This is the author manuscript accepted for publication and has undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1002/ECM.1355

Spatially-structured statistical network models for landscape genetics

Running Head: Estimating resistance parameters

Erin E. Peterson[1*], Ephraim M. Hanks[2], Mevin B. Hooten[3], Jay M. Ver Hoef[4], and Marie-Josée Fortin[5]

[1] ARC Centre for Excellence in Mathematical and Statistical Frontiers (ACEMS) and the Institute for Future Environments, Queensland University of Technology (QUT), Brisbane, QLD, 4000, Australia

[2] Department of Statistics, Pennsylvania State University, University Park, PA, 16801, USA

[3] U.S. Geological Survey, Colorado Cooperative Fish and Wildlife Research Unit, Department of Fish, Wildlife, and Conservation Biology, and Department of Statistics, Colorado State University, Fort Collins, CO, 80523, USA

[4] Marine Mammal Laboratory, NOAA-NMFS Alaska Fisheries Science Center, Seattle, WA, 98115, USA

[5] Department of Ecology & Evolutionary Biology University of Toronto, Toronto, ON, Canada

[*]Corresponding Author Email: Erin.Peterson@qut.edu.au

January 7, 2019

Abstract

A basic understanding of how the landscape impedes, or creates resistance to, the dispersal of organisms and hence gene flow is paramount for successful conservation science and management. Spatially structured ecological networks are often used to represent spatial landscape-genetic relationships, where nodes represent individuals or populations and resistance to movement is represented using non-binary edge weights. Weights are typically assigned or estimated by the user, rather than observed, and validating such weights is challenging. We provide a synthesis of current methods used to estimate edge weights and an overview of common model types, stressing the advantages and disadvantages of each approach and their ability to model landscape-genetic data. We further explore a set of spatial-statistical methods that provide ecologists with alternative approaches for modeling spatially explicit processes that may affect genetic structure. This includes an overview of spatial autoregressive models, with a particular focus on how correlation and partial correlation are used to represent neighborhood structure with the inverse of the covariance matrix (i.e., precision matrix). We then demonstrate how to model resistance by specifying an appropriate statistical model on the nodes, conditioned on the edge weights, through the precision matrix. This integration of network ecology and spatial statistics provides a practical analytical framework for landscape-genetic studies. The results can be used to make statistical inferences about the relative importance of individual landscape characteristics, such as the vegetative cover, hillslope, or the presence of roads or rivers, on gene flow. In addition, the R code we include allows readers to explore landscape-genetic

structure in their own datasets, which will potentially provide new insights into the evolutionary processes that generated ecological networks, as well as valuable information about the optimal characteristics of conservation corridors.

Key Words: spatial statistics, landscape genetics, spatially structured ecological network, resistance values, edge weights

Introduction

Landscape genetics focuses on the effects of landscape pattern, structure, composition, and quality on spatial-genetic variation and gene flow (Storfer et al. 2007). It is a relatively new field of research (Manel et al. 2003) that draws on concepts from landscape ecology, population genetics, mathematics, and statistics. However, truly integrative research is challenging in this rapidly advancing field, where useful developments are occurring simultaneously in multiple disciplines (Balkenhol et al. 2016a). This is especially true of methods used to quantitatively describe genetic-landscape structure to gain inferences about causal evolutionary and ecological processes.

In landscape genetics, microsatellite allele and multiple single nucleotide polymorphism (SNP) data collected from individuals or populations at multiple locations are often used to generate genetic distance or dissimilarity matrices, which are subsequently used to infer rates of gene flow. Many different distance metrics can be used to calculate genetic distances between individuals (e.g., Euclidean distance) or populations (e.g., Nei's genetic distance; Nei 1972), with each relying on different geometric and/or evolutionary assumptions (Dyer, 2017a). These genetic distance matrics are then used to investigate how resistance to movement facilitates/prevents the dispersal of organisms and gene flow (Holderegger and Wagner 2008). Within this context, landscape resistance represents the effects of landscape characteristics such as vegetation or roads, on movement between them (Holderegger and Wagner 2008). Evolutionary processes influencing resistance typically fall into three categories: 1) isolation-by-distance (IBD), where distances between locations are

greater than the organisms's dispersal ability (Wright 1943); 2) isolation-by-resistance (IBR), which occurs when landscape characteristics lead to inhomogeneous migration rates across space (McRae 2006); and 3) isolation-by-barrier (IBB), where landscape features such as waterbodies form non-permeable or semi-permeable barriers to movement (Smouse et al. 1986). These relationships can be represented as a spatially structured ecological network (SSEN; Dale and Fortin 2010), where nodes have a location and size, and edges have a physical location and length in geographic space. Thus, the SSEN provides a natural, spatially explicit framework used to explore patterns of landscape-genetic structure.

Although inference about the relationship between resistance and genetic structure is the focus of many studies, it is rare for resistance values to be measured directly using empirical data (Fletcher et al. 2011). When movement is measured, it is typically based on detection (i.e., sightings), relocation (i.e., mark-recapture) or pathway (i.e., global positioning system telemetry) data (Zeller et al. 2012). However, resistance is more often based on *a priori* experimental evidence (e.g., species dispersal ability based on telemetry data) or expert opinion (Beier et al. 2008; Zeller et al. 2012). A causal-modeling approach is sometimes used to compare how well the hypothesized resistance values, which are based on conceptual models of evolutionary processes, fit the data (Legendre and Troussellier 1988; Cushman et al. 2006). Resistance estimates are crucial because they define the structure of the system and underpin inferences related to dispersal, population definition, and gene flow. Yet, there are significant challenges associated with validating landscape-connectivity values, given that independent data are often lacking and many

combinations of biotic and abiotic processes could produce similar connectivity values (Whitlock and McCauley 1999; Dyer and Nason 2004).

Spatial statistical methods are specifically designed to model spatially dependent data and may be particularly suited to landscape-genetic studies. In the field of statistics, a spatial statistical model uses the spatial location of data in the probabilistic model component (i.e., spatial dependence in the residual errors is modeled as a function of space). These models are sometimes referred to as "spatial error" models in ecology (Keitt et al. 2002). Spatial autoregressive (SA) models (Lichstein et al. 2002; Ver Hoef et al. 2018) represent a broad class of spatial statistical models implemented as an SSEN. Hence, there are obvious conceptual similarities between landscape genetics and SA models. In landscape genetics, connectivity among individuals or populations can be represented using non-binary weights (i.e., resistance distance or cost-weighted distance) that may or may not incorporate a physical distance; while in SA models, relationships among measurements are represented in the precision matrix, which is often modeled as a function of Euclidean distance (i.e., relative weight) between locations.

Our goal is to describe how a spatial statistical approach can be used to model resistance in landscape-genetic studies. Specifically, we 1) provide an overview of landscape-genetic data and their representation as SSENs, 2) provide a brief summary of methods currently used to validate models of resistance, including a synthesis of their strengths and weaknesses, and 3) demonstrate how resistance distances can be estimated using SA models.

Modeling Spatially Structured Ecological Networks

CALCULATING EDGE WEIGHTS

A SSEN can be used to represent landscape-genetic relationships, where nodes represent the location of individuals or sub-populations, and edges describe the functional relationship (e.g., animal movement or gene flow) between nodes. Thus, the resistance distance between nodes may differ depending on their proximity to one another, as well as the landscape characteristics and features that lie between them. Such edge weights are usually estimated and then validated using genetic dissimilarity between nodes because data describing an organism's movement are rarely available in sufficient quantities to describe the SSEN structure (Fletcher et al. 2011).

Contiguous nodes share a boundary, thus there is no physical distance between them; therefore, covariates (i.e., predictors) representing resistance (i.e., resistance covariates) can be based on node characteristics, or the distance between node centroids (Hanks and Hooten 2013), that have been selected to represent an underlying conceptual model of evolutionary processes (e.g., IBD, IBR, and/or IBB; Figure 1). Resistance covariates for non-contiguous nodes can also be based on node characteristics (e.g., Botta et al. 2015), characteristics of the edges that join node pairs (e.g. Petkova et al. 2016), or both. Regardless of which method is used, *a priori* assumptions must be made about the neighborhood structure, the edge location, and/or the resistance values. These assumptions affect how resistance is represented in the model and the inference that can be made (Figure 1).

Estimating edge weights for non-contiguous nodes is more complicated than for contiguous nodes because the uncertainty associated with the physical edge location in geographic space increases as the distance between nodes increases (i.e., multiple potential pathways exist). Two approaches are commonly used to address this issue (Figure 1): 1) an *a priori* decision about edge location is made, which defines the area over which resistance covariates are calculated (e.g., Rioux Paquette et al. 2014); or 2) an *a priori* decision about resistance values is made and edges are delineated based on those values (e.g., Beier et al. 2009; Petkova et al. 2016). Many methods are used to parameterize resistance values and a full review is beyond the scope of this paper (see Spear et al. (2010) and Zeller et al. (2012) for in-depth reviews). However, the commonality among these methods is that *a priori* decisions must be made about the relative importance of individual covariates of resistance and/or the physical location of the edge before the edge weights are generated (Figure 1). Assigning resistance values is challenging because scientific knowledge about dispersal and habitat preferences is often lacking. Habitat and dispersal data may be unavailable or collected at an inappropriate spatio-temporal resolution (Zeller et al. 2012). Resistance values may be assigned based on expert opinion, a literature review, and/or empirical data such as species occurrence, individual animal movement, rates of interpatch movement, or genetic distance (Beier et al. 2008; Minor and Urban 2008; Zeller et al. 2012). However, there are obvious consequences in assuming that the drivers of resistance and gene flow are known (Cushman et al. 2006); if this assumption is incorrect, the conclusions of the study may be misleading and subsequent management actions may not have the desired outcome (Shirk et al. 2010; Spear et al. 2010).

COMMON MODELS

A number of approaches are used to analyze landscape genetic-data, but the most common methods generally fall into four categories: computer simulation, matrix correlation, ordination, and regression (Figure 2). These methods tend to be borrowed from other disciplines (Balkenhol et al. 2016a) and, as such, often do not meet basic modeling needs for landscape genetic studies (Figure 2). We are not the first to point this out; there have been widespread calls from landscape-genetic researchers for more robust methods of exploring relationships between genetic diversity and drivers of resistance (Storfer et al. 2007, Balkenhol et al. 2009; Cushman and Landguth 2010; Manel and Holderegger 2013; Balkenhol et al. 2016b). Some of the most common criticisms include the 1) lack of statistical power, especially for small sample sizes (Legendre and Fortin 2010); 2) parameter bias and low statistical power when tests are performed on spatially dependent data (Legendre and Fortin 2010; Wagner and Fortin 2013); 3) inability to assess individual components of resistance (e.g., vegetation cover), rather than matrices of dissimilarity, and their interactions (Storfer et al. 2007; Beier et al. 2008); and 4) need for *a priori* decisions about resistance values, which constrains the parameter space (Beier et al. 2008), regardless of how many resistance models are proposed (e.g., Cushman and Landguth 2010, Shirk et al. 2010). Despite the widespread criticisms, these methods continue to be used to gain insight into evolutionary and ecological processes because there are few alternatives in this emerging field of research. At the same time, there is a critical need for 1) suitable methods for model selection (Cushman and Landguth 2010; Wagner and Fortin 2013) and validation (Dyer and Nason 2004; Balkenhol et al. 2009); 2) statistical methods that can be

used to predict when the network is not fully observed (i.e., missing data; Hanks and Hooten 2013) or under future land-use or climate scenarios (McRae 2006; Storfer et al. 2007; Beier et al. 2008); and 3) methods that describe uncertainty in resistance parameter estimates (Beier et al. 2008; Zeller et al. 2012; Hanks and Hooten 2013). Thus, clear methodological gaps exist and new quantitative methods are needed to make inference about the suitability of these mechanistic models of connectivity and their uncertainty, as well as the underlying processes that generated the network structure.

Spatial Autoregressive Models

Spatial autocorrelation underpins numerous hypotheses in ecological studies (Legendre and Fortin 2010); if genetic data do not exhibit a spatial structure, then evolutionary-process hypotheses related to IBD, IBR, and IBB are irrelevant. Thus, an approach that makes use of spatial autocorrelation (Figure 2), rather than attempting to avoid it, is likely to provide a better understanding of landscape-genetic relationships when the data are spatially dependent (Balkenhol et al. 2009).

SA models are spatial statistical models that have been specifically designed to model areal or network data. The general form of an SA model is $\{y(\mathbf{s_i}) : \mathbf{s_i} \in D,\ i = 1, ..., M$, where y is an observed (or unobserved) random variable at node i, at location $\mathbf{s_i}$, that belongs to the spatial domain of interest, D. For example, the random variable could represent allele counts, while the domain-of-interest could be a management unit. An SA model differs from other spatial statistical models (e.g., geostatistical or spatial point process models)

because 1) D is a fixed and finite set of nodes, rather than continuous space and 2) spatial dependence is modeled as a function of network structure, rather than Euclidean distance.

MATRIX REPRESENTATIONS OF NETWORK STRUCTURE

An SSEN is defined by its graphical structure (e.g., nodes and edges connecting nodes) and, in a weighted network, by the weights assigned to edges (Figure 3a). To define this formally, let $\mathbf{G} \equiv (\mathbf{V}, \mathbf{W})$ be an SSEN with M nodes, $\mathbf{V} \equiv \{V_1, V_2, \ldots, V_M\}$, and the edges or edge weights, $\mathbf{W} \equiv \{w_{ij}\}$, between them. Note that the edge weights could potentially be directed (i.e., asymmetric) to account for processes such as source-sink dynamics or dispersal preferences (Dale and Fortin 2010). The edges of the SSEN can also be represented as an $M \times M$ matrix (Figure 3a). The element w_{ij} in the ith row and jth column of the matrix $\mathbf{W}$ is the directed or undirected edge weight connecting nodes i and j in the network. In an unweighted graph, connectivity is simply represented using a binary adjacency matrix, where $w_{ij} = 1$ and $w_{ij} = 0$ imply that an edge exists or does not exist between nodes i and j, respectively. By definition, edges do not connect nodes to themselves in SA models and therefore diagonal elements are also defined as $w_{ii} = 0$. These same rules apply in a weighted network, except that $w_{ij} > 0$ indicates that there is an edge between two nodes and the strength of connectivity between node pairs is allowed to vary (Figure 3a). If the SSEN is undirected, then $\mathbf{W}$ is a symmetric matrix and $w_{ij} = w_{ji}$.

CORRELATION AND PARTIAL CORRELATION

A key component of an SSEN is the conditional dependence (i.e., structure) implied by the edges. When an edge exists between nodes, $w_{ij} > 0$, then nodes i and j are first-order neighbors and are considered connected (e.g., V_1 and V_2, Figure 3a). If two nodes are not directly connected by an edge, $w_{ij} = 0$, a path between the nodes may still exist through intervening nodes (e.g., V_2 and V_4, Figure 3a). Thus, observations at nodes that are not first-order neighbors are conditionally independent in the precision matrix (e.g., $\mathbf{Q}_{3,2} = 0$ in Figure 3b).

A statistical concept strongly related to the network structure defined by edges is *partial correlation*. Consider the situation where a process such as genetic variation in individuals or populations, $\mathbf{y}$, is measured on nodes. The topological structure implied by the edges helps define the correlation structure on the process $\mathbf{y}$. This correlation structure is represented by $\mathbf{\Sigma}$, which is the $M \times M$ covariance matrix of $\mathbf{y}$ (Figure 3c). Thus, the i, jth element of $\mathbf{\Sigma}$ is the covariance between y_i and y_j:

$$\Sigma_{ij} = \text{cov}(y_i, y_j) = E\left[(y_i - E(y_i))(y_j - E(y_j))\right].$$

The inverse covariance matrix, or precision matrix, $\mathbf{Q} = \mathbf{\Sigma}^{-1}$, defines the partial correlation of $\mathbf{y}$ after accounting for the influence of intervening nodes (Figure 3b). For example, let $\{y_1, y_2, \ldots, y_n\}$ be Gaussian observations on an M-node network. The partial correlation between y_i and y_j is defined as $\kappa_{ij|\cdot} = \text{corr}(\varepsilon_{i|\cdot}, \varepsilon_{j|\cdot})$, where $\varepsilon_{i|\cdot}$ are the residuals from a regression with the response y_i and $\{y_k, k \neq i, j\}$ as covariates (e.g., node size or habitat

quality)(Figure 3d). If $\kappa_{ij|\cdot} = 0$, then nodes i and j are not first-order neighbors and any dependence between y_i and y_j is captured by intervening nodes $\{y_k, k \neq i, j\}$. For any precision matrix, $\mathbf{Q}_{ij} = 0$ if and only if $\kappa_{ij|\cdot} = 0$. Thus, information about local connectivity and dependence can be encoded in the precision matrix of a multivariate random variable. Note that two nodes may still be correlated through intervening nodes and this dependence is captured by the covariance matrix, $\mathbf{\Sigma} = \mathbf{Q^{-1}}$ ($\Sigma_{3,2} = 0.42$, Figure 3c), which is obtained by inverting $\mathbf{Q}$. This idea is conceptually similar to the role of stepping stones, which promote connectivity and facilitate organism movement or gene flow between isolated habitat patches (Saura et al. 2014).

Partial correlation is not a new concept in ecology; the partial-correlation structure accommodated by the precision matrix is increasingly being used to estimate network topology, which are subsequently used to understand the influence of network structure on evolutionary processes (i.e., Population Graphs; Dyer and Nason 2004). However, partial correlation and conditional independence in a SSEN can also be modeled as elements of the precision matrix in a SA model. In the next section, we provide background information about SA models for estimating edge weights using a data-driven approach.

CAR and ICAR Models

If the SA model has a Gaussian error distribution, it can be written as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\alpha} + \boldsymbol{\eta} + \boldsymbol{\varepsilon} \ , \tag{1}$$

where the "error" models are $\boldsymbol{\varepsilon} \sim \mathrm{N}(\mathbf{0}, \sigma_\varepsilon^2 \mathbf{I})$ and $\boldsymbol{\eta} \sim \mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma})$. The mean structure describes the conditional mean of the response given a set of covariates, if they are present. In Eq. 1, the mean (or first moment) structure is modeled using a regression that includes covariates, $\mathbf{X}$, as well as a latent spatial-random process, $\boldsymbol{\eta}$. The covariates are used to account for influential processes or conditions that have been measured, while the latent spatial-random process is used to describe residual spatial dependence. Thus, spatial dependence may result from a lack of understanding about the ecological process, an inability to measure influential covariates, or inherent spatial dependence in the response variable (Keitt et al. 2002). The term $\boldsymbol{\eta}$ is not directly measured and instead must be inferred using a statistical model.

This model formulation is fundamentally different from most models built to explore associations between allele prevalence and landscape features (selection), where there is often no mean structure, because typical data come from neutral regions of the genome. In contrast, it might be important to include covariates in the model mean structure in a landscape genomics study, where allele frequencies from non-neutral regions are affected by natural selection. In addition, spatial autocorrelation is not a nuisance in landscape-genetic studies, but rather the main focus of the analysis. Thus, we often assume that $\mathbf{y} = \boldsymbol{\eta} \sim \mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma})$. We will later consider other data models more appropriate for non-Gaussian genetic data, but the Gaussian model serves as a canonical model for spatial dependence in SSENs.

The precision matrix, $\mathbf{Q} \equiv \boldsymbol{\Sigma}^{-1}$, is used to describe the spatial dependence in the residual

errors and, in the case of the conditional autoregressive (CAR) model, we assume:

$$\mathbf{Q} = \mathbf{D} - \rho\mathbf{W}. \tag{2}$$

Here, $\mathbf{W}$ is a binary or non-binary edge weights matrix, $\mathbf{D}$ is a diagonal matrix with elements $D_{ii} = \sum_k W_{ik}$, and $\rho \in (0, 1)$ is a parameter affecting correlation. Other equivalent forms for the CAR precision matrix have been used in the literature (e.g., $\mathbf{Q} = \tau^2\mathbf{M}^{-1}(\mathbf{I} - \mathbf{C})$ for matrices $\mathbf{M}$ and $\mathbf{C}$; Banerjee et al. 2004). However, the formulation in Eq. 2 highlights the direct link between the edge weights in an SSEN and the precision matrix of a spatial CAR model (Figures 3a,b).

The term "conditional" in the conditional autregressive (CAR) model is used because each element of the random process is specified as conditional on those found on all first-order neighboring nodes, rather than all of the nodes (Figure 3):

$$\eta_i | \boldsymbol{\eta}_{j|j\neq i} \sim \mathrm{N}\left(\sum_j \frac{\rho W_{ij}\eta_j}{\sum_{j\neq i} W_{ij}}, \frac{\sigma^2}{\sum_{j\neq i} W_{ij}}\right). \tag{3}$$

This conditional representation shows that the conditional mean of η_i is a weighted average of its neighbors ($\eta_j : j \in N(i)$, where $N(i)$ is the set of first-order neighbors of node i), scaled by ρ. If $\rho = 0$, then each η_i is independent of all other η_i, and there is no spatial autocorrelation, while larger values of ρ yield stronger correlation. If $W_{ij} > W_{ik}$, then the mean of η_i is more strongly influenced by η_j than by η_k. Thus, proportionally larger edge weights imply that there is a stronger functional relationship between nodes. Finally, the

conditional variance of η_i is the conditional variance parameter, σ^2, over the sum of the edge weights connected to node i. Thus, the mean and variance of the spatial random process are both nonstationary, varying with node i. The conditional representation also makes it clear how to model spatial correlation in SSENs using edge weights and CAR models; increasing all edge weights decreases the marginal variance, while proportionaly larger edge weights imply stronger connectivity and correlation between nodes.

Correlation (Figure 3e) is a scaled version of covariance, which also contains information about connectivity and dependence within the SSEN. However, the covariance and correlation implied by a CAR model are sometimes counter-intuitive (Wall 2004). For example, in Figure 3c, the highest covariance is found between V_1 and V_2, but Figure 3e shows that the highest correlation is found between V_1 and V_4. This discrepancy is due to the nonstationary nature of the model; in a CAR model, the least connected nodes have high conditional variances (Eq. 3), and often have high marginal variances, which inflates the covariance. Nevertheless, a CAR model provides some intuition on the correlation and covariance implied by a SSEN. In this case, V_2 is the least connected of all nodes in the network (Figure 3a), and thus it makes sense that the correlation with other nodes would be relatively small.

An intrinsic conditional autoregressive model (ICAR; Besag and Kooperberg, 1995) is a limiting case of a CAR model, where $\rho = 1$. In this case, $\mathbf{Q} = (\mathbf{D} - \mathbf{W})$ is not invertible, but the ICAR can still be used as a prior in a Bayesian spatial model (e.g., Cressie 2015). The covariance matrix of the ICAR can also be defined as the generalized inverse, $\mathbf{Q}^-$, under the constraint that the spatial-random effects sum to a constant (e.g., $\sum_{i=1}^{n} \eta_i = 0$)

(Rue and Held, 2005).

MISSING DATA

In previous sections, we assumed that all nodes in the network were fully observed and that one observation, y_i, was obtained for each node, but this is unusual in practice. Consider the general case where there are n_{obs} total observations at m_{nodes} nodes. When multiple observations are collected on nodes (e.g., multiple individuals are genetically sampled within a population), a nugget effect, τ, can be introduced into the covariance structure (Besag et al. 1991) to account for within-node variation. Let $\mathbf{y} \equiv (y_1, y_2, \ldots, y_{n_{\text{obs}}})'$ be the vector of n_{obs} observations from the network and let $\mathbf{\Sigma}_{\text{nodes}}$ be the $m_{\text{nodes}} \times m_{\text{nodes}}$ covariance matrix of the entire network. When there are multiple observations on a node or missing data on other nodes, there is not a one-to-one relationship between nodes and observations. To account for this mismatch, an $n_{\text{obs}} \times m_{\text{nodes}}$ matrix $\mathbf{K}$ is created to "map" observations to nodes; $K_{ij} = 1$ if the ith observation (y_i) is taken at the jth node, and $K_{ij} = 0$ otherwise. The matrix $\mathbf{K}$ can then be used in the $n_{obs} \times n_{obs}$ covariance matrix $\mathbf{\Psi}$ of the observations $\mathbf{y}$, where

$$\mathbf{\Psi} = \mathbf{K}\mathbf{\Sigma}\mathbf{K}' + \tau^2\mathbf{I}.$$

Estimation of the edge weights, which define $\mathbf{\Sigma}$, can then be carried out by substituting $\mathbf{\Psi}$ for $\mathbf{\Sigma}$ in a CAR or ICAR model.

The ability to use the entire network in the modeling process has numerous advantages, even if it is partially unobserved. Nodes with missing data are usually removed from the

analysis, which equates to a loss of information (Nakagawa and Freckleton 2008). If data are not missing at random, it alters the topology of the network (Kossinets 2006; Fletcher et al. 2011), results in loss of statistical power, and produces biased parameter estimates for processes on the network (Nakagawa and Freckleton 2008). A covariance matrix that represents all of the network nodes can also be used within a SA model to make predictions, with estimates of uncertainty, at unobserved nodes. These predictions provide estimates of processes on, and the topology of a network that has not been fully observed based on the observed data. However, they can also be used for model validation in a k-fold cross-validation procedure. Another important advantage is the ability to incorporate nodes with missing data into the statistical model, which means that a contiguous data model can be used to estimate resistance; thus, removing the need to make *a priori* and potentially incorrect assumptions about the spatial location of edges between non-contiguous nodes (Figure 1). Although there may not be a partial correlation between two observed nodes separated by nodes with missing values, spatial dependence may still exist because of the intervening nodes in the path between them (Figure 3c).

Edge Weight Estimation Using Spatial Autoregressive Models

Although SA models are often used in spatial statistics, the specification of weights has received little attention. In most cases, weights are arbitrarily described using representations of adjacency with little thought devoted to the processes that drive

connectivity. When weights are specified in this manner, they are considered fixed and known, which implies that the topology of the SSEN is known exactly; however, this is almost never the case in ecology. In fact, ecological questions often *focus* on understanding the drivers of landscape connectivity. We reconcile these statistical and ecological perspectives, with the goal of gaining a better ecological understanding of resistance in SSENs.

Characteristics on the nodes (e.g., habitat quality, size, or population size) or along the edges (e.g., length, vegetation cover, or barriers to movement) may describe increases/decreases in landscape connectivity between node pairs. Thus, the resistance distance between nodes may, or may not, be solely dependent on the physical distance between them. Here we define resistance distance as the cumulative resistance between observations based on circuit theory (McRae 2006, Zeller et al. 2012). For example, if the nodes are irregularly spaced or irregular in size, it would make sense to model connectivity (i.e., edge weights) as a function of distance between nodes. The most natural approach is to treat the centroid of each node as the location, and include the distance, or log-distance (Hanks and Hooten 2013) beween nodes as a resistance covariate in Eq. (5), with or without other resistance covariates.

An SA model may include a multivariate response, y_i, such as microsatellites or multiple SNPs for individuals or populations. The CAR or ICAR model can be connected to more ecologically relevant network-based approaches when the weights matrix is constructed. Instead of defining edge weights based on conceptual models of evolutionary processes, Hanks and Hooten (2013) showed that they may be estimated by specifying an appropriate

statistical model for $\mathbf{y}$, conditioned on the edge weights through the precision matrix. For example, edge weights, w_{ij}, could be modeled as

$$w_{ij} = \begin{cases} 0 & \text{if } i \text{ and } j \text{ are not first-order neighbors,} \\ f(\mathbf{x}_{ij}, \boldsymbol{\beta}) & \text{if } i \text{ and } j \text{ are first-order neighbors,} \end{cases} \tag{4}$$

where $\mathbf{x}_{ij}$ is a vector of covariates used to model the edge weight between i and j (e.g., slope or vegetation cover), and $\boldsymbol{\beta}$ is a vector of estimated parameters. Edge weights are usually greater than zero, thus one potential model relating $\mathbf{x}_{ij}$ and w_{ij} is a log-linear model:

$$f(\mathbf{x}_{ij}, \boldsymbol{\beta}) = \exp\{\mathbf{x}_{ij}'\boldsymbol{\beta}\}. \tag{5}$$

For the IBD model, $f(\mathbf{x}_{ij}, \boldsymbol{\beta}) \equiv 1$ so we obtain an estimated distance-only decay function, with no other effects, that depends conditionally on first-order neighbors; although autocorrelation decays with distance throughout the study area (e.g., Ver Hoef et al. 2018). Many other model formulations are also possible. For example, in Ver Hoef et al. (2018), $\boldsymbol{\beta}$ was estimated as a function of categorical variables representing differences in harbor seal sub-population membership. Similarly, the matrix $\mathbf{x}_{ij}$ could contain extra resistance covariates for models representing the IBR (e.g., vegetation cover) and IBB (e.g., rivers) evolutionary-process hypotheses, in addition to an intercept.

As mentioned previously, models are often fit to genetic distance or diversity matrices in landscape-genetic studies and these matrices can be generated based on a variety of distance metrics. For example, Wright's F_{ST} (Wright 1931) and Nei's D (Nei 1972) can be

used to describe population-based genetic diversity, while the Bray-Curtis (Legendre and Legendre 2012) and other measures of relatedness (Queller and Goodnight 1989) are typically used to measure individual-level genetic diversity. However, the advantages of modeling genetic distance using an SA model as described here are only realized if there is an appropriate statistical distribution for an observed distance matrix and the covariance matrix. The generalised Wishart distribution has been used in recent landscape-genetic studies to visualise patterns of population structure (Bradburd et al. 2016) and to estimate ancestry proportions from multiple populations (Bradburd et al. 2018). McCullagh (2009) showed that a generalized Wishart distribution is the appropriate statistical model if the genetic distance matrix, $\mathbf{D}$, is based on squared-Euclidean distance of a normally distributed random variable (Appendix S1). Under these assumptions, $-\mathbf{D} \sim \mathrm{GW}_{\nu}(\mathbf{1}, \mathbf{2\Sigma})$, where $\mathbf{\Sigma} = \boldsymbol{Q}^{-1}$. However, there is no guarantee that the generalized Wishart distribution will be appropriate for all dissimilarity matrices and future research is needed to develop diagnostic tools to check the validity of these distributional assumptions. The advantage of this approach is that it provides a formal statistical likelihood for pairwise distance data. This makes the whole range of likelihood-based tools such as maximum likelihood estimation, asymptotic confidence intervals on parameters, and model selection using Akaike's information criterion (AIC; Akaike 1974) and other information criteria applicable to genetic analyses. Another major benefit is that the parameter estimates, β (Eq. 5), are comparable between different populations and studies. As a result, it is possible to fit similar models to multiple disparate populations and assess how consistent the landscape-genetic relationships are.

Under a CAR model, the edge weights $\mathbf{W}$ and the parameter ρ completely define $\mathbf{Q}$ and $\mathbf{\Sigma}$ (Figures 3b,c), and the likelihood of the data under the generalized Wishart model. Multiple conceptual models of connectivity could be specified using different formulations of $\mathbf{W}$ or $\mathbf{\Sigma}$ and compared using AIC. This provides a flexible modeling framework, where genetic data on the nodes are converted to genetic data on the edges (e.g., genetic-distance matrices), and modeled as a function of covariates on the nodes (e.g., node or neighborhood level) and/or edges of the SSEN. This method does not fit neatly into the four levels of analysis proposed by Wagner and Fortin (2013) to relate genetic data to landscape data (e.g., node, link, neighborhood, and boundary). Instead, we refer to it as a *network-based* method because it can be used to represent all four levels of analysis, depending on how the model is parameterized and the research question of interest.

SIMULATED EXAMPLE

Observed genetic patterns may be produced by the combined influence of geographic distance, resistance, and barriers, rather than a single evolutionary process (Landguth and Cushman 2010). The SA model can be used to account for proximity in terms of variables on nodes and/or edges, physical distance (e.g., Euclidean or least-cost path), and unobserved drivers of landscape connectivity. Next, we provide an example demonstrating how edge weights can be estimated within an ICAR model by incorporating resistance covariates into the off-diagonal elements of the precision matrix. We provide data (dataS1 and dataS2) and R code (dataS1 and Appendix S2) so that readers can recreate the example.

We simulated resistance surfaces for the IBD, IBR, and the IBB scenarios (Figure 4, Appendix S2). The locations for 30 subpopulations were randomly generated and the pairwise resistance distance was calculated based on the IBD, IBR, and IBB models (Appendix S2). This distance is equivalent to the cumulative resistance between population locations based on circuit theory (McRae et al. 2008).

Genetic data were simulated under the IBD, IBR, and IBB evolutionary-process models for 450 individuals (30 subpopulations x 15 individuals) using the PopGenReport package (Adamack and Gruber 2014, Appendix S2). Genetic distance matrices for individual allele counts were calculated for the simulated datasets based on Manhattan distance.

We fit three models (IBD, IBR, and IBB) to each of the genetic-distance matrices ($\mathbf{D}_{\mathrm{IBD}}$, $\mathbf{D}_{\mathrm{IBR}}$, $\mathbf{D}_{\mathrm{IBB}}$) using a generalized Wishart distribution (Appendix S1). The nine models had the form

$$-\mathbf{D} \sim \mathrm{GW}_{\nu}(\mathbf{1}, \mathbf{2\Psi}), \tag{6}$$

where GW is the generalised Wishart distribution and $\nu = 20$ represents the number of genetic loci used to compute $\mathbf{D}$.

The SA models were fit using a raster-based network representation, with contiguous nodes and edge weights (and corresponding off-diagonal elements of the ICAR precision matrix) a function of the distance between node centroids, and the resistance value at neighboring raster cells estimated from the data. The spatial covariance for the models was given by

$$\mathbf{\Psi} = \mathbf{K}\mathbf{Q}^{-}\mathbf{K}' + \tau^2\mathbf{I}, \tag{7}$$

where $\mathbf{K}$ is a design matrix linking observations to nodes (raster cells) in the SSEN, τ^2 models non-spatial variability, and $\mathbf{Q}$ is an ICAR precision matrix (Eq. 2), with edge weights a function of resistance covariates, $\mathbf{x}_{ij}$, as shown in Figure 4.

The edge weights were modeled as a log-linear function of an intercept only for the IBD model, an intercept and a continuous resistance covariate for the IBR model, and an intercept and a binary covariate representing a non-permeable barrier to movement for the IBB model (Figure 4). Notice that the IBR and IBB models account for both resistance covariates *and* the distance between individuals, while the IBD model is based purely on distance. Parameters were estimated using maximum likelihood.

We compared the models using AIC (Akaike 1974) and found that for $\mathbf{D}_{\text{IBD}}$, the data generating model (IBD) had slightly more support in the data than the IBR estimating model, and considerably more support than IBB (Table 1). This is not surprising based on the patterns observed in the simulated genetic distance versus resistance plots (Appendix S2). However, there was no question about which models had the most support in the data for $\mathbf{D}_{\text{IBR}}$ and $\mathbf{D}_{\text{IBB}}$. The AIC value for the IBR data-generating model was more than 14 units lower than the competing IBD and IBB estimating models for $\mathbf{D}_{\text{IBR}}$, while the AIC for the $\mathbf{D}_{\text{IBB}}$ data-generating model (IBB) was more than 56 units lower than alternative IBD and IBR estimating models (Table 1).

The exponentiated β parameter estimates produced by the final models describe the relationship between the conductance (i.e., 1/resistance) and the original resistance covariates (Figure 4) and this relationship can be plotted, with 95% confidence intervals. Figure 5a shows that the relationship between conductance and the resistance covariate

described by the fitted $\mathbf{D}_{\mathrm{IBR}}$ data-generating model is non-linear, which is not surprising given that a log-linear model was used. As expected, conductance through cells with low IBR resistance-covariate values is higher than those with larger values, with conductance dropping off rapidly as resistance increases from 1 to 5. The 95% confidence intervals show that there is more uncertainty about this relationship when resistance is moderate (e.g., 5 to 10) compared to when it is low or high (Figure 5a). The relatively low AIC value for this model (Table 1) indicates that the $\mathbf{D}_{\mathrm{IBR}}$ data-generating model was able to describe this relationship more accurately than the other models and thus provides greater insight into the relationship between the IBR resistance covariate and simulated gene flow. Furthermore, maps of conductance generated using the SA model (Figure 5b) could be used to define movement corridors between conservation reserves or examine scenarios of land-management impacts on gene flow (e.g., McRae et al. 2008; Landguth and Cushman 2010).

Considerations

The benefit of using SA models with SSENs is the ability to model spatially dependent data and gain statistically robust inferences. However, the advantages gained in fitting a SA model strongly depend on the genetic distance matrices containing sufficient information to estimate edge weights. In other words, there must be relatively strong spatial dependence in the data and this is affected by both the genetic and field survey design.

Genetic data are collected from individuals at multiple locations in landscape-genetic

studies, and often transformed into a genetic distance matrix prior to modelling. These matrices are usually based on a subset of alleles found on neutral loci (i.e., microsatellite alleles or SNPs) that have no known function and as such, are not believed to be involved in natural selection (Wagner and Fortin 2013). Instead, the variability in the genetic data should reflect genetic drift; highlighting the influence of landscape resistance on gene flow and population structure. There are numerous filtering steps designed to reduce the negative effects of sequencing errors, missing data, duplicated loci, linkage disequilibrium, deviations from Hardy-Weinberg equilibrium, and polymorphism (Benestan et al. 2016). As noted by the authors, these choices can affect inferences in models fit to genetic data, but filtering decisions will be dependent on the dataset and the research question of interest (Andrews et al. 2016). The initial choice of alleles was particularly important in the past, when it was often cost prohibitive to sample more than 20 loci (Waits and Storfer 2016). However, with the advent of next generation sequencing, it is not uncommon to obtain genetic data at tens of thousands of loci. As a result, genetic sampling is expected to be the least limiting factor in future landscape-genetic studies (Balkenhol and Fortin 2016).

The field survey design is another important consideration, but the optimal design is expected to differ depending on the environment and species-of-interest (Balkenhol and Fortin 2016). The number of individuals must be sufficient to represent the genetic diversity in the population and appropriate for the research question (Waits and Storfer 2016). If the genetic diversity is low, then it may be captured with a relatively small number of individuals and alleles; while more individuals and alleles will be required when genetic diversity is high. In rare cases, power analysis is used to identify the minimum

sample size needed (Ryman and Palm 2006). Simulation studies can also help identify the minimum number of individuals and sub-populations needed to detect the effects of distance and landscape resistance on gene flow (Manel et al. 2012). General rules of thumb have been proposed, suggesting that 20 to 30 individuals are needed when using microsatellite data (Hale et al. 2012). However, these numbers are insufficient for the SA models described here. Instead, larger minimum sample sizes are needed (>100 observations in our experience) due to the additional parameters being estimated and the loss of effective degrees of freedom. Larger sample sizes may also be needed as the complexity of the edge-weights model increases. Nevertheless, sample size may not be an issue in many studies, where researchers have artificially decreased the sample size by aggregating genetic data from individuals to the sub-population level. Aggregation is not necessary using this approach, which implies that researchers can make use of all of their genetic data. Although estimating the edge weights within a SA modeling framework may not be possible for every existing dataset, future studies could be designed to meet these requirements.

Finally, it is important to keep in mind that correlation does not equal causation. Many different environmental and biological processes can affect genetic dissimilarity between individuals and populations, and it is possible that patterns in resistance covariates and distance measures mimic patterns produce by the true causal factor (Rellstab et al. 2015). For example, if alleles are incorrectly assumed to be neutral, selection may be causing a particular pattern in genetic differentiation rather than resistance to gene flow (Whitlock and McCauley 1999). Alternatively, migration and drift may not have reached equilibrium

for populations that are currently expanding and as a result, patterns in genetic differentiation would not necessarily reflect current patterns in gene flow (Whitlock and McCauley 1999). Even when assumptions such as these are correct, multiple landscape genetic hypotheses are often highly correlated (Murphy et al. 2008); as was the case here, where we observed similar correlations between genetic data generated using an IBD model ($\mathbf{D}_{\mathrm{IBD}}$) and an IBR estimating model (Appendix S2). Thus, it is important that *a priori* hypotheses describing the effects of landscape resistance on gene flow are carefully constructed based on current scientific knowledge, and tested using sophisticated and robust modelling approaches (Cushman and Landguth 2010), such as those described here.

Conclusions

There is an undeniable need for quantitative methods in landscape genetics that can be used to explore questions about spatial structure in genetic datasets. SA models provide a natural framework to investigate those questions. Spatial autocorrelation underpins common evolutionary-process hypotheses in landscape-genetic studies and thus it is sensible to use a statistical method that incorporates spatial autocorrelation (Balkenhol et al. 2009). SA models are designed to describe the neighborhood structure in spatially correlated network data and provide a flexible probabilistic framework used to make inferences about the effects of habitat selection and movement preferences on gene flow. The data model for these *network-level* analyses may include raw genetic data or genetic distance matrices, as well as covariates on nodes and edges. Covariates representing

multiple evolutionary-process hypotheses can also be assessed within a single modeling framework, which produces interpretable parameter estimates for resistance components, with uncertainty estimates, so that inferences can be made about their relative influence within and between populations. In addition, standard model selection methods, such as regularization or information-theoretic-based approaches, may be used to compare and select among models (Hooten and Hobbs 2015); while predictions, with estimates of uncertainty, can be made at unobserved locations or under different land-use or climate scenarios. The ability to predict provides management benefits (Storfer et al. 2007), but can also be used to validate models using k-fold cross-validation. Most notably, the ability to account for missing data within the SA model means that a contiguous data model can be used when resistance values are estimated. Thus, *a priori* assumptions about the spatial location of edges between non-contiguous nodes, the relative influence of individual resistance covariates, and the overall resistance between nodes are avoided. Closer collaboration between ecologists and spatial statisticians will lead to new methods that are specifically designed to answer spatial and spatio-temporal questions about connectivity in landscape-genetic studies.

Acknowledgments

We would like to thank two anonomous reviewers and Dr Paul Conn for the constructive comments and suggestions they provided on a previous version of this manuscript. This research began from a network-model working group at the Statistics and Applied

Mathematical Sciences (SAMSI) 2014-15 Program on Mathematical and Statistical Ecology. The National Science Foundation, Division of Mathematical Sciences, collaborative research project 1614392 also provided support for this research. Any use of trade, firm, or product names is for descriptive purposes only and does not imply endorsement by the U.S. Government.

References

Adamack, A.T., & Gruber, B. (2014) PopGenReport: simplifying basic population genetic analyses in R. Methods in Ecology and Evolution, 5(4), 384–387.

Akaike, H. 1974. A new look at the statistical model identification, IEEE Transactions on Automatic Control, 19(6): 716–723.

Andrews K.R., Good, J.M., Miller, M.R., Luikart, G., Hohenlohe, P.A. (2016) Harnessing the power of RADseq for ecological and evolutionary genomics. Nature Reviews Genetics, 17: 81–92.

Balkenhol, N., Waits, L.P. & Dezzani, R.J. (2009) Statistical approaches in landscape genetics: An evaluation of methods for linking landscape and genetic data. Ecography, 32: 818–830.

Balkenhol, N., Cushman, S.A., Storfer, A. & Waits, L.P. (2016a) Introduction to landscape genetics: Concepts, methods, and applications, Chapter 1 *In:* Landscape Genetics: Concepts, Methods, and Applications. (eds.) Balkenhol N., Cushman S.A., Storfer A.T., &

Waits L.P., John Wiley & Sons, Hoboken, NJ. pg. 247–255.

Balkenhol, N., Cushman, S.A., Storfer, A. & Waits, L.P. (2016b) Opportunities, and remaining challenges in landscape genetics, Chapter 14 *In:* Landscape Genetics: Concepts, Methods, and Applications. (eds.) Balkenhol N., Cushman S.A., Storfer A.T., & Waits L.P., John Wiley & Sons, Hoboken, NJ. pg. 1–7.

Balkenhol, N., & Fortin M.–J. (2016) Basics of study design: Sampling landscape heterogeneity and genetic variation for landscape genetic studies. Chapter 4 In: Landscape Genetics: Concepts, Methods, Applications, First Edition. Eds: Balkenhol, N., Cushman, S.A., Storfer, A.T., and Waits, L.P. John Wiley & Sons, Ltd. 58–76.

Banerjee, S., Carlin, B.P. & Gelfand, A.E. 2004. Hierarchical modeling and analysis for spatial data. Boca Raton, FL, USA: Chapman & Hall/CRC.

Beier, P., Majka, D.R. & Spencer W.D. (2008) Forks in the road: Choices in procedures for designing wildland linkages. Conservation Biology, 22(4): 836 to 851.

Beier, P., Majka, D.R., & Newell, S.L. (2009) Uncertainty analysis of least-cost modeling for designing wildlife linkages. Ecological Applications, 19(8): 2067–2077.

Benestan, L.M., Ferchaud, A.–L., Hohenlohe, P.A., Garner, B.A., Naylor, G.J., Baums, I.B., Schwartz, M.K., Kelley, J.L., & Luikart, G. (2016) Conservation genomics of natural and managed populations: building a conceptual and practical framework. Molecular Ecology, 25: 2967–2977.

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

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

Botta, F., Eriksen, C., Fontaine, M.C., & Guillot, G. (2015) Enhanced computational methods for quantifying the effect of geographic and environmental isolation on genetic differentiation. Methods in Ecology and Evolution, 6: 1270–1277.

Bradburd, G.S., Ralph, P.L., and Coop, G.M. (2016) A spatial framework for understaning population structure and admixture. PLoS Genetics, 12(1): e1005703. doi:10.1371/journal.pgen.1005703.

Bradburd, G.S., Coop, G.M., & Ralph, P.L. (2018) Inferring continuous and discrete population genetic structure across space. Genetics, 210: 33–52.

Cressie, N. (2015) Statistics for Spatial Data, Revised Edition. John Wiley & Sons, Inc., Hoboken, USA. 928p.

Cushman, S.A. & Landguth, E.L. (2010) Spurious correlations and inference in landscape genetics. Molecular Ecology, 19: 3592–3602.

Cushman, S.A., McKelvey, K.S., Hayden, J. & Schwartz, M.K. (2006) Gene flow in complex landscapes: Testing multiple hypotheses with causal modeling. The American Naturalist, 168(4): 486–499.

Dale, M.R.T. & Fortin, M.-J. (2010) From graphs to spatial graphs. Annual Review of Ecology, Evolution, and Systems, 41, 21–38.

Dyer, R.J. & Nason, J.D. (2004) Population graphs: the graph-theoretic shape of genetic

structure. Molecular Ecology, 13, 1713–1728.

Dyer, R.J. (2017) Genetic Distances, In: Applied Population Genetics, Chapter 20. Retrieved on October 10, 2018 from: https://dyerlab.github.io/applied_population_genetics/genetic-distances.html

Epperson, B.K., McRae, B.H., Scribner, K., Cushman, S.A., Rosenberg, M.S., Fortin, M.-J., James, P.M.A., Murphy, M., Manel, S., Legendre, P., & Dale, M.R.T. (2010) Utility of computer simulations in landscape genetics. Molecular Ecology, 19, 3549–3564.

Fletcher, R.J., Acevedo, M.A., Reichert, B.E., Pias, K.E. & Kitchens, W.M. (2011) Social network models predict movement and connectivity in ecological landscapes. Proceedings of the National Academy of Sciences, 108(48): 19282–19287.

Fotheringham, A.S. & O'Kelly, M.E. (1989) Spatial interaction models: Formulation and applications. Kluwer Academic Publishers, Dordrecht.

Fortin, M.-J. & Dale, M.R.T. (2014) Spatial Analysis: A guide for ecologists. Second Edition. Cambridge University Press, Cambridge, UK. 438p.

Hale, M.L., Burg, T.M., & Steeves, T.E. (2012) Sampling for microsatellite-based population genetic studies: 25 to 30 individuals per population is enough to accurately estimate allele frequencies. PloS One 7, e45170.

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

Hanks, E.M. (2017) rwc: Random Walk Covariance Models. R pacage verion 1.1,

https://CRAN.R-project.org/package=rwc.

Holderegger, R. & Wagner, H.H. (2008) Landscape Genetics. Bioscience, 58(3): 199–207.

Hooten, M.B. & Hobbs, N.T. (2015) A guide to Bayesian model selection for ecologists. Ecological Monographs, 85(1): 3–28.

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

Jombart, T., Pontier, D. & Dufour, A.-B. (2009) Genetic markers in the playground of multivariate analysis. Heredity, 102: 330–341.

Keitt, T.H., Bjornstad, O.N., Dixon, P.M. & Citron-Pousty, S. (2002) Accounting for spatial pattern when modeling organism-environment interactions. Ecography, 25(5): 616–625.

Kossinets, G. (2006) Effects of missing data in social networks. Social Networks, 28: 247–268.

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

Legendre, P., Lapointe, F. & Casgrain, P. (1994) Modeling brain evolution from behavior: A permutational regression approach. Evolution, 48: 1487–1499.

Legendre, P. & Fortin, M.-J. (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.

Legendre, P. & Legendre, L.F.J. (2012) Numerical Ecology. Developments in Environmental Modelling Series, 24. Elsevier, New York, USA. 1006p.

Legendre, P. & Trousselier, M. (1988) Aquatic heterotrophic bacteria: modeling in the presence of spatial autocorrelation. Limnology and Oceanography, 33:1055–1067.

Lichstein, J.W., Simons, T. R., Shriner, S.A., and Franzreb, K.E. 2002. Spatial autocorrelation and autoregressive models in ecology. Ecological Monographs 72: 445–463.

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

Manel, S., Albert, C.H., &Yoccoz, N.G. (2012) Sampling in landscape genomics. In: Data Production and Analysis in Population Genomics. Eds. Pompanon, F., Bonin, A. Human Press, New York, Pages 3–12.

Manel, S. & Holderegger, R. (2013) Ten years of landscape genetics. Trends in Ecology & Evolution, 28(10): 614–621.

Mantel, N. (1967) The detection of disease clustering and a generalized regression approach. Cancer Research, 27: 209–220.

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

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

McRae, B.H., Dickson, B.G., Keitt, T. & Shah, V.B. (2008) Using circuit theory to model

connectivy in ecology, evolution, and conservation. Ecology, 89(10): 2712–2724.

Minor, E.E. & Urban, D.L. (2008) A graph-theory framework for evaluating landscape connectivity and conservation planning. Conservation Biology, 22: 297–307.

Murphy, M.A., Evans, J.S., Cushman, S.A., & Storfer, A. (2008) Representing genetic variation as continuous surfaces: An approach for identifying spatial dependency in landscape genetic studies. Ecography, 31:685–697.

Murphy, M.A., Dezanni, R.J., Pilliod, D. & Storfer, A. (2010) Landscape genetics of high mountain frog metapopulations. Molecular Ecology, 19: 3634–3649.

Murphy, M.A., Dyer, R. & Cushman, S.A. (2016) Graph theory and network models in landscape genetics, Chapter 10. *In:* Landscape Genetics: Concepts, Methods, and Applications. (eds.) Balkenhol N., Cushman S.A., Storfer A.T., & Waits L.P., John Wiley & Sons, Hoboken, NJ. pg. 165–179.

Nakagawa, S. & Freckleton, R.P. (2012) Misisng inaction: the dangers of ignoring missing data. Trends in Ecology and Evolution, 23(11): 592–596.

Nei, M. (1972) Genetic distance between populations. American Naturalist, 106: 283–292.

Petkova, D., Novembre, J. & Stephens, M. (2016) Visualizing spatial population structure with estimated effective migration surfaces. Nature Genetics, 48(1): 94–100.

Queller, D.C. & Goodnight, K.F. (1989) Estimating relatedness using genetic markers. Evolution, 43(2): 258–275.

R Core Team (2016) R: A language and environment for statistical computing. R

Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org.

Rellstab, C., Gugerli, F., Eckert, a.J., Hancock, A.M., & Holderegger, R. (2015) A practical guide to environmental association analysis in landscape genomics. Molecular Ecology, 24: 4348–4370.

Rioux Paquette, S.R., 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., and Held, L. (2005) Gaussian Markov Random Fields: Theory and Applications, in *Monographs on Statistics and Applied Probability* (Vol. 104), Boca Raton, FL: Chapman & Hall.

Ryman, N. & Palm, S. (2006) POWSIM: a computer program for assessing statistical power when testing for genetic differentiation. Molecular Ecology Notes, 6: 600–602.

Saura, S., Bodin, Ö & Fortin, M.-J. (2014) Stepping stones are crucial for species' long-distance dispersal and range expansion through habitat networks. Journal of Applied Ecology, 51(1): 171–182.

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

Smouse, P.E., Long, J.C. & Sokal, R.R. (1986) Multiple regression and correlation extensions of the Mantel test of matrix correspondence. Systematic Zoololgy, 35, 627–632.

Spear, S.F., Balkenhol, N., Fortin, M.-J., McRae, B.H. & Scribner, K. (2010) Use of resistance surfaces for landscape genetic studies. Molecular Ecology, 19: 3576–3591.

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

ter Braak, C.J.F. (1986) Canonical correspondence analysis: a new eigenvector technique for multivariate direct gradient analysis. Ecology, 67(5): 1167–1179.

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

Wagner, H.H. & Fortin, M.-J. (2013) A conceptual framework for the spatial analysis of landscape genetic data. Conservation Genetics, 14: 253–261.

Waits, L.P. & Storfer, A. (2016) Basics of population genetics, Chapter 3 In: Landscape Genetics: Concepts, Methods, Applications, First Edition. Eds: Balkenhol, N., Cushman, S.A., Storfer, A.T., and Waits, L.P. John Wiley & Sons, Ltd. Pages 35–57.

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

Whitlock, M.C. & McCauley, D.E. (1999) Indirect measures of gene flow and migration: $F_{ST} \neq 1/(fNm + 1)$. Heredity, 82:117–125.

Wright, S. (1931) Evolution in Mendelian populations. Genetics, 16(2): 97–159.

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

Zeller, K.A., McGarigal, K. & Whiteley, A.R. (2012) Estimating landscape resistance to movement: A review. Landscape Ecology, 27: 777–797.

Data Availability

Data and R code associated with this study are available on Data Dryad.

Tables

Table 1: The Akaike Information Criteria (AIC) values for the models based on simulated genetic distance ($\mathbf{D}_{\mathrm{IBD}}$, $\mathbf{D}_{\mathrm{IBR}}$, $\mathbf{D}_{\mathrm{IBB}}$) and the three resistance models: isolation by distance (IBD), isolation by resistance (IBR), and isolation by barrier (IBB).

Genetic Distance	**Resistance Model**	**AIC**
$\mathbf{D}_{\mathrm{IBD}}$	IBD	**22518.94**
$\mathbf{D}_{\mathrm{IBD}}$	IBR	22520.23
$\mathbf{D}_{\mathrm{IBD}}$	IBB	22527.61
$\mathbf{D}_{\mathrm{IBR}}$	IBD	22589
$\mathbf{D}_{\mathrm{IBR}}$	IBR	**22573.6**
$\mathbf{D}_{\mathrm{IBR}}$	IBB	22588.43
$\mathbf{D}_{\mathrm{IBB}}$	IBD	20414.75
$\mathbf{D}_{\mathrm{IBB}}$	IBR	20399.52
$\mathbf{D}_{\mathrm{IBB}}$	IBB	**20342.68**

Figure Captions

Figure 1. The data format of the spatially structured ecological network affects the way that edge weights are generated. Resistance covariates for contiguous nodes are based on node characteristics, but can be node- and/or edge-based for non-contiguous nodes. Regardless of the data format, *a priori* assumptions about the neighborhood structure, edge location, and/or resistance values are required and these assumptions influence how resistance is calculated and represented in the model. When *a priori* assumptions are made about the importance of resistance values, they must be aggregated to produce an overall resistance value before model-based assessment takes place. This is not the case for resistance covariates, where importance is assessed for each covariate within a model-based framework.

Figure 2. A summary of common model types and their ability to meet modeling needs for a typical landscape genetics study.

Figure 3. Spatially structured ecological networks contain nodes and edge weights represented in network or matrix format (a). The matrix $\mathbf{W}$ represents edge weights between node pairs, while $\mathbf{D_W}$ is a diagonal matrix containing the sum of the edge weights for each node's first-order neighbors (e.g., $(D_w)_{1,1} = 1 + 4 + 3 = 8$). These two matrices contain information about the conditional structure implied by the edges and is used to generate the precision matrix, $\mathbf{Q}$, in a conditional autoregressive (CAR) model (b). Two nodes that are conditionally independent in the precision matrix (e.g., $Q_{3,2} = 0$) may still be spatially dependent (i.e., correlated) through intervening nodes (e.g., $\Sigma_{3,2} \neq 0$) in the

covariance matrix, $\mathbf{\Sigma} = \mathbf{Q^{-1}}$ (c) and the correlation matrix (e). The precision matrix defines the partial correlation among measurements on nodes (d) after accounting for the influence of intervening nodes.

Figure 4. Resistance surfaces for the isolation by distance (IBD), isolation by resistance (IBR), and isolation by barrier (IBB) evolutionary-process hypotheses.

Figure 5. (a) The isolation-by-resistance (IBR) model ($\mathbf{D}_{\text{IBR}} \sim \text{IBR}$) shows that conductance (inverse resistance) has a non-linear relationship with the IBR resistance covariate (solid black line). The dotted lines denote the 95% confidence intervals. (b) Similar patterns are observed in a map of mean conductance, which is highest in areas with low resistance covariate values.

781

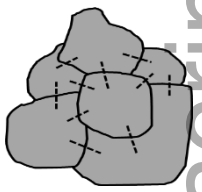
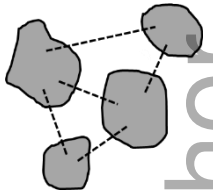
Figure 1:

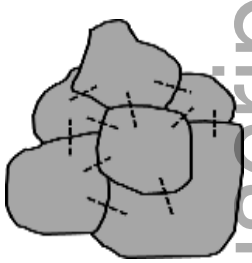
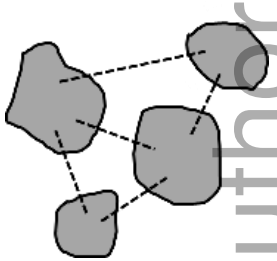
Figure 2:

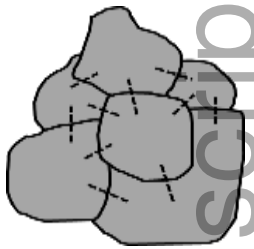
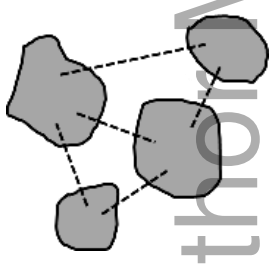
Figure 3:

Figure 4:

Figure 5:

Data Format	Resistance Covariates	*A priori* Assumptions	Assumption-based Outcomes	Resistance Aggregation	Model
Contiguous 	**Node-based** Size, habitat quality, vegetation cover, distance between centroids	**1st Order Neighborhood Structure** Rook, Queen, % Shared boundary	**Resistance Covariates**	**None**	**Model-based** Estimates for individual resistance covariates
Non-contiguous 	**Node- or/and Edge-based** Habitat quality, vegetation cover, distance between node boundaries or centroids	**Edge Location** Euclidean distance, buffered Euclidean distance	**Resistance Covariates**		
		Resistance Values Expert opinion, literature review, empirical data, animal movement rates	**Edge Location** Least-cost path, buffered least-cost path, multiple-least-cost path, Circuit-scape	Sum, difference, weighted product, mean, geometric mean, median of resistance values	**Model-based** Estimate for overall resistance values

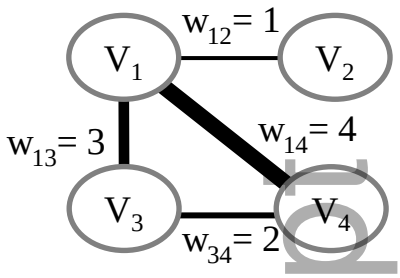
Data Format	Resistance Covariates	A priori Assumptions	Assumption-based Outcomes	Resistance Aggregation
	**Node-based** Size, habitat quality, vegetation cover, distance between centroids	**1st Order Neighborhood Structure** Rook, Queen, % Shared boundary		
		Edge Location Euclidean distance, buffered Euclidean distance	**Resistance Values** Model estimates	
	**Node- or/and Edge-based** Habitat quality, vegetation cover, distance between node boundaries or centroids	**Resistance Values** Expert opinion, literature review, empirical data, rates of animal movement	**Edge Location** least-cost path, buffered least-cost path, multiple-least-cost path, Circuit-scape	
		Edge Location + Resistance Values		

Data Format	Resistance Covariates	*A priori* Assumptions	Assumption-based Outcomes	Resistance Aggregation
	**Node-based** Size, habitat quality, vegetation cover, distance between centroids	**1st Order Neighborhood Structure** Rook, Queen, % Shared boundary	**Resistance Values** Model estimates	Sum, difference, weighted product, mean, geometric mean, median
	**Node- or/and Edge-based** Habitat quality, vegetation cover, distance between node boundaries or centroids	**Edge Location** Euclidean distance, buffered Euclidean distance	**Resistance Values** Model estimates	
		Resistance Values Expert opinion, literature review, empirical data, animal movement rates	**Edge Location** Least-cost path, buffered least-cost path, multiple-least-cost path, Circuit-scape	
		Edge Location + Resistance Values		

Type	Method	Probabilistic Distribution	No *a priori* Resistance Assumptions	Estimated Resistance Component Parameters	Spatially Correlated Residuals Permitted	Model Selection	Missing Data	Prediction	Sources and Applications
Simulation	Computer simulation	✗	✗	✗	✓	✗	✓	✓	Epperson et al. (2010) Cushman & Landguth (2010)
Matrix Correlation	Mantel/Partial Mantel tests	✗	✗	✗	✗	[1]PT	✗	✗	Mantel (1967), Smouse et al. (1986), Cushman et al. (2006), Legendre & Fortin (2010)
Ordination	Multi-dimensional scaling	✗	✗	✗	✗	[1]PT	✗	✗	Legendre & Legendre (2012), Legendre & Fortin (2010)
	Spatial principle components analysis	✗	✗	✗	✓	[1]PT	✗	✗	Jombart et al. (2008)
	Correspondence analysis, Redundancy analysis, Canonical correlation analysis	✗	✗	✗	✗	[1]PT	✗	✗	ter Braak (1986), Balkenhol et al. (2009), Jombart et al. (2009), Legendre & Fortin (2010), Fortin & Dale (2014)
Regression	Multiple regression on distance matrices	✗	✗	✓	✗	[1]PT	✗	✗	Legendre et al. (1994), Legendre et al. (2015)
	Gravity model: Unconstrained & Singly Constrained	✓	✗	✗	✗	[2]DA, ITC, CV	✗	✓	Fotheringham & O'Kelly (1989), Murphy et al. (2010), Murphy et al. (2016)
	Spatial Autoregressive model	✓	✓	✓	✓	[2]DA, ITC, CV	✓	✓	Hanks and Hooten (2013), Ver Hoef et al. (2017)

[1]Permutation test (PT), [2]Distributional assumptions, Information theoretic criteria, cross-validation (DA, ITC, CV)

Network and Matrix Representation



$\mathbf{W} =$

	V_1	V_2	V_3	V_4
V_1	0	1	3	4
V_2	1	0	0	0
V_3	3	0	0	2
V_4	4	0	2	0

$\mathbf{D}_W =$

	V_1	V_2	V_3	V_4
V_1	8	0	0	0
V_2	0	1	0	0
V_3	0	0	5	0
V_4	0	0	0	6

(a)

Precision Matrix

$\mathbf{Q} = (\mathbf{D}_W - \rho\mathbf{W}), \quad \rho = 0.90$

$\mathbf{Q}_{CAR} =$

	V_1	V_2	V_3	V_4
V_1	8.0	-0.9	-2.7	-3.6
V_2	-0.9	1.0	0	0
V_3	-2.7	0	5.0	-1.8
V_4	-3.6	0	-1.8	6.0

(b)

Covariance Matrix

$\mathbf{\Sigma} = \mathbf{Q}^{-1}$

$\mathbf{\Sigma}_{CAR} =$

	V_1	V_2	V_3	V_4
V_1	0.55	0.49	0.46	0.47
V_2	0.49	1.44	0.42	0.42
V_3	0.46	0.42	0.62	0.46
V_4	0.47	0.42	0.46	0.59

(c)

Partial Correlation Matrix

$\mathbf{k}_{i,j}|. = \mathrm{corr}(\varepsilon_{i|.}, \varepsilon_{j|.}) = -\mathbf{Q}_{i,j}/\sqrt{\mathbf{Q}_{i,i}\mathbf{Q}_{j,j}}$

$\mathbf{k}_{CAR} =$

	V_1	V_2	V_3	V_4
V_1	1	0.32	0.43	0.52
V_2	0.32	1	0	0
V_3	0.43	0	1	0.33
V_4	0.52	0	0.33	1

(d)

Correlation Matrix

$\mathbf{Corr}_{i,j} = \mathbf{\Sigma}_{i,j}/\sqrt{\mathbf{\Sigma}_{i,i}\mathbf{\Sigma}_{j,j}}$

$\mathbf{Corr}_{CAR} =$

	V_1	V_2	V_3	V_4
V_1	1	0.55	0.80	0.83
V_2	0.55	1	0.44	0.46
V_3	0.80	0.44	1	0.77
V_4	0.83	0.46	0.77	1

(e)

ecm_1355_f4.pdf