* **101,** 92–103.

Jonsen, I., Flemming, J., and Myers, R. (2005). Robust state-space modeling of animal movement data. *Ecology* **86,** 2874–2880.

Klein, D., Palacios, J., Randic, M., and Trinajstic, N. (2004). Random walks and chemical graph theory. *Journal of Chemical Information and Computer Sciences* **44,** 1521–1525.

Klein, D. and Randić, M. (1993). Resistance distance. *Journal of Mathematical Chemistry* **12,** 81–95.

Knopff, K., Knopff, A., Warren, M., and Boyce, M. (2009). Evaluating global positioning system telemetry techniques for estimating cougar predation parameters. *The Journal of Wildlife Management* **73,** 586–597.

Kuhn, C. E., Johnson, D. S., Ream, R. R., and Gelatt, T. S. (2009). Advances in the tracking of marine species: using GPS locations to evaluate satellite track data and a continuous-time movement model. *Marine Ecology Progress Series* **393,** 97–109.

Kunegis, J., Lommatzsch, A., and Bauckhage, C. (2009). The slashdot zoo: mining a social network with negative edges. In *Proceedings of the 18th International Conference on World Wide Web*, pages 741–750. ACM.

Landguth, E. and Cushman, S. (2010). cdpop: A spatially explicit cost distance population genetics program. *Molecular Ecology Resources* **10,** 156–161.

Legendre, P. and Fortin, M. (2010). Comparison of the Mantel test and alternative approaches for detecting complex multivariate relationships in the spatial analysis of genetic data. *Molecular Ecology Resources* **10,** 831–844.

Lele, S. R., Nadeem, K., and Schmuland, B. (2011). Estimability and likelihood inference for generalized linear mixed models using data cloning. *Journal of the American Statistical Association* **105,** 1617–1625.

Lima, S. (2002). Putting predators back into behavioral predator–prey interactions. *Trends in Ecology & Evolution* **17,** 70–75.

Lindgren, F., Rue, H., and 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,** 423–498.

Lindström, J. and Lindgren, F. (2008). A Gaussian Markov random field model for total yearly precipitation over the African Sahel. *Preprints in Mathematical Sciences* **8,** Lund University.

Lookingbill, T., Gardner, R., Ferrari, J., and Keller, C. (2010). Combining a dispersal model with network theory to assess habitat connectivity. *Ecological Applications* **20,** 427–441.

Manel, S., Schwartz, M., Luikart, G., and Taberlet, P. (2003). Landscape genetics: combining landscape ecology and population genetics. *Trends in Ecology & Evolution* **18,** 189–197.

Manly, B. F., McDonald, L., and Thomas, D. L. (2002). *Resource selection by animals: statistical design and analysis for field studies.* Chapman & Hall.

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

McCullagh, P. (2009). Marginal likelihood for distance matrices. *Statistica Sinica* **19,** 631–649.

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

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

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

Merrill, E., Sand, H., Zimmermann, B., McPhee, H., Webb, N., Hebblewhite, M., Wabakken, P., and Frair, J. L. (2010). Building a mechanistic understanding of predation with GPS-based movement data. *Philosophical Transactions of the Royal Society B: Biological Sciences* **365,** 2279–2288.

Morales, J. M., Haydon, D. T., Frair, J., Holsinger, K. E., and Fryxell, J. M. (2004). Extracting more out of relocation data: building movement models as mixtures of random walks. *Ecology* **85,** 2436–2445.

Morales, J. M., Moorcroft, P. R., Matthiopoulos, J., Frair, J. L., Kie, J. G., Powell, R. A., Merrill, E. H., and Haydon, D. T. (2010). Building the bridge between animal movement and population dynamics. *Philosophical Transactions of the Royal Society B: Biological Sciences* **365,** 2289–2301.

Nathan, R., Getz, W., Revilla, E., Holyoak, M., Kadmon, R., Saltz, D., and Smouse, P. (2008). A movement ecology paradigm for unifying organismal movement research. *Proceedings of the National Academy of Sciences* **105,** 19052–19059.

Oksanen, J., Blanchet, F. G., Kindt, R., Legendre, P., Minchin, P. R., O'Hara, R. B., Simpson, G. L., Solymos, P., Stevens, M. H. H., and Wagner, H. (2012). *vegan: Community Ecology Package.* R package version 2.0-3.

Owen-Smith, N., Fryxell, J. M., and Merrill, E. H. (2010). Foraging theory upscaled: the behavioural ecology of herbivore movement. *Philosophical Transactions of the Royal Society B: Biological Sciences* **365,** 2267–2278.

Park, T. and Casella, G. (2008). The Bayesian lasso. *Journal of the American Statistical Association* **103,** 681–686.

Polansky, L., Wittemyer, G., Cross, P., Tambling, C., and Getz, W. (2010). From moonlight to movement and synchronized randomness: Fourier and wavelet analyses of animal location time series data. *Ecology* **91(5),** 1506–1518.

Preisler, H. K., Ager, A. A., Johnson, B. K., and Kie, J. G. (2004). Modeling animal movements using stochastic differential equations. *Environmetrics* **15,** 643–657.

Preisler, H. K., Ager, A. A., and Wisdom, M. J. (2013). Analyzing animal movement patterns using potential functions. *Ecosphere* **4(3),** 32.

R Core Team (2013). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Rayfield, B., Fortin, M., and Fall, A. (2010). The sensitivity of least-cost habitat graphs to relative cost surface values. *Landscape Ecology* **25,** 519–532.

Royle, J. (1998). An algorithm for the construction of spatial coverage designs with implementation in SPLUS. *Computers & Geosciences* **24,** 479–488.

Royle, J. and Young, K. (2008). A hierarchical model for spatial capture-recapture data. *Ecology* **89,** 2281–2289.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys.* John Wiley and Sons, New York, New York, USA.

Rubin, D. B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association* **91,** 473–489.

Rue, H. (2001). Fast sampling of Gaussian Markov random fields. *Journal of the Royal Statistical Society. Series B, Statistical Methodology* **63,** 325–338.

Rue, H. and Held, L. (2005). *Gaussian Markov random fields: theory and applications*, volume 104 of *Monographs on Statistics and Applied Probability.* Chapman & Hall.

Rue, H., Martino, S., and 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,** 319–392.

Sahlsten, J., Thörngren, H., and Höglund, J. (2008). Inference of hazel grouse population structure using multilocus data: a landscape genetic approach. *Heredity* **101,** 475–482.

Saura, S. and Rubio, L. (2010). A common currency for the different ways in which patches and links can contribute to habitat availability and connectivity in the landscape. *Ecography* **33,** 523–537.

Schmidt, A. M. and O'Hagan, A. (2003). Bayesian inference for non-stationary spatial covariance structure via spatial deformations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **65,** 743–758.

Schroeder, M. A., Aldridge, C. L., Apa, A. D., Bohne, J. R., Braun, C. E., Bunnell, S. D., Connelly, J. W., Deibert, P. A., Gardner, S. C., Hilliard, M. A., et al. (2004). Distribution of sage-grouse in North America. *The Condor* **106,** 363–376.

Seber, G. (2008). *A Matrix Handbook for Statisticians*, volume 746. John Wiley & Sons.

Shirk, A., Wallin, D., Cushman, S., Rice, C., and Warheit, K. (2010). Inferring landscape effects on gene flow: a new model selection framework. *Molecular Ecology* **19,** 3603–3619.

Smouse, P. and Peakall, R. (1999). Spatial autocorrelation analysis of individual multiallele and multilocus genetic structure. *Heredity* **82,** 561–573.

Spear, S., Balkenhol, N., Fortin, M., McRae, B., and Scribner, K. (2010). Use of resistance surfaces for landscape genetic studies: considerations for parameterization and analysis. *Molecular Ecology* **19,** 3576–3591.

Spear, S. F., Peterson, C. R., Matocq, M. D., and Storfer, A. (2005). Landscape genetics of the blotched tiger salamander (Ambystoma tigrinum melanostictum). *Molecular Ecology* **14,** 2553–2564.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **64,** 583–639.

Stein, M. L. (1999). *Interpolation of spatial data: some theory for kriging.* Springer Verlag.

Stephens, M. (2000). Bayesian analysis of mixture models with an unknown number of components - an alternative to reversible jump methods. *The Annals of Statistics* **28,** 40–74.

Storfer, A., Murphy, M. A., Evans, J. S., Goldberg, C. S., Robinson, S., Spear, S. F., Dezzani, R., Delmelle, E., Vierling, L., and Waits, L. P. (2007). Putting the "landscape" in landscape genetics. *Heredity* **98,** 128–142.

Taylor, P. D., Fahrig, L., Henein, K., and Merriam, G. (1993). Connectivity is a vital element of landscape structure. *Oikos* **68,** 571–573.

Theobald, D., Crooks, K., and Norman, J. (2011). Assessing effects of land use on landscape connectivity: loss and fragmentation of western US forests. *Ecological Applications* **21,** 2445–2458.

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)* pages 267–288.

Tizghadam, A. and Leon-Garcia, A. (2010). On traffic-aware betweenness and network criticality. In *INFOCOM IEEE Conference on Computer Communications Workshops, 2010*, pages 1–6. IEEE.

Tizghadam, A. and Leon-Garcia, A. (2011). Robust network planning in nonuniform traffic scenarios. *Computer Communications* **34,** 1436–1449.

Tomkiewicz, S. M., Fuller, M. R., Kie, J. G., and Bates, K. K. (2010). Global positioning system and associated technologies in animal behaviour and ecological research. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* **365,** 2163–2176.

Tracey, J. A., Zhu, J., and Crooks, K. (2005). A set of nonlinear regression models for animal movement in response to a single landscape feature. *Journal of Agricultural, Biological, and Environmental Statistics* **10,** 1–18.

Turchin, P. (1998). *Quantitative Analysis of Movement.* Sinauer Associates, Inc.

Urban, D., Minor, E., Treml, E., and Schick, R. (2009). Graph models of habitat mosaics. *Ecology Letters* **12,** 260–273.

Volchenkov, D. (2011). Random walks and flights over connected graphs and complex networks. *Communications in Nonlinear Science and Numerical Simulation* **16,** 21–55.

Wakkinen, W. L., Reese, K. P., and Connelly, J. W. (1992). Sage-grouse nest locations in relation to leks. *The Journal of Wildlife Management* pages 381–383.

Wang, H. and Xia, Y. (2009). Shrinkage estimation of the varying coefficient model. *Journal of the American Statistical Association* **104,** 747–757.

Wang, I., Savage, W., and Shaffer, H. (2009). Landscape genetics and least-cost path analysis reveal unexpected dispersal routes in the california tiger salamander (ambystoma californiense). *Molecular Ecology* **18,** 1365–1374.

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

Wheeler, D. C. and Waller, L. A. (2010). Spatial analysis of genetic population structure in cougars. Technical Report 10-01, Department of Biostatistics and Bioinformatics, Rollins School of Public Health.

Wheeler, D. C., Waller, L. A., and Biek, R. (2010). Spatial analysis of feline immunodeficiency virus infection in cougars. *Spatial and Spatio-temporal Epidemiology* **1,** 151–161.

Whittle, P. (1954). On stationary processes in the plane. *Biometrika* **3,** 434–449.

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

Wikle, C. K. and Cressie, N. (1999). A dimension-reduced approach to space-time Kalman filtering. *Biometrika* **86,** 815–829.

Wikle, C. K. and Royle, J. A. (2005). Dynamic design of ecological monitoring networks for non-Gaussian spatio-temporal data. *Environmetrics* **16,** 507–522.

Wiley, R. (1973). Territoriality and non-random mating in sage-grouse, Centrocercus urophasianus. *Animal Behaviour Monographs* **6,** 87–169.

Wright, S. (1943). Isolation by distance. *Genetics* **28,** 114–138.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **68,** 49–67.

Zhu, H. and Klein, D. (1996). Graph-geometric invariants for molecular structures. *Journal of Chemical Information and Computer Sciences* **36,** 1067–1075.

Zimmerman, D. (2006). Optimal network design for spatial prediction, covariance parameter estimation, and empirical prediction. *Environmetrics* **17,** 635–652.