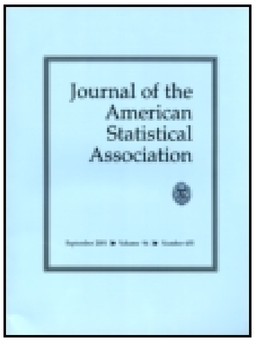


Taylor & Francis
Taylor & Francis Group

Journal of the American Statistical Association

ISSN: 0162-1459 (Print) 1537-274X (Online) Journal homepage: http://www.tandfonline.com/loi/uasa20

Modeling spatial covariance using the limiting distribution of spatio-temporal random walks

Ephraim M. Hanks

To cite this article: Ephraim M. Hanks (2016): Modeling spatial covariance using the limiting distribution of spatio-temporal random walks, Journal of the American Statistical Association, DOI: 10.1080/01621459.2016.1224714

To link to this article: http://dx.doi.org/10.1080/01621459.2016.1224714

Accepted author version posted online: 26 Aug 2016.

Modeling spatial covariance using the limiting distribution of spatio-temporal random walks

Ephraim M. Hanks*
Department of Statistics, The Pennsylvania State University

Abstract

We present an approach for modeling areal spatial covariance in observed genetic allele data by considering the stationary (limiting) distribution of a spatio-temporal Markov random walk model for gene flow. This stationary distribution corresponds to an intrinsic simultaneous autoregressive (SAR) model for spatial correlation, and provides a principled approach to specifying areal spatial models when a spatio-temporal generating process can be assumed. We apply the approach to a study of spatial genetic variation of trout in a stream network in Connecticut, USA.

Keywords: Spatial autocorrelation, SAR models, Diffusion, Autoregressive models.

*The author gratefully acknowledges conversations with John Fricks, whose suggestions were instrumental in the analysis of Section 3.1

1 Introduction

Almost all spatial data can be viewed as arising from a spatio-temporal generating process. For example, a spatial survey of infectious disease prevalence is a snapshot of a dynamic epidemic process occuring in space and time. Similarly, spatial genetic data are the result of spatio-temporal dispersal, mating, and survival processes at the population level. When these spatial processes are observed at multiple successive time points, the known science behind the spatio-temporal process is often used to motivate a spatio-temporal statistical model (e.g., Wikle & Hooten 2010, Cressie & Wikle 2011).

In contrast, consider the case of "spatial" data, where only one temporal realization of the spatio-temporal process is observed. In this case, spatial autocorrelation is often modeled by including a spatial random effect (e.g., Diggle & Ribeiro 2007) in the fitted statistical model. The prior distribution for this spatial random effect is almost always modeled semiparametrically using a Gaussian process model with covariance function chosen based on the support of the data, irrespective of the spatio-temporal generating process. For example, when the spatial data are point-referenced, the Matern class of covariance functions (e.g., Cressie 1993) are often used, while if the spatial data have areal or lattice support, then either conditional autoregressive (CAR; e.g., Besag 1974, Besag & Kooperberg 1995, Rue & Held 2005) or simultaneous autoregressive (SAR: e.g., Wall 2004, Cressie & Wikle 2011) models are common. In either case, the choice of prior distribution for the spatial random effect is almost always made based solely on the support of the data, without consideration of an underlying generating process.

The analysis of spatial genetic data stands out as a field where the analysis of spatial data has relied on explicit spatio-temporal models for what are essentially spatial data. Genetic or genomic data such as microsatellite alleles or single nucleotide polymorphisms (SNPs) are collected from multiple individuals in spatially-referenced locations. Common goals in the analysis of this spatial data are to delineate spatial boundaries separating genetically distinct sub-populations (e.g., Guillot et al. 2005) and to understand spatial gene flow (or migration rates) between spatial regions (e.g., Manel et al. 2003, Wilson & Rannala 2003, Beerli 2006, Beerli & Palczewski 2010).

While some spatial genetic data are collected across multi-generational time scales (e.g., the analysis of museum specimans or historical data), most spatial genetic studies rely on spatial data that are essentially a snapshot in time of the spatio-temporal genetic process. However, the goal of understanding gene flow and migration has led to models for spatial genetic data that have an explicit spatio-temporal motivation.

Multiple approaches (e.g., Wilson & Rannala 2003, Beerli 2006, Beerli & Palczewski 2010) consider explicit spatio-temporal population models using coalescent theory (Kingman 1982). These approaches allow for asymmetric migration between populations (or spatial locations), making it possible to identify source populations and directional gene flow from spatial genetic data. However, these models do not typically explicitly incorporate space, distance between demes, or different migration rates through different terrain. The result is that these models provide a flexible model for gene flow in a specific system, but the results of any analysis are not easily generalized outside of the system.

In contrast to this are models for gene flow that attempt to be generalizable by modeling gene flow or genetic distance as functions of the distance between demes and the characteristics of the habitat in the space separating demes. The simplest of these approaches is an isolation by distance (IBD) approach, in which genetic distance between individuals or populations is correlated with geographic distance between the spatial locations where genetic samples are obtained (Wright 1943). More recent approaches consider spatio-temporal migration to be modeled as a symmetric random walk on a landscape grid (McRae 2006, Hanks & Hooten 2013), or on a graph, with migration rates on the edges defined by the landscape between graph nodes (Rioux Paquette et al. 2014). These approaches allow for generalizability of results through parametric modeling of the effects of landscape features on genetic correlation, but do not allow for directional gene flow.

The goal of this paper is to propose a general constructive approach to modeling spatial correlation based on considering the stationary distribution of a spatio-temporal generating process. For genetic data, we consider an asymmetric random walk model for gene flow, with random walk rates a function of local landscape features. Considering the stationary distribution of this spatio-temporal random walk results in a spatial model for genetic correlation that captures asymmetric gene flow within a parametric framework.

In Section 2, we introduce a motivating example of trout gene flow in a stream network, and propose a model for the observed genetic samples that relies on latent spatially-correlated random effects. In Section 3, we review existing models for spatial data motivated by spatio-temporal processes. In Section 4, we develop a population-level spatio-temporal random walk model and show that its stationary distribution corresponds to a particular intrinsic simultaneous autoregressive model for spatial covariance. In Section 5, we use this model to make inference on asymmetric migration rates of trout gene flow. Finally, in Section 6 we discuss possible extensions to this constructive spatio-temporal approach to modeling spatial covariance.

2 Modeling spatial genetic variation on stream networks

We will focus our analysis on the system studied by Kanno et al. (2011*b*), consisting of trout in the Jefferson-Hill Spruce Brook in Connecticut, USA. 470 trout were captured at 173 spatial locations along the brook (Figure 1) and genotyped, with microsatellite allele data obtained at 8 loci. All data are available online (Kanno et al. 2011*a*). Of interest in this system is understanding the influence that the characteristics of the stream network have on gene flow and migration. Kanno et al. (2011*b*) identify two seasonal blockages of the brook - two locations where the brook dries up in the late Summer and Fall and is seasonally impassible to the trout (indicated by the solid lines at “a” and “b” in Figure 1). The hypothesized drivers of gene flow and genetic connectivity among the trout population on Jefferson Hill Spruce Brook are both directional (differential rates of movement upstream and downstream) and non-directional (decreased connectivity between stream locations on opposite sides of the seasonal blockages).

Kanno et al. (2011*b*) employed the genetic clustering algorithm of Pritchard et al. (2000) to identify genetically distinct sub-populations of trout in the stream network. They then used the approach of (Beerli 2006) to estimate migration rates between these sub-populations. The results of this analysis indicated that there may be asymmetric migration rates between some subpopulations, with potentially greater migration downstream than upstream. In addition, some genetic sub-population boundaries corresponded with the locations of the seasonal blockages. The analysis done by Kanno et al. (2011*b*) provides detailed information about the genetic landscape of trout in the Jefferson Hill Spruce Brook; however, it is unclear how to generalize their results to other similar systems. In particular, the analysis of Kanno et al. (2011*b*) does not explicitly model the spatial locations of the samples, the directionality of stream flow, or the known locations of the seasonal blockages, making it difficult to quantitatively predict how such features might impact gene flow in another stream network.

An alternative approach to analyzing these spatially-referenced genetic data could be based on modeling second-order structure (pairwise genetic distance or correlation). Under an isolation by resistance (IBR) approach (McRae 2006, Hanks & Hooten 2013), we could discretize the stream network into many nodes (for example one node at each spatial location where trout genetic samples were obtained). The rate of movement between neighboring nodes could then be modeled parametrically as a function of the spatial characteristics of the stream network. This would allow for the quantification of the effect of the seasonal blockages on gene flow by modeling different movement rates between nodes separated by the seasonal blockages, relative to movement rates between nodes not separated by the seasonal blockages. However, existing distance-based or correlative approaches to gene flow do not allow for asymmetry in migration rates between neighboring nodes (McRae 2006, McRae et al. 2008, Cushman & Lewis 2010, Guillot et al. 2013, Hanks & Hooten 2013), making it impossible to model asymmetric gene flow using existing methods.

It may seem counter-intuitive to attempt to model any sort of asymmetry using symmetric measures such as spatial correlation (Hanks & Hooten 2013) or genetic distance (Kanno et al. 2011*b*). However, we will show that it is indeed possible to model the symmetric spatial correlation of a spatio-temporal process with asymmetric (directional) movement rates, and that such directional asymmetry (drift) is identifiable from spatial data in all but pathological scenarios. Before doing so, we present a model for the observed trout genetic data.

2.1 A Spatial Categorical Model for Microsatellite Allele Data

Microsatellite allele data were observed at $L = 8$ distinct loci for each of the $N = 470$ spatially referenced trout captured at $S = 173$ distinct spatial locations (Figure 1) in the Jefferson Hill Spruce Brook stream network. At the ℓ^{th} locus, $\ell = 1, 2, \ldots, L$, denote the list of of K_ℓ distinct observed alleles from all individuals in the study as $\{a_{\ell 1}, a_{\ell 2}, \ldots, a_{\ell K_\ell}\}$. Following Guillot et al. (2005) and others, we model the two observed alleles for each (diploid) individual as arising 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.

We follow Hanks et al. (2016) and specify a multinomial probit model (e.g., Albert & Chib 1993) for the categorical allele data in terms of latent variables $\{z_{sip\ell k}\}$. 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. Following the multinomial probit model of Albert & Chib (1993), let

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

where

$$z_{sip\ell k} \sim N(\mu_{\ell k} + \eta_{s\ell k}, 1). \tag{2}$$

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 $z_{sip\ell k}$ in (2) consists of the sum of $\mu_{\ell k}$, an allele specific intercept, and $\eta_{s\ell k}$, which is a spatially varying random effect that allows the allele frequencies $\mathbf{p}_{s\ell}$ to vary over the stream network. 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 in this case the k^{th} allele will be more prevalent than the $(k')^{\text{th}}$ allele. The model (1)-(2) is invariant to a constant increase or decrease in all $\mu_{\ell k}$, as the likelihood is a function of the contrasts $z_{sip\ell k} - z_{sip\ell k'}$, rather than the values of $z_{sip\ell k}$. That is, the likelihood of the observed allele data is invariant to a shift a to all latent means (replacing $\mu_{\ell k}$ with $\mu_{\ell k} + a$ for $k = 1, 2, \ldots, K_\ell$). We thus fix $\mu_{\ell 1} = 0$ for $\ell = 1, 2, \ldots, L$, as only the relative differences (contrasts) in $\mu_{\ell k}$ are identifiable (Hanks et al. 2016).

We now consider the spatially varying random effect $\eta_{s\ell k}$ in (2). Let

$$\boldsymbol{\eta}_{\ell k} = \begin{bmatrix} \eta_{1\ell k} \\ \eta_{2\ell k} \\ \vdots \\ \eta_{n\ell k} \end{bmatrix} \sim N(\mathbf{0}, \boldsymbol{\Sigma}), \quad \mathbf{1}'\boldsymbol{\eta}_{\ell k} = 0 \tag{3}$$

where $\mathbf{\Sigma}$ is a spatial covariance matrix. Note that we are constraining each $\boldsymbol{\eta}_{\ell k}$ to be centered at zero to avoid confounding with the intercept $\mu_{\ell k}$ (Hodges & Reich 2010, Paciorek 2010, Hughes & Haran 2013, Hanks, Schliep, Hooten & Hoeting 2015). In a similar model to (1)-(3), Hanks & Hooten (2013) parameterized a covariance matrix $\mathbf{\Sigma}$ using landscape covariates embedded in a conditional autoregressive (CAR, Besag 1974, Besag & Kooperberg 1995, Rue & Held 2005) model corresponding to an IBR approach to modeling genetic distance. However, Hanks & Hooten (2013) were unable to model asymmetry in gene flow - which is one of the main goals of our analysis of gene flow in the Jefferson Hill Spruce Brook system.

3 Using Spatio-Temporal Processes to Model Spatial Covariance

To model asymmetry in gene flow, we will consider spatio-temporal generating processes that directly model asymmetry, and use the stationary distribution of these processes as a spatial model that captures the asymmetry in the spatio-temporal process. We first review similar existing approaches in the literature, and then present a spatio-temporal random walk model of asymmetric migration and gene flow for the trout system of Kanno et al. (2011*b*). Deriving the spatial covariance function of the stationary distribution of this spatio-temporal process results in a parametric spatial covariance for the latent allelic processes in (3) that is based on an asymmetric spatio-temporal random walk model of gene flow.

3.1 Continuous Spatio-Temporal and Spatial Processes

Early work on modeling spatial correlation focused on extending one-dimensional models for time-series autocorrelation to two (and higher) dimensional models for spatial autocorrelation (e.g., Whittle 1954). In continuous space, an early model for spatially autocorrelated random variables is the solution to the following stochastic partial differential equation (SPDE) in two spatial dimensions (denoted here by x_1 and x_2)

$$(\kappa^2 - \Delta)^{\alpha/2}\boldsymbol{\pi}(x_1, x_2) = \mathbf{W}(x_1, x_2), \tag{4}$$

where $\Delta = \frac{\partial^2}{\partial x_1^2} + \frac{\partial^2}{\partial x_2^2}$ is the Laplacian and $\mathbf{W}(x_1, x_2)$ is spatial Gaussian white noise that is constant in time. The solution to this equation is a random field of the Matern class (Whittle 1954, Lindgren et al. 2011), with covariance function given by

$$\mathrm{cov}(\mathbf{s}_i, \mathbf{s}_j) = \sigma^2 \frac{1}{\Gamma(\nu)2^{\nu-1}} \left(\sqrt{2\nu} \frac{d_{ij}}{\phi} \right)^{\nu} K_\nu \left(\sqrt{2\nu} \frac{d_{ij}}{\phi} \right)$$

where $d_{ij} = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2}$ is the Euclidean distance between the spatial locations of the i-th and j-th observations, σ^2 is the partial sill parameter, ν is the Matern smoothness parameter, ϕ is a range parameter, and $K_\nu(\cdot)$ is the modified Bessel function of the second kind (e.g., Cressie 1993).

While this formulation of a spatial process has no explicit temporal variation, it is easily seen (but has not to our knowledge been noted in the literature) that (4) specifies the stationary distribution $\boldsymbol{\pi}$ of a spatio-temporal process defined by the spatio-temporal SPDE

$$\frac{\partial}{\partial t} \mathbf{y}(x_1, x_2, t) = (\Delta - \kappa^2)^{(\nu+1)/2} \mathbf{y}(x_1, x_2, t) + \mathbf{W}(x_1, x_2). \tag{5}$$

In the case where $\kappa = 0$ and $\nu = 1$, (5) defines a spatio-temporal diffusion process with spatial sources and sinks defined by $\mathbf{W}(x_1, x_2)$. Thus the most commonly used form for spatial autocorrelation in continuous-space can be seen as the stationary distribution of a spatio-temporal process driven by temporally-persistent white noise.

For modeling trout gene flow, We could consider extending the spatio-temporal model (5) by adding a drift term to model increased gene flow downstream, and model nonstationarity by specifying a spatially heterogeneous diffusion rates, as is done in spatial deformation approaches to modeling nonstationary spatial covariance (e.g., Schmidt & O'Hagan 2003, Lindgren et al. 2011). While this approach would be appealing to model gene flow in continuous two-dimensional space, the topology of a stream network does not allow for a straightforward application of Matern-class covariance (Ver Hoef & Peterson 2010).

Instead of a continuous-space model, we will consider discretizing the stream network into a large set of discrete nodes on the stream network, and consider areal (graphical) models for gene flow and spatial covariance.

3.2 Links Between Spatio-Temporal and Spatial Areal Processes

There have been multiple links made in the literature between areal spatial models, in particular simultaneous autoregressive (SAR) models, and spatio-temporal processes. The standard SAR model can be written (see e.g., Section 4.2.7 of Cressie & Wikle 2011) as

$$\mathbf{y} \sim N(\mathbf{0}, (\mathbf{I} - \mathbf{B})^{-1}\mathbf{\Lambda}(\mathbf{I} - \mathbf{B}')^{-1}) \tag{6}$$

where $\mathbf{B}$ has zeroes on the diagonal and $\mathbf{\Lambda}$ is a diagonal matrix with i-th diagonal Λ_{ii}.

In the spatial econometrics literature, LeSage & Pace (2009) consider a discrete-time lagged autoregressive process

$$\mathbf{y}_t = \rho\mathbf{B}\mathbf{y}_{t-1} + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}_t \tag{7}$$

where $\rho \in (0, 1)$, $\boldsymbol{\epsilon}_t$ is a vector of independent, zero-mean Gaussian random variables with variance σ^2, and $\mathbf{B}$ is a row-stochastic matrix describing the time-lagged dependence in $\mathbf{y}_t$ on the values at the previous time step ($\mathbf{y}_{t-1}$). The idea behind this formulation is that at time t, the mean of y_{kt}, the k-th element of $\mathbf{y}_t$, is a function of covariates $\mathbf{x}_k'\boldsymbol{\beta}$ and the process at neighboring locations, with neighborhood weights given by the off-diagonal elements of the $k-th$ row of $\mathbf{B}$ (see LeSage & Pace 2009, pp.25-26 for additional details and motivation). Recursive expansion of $\{\mathbf{y}_{t-q}, q = 1, 2, \ldots\}$ results in the following model for $\mathbf{y}_t$

$$\mathbf{y}_t \sim N((\mathbf{I} - \rho\mathbf{B})\mathbf{X}\boldsymbol{\beta}, (\mathbf{I} - \rho\mathbf{B})(\mathbf{I} - \rho\mathbf{B})' + \sigma^2\mathbf{I}). \tag{8}$$

If $\mathbf{X}\boldsymbol{\beta} = \mathbf{0}$, then $\mathbf{y}_t$ follows a SAR model (6) with an additional nugget σ^2 to model small-scale spatial variability. When $\mathbf{X}\boldsymbol{\beta} \neq \mathbf{0}$, the term $(\mathbf{I} - \rho\mathbf{B})$ jointly smooths the marginal mean, modeled by $\mathbf{X}\boldsymbol{\beta}$ in (8), and models areal spatial covariance. A similar form was found by Hooten et al. (2013) in the modeling of pre-smoothed resource utilization functions. This joint smoothing of mean and residual structure is not directly useful in our trout gene flow system, as the observed alleles are chosen from locations on the genome where selective forces are not at work, and variation in genotype is assumed to be a result of genetic drift alone. However, this joint modeling of mean and covariance common in the spatial econometrics literature holds strong potential for use in spatial statistical models, which typically model mean and covariance individually. Recent work on spatial confounding (Hodges & Reich 2010, Hughes & Haran 2013, Hanks, Schliep, Hooten & Hoeting

2015) has shown that spatial covariance can strongly influence the estimation and interpretation of mean structure, and joint modeling of mean and covariance, as in (8), provides a principled approach to disentangling the effects of mean structure (as modeled by $\mathbf{X}\boldsymbol{\beta}$) and spatial covariance (as modeled by $\mathbf{B}$).

As $\mathbf{B}$ in (7) is a row-stochastic matrix, we could consider modeling the latent allelic processes in (3) using a discrete time Markov random walk with probability transition matrix $\mathbf{B}$, where B_{ij} is the probability of a trout being in spatial location j at time t, given that it was at spatial location i at time $t-1$. Here t could represent generations, years, or some other fixed unit of time. We could then model asymmetry and nonstationarity in this spatio-temporal random walk by modeling B_{ij}. Assunção & Krainski (2009) note links between proper SAR and discrete-time Markov chains, though they do not explore the spatio-temporal implications further. This general approach is appealing, but offers some difficulties in the trout system of Kanno et al. (2011*b*). If we divide the stream network (Figure 1) into a discrete set of locations, it isn't clear a priori how to model sparsity in $\mathbf{B}$ - that is, if $\mathbf{B}$ is the 1-year transition probability matrix for trout moving between locations in the stream network, it isn't immediately clear which transitions should be impossible in that time frame. Instead of focusing on a discrete-time random walk, we will later consider a related continuous-time random walk model for trout movement and gene flow. The continuous-time process is appealing in the case of a stream network as it provides a natural approach for modeling sparsity, as in continuous time a trout must move to a neighboring location upstream or downstream before moving to locations further away. Thus, by assuming a continuous-time random walk, we only need to model the transition rates to the nearest neighboring locations on the stream network. In addition, the continuous-time framework will allow us to prove in Section 4.3 that directional bias in a spatio-temporal random walk is identifiable from spatial data alone, allowing us to infer the effect of stream flow on trout gene flow.

4 Spatial models from spatio-temporal random walks

For our trout genetic system, we will consider a large population spatio-temporal random walk process, and use the spatial covariance of the stationary distribution of the spatio-temporal random walk as the spatial covariance $\mathbf{\Sigma}$ of our latent spatial allele random effects in (3). The approach we consider is analogous to the links between continuous spatial models and diffusive spatio-temporal processes described in Section 3.1, and consists of the following general approach.

1. Define a stochastic spatio-temporal generating model for the spatio-temporal process $\mathbf{y}(s,t)$, where s indexes space and t indexes time in which the process is driven with time-persistent spatial noise $\mathbf{W}(s)$

   $$\frac{\partial}{\partial t}\mathbf{y}(s,t) = \mathcal{F}\left(\mathbf{y}(s,t)\right) + \mathbf{W}(s) \quad , \quad \mathbf{W}(s) \sim N(\cdot,\cdot).$$

2. Finding the distribution of the limiting random field $\boldsymbol{\pi}(s) = \lim_{t\to\infty}\mathbf{y}(s,t)$ of (5), when it exists, provides a spatial model capturing the dynamics of the spatio-temporal process.

   $$\frac{\partial}{\partial t}\mathbf{y}(s,t) = 0 \quad \Rightarrow \quad \mathcal{F}\left(\boldsymbol{\pi}(s)\right) = -\mathbf{W}(s)$$

4.1 Continuous-time Markov random walks

Continuous-time random walk models are among the most common models for gene flow. McRae (2006) showed that under a random walk model for migration, a common formulation of genetic dissimilarity (the linearized fixation index) was proportional to the circuit resistance distance (Klein & Randić 1993) between the nodes in question. Under the formulation of McRae (2006), the spatial domain is envisioned as a graph of spatial nodes with symmetric edge weights α_{ij} proportional to the (symmetric) rate of random walkers between nodes (Figure 2). The resistance distance is the effective resistance in an electric circuit where the nodes are connected by resistors with resistance $1/\alpha_{ij}$ equal to the inverse of the migration rate (Klein & Randić 1993). This approach to studying gene flow is known as the isolation by resistance (IBR) approach, and is often used to explore the relationship between landscape features and gene flow. The IBR approach assumes symmetric edge weights (and thus symmetric migration rates), and thus is incapable of modeling asymmetric gene flow upstream and downstream in the Jefferson Hill Spruce Brook stream network.

Here we will consider a Markov random walk on a discrete spatial support, which leads to a population-level diffusion SPDE based on a large-population approximation to the spatial movements of many individuals. The stationary distribution of this SPDE will provide a spatial covariance model based on a population-level spatial random walk or diffusion process. We will use this resulting covariance matrix in our latent spatial allele process model (3).

Consider a Markov random walk (Markov jump process) on a discrete spatial support. Let $\mathbf{G} = (\mathbf{V}, \mathbf{E})$ be a graph with vertices $\mathbf{V} = \{V_1, V_2, \ldots, V_M\}$ and nonnegative directed edges $\mathbf{E} = \{\alpha_{ij}, i = 1, 2, \ldots, M; j = 1, 2, \ldots, M\}$. In particular, consider the case where α_{ij} is the exponential rate at which a random walker in node i transitions to node j. As in a standard continuous-time Markov chain (CTMC) model for a random walk, the time T_i spent by a random walker in node i before transitioning to any other node is exponentially-distributed with rate $\alpha_i = \sum_{k=1}^{n} \alpha_{ik}$.

Now assume a population-level process on the graph $\mathbf{G}$ in which there are N members of the population, all moving independently based on the CTMC model defined by $\mathbf{G}$. If there are $n_i(t)$ individuals at node i and time t, then the rate at which individuals move from node i to node j is given by $n_i\alpha_{ij}$. In an open population model, individuals may enter (births) or leave (deaths) the system continuously in time at any node (Figure 2). It is common to model the birth and death rates at node i as being density dependent, with birth rate of n_ib and death rate of n_id, for constant rates b and d shared across space. Instead, we will allow the birth and death rates to vary spatially, as this will provide a convenient mechanism for accounting for unmodeled spatial variation. To this end, consider birth and death rates that scale with the total population size (N). Let Nb_i be the rate at which individuals are introduced into node i and let Nd_i be the rate at which individuals in node i are removed from the system.

Given an initial population state $\mathbf{n}(0)$ at time zero, the transient distribution $\mathbf{n}(t)$ is given by (e.g., Baxendale & Greenwood 2011)

$$\mathbf{n}(t) = \mathbf{n}(0) + \sum_{ij\neq 0}(\mathbf{e}_j - \mathbf{e}_i)P_{ij}\left[\int_0^t n_i(s)\alpha_{ij}\mathrm{d}s\right] + \sum_i\left(\mathbf{e}_iP_{0i}\left[\int_0^t Nb_i\mathrm{d}s\right] - \mathbf{e}_iP_{i0}\left[\int_0^t Nd_i\mathrm{d}s\right]\right)$$

where

$$P_{ij}(a) \sim Pois(a), \quad i = 0, 1, \ldots, M;\ j = 0, 1, \ldots, M;\ i \neq j$$

and $\mathbf{e}_i$ is the canonical vector with M componants, all of which are zero except for the i-th element, which is equal to 1. Following Kurtz (1978) and Baxendale & Greenwood (2011), the normalized population process $\mathbf{z}(t) = [z_1(t)\ z_2(t)\ \ldots z_M]$ can be defined as $z_i(t) = n_i(t)/N$. The transient distribution for this normalized density is given by

$$\mathbf{z}(t) = \mathbf{z}(0) + \sum_{ij\neq 0}(\mathbf{e}_j - \mathbf{e}_i)\frac{1}{N}P_{ij}\left[\int_0^t n_i(s)\alpha_{ij}\mathrm{d}s\right] + \sum_i \mathbf{e}_i\left(\frac{1}{N}P_{0i}\left[Nb_it\right] - \frac{1}{N}P_{i0}\left[Nd_it\right]\right). \tag{9}$$

Taking the large population limit as $N \to \infty$ (Kurtz 1978, Baxendale & Greenwood 2011) gives the integral equation for the normalized density

$$\mathbf{z}(t) = \mathbf{z}(0) + \sum_{i\neq j}(\mathbf{e}_j - \mathbf{e}_i)\int_0^t z_i(s)\alpha_{ij}\mathrm{d}s + \sum_i \mathbf{e}_i(b_i - d_i)t. \tag{10}$$

Details of this calculation are given in Appendix A.

The differential equation associated with (10) is

$$\frac{\partial \mathbf{z}(t)}{\partial t} = \sum_{i\neq j}\alpha_{ij}(\mathbf{e}_j - \mathbf{e}_i)z_i(t) + \sum_i \mathbf{e}_i(b_i - d_i)$$

which has vectorized form

$$\frac{\partial \mathbf{z}(t)}{\partial t} = -\mathbf{Q}'\mathbf{z}(t) + (\mathbf{b} - \mathbf{d}) \tag{11}$$

where $\mathbf{b} = [b_1\ b_2\ \ldots\ b_M]'$, $\mathbf{d} = [d_1\ d_2\ \ldots\ d_M]'$, and $\mathbf{Q}$ is the infinitessimal generator of the CTMC or the Laplacian matrix of the graph

$$\mathbf{Q} = \begin{pmatrix} \sum_k \alpha_{1k} & -\alpha_{12} & -\alpha_{13} & \cdots & -\alpha_{1m} \\ -\alpha_{21} & \sum_k \alpha_{2k} & -\alpha_{23} & \cdots & -\alpha_{2m} \\ -\alpha_{31} & -\alpha_{32} & \sum_k \alpha_{3k} & \cdots & -\alpha_{3m} \\ \vdots & & & \ddots & \vdots \\ -\alpha_{m1} & -\alpha_{m2} & -\alpha_{m3} & \cdots & \sum_k \alpha_{mk} \end{pmatrix}. \tag{12}$$

Equation (11) specifies a graph diffusion process where $\mathbf{b} - \mathbf{d}$ is a vector of net inputs and outputs to the system and $-\mathbf{Q}'$ is a matrix derived from proportional rates of transfer between spatial locations.

4.2 Spatial Models From Random Walks

To specify a spatial model motivated by the differential equation (11), consider modeling the spatial birth and death rates as time-persistent Gaussian spatial white noise

$$\boldsymbol{\gamma} = \mathbf{b} - \mathbf{d} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$$

subject to the constraint that $\mathbf{1}'\boldsymbol{\gamma} = 0$. This sum-to-zero constraint on $\boldsymbol{\gamma}$ is necessary to ensure the existence of a stationary distribution $\boldsymbol{\pi}$ for (11). The spatio-temporal differential equation (11) is now stochastic and can be written as

$$\frac{\partial}{\partial t}\mathbf{z}(t) = -\mathbf{Q}'\mathbf{z}(t) + \boldsymbol{\gamma}, \quad \boldsymbol{\gamma} \sim N(\mathbf{0}, \sigma^2\mathbf{I}). \tag{13}$$

The limiting field $\boldsymbol{\pi} = \lim_{t\to\infty} \mathbf{z}(t)$ for the normalized population process satisfies the balance equation that $\frac{\partial}{\partial t}\mathbf{z}(t) = \mathbf{0}$, which implies that

$$\mathbf{Q}'\boldsymbol{\pi} = \boldsymbol{\gamma}, \quad \boldsymbol{\gamma} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$$

and thus the distribution for this limiting field, which is the stationary distribution of (13), is given by

$$\boldsymbol{\pi} \sim N(\mathbf{0}, \sigma^2(\mathbf{Q}\mathbf{Q}')^{-}), \text{ with } \mathbf{1}'\boldsymbol{\pi} = 0. \tag{14}$$

This stationary distribution is a random field on the discrete spatial support of the population process $\mathbf{z}(t)$ with spatial covariance defined by the spatio-temporal CTMC random walk with infinitessimal generator $\mathbf{Q}$ (12).

The random field $\boldsymbol{\pi}$ in (14) is an intrinsic random field, in that only linear combinations are proper (Besag & Kooperberg 1995). An alternative formulation is that the density for $\boldsymbol{\pi}$ is proper under the constraint that $\boldsymbol{\pi}$ sums to zero over the spatial domain. Intrinsic random fields are often used as prior distributions, where the posterior distribution is proper. For example, consider modeling a Gaussian response $\mathbf{y}$ as

$$\mathbf{y} = \mu\mathbf{1} + \boldsymbol{\pi} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim N(\mathbf{0}, \tau^2\mathbf{I})$$

where $\boldsymbol{\pi} \sim N(\mathbf{0}, (\mathbf{Q}\mathbf{Q}')^{-})$, with $\mathbf{1}'\boldsymbol{\pi} = 0$. Under this formulation, $\boldsymbol{\pi}$ is constrained to sum to zero, but $\mu\mathbf{1} + \boldsymbol{\pi}$ is not. This formulation can be seen as a form of restricted spatial regression (Hughes

& Haran 2013, Hanks, Schliep, Hooten & Hoeting 2015) where the spatial random effect $\boldsymbol{\pi}$ is constrained to be orthogonal to the intercept $\mu\mathbf{1}$.

The random field in (14) corresponds to an intrinsic SAR model (6) for spatial correlation. Setting

$$B_{ij} = \frac{\alpha_{ji}}{\sum_k \alpha_{ik}} \text{ and } \Lambda_{ii} = \frac{1}{(\sum_k \alpha_{ik})^2}$$

in (6) expresses (14) as an intrinsic SAR model. As noted by Lindgren et al. (2011), any SAR model can be written as a CAR model with a different neighborhood structure. Thus, these results for SAR models can also apply to corresponding spatial random fields with CAR structure. As in standard SAR models, the matrix $\mathbf{Q}$ from (10) does not have to be symmetric, but rather can incorporate models for asymmetric random walks. Additionally, if $\mathbf{Q}$ is sparse (many of the $\{\alpha_{ij}\}$ are zero), then sparse matrix methods (e.g., Rue 2001, Rue & Held 2005) can be employed to sample from and evaluate the density in (14).

4.3 Identifiability

We have proposed the model in (14) as an appropriate distribution for spatial data arising from a spatio-temporal CTMC random walk with infinitessimal generator $\mathbf{Q}$, with the goal of using this model to quantify potential asymmetry in spatio-temporal trout gene flow using only spatial (not spatio-temporal) data. While $\mathbf{Q}$ is not symmetric, the precision matrix $\mathbf{QQ}'$ is, and it isn't immediately clear that directional bias in $\mathbf{Q}$ is identifiable from spatial data alone.

The likelihood of (14)

$$f(\boldsymbol{\pi}|\mathbf{Q}) \propto |\mathbf{QQ}'|^{-1/2} \exp\left\{-\frac{1}{2\sigma^2}\boldsymbol{\pi}'\mathbf{QQ}'\boldsymbol{\pi}\right\}$$

is a function of $\mathbf{QQ}'$, rather than purely a function of the infinitessimal generator $\mathbf{Q}$. Thus, if there are two generator matrices $\mathbf{Q}$ and $\mathbf{W}$ such that $\mathbf{QQ}' = \mathbf{WW}'$, then $\mathbf{Q}$ is not identifiable. However, the special structure required for a generator matrix of a CTMC allows us to prove that $\mathbf{Q}$ is identifiable in all but pathological situations.

Theorem 4.1 *If $\mathbf{Q}$ and $\mathbf{W}$ are both generator matrices (12) for irreducible M-state CTMCs, and at least one row of $\mathbf{Q}$ has more than one nonzero off-diagonal entry, then $\mathbf{QQ}'$=$\mathbf{WW}'$ if and only if $\mathbf{Q} = \mathbf{W}$.*

The proof is given in Appendix B. The significance of this result is that the only forms for **Q** that are unidentifiable come when the embedded chain of the irreducible CTMC governed by **Q** is deterministic and topologically the graph given by **Q** is a single closed loop, with flow only possible in one direction (either clockwise or counter-clockwise). In all other graph topologies, identifiability is guaranteed.

The multiplicative relationship between the variance of the spatial noise σ^2 and $\mathbf{QQ}'$ in the likelihood of (14) means that only proportional movement rates α_{ij} are identifiable. Taking this together with the result of Theorem 3.1, it is clear that *it is possible in almost all areal spatial cases to infer proportional spatio-temporal movement rates from spatial data alone*.

5 Analysis of Trout Genetics

This spatio-temporal model for spatial data can now be applied to the trout genetic data of Kanno et al. (2011*b*). The hypothesized drivers of gene flow and genetic connectivity among the trout population on Jefferson Hill Spruce Brook are both directional (differential rates of movement upstream and downstream) and non-directional (decreased connectivity between stream locations on opposite sides of the seasonal blockages). If spatio-temporal trout movement data were available, modeling these directional and non-directional responses to covariates would be straightforward (Hooten et al. 2010, Hanks, Hooten & Alldredge 2015). For example, movement could be modeled as a CTMC occuring on a graph with a node at each spatial location where trout were sampled, and edge weights equal to random walk transition rates between nodes could be modeled as

$$\alpha_{ij} = \begin{cases} \frac{1}{d_{ij}} \exp\left\{\beta_0 + \beta_1 u_{ij} + \beta_2 v_{ij}\right\} & \text{if nodes } i \text{ and } j \text{ are neighbors} \\ 0 & \text{otherwise} \end{cases}$$

where $\{u_{ij}\}$ and $\{v_{ij}\}$ are indicator variables with $u_{ij} = 1$ if node j is downstream from node i and $v_{ij} = 1$ if a seasonal blockage is located between nodes i and j. In this formulation, each node on a branch of the stream network has two neighbors, one upstream and one downstream, and edge weights α_{ij} are zero for all other non-neighboring nodes. Each node at a confluence of two stream branches will have three neighbors, one downstream and two upstream. The rate at which a random walker at a node i on a branch of the stream network transitions to the nearest upstream node j is $\alpha_{ij} = 1/d_{ij}\exp\{\beta_0\}$ if there is not a seasonal blockage between nodes i and j. Similarly, the rate at which the random walker transitions from i to the nearest downstream node k is $\alpha_{ik} = 1/d_{ik}\exp\{\beta_0 + \beta_1\}$. The parameter β_2 models the additive effect that a seasonal blockage has on log-transition rates. Together, this simple random walk model allows for transition rates that vary with direction and location based on known spatial stream characteristics.

The infinitessimal generator of this CTMC random walk is $\mathbf{Q}$ from (12), and the stationary distribution of the related spatio-temporal process in (11) is given by (14). We can thus specify a complete hierarchical statistical model for spatially-referenced microsatellite allele data based on an underlying spatio-temporal asymmetric random walk model for gene flow. We repeat (1)-(2) here for clarity. We model the observed alleles using a multinomial probit model

$$y_{sip\ell k} = \begin{cases} 1 & , \ z_{sip\ell k} = \max\{z_{sip\ell a},\ a = 1, \ldots, K_\ell\} \\ 0 & , \ \text{otherwise} \end{cases} \tag{15}$$

$$z_{sip\ell k} \sim N(\mu_{\ell k} + \eta_{s\ell k}, 1). \tag{16}$$

Each latent spatial allelic process is modeled using the random walk covariance model developed in Section 3.2

$$\boldsymbol{\eta}_{\ell k} \sim N(\mathbf{0}, (\mathbf{QQ}')^{-}),\ \text{with } \mathbf{1}'\boldsymbol{\eta}_{\ell k} = 0. \tag{17}$$

$$\mathbf{Q} = \begin{pmatrix} \sum_k \alpha_{1k} & -\alpha_{12} & -\alpha_{13} & \cdots & -\alpha_{1m} \\ -\alpha_{21} & \sum_k \alpha_{2k} & -\alpha_{23} & \cdots & -\alpha_{2m} \\ -\alpha_{31} & -\alpha_{32} & \sum_k \alpha_{3k} & \cdots & -\alpha_{3m} \\ \vdots & & & \ddots & \vdots \\ -\alpha_{m1} & -\alpha_{m2} & -\alpha_{m3} & \cdots & \sum_k \alpha_{mk} \end{pmatrix} \tag{18}$$

with the random walk rates $\{\alpha_{ij}\}$ given by (13)

$$\alpha_{ij} = \begin{cases} \frac{1}{d_{ij}}\exp\left\{\beta_0 + \beta_1 u_{ij} + \beta_2 v_{ij}\right\} & \text{if nodes } i \text{ and } j \text{ are neighbors} \\ 0 & \text{otherwise} \end{cases} \tag{19}$$

The model is completed by specifying diffuse Gaussian priors for the random walk parameters $\beta_0,\ \beta_1,\ \beta_2$ and the allele specific intercepts

$$\beta_j \sim N(0, 10^2), \quad j = 0, 1, 2 \tag{20}$$

$$\mu_{\ell k} \sim N(0, 10^2), \quad \ell = 1, 2, \ldots, L; \quad k = 2, 3, \ldots, K_\ell. \tag{21}$$

Note that (17) differs from (14) by the omission of the variance parameter σ^2, which was omitted as it is unidentifiable from β_0 in (19). Similarly, note that (16)-(17) can be written in a combined form, marginalizing over $\boldsymbol{\eta}_{\ell k}$

$$\mathbf{z}_{\ell k} \sim N(\mu_{\ell k}\mathbf{1}, \mathbf{C}(\mathbf{Q}\mathbf{Q}')^{-}\mathbf{C} + \mathbf{I}) \tag{22}$$

where $\mathbf{z}_{\ell k} = [z_{111\ell k} \; z_{112\ell k} \; \cdots \; z_{211\ell k} \; \cdots]'$ is a vector of all $z_{sip\ell k}$ corresponding to allele k at locus ℓ, and $\mathbf{C}$ is a design matrix linking the p-th entry in $\mathbf{z}_{\ell k}$ to the correct spatial location ($C_{ps} = 1$ if the p-th entry of $\mathbf{z}_{\ell k}$ is from spatial location s, and $C_{ps} = 0$ otherwise). In this formulation, the likelihood is invariant to the choice of generalized inverse $(\mathbf{Q}\mathbf{Q}')^{-}$ as long as at least one node in the graph defined by $\mathbf{Q}$ does not have any observations in $\mathbf{z}_{\ell k}$ (see e.g., pp.119 of Harville 2008). To ensure this, we add one node to the end of each reach in the Jefferson Brook Spruce Hill stream network. This results in a spatial random field $\{\boldsymbol{\eta}_{\ell k}\}$ on a slightly larger spatial domain that the data, which has the additional benefit of minimizing edge effects which are common at the borders of spatial random fields.

We conducted a simulation example in which each μ_{lk} was drawn randomly from a standard normal distribution, and the random walk parameters were fixed as $[\beta_0, \beta_1, \beta_2] = [-1.2, 7, -.1]$, which are similar to the posterior mean estimates obtained below in Section 5.1. Using these values, we simulated genetic data from the model (15)-(19), with observation locations and the number of simulated alleles chosen to match the observed trout genetic data (8 alleles for each of 470 simulated trout at the locations shown in Figure 1). We then fit the simulated data using the approach described below in Section 5.1. The 95% equal-tailed credible intervals for all parameters except the intercept β_0 overlapped the true values, and the 95% equal-tailed interval for β_1 did not overlap zero. This simulation confirms the potential to identify proportional spatio-temporal random walk rates from spatial genetic data.

5.1 Inference

A Markov chain Monte-Carlo (MCMC) algorithm was constructed to sample from the posterior distribution of model parameters, given the observed microsatellite allele data. Full-conditional distributions are available for all parameters in the model (15)-(21) except for the random walk covariance parameters $\{\beta_j, j = 0, 1, 2\}$. Updates for these parameters were obtained using a random walk Metropolis step with the prior (20) and likelihood (22) which marginalizes over the spatial random effects $\{\boldsymbol{\eta}_{\ell k}\}$. This results in vastly better mixing than a similar random walk Metropolis step using the likelihood in (17). Ten chains were run with different starting values. Each chain was run for $58,000$ iterations, with the first $8,000$ samples discarded as burn in. Convergence was assessed by comparing posterior histograms obtained from only the first half of each chain with posterior histograms obtained from only the second half of each chain. Histograms of the marginal posterior distributions of the random walk parameters are given in Figure 3. The posterior distribution for β_1 is greater than zero, indicating that the data support the directional anisotropic hypothesis that gene flow is more rapid downstream than upstream. The posterior distribution for β_2, which captures the effect of the seasonal blockages, overlaps zero (Figure 3(c)), with the 95% equal-tailed credible interval being bounded by $(-0.49, 0.55)$. This indicates little or no support for the hypothesis that the seasonal blockages affect gene flow. A separate model was fit without the seasonal barrier (DIC=14003), and was found to be inferior to the model containing the barrier (DIC=13897) based on deviance information criterion (DIC; Spiegelhalter et al. 2002).

Posterior mean values for the allele specific intercepts $\mu_{\ell k}$ ranged from -2.2 to 0.7. These parameters correlate with the general prevalence of an allele in the trout population, but do not affect spatial gene flow, as modeled by the random walk process defined in Section 4.1.

To qualitatively illustrate the genetic correlation structure implied by the estimated random walk parameters $\boldsymbol{\beta}$, four realizations of random fields on the stream network were simulated using the posterior mean parameter values. These random fields are shown in Figure 4. The constructive spatio-temporal approach proposed here provides a valid autoregressive spatial model for data collected on a stream network. In contrast, Ver Hoef & Peterson (2010) present a moving average (convolution) approach to modeling spatial autocorrelation on stream networks.

6 Discussion

Current standard approaches to modeling spatial correlation focus on semiparametric random effect models. This work proposes a parametric constructive approach to modeling spatial random effects based on an assumed spatio-temporal generating process. We showed how existing scientific knowledge about the system (gene flow on a stream network) could be used to specify a spatio-temporal generating model (a population-level random walk). The stationary distribution of this spatio-temporal process then defines the distribution of the spatial random effect used to model genetic correlation.

While we have focused on discrete space models, this general approach has potential for application in continuous space as well. Spatial deformation approaches to nonstationary covariance (e.g., Schmidt & O'Hagan 2003, Lindgren et al. 2011) can be viewed as stationary distributions of diffusion processes with spatially heterogeneous diffusion rates. Reaction-diffusion models are common in ecology and other fields (e.g., Keeling et al. 2004, Hu et al. 2013) and would provide a natural spatio-temporal generating process basis for spatial random effect models in a wide variety of systems. Finite element basis and grid-based approaches to approximating continuous spatial fields have a long history in spatio-temporal (e.g., Wikle & Hooten 2010) and spatial (e.g., Lindgren et al. 2011) analysis, and could be used to approximate the stationary distribution of a continuous (infinite-dimensional) spatio-temporal generating process with a finite number of basis functions.

Modeling spatial random effects semiparametrically using CAR, SAR or Matern models is the current standard practice; however, there are benefits to parametric modeling of spatial random effects when the existing science can suggest a spatio-temporal generating mechanism. We have illustrated this by considering a spatio-temporal random walk model for gene flow, and modeling observed spatial correlation using the stationary distribution of of the random walk. Considering spatial random fields as limiting distributions of spatio-temporal processes will allow results and intuition for spatio-temporal processes to be applied to the modeling and analysis of spatial data in a wide variety of situations.

Appendix A: Large population limits of population processes

The interested reader is referred to Kurtz (1981) for a full treatment of stochastic population processes. This derivation follows the spirit of Kurtz (1981) and Baxendale & Greenwood (2011), but with the novelty of birth and death rates that are not density dependent.

Following from (6) in Section 3.1, the transient distribution for the normalized density $\mathbf{z} = \mathbf{n}/N$ is given by

$$\mathbf{z}(t) = \mathbf{z}(0) + \sum_{ij\neq 0}(\mathbf{e}_j - \mathbf{e}_i)\frac{1}{N}P_{ij}\left[\int_0^t n_i(s)\alpha_{ij}\mathrm{d}s\right] + \sum_i \mathbf{e}_i\left(\frac{1}{N}P_{0i}\left[Nb_it\right] - \frac{1}{N}P_{i0}\left[Nd_it\right]\right)$$

where

$P_{ij}(a) \sim Pois(a), \quad i = 0, 1, \ldots, M; \; j = 0, 1, \ldots, M; \; i \neq j.$

Note that

$$\begin{aligned} P_{ij}(a) &= a + (P_{ij}(a) - a) \\ &= a + W_{ij}(a)\,, \qquad W_{ij}(a) \sim (0, a) \end{aligned}$$

where each W_{ij} has mean zero on constant variance. Applying this to the transient distribution gives

$$\begin{aligned} \mathbf{z}(t) = \mathbf{z}(0) &+ \sum_{ij\neq 0}(\mathbf{e}_j - \mathbf{e}_i)\frac{1}{N}\left[\int_0^t n_i(s)\alpha_{ij}\mathrm{d}s\right] + \sum_i \mathbf{e}_i\left(b_it - d_it\right) \\ &+ \frac{1}{N}\left(\sum_{i\neq j}(\mathbf{e}_j - \mathbf{e}_i)W_{ij}\left[\int_0^t n_i(s)\alpha_{ij}\mathrm{d}s\right] + \sum_i \mathbf{e}_i\left(W_{0i}\left[Nb_it\right] - W_{i0}\left[Nd_it\right]\right)\right). \end{aligned}$$

Consider a fixed $t > 0$ and note that $N \geq n_i(s)$ for all $s \in (0, t)$. This gives the result that

$\int_0^t n_i(s)\alpha_{ij}\mathrm{d}s \leq N\alpha_{ij}t.$

Then to show that all terms above including random variables W_{ij} disappear in the limit as $N \to \infty$, it is enough to consider the behavior of

$\frac{1}{N}W(Na), \quad W(a) \sim (0, a)$

for a constant $a > 0$. It is trivial to note that

$$E\left[\frac{1}{N}W(Na)\right] = 0$$

and that

$$Var\left[\frac{1}{N}W(Na)\right] = \frac{1}{N^2}Na$$

which vanishes in the limit as $N \to \infty$.

Then, in the large population limit, the transient distribution of the normalized population $\mathbf{z}(t)$ will be given by

$$\mathbf{z}(t) = \mathbf{z}(0) + \sum_{i \neq j}(\mathbf{e}_j - \mathbf{e}_i)\frac{1}{N}\left[\int_0^t n_i(s)\alpha_{ij}\mathrm{d}s\right] + \sum_i \mathbf{e}_i\left(b_i t - d_i t\right).$$

Appendix B: Proof of Theorem 3.1

In this appendix, we prove Theorem 3.1. The proof follows from the fact that $\mathbf{QQ}'$ is a Gramian matrix (e.g., Gentle 2007) and thus $\mathbf{QQ}' = \mathbf{WW}'$ if and only if $\mathbf{W} = \mathbf{QU}'$ for a real unitary matrix $\mathbf{U}'$. As $\mathbf{W}$ and $\mathbf{Q}$ are both generators for CTMC random walks, their rows sum to zero ($\mathbf{Q1} = \mathbf{W1} = \mathbf{0}$), with negative diagonal entries ($q_{ii} < 0$, $w_{ii} < 0$) and non-negative off-diagonal entries ($q_{ij} \geq 0$, $w_{ij} \geq 0$ for $i \neq j$). If $\mathbf{Q}$ and $\mathbf{W}$ are both generators for irredicible CTMCs, then both matrices have rank $n - 1$ and their null spaces are both spanned by the $\mathbf{1}$ vector. As $\mathbf{W1} = \mathbf{0}$, it follows that $\mathbf{QU}'\mathbf{1} = \mathbf{0}$ and thus $\mathbf{U}'\mathbf{1} = \lambda\mathbf{1}$ for some λ. The eigenvalues of any unitary matrix $\mathbf{U}'$ have absolute value equal to 1, so λ either equals 1 or -1. If $\mathbf{u}_i'$ is the i-th row of $\mathbf{U}'$, then $\mathbf{u}_i'\mathbf{1}$ equals either 1 or -1, but since $\mathbf{U}$ is unitary, $\mathbf{u}_i'\mathbf{u}_i = 1$. These requirements both hold if and only if $\mathbf{u}_i = \lambda\mathbf{e}_k$, where $\mathbf{e}_k$ is the canonical vector with k-th element equal to 1 and all other elements equal to zero. As $\mathbf{U}$ is of full rank, the rows of $\mathbf{U}'$ must contain a full set of canonical vectors spanning $\mathcal{R}^n$.

First consider the case where $\lambda = 1$. Then $\mathbf{U}'$ is a permutation matrix, with the columns of $\mathbf{W}$ being permuted columns of $\mathbf{Q}$. However, as $\mathbf{W}$ and $\mathbf{Q}$ are generator matrices, each diagonal entry of $\mathbf{W}$ and $\mathbf{Q}$ must be negative, while all off-diagonal entries are non-negative. This can only hold for $\mathbf{W}$ if the permutation matrix $\mathbf{U}'$ is the identity matrix, and thus $\mathbf{W} = \mathbf{Q}$.

Now consider the case where $\lambda = -1$. Again $\mathbf{U}'$ permutes the columns of $\mathbf{Q}$, but now the sign of all entries is changed through multiplication by $\lambda = -1$. So $w_{ii} = -q_{ik}$ and $w_{ik} = -q_{ii}$ for some k. As $\mathbf{W}$ is a generator matrix, $w_{ii} = -\sum_{j \neq i} w_{ij}$, which is only possible if q_{ik} is the only non-zero off-diagonal entry in the i-th row of $\mathbf{Q}$. This completes the proof.

References

Albert, J. & Chib, S. (1993), ‘Bayesian analysis of binary and polychotomous response data’, *Journal of the American Statistical Association* **88**(422), 669–679.

Assunção, R. & Krainski, E. (2009), ‘Neighborhood dependence in Bayesian spatial models’, *Biometrical Journal* **51**(5), 851–869.

Baxendale, P. H. & Greenwood, P. E. (2011), ‘Sustained oscillations for density dependent Markov processes’, *Journal of mathematical biology* **63**(3), 433–457.

Beerli, P. (2006), ‘Comparison of Bayesian and maximum-likelihood inference of population genetic parameters’, *Bioinformatics* **22**(3), 341–345.

Beerli, P. & Palczewski, M. (2010), ‘Unified framework to evaluate panmixia and migration direction among multiple sampling locations’, *Genetics* **185**(1), 313–326.

Besag, J. (1974), ‘Spatial interaction and the statistical analysis of lattice systems’, *Journal of the Royal Statistical Society. Series B (Methodological)* **36**(2), 192–236.

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

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

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

Cushman, S. A. & Lewis, J. S. (2010), ‘Movement behavior explains genetic differentiation in American black bears’, *Landscape Ecology* **25**(10), 1613–1625.

Diggle, P. & Ribeiro, P. J. (2007), *Model-based geostatistics*, Springer.

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

Guillot, G., Mortier, F. & Estoup, A. (2005), ‘Geneland: a computer package for landscape genetics’, *Molecular Ecology Notes* **5**(3), 712–715.

Guillot, G., Vitalis, R., le Rouzic, A. & Gautier, M. (2013), 'Detecting correlation between allele frequencies and environmental variables as a signature of selection. A fast computational approach for genome-wide studies', *Spatial Statistics* **8**, 145–155.

Hanks, E. M. & Hooten, M. B. (2013), 'Circuit Theory and Model-Based Inference for Landscape Connectivity', *Journal of the American Statistical Association* **108**, 22–33.

Hanks, E. M., Hooten, M. B. & Alldredge, M. W. (2015), 'Continuous-time Discrete-space Models for Animal Movement', *The Annals of Applied Statistics* **9**(1), 145–165.

Hanks, E. M., Hooten, M. B., Knick, S. T., Oyler-McCance, S. J., Fike, J. A., Cross, T. B. & Schwartz, M. K. (2016), 'Latent spatial models and sampling design for landscape genetics', *The Annals of Applied Statistics* **10**(2), 1041–1062.

Hanks, E. M., Schliep, E. M., Hooten, M. B. & Hoeting, J. A. (2015), 'Restricted spatial regression in practice: geostatistical models, confounding, and robustness under model misspecification', *Environmetrics* **26**(4), 243–254.

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

Hodges, J. S. & Reich, B. J. (2010), 'Adding spatially-correlated errors can mess up the fixed effect you love', *The American Statistician* **64**(4), 325–334.

Hooten, M. B., Hanks, E. M., Johnson, D. S. & Alldredge, M. W. (2013), 'Reconciling resource utilization and resource selection functions', *Journal of Animal Ecology* **82**, 1146–1154.

Hooten, M. B., Johnson, D. S., Hanks, E. M. & Lowry, J. H. (2010), 'Agent-Based Inference for Animal Movement and Selection', *Journal of Agricultural, Biological, and Environmental Statistics* **15**(4), 523–538.

Hu, J., Kang, H.-W. & Othmer, H. G. (2013), 'Stochastic analysis of reaction-diffusion processes', *Bulletin of Mathematical Biology* .

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

Kanno, Y., Vokoun, J. C. & Letcher, B. H. (2011*a*), 'Data from: Fine-scale population structure and riverscape genetics of brook trout (Salvelinus fontinalis) distributed continuously along headwater channel networks'.

Kanno, Y., Vokoun, J. C. & Letcher, B. H. (2011*b*), 'Fine-scale population structure and riverscape genetics of brook trout (Salvelinus fontinalis) distributed continuously along headwater channel networks', *Molecular Ecology* **20**(18), 3711–3729.

Keeling, M. J., Brooks, S. P. & a Gilligan, C. (2004), 'Using conservation of pattern to estimate spatial parameters from a single snapshot.', *Proceedings of the National Academy of Sciences of the United States of America* **101**(24), 9155–60.

Kingman, J. F. C. (1982), 'The coalescent', *Stochastic Processes and their Applications* **13**(3), 235–248.

Klein, D. & Randić, M. (1993), 'Resistance distance', *Journal of Mathematical Chemistry* **12**(1), 81–95.

Kurtz, T. G. (1978), 'Strong approximation theorems for density dependent Markov chains', *Stochastic Processes and Their Applications* **6**(3), 223–240.

Kurtz, T. G. (1981), *Approximation of population processes*, Vol. 36, SIAM.

LeSage, J. & Pace, R. K. (2009), *Introduction to Spatial Econometrics*, Chapman and Hall/CRC, Boca Raton, FL, USA.

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

Manel, S., Schwartz, M., Luikart, G. & Taberlet, P. (2003), 'Landscape genetics: combining landscape ecology and population genetics', *Trends in Ecology & Evolution* **18**(4), 189–197.

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

McRae, B., Dickson, B., Keitt, T. & Shah, V. (2008), 'Using circuit theory to model connectivity in ecology, evolution, and conservation', *Ecology* **89**(10), 2712–2724.

Paciorek, C. J. (2010), 'The importance of scale for spatial-confounding bias and precision of spatial regression estimators', *Statistical Science* **25**(1), 107–107.

Pritchard, J. K., Stephens, M. & Donnelly, P. (2000), 'Inference of Population Structure Using Multilocus Genotype Data', *Genetics* **155**(2), 945–959.

Rioux Paquette, S., Talbot, B., Garant, D., Mainguy, J. & Pelletier, F. (2014), 'Modelling the dispersal of the two main hosts of the raccoon rabies variant in heterogeneous environments with landscape genetics', *Evolutionary Applications* **7**(7), 734–749.

Rue, H. (2001), 'Fast sampling of Gaussian Markov random fields', *Journal of the Royal Statistical Society. Series B, Statistical Methodology* **63**(2), 325–338.

Rue, H. & Held, L. (2005), *Gaussian Markov random fields: theory and applications*, Vol. 104, Chapman & Hall.

Schmidt, A. M. & 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**(3), 743–758.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P. & van der Linde, A. (2002), 'Bayesian measures of model complexity and fit', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **64**(4), 583–639.

Ver Hoef, J. M. & Peterson, E. E. (2010), 'A moving average approach for spatial statistical models of stream networks', *Journal of the American Statistical Association* **105**(489).

Wall, M. (2004), 'A close look at the spatial structure implied by the CAR and SAR models', *Journal of Statistical Planning and Inference* **121**(2), 311–324.

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

Wikle, C. & Hooten, M. (2010), ‘A general science-based framework for dynamical spatio-temporal models’, *Test* **19**(3), 417–451.

Wilson, G. A. & Rannala, B. (2003), ‘Bayesian inference of recent migration rates using multilocus genotypes’, *Genetics* **163**(3), 1177–1191.

Wright, S. (1943), ‘Isolation by distance’, *Genetics* **28**(2), 114–138.

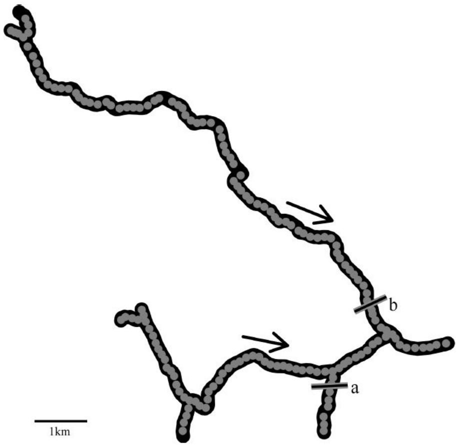


Figure 1: Trout sampling locations (circles) on the Jefferson Hill Spruce Brook. Arrows indicate the direction of stream flow. The location of two seasonal blockages are shown by solid lines crossing the stream network at “a” and “b”. Data are from Kanno et al. (2011*b*).

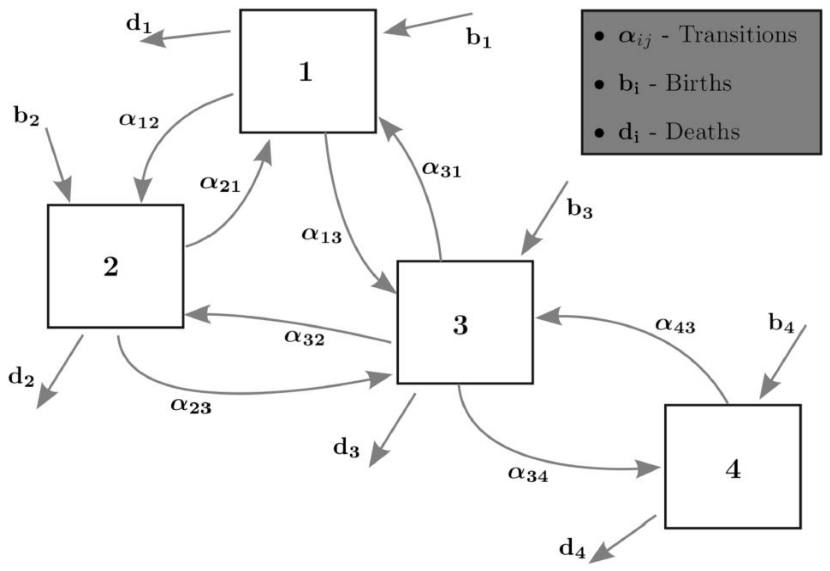


Figure 2: Continuous-time Markov random walk model example. α_{ij} is the transition rate from node *i* to node *j* and may be zero, indicating direct migration is impossible without traversing other nodes. b_i is the rate at which individuals are introduced into the system at node *i*, and d_i is the rate at which individuals in node *i* are removed from the system.

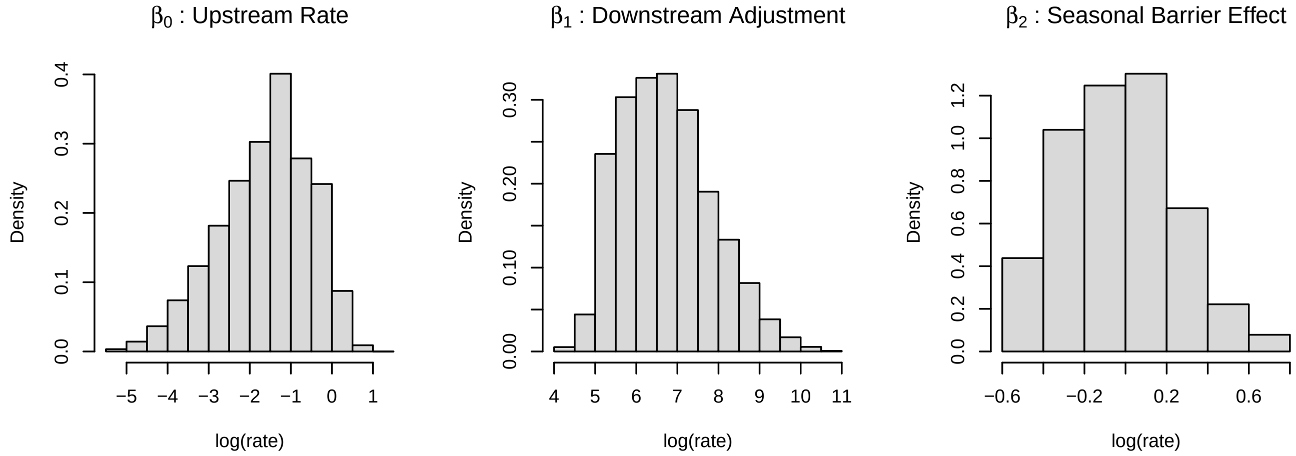
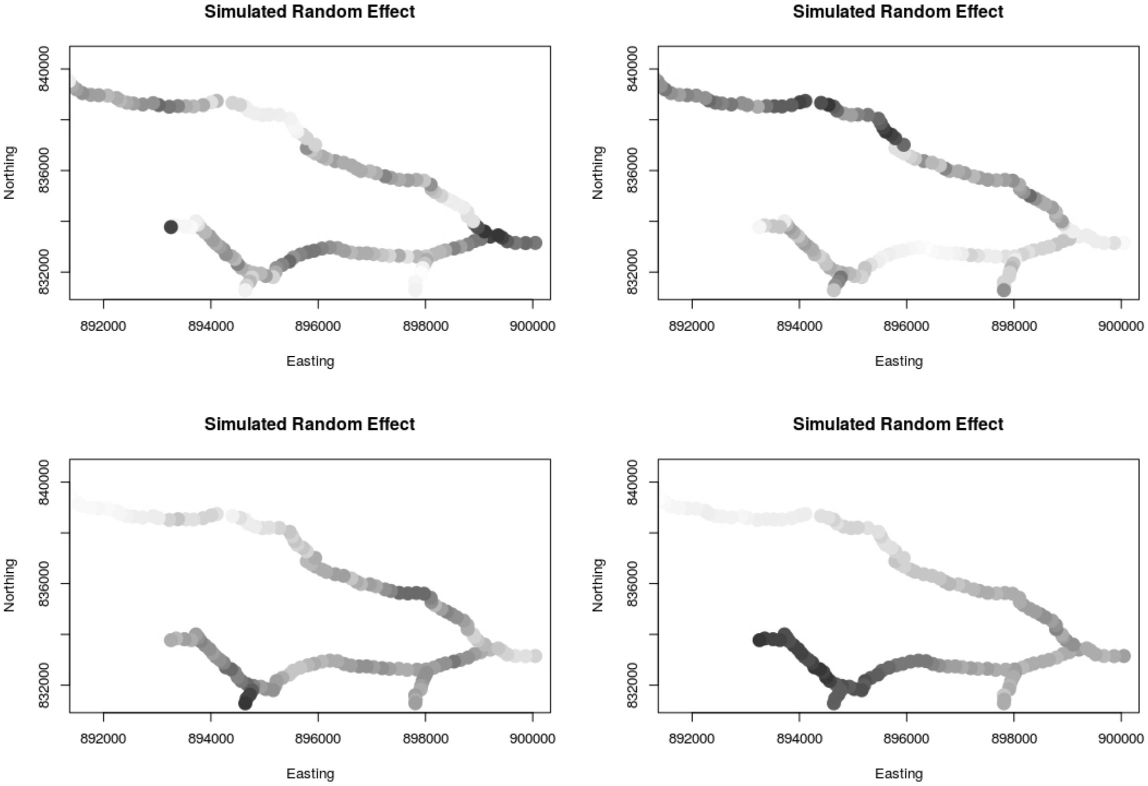


Figure 3: Posterior histograms of random walk model parameters in the spatial genetic analysis of trout in the JeffersonHill Spruce Brook.



Simulated Random Effect

Simulated Random Effect

Simulated Random Effect

Simulated Random Effect

Figure 4: Four realizations of random fields on the Jefferson-Hill Spruce Brook simulated using posterior mean parameter values of the random walk covariance model.