Chapter 5

BASICS OF SPATIAL DATA ANALYSIS: LINKING LANDSCAPE AND GENETIC DATA FOR LANDSCAPE GENETIC STUDIES

Helene H. Wagner and Marie-Josée Fortin

Department of Ecology and Evolutionary Biology, University of Toronto, Canada

5.1 INTRODUCTION

The research questions and data analysis methods in landscape genetics stem from a broad range of fields, including population genetics, numerical ecology, metapopulation, landscape ecology, and spatial statistics. Yet landscape genetics has unique research questions (see Chapters 1 to 4 in this book) that can only be addressed through a combination of statistical methods originating from different disciplines. A better integration and unification of these statistical methods may thus be crucial for advancing landscape genetics (Balkenhol et al. 2009a, 2009b).

A core issue is how to explicitly account for the inherent spatial structure of the genetic and landscape data while analyzing their relationship. Indeed, geophysical processes and land-use practices create spatial structure in environmental factors to which the organisms respond, and biotic processes such as spatially restricted mating and dispersal create further spatial structure. Statistically speaking, these processes often create positive ***spatial autocorrelation*** in the genetic data, which means that, on average, nearby observations are more similar than distant ones. Negative spatial autocorrelation, on the other hand, occurs when similar observations are spaced more regularly than random, so that nearby observations are on average more dissimilar than more distant ones. It is helpful to distinguish between types of spatial autocorrelation based on the generating process. ***Induced spatial dependence*** results from the response of organisms to environmental gradients, whereas ***inherent spatial autocorrelation*** arises from spatial ecological or evolutionary processes such as mating, dispersal, and resulting gene flow. Unfortunately, spatial analysis methods cannot discriminate between induced spatial dependence and inherent

Landscape Genetics: Concepts, Methods, Applications, First Edition. Edited by Niko Balkenhol, Samuel A. Cushman, Andrew T. Storfer, and Lisette P. Waits.

spatial autocorrelation. This means that spatial autocorrelation in regression residuals may be due to biological processes or a missing landscape predictor. Therefore prior knowledge and hypotheses about the underlying processes responsible for the spatial pattern should be used to design an effective sampling design (see Chapter 4) and to apply the appropriate levels of analysis, which in turn will allow researchers to differentiate between the origins of the spatial patterns.

The presence of spatial autocorrelation in sampled data can be assessed using spatial statistics (see Box 5.1). Like any parametric statistics, spatial statistics are based on assumptions. The main assumption of spatial statistics is that the underlying process that generated the spatial pattern is stationary (i.e., constant mean and variance over the study area) (Fortin and Dale 2005). Non-stationarity can be due to several aspects: (i) large-scale spatial trend in the data such that the mean changes approximately linearly throughout the studied area; (ii) there are spatial changes in mean, variance, or both according to location in the studied area; and (iii) there is directionality (e.g., due to wind direction) in the spatial pattern of the data. When this stationarity assumption is valid, however, spatial statistics (e.g., Moran's *I*, Geary's *c*, semivariance) that measure the average degree of spatial autocorrelation at different spatial lags for the entire study area can be used. These spatial statistics estimate the degree of spatial autocorrelation of a single variable (univariate methods) for the entire study area (i.e., magnitude of spatial autocorrelation, spatial range of the pattern). However, the multivariate nature of genetic data and the many factors that contribute to their spatial structure require multivariate spatial analysis methods (Dray et al. 2012).

Moran's I is the most widely used statistic to estimate spatial autocorrelation. Moran's I mostly varies between -1 and $+1$ (deviations may occur for small samples; de Jong et al. 1984), with an expected value of $-1/(n-1)$ for a sample of n spatially independent observations (Moran 1950). Positive values of Moran's I indicate positive spatial autocorrelation, where nearby observations are more similar on average than distant ones, whereas negative values indicate negative spatial autocorrelation, where nearby samples are more dissimilar than distant ones. A Moran's I correlogram for a single variable $\mathbf{y}$ is constructed by calculating Moran's $I(d)$ for each spatial lag d, with weights $w_{ij(d)} = 1$ if the pair of sites i and j fall into spatial lag d and $w_{ij(d)} = 0$ otherwise. Note that a global Moran's I index refers to the value of a Moran's $I(d)$ correlogram for the first lag, $d = 1$.

A spatial pattern is typically quantified in one of two ways that differ in the interpretation of spatial lags: (1) as a function of the *distance lag* between pairs of observations or (2) through a *neighbor matrix* and weights associated with these neighbors (see Box 5.1). The distance lag approach is compatible with an isolation-by-distance model of gene flow, where rates of gene flow depend largely on the total distance between sampling locations. The neighbor matrix approach is compatible with a stepping stone model, where organisms are expected to disperse to neighboring patches only, and gene flow over larger distances is the result of such stepwise dispersal (migration) events over multiple generations. In each case, we need to be clear about the null and alternative hypothesis we aim to test. The conceptual difference between these two approaches is best illustrated with the calculation and interpretation of a Moran's I correlogram (Epperson 2003; Guillot et al. 2009; see Box 5.1).

When the stationarity assumption of spatial statistics is not valid, spatial autocorrelation can be measured locally at each sampling location with "local spatial statistics" (e.g., local Moran, local Getis; Anselin 1995; Sokal et al. 1998), using only neighboring samples (i.e., first neighbors, $d = 1$). As landscape genetics studies are often designed over large regions, the likelihood that several ecological and environmental factors are acting on the genetic spatial structure is high, which makes the assumption of stationarity unlikely. With such genetic data, local spatial statistics can be used to detect local spatial heterogeneity (Sokal et al. 1998).

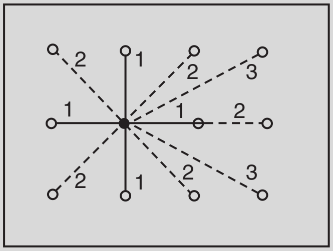
Quantification of the spatial structure in the respective data sets (landscape and genetic), on the one hand, provides key information on the potential generating processes and the scales at which they affect the genetic structure of the data (e.g., isolation-by-distance (IBD), Wright 1943; isolation-by-resistance (IBR), McRae 2006; or isolation-by-barrier (IBB), Vignieri 2005). On the other hand, not accounting for the presence of spatial autocorrelation in the data may invalidate inferential statistical tests of the relationship between genetic and landscape data, as these statistical tests assume that the data are independent (see Box 5.1).

Most ecological and genetic data show inherent structure in space (e.g., nearby samples usually have similar values; see Box 5.1), time (e.g., population fluctuations), or phylogeny (e.g., species relatedness) (Fortin & Dale 2005; Peres-Neto 2006). In this chapter we focus solely on spatial structure, acknowledging that the other types of dependency occur. Also, the power of landscape

Box 5.1 Spatial statistics

Distance lag approach

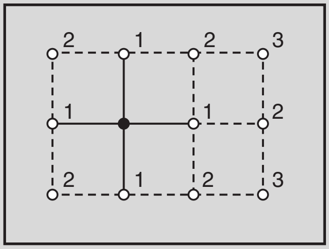
A correlogram based on distance lags is constructed by dividing all unique pairs of observations into spatial distance classes (i.e., spatial lags). The null hypothesis is the absence of spatial autocorrelation. The alternative hypothesis is that under isolation-by-distance, we expect increased rates of gene flow among nearby sampling locations, resulting in positive autocorrelation for the first spatial lags. Technically, we can assign each pair of observations at locations *i* and *j* a weight $w_{ij(d)} = 1$ if it falls into distance class or lag *d* and $w_{ij(d)} = 0$ otherwise. For each lag, the autocorrelation is estimated by dividing the (weighted) mean covariance for all pairs within the distance class by the mean covariance among all pairs in the data set. The spatial correlogram is a plot of these autocorrelation estimates against lag distance. In this case, a progressive, one-sided test is appropriate, where we test the first lag and only progress to further lags if all previous lags showed significant positive spatial autocorrelation (Legendre & Legendre 2012). In contrast, testing each lag individually may provide a single lag with positive or negative spatial autocorrelation at a larger distance, which is difficult to interpret in biological terms. Similarly, negative autocorrelation, where nearby observations are genetically more different than distant ones, will rarely be expected. The shape of the correlogram may be used to select and fit an appropriate function for modeling the correlation among regression errors in generalized least squares (GLS) regression or generalized linear mixed models (GLMMs).



Distance lag approach. Each line indicates a link between the focal site *i* (filled circle) and a nearby site *j*. Numbers indicate distance classes (lags) and depend on lag definition. Solid lines indicate links in the first lag.

Neighbor matrix approach

Alternatively, we can start by constructing an $n \times n$ neighbor matrix, where "1" indicates that the row and column observations are neighbors and "0" that they are not neighbors. Neighbors can be defined using a distance threshold or based on a specific graph model (see Chapter 9; Dale & Fortin 2010; Spear et al. 2010). The first spatial lag is defined by first neighbors (those indicated by "1" in the neighbor matrix). The second lag is defined by second neighbors (i.e., locations that could be reached in two generations), etc. If using binary weights, each pair of observations receives the same weight and we proceed as above. However, other weights may be used: we may want to adjust for the number of neighbors *j* of each observation *i*, so that the weights w_{ij} of all neighbors of *i* sum to one (such row-standardized weights should be used for spatial regression). For an irregular spatial sampling design, we may want to account for the physical distance between neighbors, so that close neighbors receive more weight than distant ones. In a stepping stone model, we are typically interested in the autocorrelation among first neighbors only (i.e., global Moran's *I* index calculated for the first spatial lag, $d = 1$), as rates of gene flow among neighbors that are further remote follow from the connectivity among first neighbors. The exact estimate and *p*-value of global Moran's *I* index will vary with different definitions of first neighbors and weights. Note that when modeling spatial dependence, e.g., using spatial regression methods, the ideal number of neighbors is 4 – 6, whereas higher numbers of neighbors are inefficient (Florax & Ray 1995; Griffith 1996). The dependence between a sampling location and its second, third, etc., neighbors is indirectly modeled through the dependence on their common neighbours.



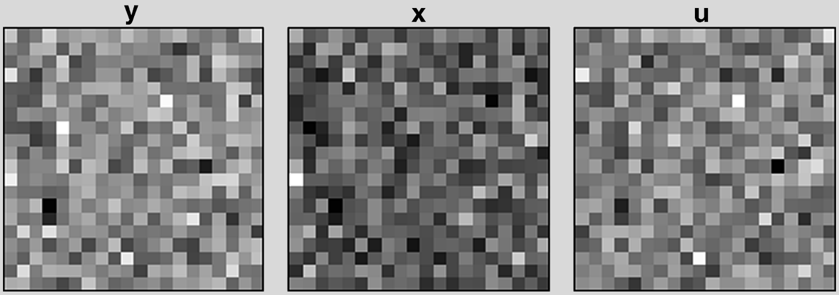
Neighbor matrix approach. Each line indicates a connection between nearest neighbors, thus depending on the definition of neighbors. Numbers indicate the lag, defined by the minimum number of steps between the focal site *i* (filled circle) and a nearby site *j*. Solid lines indicate connections between site *i* and its nearest neighbors.

genetics analyses to detect significant spatial relationships between genetic and landscape data is directly linked to the sampling design (Muirhead et al. 2008; Chapter 4), the spatial analysis methods (Fortin & Dale 2005; Epperson 2003), and the genetic markers used (Ryman et al. 2006). Here we focus only on the statistical aspects; discussion about the importance of the sampling design and genetic markers can be found elsewhere (e.g., Selkoe & Toonen 2006; Chapters 2 and 4).

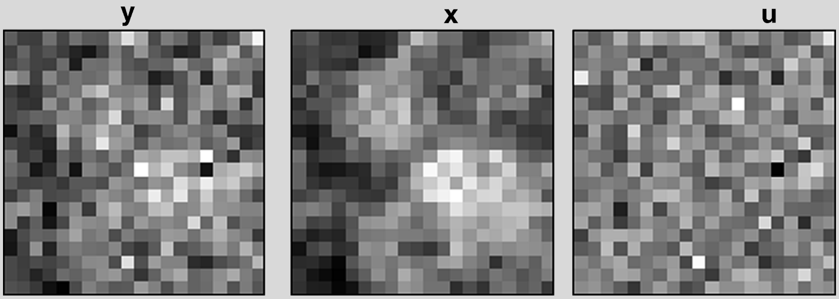
While we may gain new insights through characterizing spatial structure in the data with spatial statistics, the presence of spatial autocorrelation may invalidate statistical results (Legendre 1993). The example in Box 5.2 illustrates that spatial autocorrelation in either the

Box 5.2 Spatial autocorrelation and regression inference

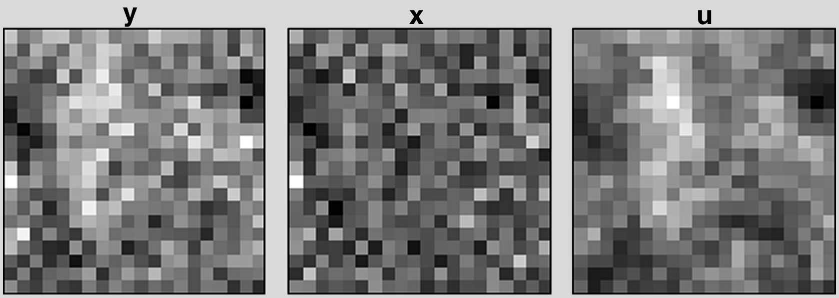
Our example for illustrating the effects of spatial autocorrelation on regression inference uses a linear simple regression with a single response variable **y**, a single predictor **x**, and an error **u**, simulated on a 20 × 20 grid ($n = 400$) under four scenarios. The scenarios differ in the spatial structure of predictor **x** and error **u** and consequently of the response **y** simulated as $\mathbf{y} = 0.5\mathbf{x} + 0.5\mathbf{u}$. In all simulations, the true slope is $b = 0.5$. Under each scenario, we calculated Moran's *I* of the residuals of a regression of **y** on **x** and estimated the slope (*b*) and its standard error (*SE*). We repeated the simulation and regression analysis 10,000 times under each scenario to estimate the true *SE* of slope *b* from the standard deviation of all replicate slope estimates. For each simulated data set we also regressed **u** on **x** to determine type I error rates. Values for Moran's *I* of residuals, slope *b*, and its estimated *SE* are means over 10,000 replicates.



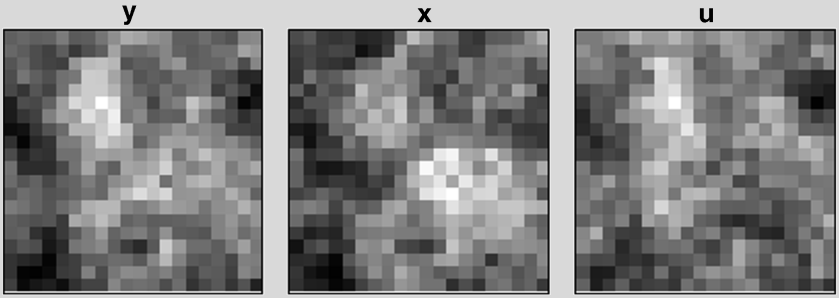
Scenario 1: **x** random, **u** random. The landscape predictor and the response are spatially independent and linear regression results are correct: Moran's *I* of residuals = –0.002, slope estimate ($b \pm SE$) = 0.500 ± 0.025, true *SE* of slope: 0.025, type I error rate: 0.050.



Scenario 2: **x** autocorrelated, **u** random. The landscape predictor is spatially structured, but the response is spatially independent. Linear regression results are correct: Moran's *I* of residuals = –0.003, slope estimate ($b \pm SE$) = 0.500 ± 0.025, true SE of slope: 0.025, type I error rate: 0.048.



Scenario 3: **x** random, **u** autocorrelated. The landscape predictor is spatially independent, but the response is spatially structured, which may be due to a biotic process such as dispersal or a missing landscape factor that is spatially structured. The linear regression results are correct, although the residuals are spatially autocorrelated: Moran's *I* of residuals = 0.640, slope estimate ($b \pm SE$) = 0.501 ± 0.025, true SE of slope: 0.025, type I error rate: 0.054.



Scenario 4: **x** autocorrelated, **u** autocorrelated. The landscape predictor is spatially structured and the response to it shows further spatial dependence. Linear regression results are incorrect: Moran's *I* of residuals = 0.629, slope estimate ($b \pm SE$) = 0.501 ± 0.025, true *SE* of slope: 0.071, type I error rate: 0.504.

predictor or the error of the response (i.e., residuals) does not necessarily invalidate regression results. If both the landscape predictor and the error are spatially autocorrelated, however, the effect on regression results can be severe, biasing the estimation of the parameters of the regression and their significance. Indeed, type I error rates may be considerably inflated (in the example ranging from 0.05 to 0.5), which means that on average one in two significant slope coefficients may be spurious results. In addition, the standard error of the slope coefficient *b* may be underestimated (in the example in Box 5.2 by a factor of 0.071/0.025 = 2.84). This means that confidence intervals for the slope constructed with the estimated standard error would be almost three times too narrow on average. Thus, both statistical hypothesis testing and parameter estimation in simple linear regression are invalid in this situation (Bini et al. 2009).

This chapter discusses the main issues of relating genetic variation to landscape predictors (Foll & Gaggiotti 2006; Bradburd et al. 2013) in the familiar context of regression analysis or, more generally, in the framework of the linear model. We start with considerations for choosing an appropriate model depending on assumptions about the main underlying evolutionary process, i.e., selection or gene flow (Fig. 5.1). We then present different approaches through which space can be incorporated into the linear model (Fig. 5.2). Finally, we focus on two common goals of spatial analysis in landscape genetics: first, how to test for

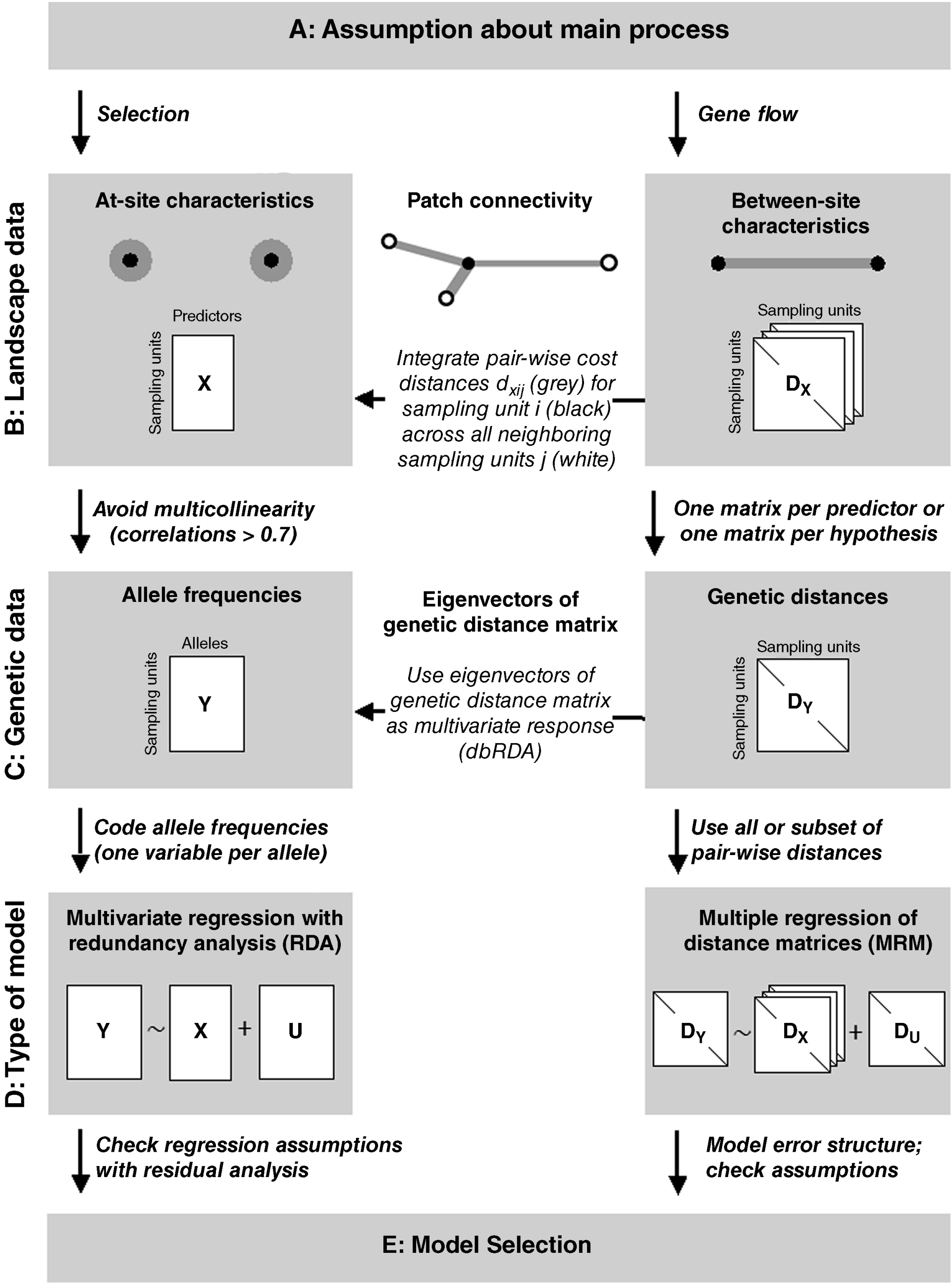

A: Assumption about main process
Selection
Gene flow
B: Landscape data
At-site characteristics
Predictors
Sampling units
X
Patch connectivity
Integrate pair-wise cost distances d_{xij} (grey) for sampling unit i (black) across all neighboring sampling units j (white)
Between-site characteristics
Sampling units
Sampling units
D_X
Avoid multicollinearity (correlations > 0.7)
One matrix per predictor or one matrix per hypothesis
C: Genetic data
Allele frequencies
Alleles
Sampling units
Y
Eigenvectors of genetic distance matrix
Use eigenvectors of genetic distance matrix as multivariate response (dbRDA)
Genetic distances
Sampling units
Sampling units
D_Y
Code allele frequencies (one variable per allele)
Use all or subset of pair-wise distances
D: Type of model
Multivariate regression with redundancy analysis (RDA)
Y ~ X + U
Multiple regression of distance matrices (MRM)
D_Y ~ D_X + D_U
Check regression assumptions with residual analysis
Model error structure; check assumptions
E: Model Selection

Fig. 5.1 Flowchart of the statistical model that can be used to relate genetic to landscape data depending on whether one assumes selection or gene flow to be the main underlying evolutionary process. In either case five steps are needed. (A) Determining implicitly or explicitly the main assumptions of the processes. (B) Determining how the landscape data will be analyzed. (C) Determining how the genetic data will be analyzed. (D) Selecting the appropriate regression framework. (E) Selecting the appropriate model.

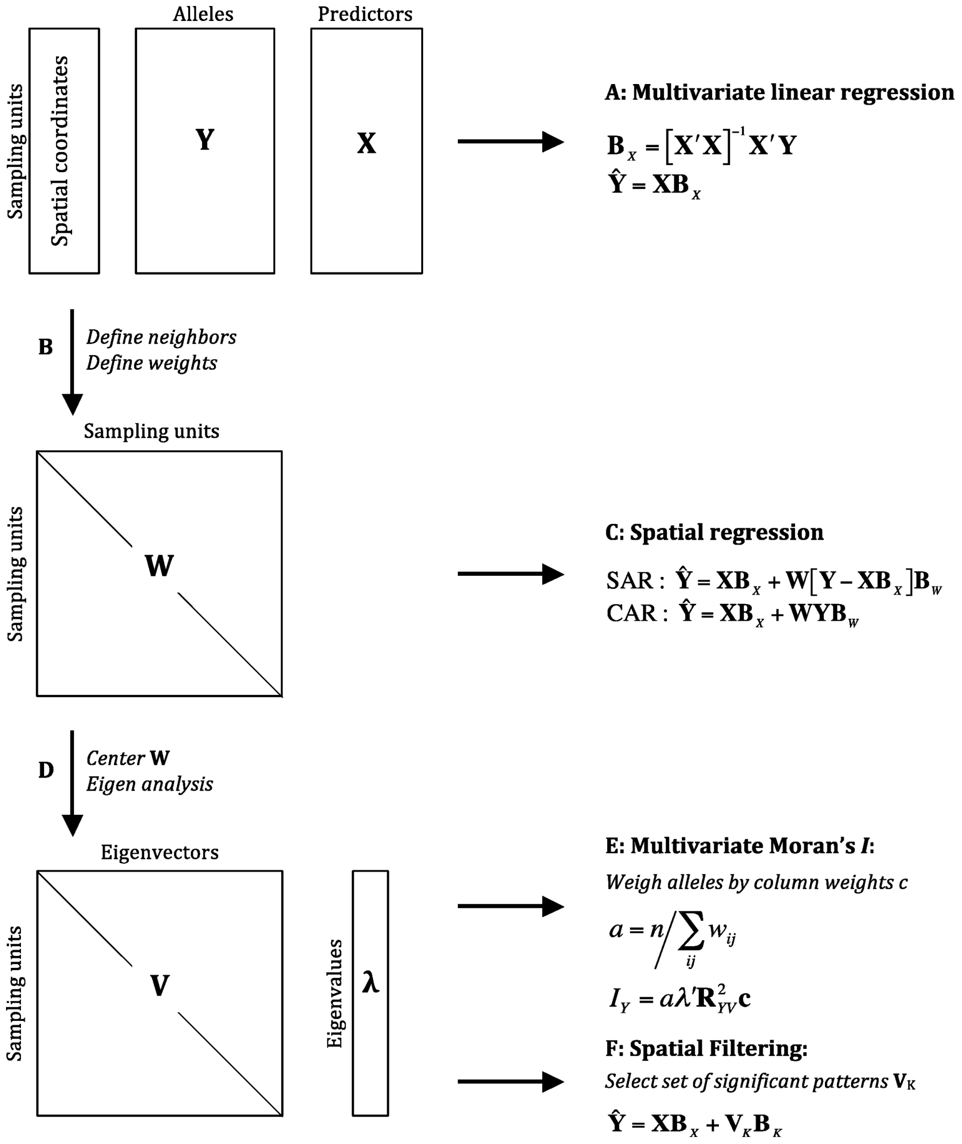

Sampling units
Spatial coordinates
Alleles
Y
Predictors
X
A: Multivariate linear regression
$\mathbf{B}_X = [\mathbf{X}'\mathbf{X}]^{-1}\mathbf{X}'\mathbf{Y}$
$\hat{\mathbf{Y}} = \mathbf{X}\mathbf{B}_X$
B
Define neighbors
Define weights
Sampling units
Sampling units
W
C: Spatial regression
SAR : $\hat{\mathbf{Y}} = \mathbf{X}\mathbf{B}_X + \mathbf{W}[\mathbf{Y} - \mathbf{X}\mathbf{B}_X]\mathbf{B}_W$
CAR : $\hat{\mathbf{Y}} = \mathbf{X}\mathbf{B}_X + \mathbf{W}\mathbf{Y}\mathbf{B}_W$
D
Center W
Eigen analysis
Eigenvectors
Sampling units
V
Eigenvalues
λ
E: Multivariate Moran's I:
Weigh alleles by column weights c
$a = n \Big/ \sum_{ij} w_{ij}$
$I_Y = a\boldsymbol{\lambda}'\mathbf{R}^2_{YV}\mathbf{c}$
F: Spatial Filtering:
Select set of significant patterns $\mathbf{V}_K$
$\hat{\mathbf{Y}} = \mathbf{X}\mathbf{B}_X + \mathbf{V}_K\mathbf{B}_K$

Fig. 5.2 Flowchart illustrating the steps involved when incorporating space into the multiple linear regression model. (A) Non-spatial regression predicts response **Y** from predictors **X** without reference to spatial coordinate information. (B) Many methods for incorporating space into the regression model start with creating a spatial weight matrix **W**. (C) Spatial regression methods use **W** directly to add a spatial neighborhood term for the response **Y** (conditional autoregressive model, CAR) or the error **U** (simultaneous autoregressive model, SAR). (D) Alternatively, the spatial weight matrix **W** can be centered and subjected to eigenanalysis to extract matrix **V** of spatial eigenvectors. (E) The eigenvalues **λ** corresponding to the spatial eigenvectors **V** can be used to calculate univariate or multivariate Moran's *I* as a measure of the degree of spatial autocorrelation in the data. (F) Spatial filtering methods add a set $\mathbf{V}_K$ of spatial eigenvectors that are significantly associated with the response **Y** to the regression model to control for spatial autocorrelation.

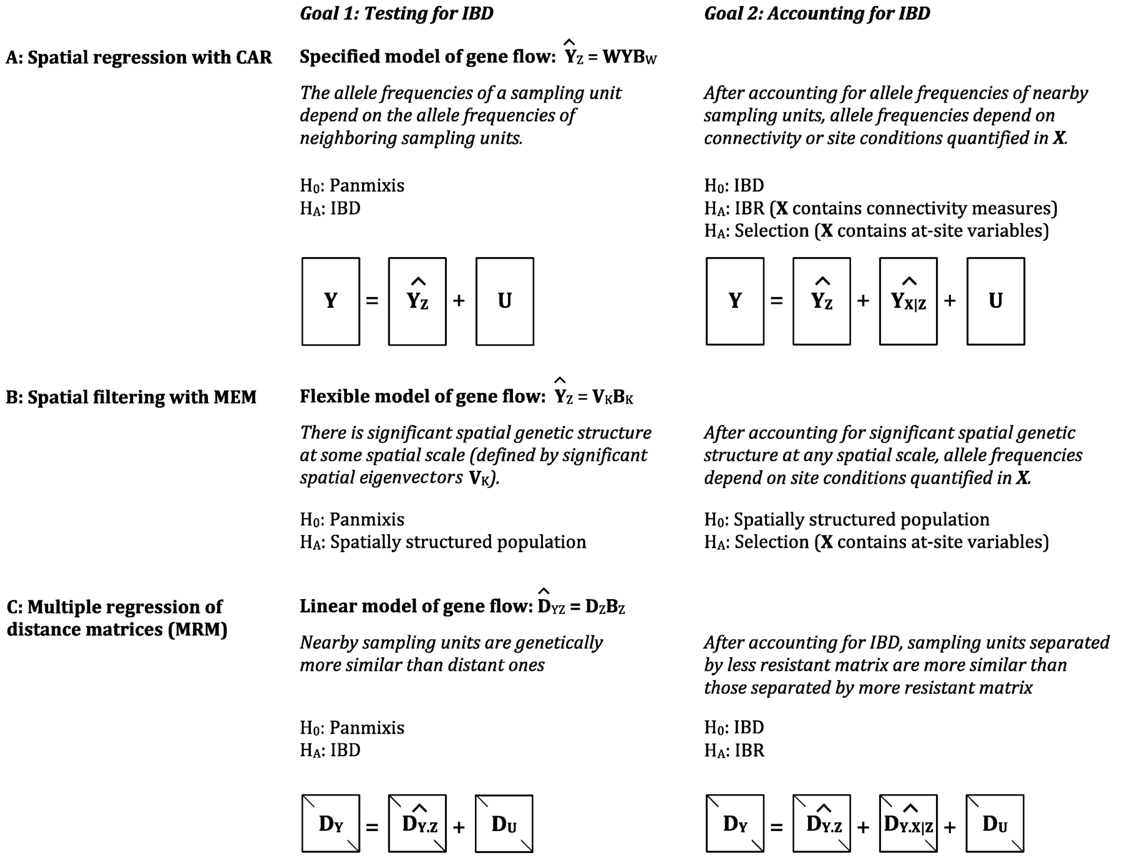

Goal 1: Testing for IBD
Goal 2: Accounting for IBD
A: Spatial regression with CAR
Specified model of gene flow: $\hat{Y}_Z = WYB_W$
The allele frequencies of a sampling unit depend on the allele frequencies of neighboring sampling units.
After accounting for allele frequencies of nearby sampling units, allele frequencies depend on connectivity or site conditions quantified in X.
H_0: Panmixis
H_A: IBD
H_0: IBD
H_A: IBR (X contains connectivity measures)
H_A: Selection (X contains at-site variables)
$Y = \hat{Y}_Z + U$
$Y = \hat{Y}_Z + \hat{Y}_{X|Z} + U$
B: Spatial filtering with MEM
Flexible model of gene flow: $\hat{Y}_Z = V_K B_K$
There is significant spatial genetic structure at some spatial scale (defined by significant spatial eigenvectors V_K).
After accounting for significant spatial genetic structure at any spatial scale, allele frequencies depend on site conditions quantified in X.
H_0: Panmixis
H_A: Spatially structured population
H_0: Spatially structured population
H_A: Selection (X contains at-site variables)
C: Multiple regression of distance matrices (MRM)
Linear model of gene flow: $\hat{D}_{YZ} = D_Z B_Z$
Nearby sampling units are genetically more similar than distant ones
After accounting for IBD, sampling units separated by less resistant matrix are more similar than those separated by more resistant matrix
H_0: Panmixis
H_A: IBD
H_0: IBD
H_A: IBR
$D_Y = \hat{D}_{Y.Z} + D_U$
$D_Y = \hat{D}_{Y.Z} + \hat{D}_{Y.X|Z} + D_U$

Fig. 5.3 Summary of how the null and alternative hypotheses change when the goal is either to test for IBD (left column) or to account for IBD (right column) when testing other landscape predictors, depending on the statistical approach used. (A) Spatial regression using the conditional autoregressive (CAR) model. (B) Spatial filtering using spatial eigenvectors (MEM). (C) Multiple regression of distance matrices (MRM).

the presence of significant IBD and then how to account for IBD by incorporating it into the null model when testing for other landscape effects (Fig. 5.3). We hope that this presentation will aid the current efforts of landscape geneticists to develop new and more integrated methods for linking genetic and landscape data to address the complexity of landscape genetic research questions (Wagner & Fortin 2013).

5.2 HOW TO MODEL LANDSCAPE EFFECTS ON GENETIC VARIATION

To describe the relationship between landscape predictors and genetic data, we need an appropriate statistical model. The type of model depends on the data types, which again will depend on how we think about the underlying processes (Fig. 5.1A).

5.2.1 Type of landscape data

Organism behavior may depend on (i) the local site conditions at the sampling locations (*at-site characteristics*, i.e., conditions at the grid cell where an individual was sampled or the patch where a discrete local population was sampled), (ii) the local neighborhood (e.g., proximity of a road or resource availability within a distance threshold), or (iii) the intervening landscape matrix between locations with suitable habitat (*between-site characteristics*, e.g., the presence of barriers or the mortality and energetic cost of movement associated with different cover types).

If selection is the main evolutionary process that structures the population of interest, we would expect that genetic variation depends mostly on at-site characteristics, although between-site characteristics may affect the rate of spread of adaptive genetic variation across the landscape. If, however, gene flow is the dominant process, genetic variation would depend to a large degree on between-site characteristics that are likely to affect rates of gene flow, although the at-site characteristics may also affect the probability of individuals leaving a patch, finding a patch, and settling in it (Fig. 5.1B).

In the case of selection, we will represent a set of p landscape predictors observed for n sampling units (representing n individuals in a continuous population or n demes as spatially discrete populations) in a predictor matrix $\mathbf{X}$ with n rows and p columns (*node-level analysis*, Box 5.3; Wagner & Fortin 2013). In the case of gene flow, the values of the landscape predictors refer to the pairwise distances between sampling units and are best represented as a set of distance matrices $\mathbf{D_X}$, each with n rows and n columns (*link-level analysis*, Box 5.3). The terminology is borrowed from graph theory (see Chapter 10), where one would describe a set of habitat patches as *nodes* connected by links along which organisms may move across the intervening matrix (Wagner & Fortin 2013).

For link-level analysis, either each predictor is represented by its own distance matrix (e.g., one matrix for roads, one for forest), leading to a set of p matrices $\mathbf{D_{X1}}$ to $\mathbf{D_{X\mathit{p}}}$, or each matrix represents a complex hypothesis of landscape resistance that assigns a specific set of resistance values to all landscape features (e.g., high resistance to roads and low resistance to forest). The main difference in terms of statistical analysis is that, if the set of p matrices $\mathbf{D_X}$ represents p landscape predictors, more than one may be required to explain the genetic data (e.g., roads and forest). However, if $\mathbf{D_X}$ represents p competing hypotheses of landscape resistance, any one hypothesis would exclude all the others; hence we would not want to use more than one matrix $\mathbf{D_X}$ as the predictor in the same regression model.

If we are going to use multiple predictors in a multiple regression model, we need to avoid *multicollinearity*. In the strict sense, perfect multicollinearity refers to the case where one predictor variable is a linear combination of some other predictors. An example would be coding a categorical factor with four levels A to D, corresponding to different cover types, into $q=4$ dummy variables $\mathbf{X}_\mathrm{A}$ to $\mathbf{X}_\mathrm{D}$, one for each factor level. A value of $\mathbf{X}_\mathrm{D}=1$ thus indicates that the sampling unit was classified as cover type D and $\mathbf{X}_\mathrm{D}=0$ that it was not classified as type D. However, if we know variables $\mathbf{X}_\mathrm{A}$ to $\mathbf{X}_\mathrm{C}$, we know the value of $\mathbf{X}_\mathrm{D}$ because the four dummy variables must sum to one for each sampling unit; hence $\mathbf{X}_\mathrm{D}$ is a linear combination of $\mathbf{X}_\mathrm{A}$ to $\mathbf{X}_\mathrm{C}$. This problem can be avoided by representing q factor levels by $q-1$ dummy variables, thus omitting one level.

In the broader sense, the term (multi-)collinearity refers to a high degree of linear correlation among the p predictors. This can be screened by checking a matrix of pairwise correlations among predictors, where, as a rule of thumb, correlations above 0.7 are regarded as problematic (Dormann et al. 2013). If there are two or more highly correlated predictors (e.g., $\mathbf{X}_1$, $\mathbf{X}_2$, and $\mathbf{X}_3$), they should not be used in the same regression model. This can be avoided, either by retaining only one of the intercorrelated predictors (e.g., $\mathbf{X}_3$) or with latent variable methods, where one or more eigenvectors (Box 5.4) are extracted from the set of intercorrelated predictors $\mathbf{X}_1$, $\mathbf{X}_2$, and $\mathbf{X}_3$, thus capturing their joint variation, and the eigenvectors are then used as predictors (Dormann et al. 2013).

Neighborhood-level analysis (Box 5.2; Wagner & Fortin 2013) presents alternative ways of analyzing between-site characteristics, which allows them to be combined with at-site characteristics in the same statistical model (Balkenhol et al. 2009a; James et al. 2011). Patch connectivity indices developed in metapopulation ecology can be used to transform pairwise distances of landscape predictors into node-level measures of potential functional connectivity. In this approach, instead of focusing on individual links, connectivity indices are computed that integrate across all links connecting the focal patch with any of its neighbors (Fig. 5.1B). In metapopulation ecology, patch connectivity S_i (representing the unknown number of migrants from all neighboring source patches j into focal patch i) is commonly modeled with an incidence function model (Hanski 1994; Moilanen & Nieminen 2002):

$$S_i = \sum\nolimits_j oj\ \mathbf{A}_j^b \exp(-\alpha d_{ij})$$

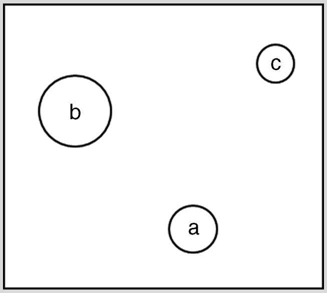
where o_j is a binary indicator whether the species (or allele) being modeled is present in source patch j, A_j refers to the source patch area or another patch characteristic,

Box 5.3 Analytical levels

The main approaches of landscape genetics studies can be classified into four analytical levels. The following illustrations are adapted from Wagner and Fortin (2013).

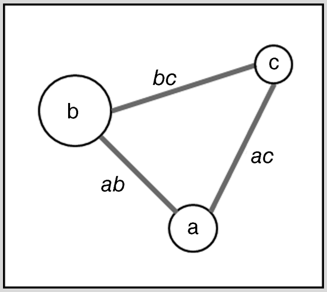
1 Node-level analysis

This relates adaptive variation to local landscape factors at sites *a*, *b*, and *c* while accounting for isolation-by-distance (Schoville et al. 2012). The node-level methods include multivariate ordination methods (e.g., RDA; Dray et al. 2012; Manel et al. 2012) and general linear models (Bolker 2008).



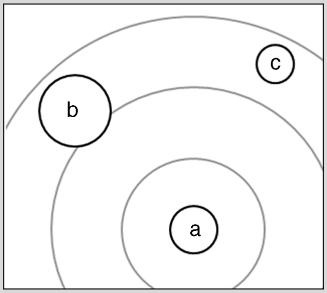
2 Link-level analysis

This relates neutral variation between sites *a*, *b*, and *c* to between-site landscape factors observed along links *ab*, *ac*, and *bc* to test hypotheses on isolation-by-distance (IBD), isolation-by-resistance (IBR), or isolation-by-barrier (IBB). The most commonly used link-level method is the Mantel test (Mantel 1967; Smouse et al. 1986; Cushman & Landguth 2010), which for multiple predictors extends to multiple regression on distance matrices (MRMs) (Smouse et al. 1986). Partial Mantel tests (Smouse et al. 1986) and causal modeling (Cushman et al. 2006) have been used to account for one process (e.g., IBD) while testing for another process (e.g., IBR). However, several studies have showed the relative lower power of the Mantel test to detect significant relationships and other inferential problems (Dutilleul et al. 2000: Legendre & Fortin 2010: Guillot & Rousset 2013).



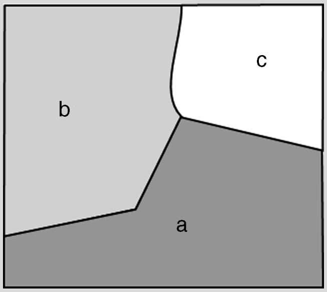
3 Neighborhood-level analysis

This relates the relative contribution of all neighboring sampled locations (here *b* and *c*) to the genetic variation observed at a given sampling location *a*. Connectivity measures (Keyghobadi et al. 2005; James et al. 2011) and gravity models (Murphy et al. 2010) can be used in neighborhood-level analyses to assess neighborhood effects on spatial genetic structure.

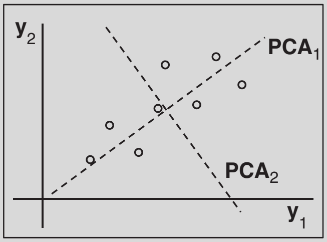


4 Boundary-level analysis

This relates genetic groups *a*, *b*, and *c* to landscape barriers. Once spatial groups are identified based on either Bayesian clustering algorithms or edge-detection techniques (see Chapter 7; Guillot et al. 2005; François & Durand 2010; Safner et al. 2011), the next step is to relate these genetic barriers to environmental and landscape barriers using spatial boundary overlap methods (Fortin et al. 1996) or POPS (Prediction of Population genetic Structure Program) (Jay 2011).

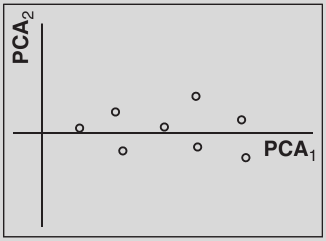


Box 5.4 Eigenanalysis



Similar to ecological species composition data, genetic allele frequency data often have a large number *m* of variables (one variable per allele) that were observed at the same *n* sampling locations (individuals or demes). Analysis of such data is difficult because the signal of common structure in the data set is obscured by noise related to random variation in each variable. Also, the variables may be correlated among themselves so they should not be analyzed independently. Eigenanalysis methods such as ***principal component analysis*** (PCA) can help to reduce data, thus separating signal from noise, and to replace the original, intercorrelated variables with a set of synthetic variables (eigenvectors) that are orthogonal and uncorrelated among themselves.

The scatterplot (top) shows a simple example of two correlated variables $\mathbf{y}_1$ and $\mathbf{y}_2$ (solid lines), such as the frequencies of two alleles, with values observed at eight sampling locations (circles). Eigenanalysis with PCA defines a set of new synthetic variables, known as PCA axes (dashed lines). The first PCA axis, called PCA_1, is defined to capture a maximum of the variation in the original variables $\mathbf{y}_1$ and $\mathbf{y}_2$. The next PCA axis, PCA_2, is chosen so that it is orthogonal to PCA1 and captures a maximum of the remaining variation in the data.



In this example with only two variables $\mathbf{y}_1$ and $\mathbf{y}_2$, two PCA axes are sufficient to fully capture the variation in the data. A scatterplot (bottom) of the values (or scores) for PCA_1 and PCA_2 recreates the same point cloud as the scatterplot (top) of the values of $\mathbf{y}_1$ and $\mathbf{y}_2$ for each sampling location, except for a rotational shift. Note that adding a third variable $\mathbf{y}_3$ would result in a third axis PCA_3 orthogonal to both PCA_1 and PCA_2, and so forth, so that *m* variables result in an *m*-dimensional PCA space.

More generally, a set of *m* PCA axes PCA_1 to PCA_m will capture all variation in a data set with *m* variables $\mathbf{y}_1$ to $\mathbf{y}_m$, but contrary to the original variables $\mathbf{y}_1$ to $\mathbf{y}_m$, the new synthetic variables PCA_1 to PCA_m are orthogonal and their pairwise correlation is exactly zero. The first axis, PCA_1, will have the highest variance as it contains the largest fraction of the variance in the original data, and each further PCA axis will have a lower variance than the previous ones. In fact, every eigenvector (i.e., PCA axis) has an associated eigenvalue λ that is proportional to the variance in the original data that the eigenvector represents. Note that, if the original variables have been centered so that they each have a mean of zero, the last PCA axis, PCA_m, will have zero variance and an eigenvalue of $\lambda_m = 0$. Data reduction with PCA is based on the idea that the first few PCA axes contain the multivariate signal, i.e., the variance shared among the variables in the data set, whereas the remaining PCA axes contain largely noise.

PCA is the basic and most common method of eigenanalysis, also referred to as ordination methods. While PCA extracts eigenvectors of a variance–covariance matrix or a matrix of Euclidean distances between observations, principal coordinate analysis (PCoA) and non-metric multidimensional scaling (NMDS) will extract eigenvectors from any measure of resemblance and thus can be used with various measures of genetic distance (Legendre & Legendre 2012). Constrained ordination methods (also known as direct ordination methods) such as redundancy analysis (RDA) extract eigenvectors separately for the fitted values $\hat{\mathbf{Y}}$ and for the residuals $\mathbf{U}$ of a regression-type model, where the variation in a multivariate response $\mathbf{Y}$ is explained by a set of predictors $\mathbf{X}$ (Legendre & Legendre 2012). A special case is distance-based redundancy analysis (dbRDA) (Legendre & Legendre 2012), where, in a first step, a genetic distance matrix $\mathbf{D}_\mathbf{Y}$ is subjected to PCoA to extract a matrix of eigenvectors, which in a second step serves as the response matrix $\mathbf{Y}$ in redundancy analysis RDA (Fig. 5.1C).

such as local population size or habitat quality, d_{ij} is the distance between patches i and j, and b and α (with $\alpha > 0$) are scaling constants related to emigration and dispersal distance. Other connectivity indices can include some particular attributes of the patches (Moilanen & Nieminen 2002; Saura & Rubio 2010).

Such connectivity indices quantify how connected (as opposed to isolated) each patch is, based on an explicit model of connectivity. Connectivity indices have been used to model univariate genetic data such as genetic diversity (Keyghobadi et al. 2005; Rico et al. 2014) and multivariate allele frequencies (James et al. 2011). Gravity models (Murphy et al. 2010) present an alternative way of combining at-site and between-site variables, and are discussed in Chapter 10.

5.2.2 Type of genetic data

Once we have identified the best landscape data and analytical level for our analyses, we need to consider the different types of genetic data (Fig. 5.1C). The original data consist of a set of loci genotyped for all sampled individuals. In most cases, however, analysis will be based either on a table of allele frequencies or a matrix of genetic distances (see Chapter 3) derived from the genotype data.

For multivariate analysis, genotype data need to be coded in a table of allele frequencies for n sampling units (rows) observed at m alleles (columns), thus using one variable per allele (Smouse & Peakall 1999). Technically this will introduce multicollinearity (in the strict sense, see above) in **Y**, which some analysis methods can handle, while for others, one allele per locus may need to be dropped.

If genotype data are converted to a matrix of pairwise genetic distances (i.e., dissimilarities), analysis may be based on all links or on a subset of meaningful links only. Such a subset may be based on different types of graph models (Chapter 10), which in essence are algorithms to define which sampling units are neighbors (Dale & Fortin 2010). Alternatively, Dyer et al. (2010) proposed testing each link to evaluate whether it shows statistically significant higher similarity (and thus lower genetic distance) than expected, given all other indirect paths connecting the two sampling units (conditional genetic distance). For the remainder of this chapter, however, we will assume that all links are retained.

A genetic distance matrix can be converted into a node-based framework by subjecting it to principal coordinate analysis (PcoA) (equivalent to non-linear multidimensional scaling; Legendre & Legendre 2012), which is similar to principal component analysis (PCA) (see Glossary and Box 5.4) but allows for non-Euclidean distance measures and thus can accommodate any measure of genetic distance. PCoA will return a table of n rows and up to m columns (eigenvectors of the genetic distance matrix), which in essence are a set of perfectly uncorrelated and orthogonal synthetic variables. This table of synthetic variables contains all the variation in the genetic data, which means that it can be used as response matrix **Y** in subsequent analysis (see distance-based redundancy analysis, dbRDA, in Box 5.4).

5.2.3 Type of statistical model

When modeling landscape effects on genetic variation, we typically assume a *directed relationship* (regression-type model), where the genetic data (response **Y**) are constrained (i.e., determined) by the landscape predictors **X**, with potentially high stochasticity, which we will represent by an error **U** (i.e., residuals; Fig. 5.1D). It is important to note that we are not talking about establishing a causal mechanism (which can only be done in a properly controlled experiment, which will rarely apply to landscape genetic data), but we have a clear hypothesis that **X** affects **Y**. This is opposed to an *undirected relationship* (correlation-type model), where we would merely assume an association between genetic and landscape data without presuming that one depends on the other.

Ordinary least squares (OLS) regression relies on a set of assumptions. First, the sample must be representative of the population. If this is not the case, statistical tests are not applicable due to unknown sampling bias; hence the sampling design is of key importance (Chapter 4). Second, the predictors **X** should be measured without error and there must not be any multicollinearity in the strict sense (see above) among predictors. Third, each of the n residuals in $\mathbf{u}_i = \{u_{i1,} u_{i2,\ldots,} u_{in}\}$ of a single response variable $\mathbf{y}_i$ is itself an outcome of a random variable, and these random variables must be independent of each other and follow the same distribution (often a normal distribution is assumed) with a mean of zero and constant variance (homoscedasticity). A range of ***residual analysis*** tools facilitates checking for violations of regression assumptions for a single response variable **y**. A normal probability plot shows whether the residuals follow a normal distribution; a plot of residuals against the predicted values may reveal problems with non-constant variance (heteroscedasticity) or a non-linear response, and other plots may help identify ***influential points*** that may have a strong influence on parameter estimates and model fit. Transformations such as log(**y**) may be used to stabilize the error variance. For binary response data, such as dominant alleles coded as $1 =$ present and $0 =$ absent, we may expect different error distributions than the normal distribution, which can be accommodated by extensions of the basic OLS model to the generalized linear model (GLM).

If the residuals are not independent of each other, with nearby residuals tending to be more similar than distant ones, this indicates positive spatial autocorrelation, such as generated by isolation-by-distance. A correlogram or a test of Moran's I can be used to check for spatial autocorrelation in the residuals (Box 5.1). If the autocorrelation structure is constant across the study area (i.e., stationary), it may be modeled with spatial regression (see below). Genetic data typically consist of many response variables (e.g., one for each allele of each locus), which calls for ***multivariate regression***. Note that *multivariate regression* means multiple response variables (**Y**), whereas *multiple regression* means multiple predictor variables (**X**). Multivariate regression involves fitting a regression model to each response variable, i.e., each allele. The regression assumptions apply for each allele; however, they are rarely checked individually and there is a lack of multivariate residual analysis tools. To avoid multiplying error rates and to account for correlation among response variables, multivariate significance tests should be used instead of testing each allele individually. Similarly, regression results are not necessarily interpreted individually but across all alleles, which means that the results are a weighted average across alleles. Model fit can be summarized by the canonical R^2, the proportion of variance of **Y** (alleles) explained by a linear model of the variables in **X** (landscape predictors), which is a weighted mean of the coefficient of determination R^2 of each allele weighted by its variance.

Constrained ordination methods such as redundancy analysis (RDA) facilitate the interpretation of multivariate regression analysis. RDA combines multivariate linear regression ($\mathbf{Y} = \mathbf{XB} + \mathbf{U}$, where **U** is a table of residuals) with principal component analysis (PCA) (see Box 5.3) of the table of fitted values $\hat{\mathbf{Y}} = \mathbf{XB}$ (Legendre & Legendre 2012) to quantify the variation in **Y** that is related to **X**. Permutation tests can be performed overall, with the null hypothesis that **X** does not explain more variation in **Y** than expected by chance, or for individual canonical axes (i.e., eigenvectors of a PCA of the fitted values). Note that significance tests are not performed on individual predictors, which makes RDA relatively robust towards colinearity in predictors. A PCA of the residuals **U** may reveal further genetic structure not explained by **X**, e.g. by mapping the scores of the first PCA axis in space (residual analysis).

Link-level analysis has been used to assess whether spatial genetic structure can be related to isolation-by-distance (Wright 1943; Epperson 2003), isolation-by-resistance (McRae 2006; Spear et al. 2010), or isolation-by-barrier (IBB) (Vignieri 2005). As these questions are similar to those addressed in evolutionary biology and spatial genetics, landscape genetics uses the same suite of link-based methods as in spatial genetics (Epperson 2003; Sokal & Wartenberg 1983; Guillot et al. 2009): Mantel test (Mantel 1967), partial Mantel test (Smouse et al. 1986), and multiple regression based on distance matrices (MRM) (Legendre & Legendre 2012; Smouse et al. 1986; Lichstein 2007).

Mantel tests have been widely used to assess the association between two distance matrices $\mathbf{D_X}$ and $\mathbf{D_Y}$. The lower (or upper) triangle of each distance matrix of size $n \times n$ is extracted as a vector of length $N = n(n-1)/2$, so that each link is included only once. In the following, we will use "vector $\mathbf{D_Y}$" to refer to a vector of unique pairwise distances and "matrix $\mathbf{D_Y}$" to refer to the corresponding distance matrix. As an example, a sample of size $n = 100$ would result in a distance vector $\mathbf{D_Y}$ of length $N = 4950$, thus heavily inflating sample size. The Mantel statistic quantifies the linear (r_M) or rank correlation (rho_M) between the distance vectors $\mathbf{D_Y}$ and $\mathbf{D_X}$. However, the $N = 4950$ pair-wise distance values in a distance vector are not independent of each other. In our example, each of the $n = 100$ observed values is compared to all $n - 1 = 99$ other values. As a consequence, the statistical significance of r_M or rho_M (Mantel test) must be assessed with a permutation test where rows and columns of the $n \times n$ matrix $\mathbf{D_Y}$ are permuted together before extracting distance vector $\mathbf{D_Y}$. Note that it is sufficient to permute either matrix $\mathbf{D_Y}$ or matrix $\mathbf{D_X}$.

In a partial Mantel test, the Mantel statistic is calculated after partialling out the linear effect of a third distance vector $\mathbf{D_{X2}}$, thus assessing the residual correlation between distance vectors $\mathbf{D_Y}$ and $\mathbf{D_{X1}}$ after accounting for $\mathbf{D_{X2}}$. Causal modeling with distance matrices uses all Mantel and partial Mantel statistics calculated between pairs of distance vectors to discriminate between competing hypotheses (Legendre & Trousselier 1988; Cushman et al. 2006).

The main difference between a Mantel test of two distance vectors $\mathbf{D_Y}$ and $\mathbf{D_X}$ and a simple linear regression of vectors $\mathbf{D_Y}$ on $\mathbf{D_X}$ is the postulation of a directed relationship (see above), which is appropriate if we want to explain genetic structure by landscape predictors. In addition, multiple regression with distance matrices (MRM) allows the simultaneous consideration of multiple explanatory distance vectors $\mathbf{D_X}$ as in a multiple regression. However, as in the Mantel test, the residuals in MRM are not independent of each other, which violates an important assumption of linear regression. To test the slope coefficients $\mathbf{b}$, $\mathbf{D_Y}$ needs to be permuted as described above. If there are no further problems, such as spatial autocorrelation in the residuals, and the goal is significance testing of the regression model and its parameters, this permutation test should be valid.

However, the interpretation of the regression model is rather different if we use vector $\mathbf{D_Y}$ instead of $\mathbf{Y}$, and tests based on distance vectors $\mathbf{D_Y}$ and $\mathbf{D_X}$ are notoriously less powerful than those based on $\mathbf{Y}$ and $\mathbf{X}$ (Legendre & Fortin 2010). This is true both for correlation-type Mantel tests and regression-type MRM. Despite the fact that simple and partial Mantel tests and their extension to MRM are the most commonly applied statistics in landscape genetics (Storfer et al. 2010), multiple studies have identified various shortcomings of Mantel-based approaches (e.g., Balkenhol et al. 2009a; Cushman & Landguth 2010; Legendre & Fortin 2010; Graves et al. 2013; Guillot & Rousset 2013). Throughout this chapter, we highlight several of these issues and discuss alternative statistical approaches.

5.2.4 Model selection

Model selection (Johnson & Omland 2004) refers to the problem of selecting, from a set of candidate models, either a single best model or models weighed according to how well they fit the data and perform weighted averaging. The problem is that (i) more than one model may be statistically significant, (ii) the parameter estimate of the regression slope (and its statistical significance) for a given landscape predictor may depend on which other predictors are included in the model (unless predictors are uncorrelated), and (iii) everything else being equal, we expect a model with more predictors to fit the data better than a model with fewer predictors. In fact, any regression model of n observations of response $\mathbf{Y}$ with n predictors will explain 100% of the variation in $\mathbf{Y}$, even if the predictors were sampled at random. Several methods have been proposed to penalize models for the number of predictors, including adjusted R^2, Akaike information criterion (AIC, or AICc with small-sample correction), and Bayes information criterion (BIC), where the best model would have the highest adjusted R^2 or the lowest AIC, AICc, or BIC (Johnson & Omland 2004).

In MRM, the errors in $\mathbf{D}_\mathrm{U}$ are not independent. In statistical hypothesis testing, this problem can be addressed with an appropriate permutation test (see above), but the pairwise nature of the data complicates residual analysis and model selection. Indeed, AIC and similar indices as listed above should not be applied to MRM models (Van Strien et al. 2012), as model rankings are highly unreliable unless sample size is corrected and, when correcting for sample size, their power to detect meaningful predictors is very low (Franckoviak et al. submitted). This is a pressing problem and more research is needed to show how the correlation structure among the errors $\mathbf{D}_\mathrm{U}$ may be explicitly modeled to allow the use of AIC or similar measures in model selection. Clarke et al. (2002) proposed maximum-***likelihood*** population effects (MLPEs) to explicitly model the dependence among pairwise observations. As MLPE models are fitted by residual maximum-likelihood (REML) methods, AIC, AICc, and BIC are not applicable, but a marginal R^2 statistic may be used to identify the most parsimonious model (Orelien & Edwards 2008; Van Strien et al. 2012).

5.2.5 How to put space into the multivariate regression model

Thus far, we have assumed that residuals of a regression model do not show any spatial autocorrelation. However, as spatial autocorrelation in the residuals violates the assumption of independent errors, we need to test for spatial autocorrelation in the residuals and remove it before interpreting a regression model. The methods for testing and accounting for spatial autocorrelation are similar and will be discussed jointly in this section. We will only present a subset of methods and refer to the ongoing debate about valid regression-type analysis of spatial ecological data (e.g., Beale et al. 2010; Kühn & Dormann 2012). Note that while generalized linear mixed models (GLMMs) (Zuur et al. 2009; Gałecki & Burzykowski 2013) and generalized least squares (GLS) regression as a special case of a GLMM are gaining importance in ecological data analysis, their application to multivariate genetic data has not been fully developed yet and their coverage is beyond the scope of this chapter.

5.2.6 Multivariate linear regression with OLS

The basic regression model used in ordinary least-squares (OLS) regression is spatially implicit and the observations are assumed to be spatially independent. Thus, although observations were taken at specific sampling locations, this spatial information is not considered in the analysis. Consequently, each observation receives the same weight in the estimation of regression coefficients **B**, which are assumed to be constant across the study area. The predicted values $\hat{\mathbf{Y}}$ are modeled as a linear combination **XB** of predictors in **X** (Fig. 5.2A).

5.2.7 Spatial weights matrix W

Spatial analysis is based on the expectation that nearby observations are on average more similar than distant ones. The first step in building such spatial relationships into the regression model is to define a spatial weights matrix **W** (Fig. 5.2B). Based on the spatial coordinates, we first define for each sampling unit *i* whether other sampling units *j* are its neighbors. Then we assign weights w_{ij} to each pair, either as binary weights with $w_{ij} = 1$ for neighbors and $w_{ij} = 0$ for non-neighbors, or as a decreasing function of distance, often up to a threshold distance beyond which all weights are $w_{ij} = 0$. Finally, we may want to adjust for the number of neighbors *j* of each observation *i* so that the weights w_{ij} of all neighbors of *i* sum to one.

5.2.8 Spatial regression

Different types of spatial regression (Fig. 5.2C) associate the spatial weights matrix **W** with different terms of the regression model, which implies different assumptions about the origin of the spatial autocorrelation. (i) Spatial lag models (e.g., conditional autoregressive models, CAR) add a weighted mean of the response **Y** at neighboring locations as a predictor to the model. The assumption is that the presence or abundance of an allele at location *i* depends on its presence or abundance in the neighborhood, which is a reasonable assumption under IBD. (ii) Spatial error models (e.g., simultaneous autoregressive models, SAR) add a weighted mean of the error **U** at neighboring locations as a predictor to the model. This is indicated if we attribute spatial autocorrelation to an unmeasured nuisance factor. Spatial regression methods are commonly applied to a univariate response **y**, although multivariate CAR models have been proposed (Gelfand & Vounatsou 2003). More research is needed to show their applicability to model IBD in landscape genetic data.

If the process is ***non-stationary*** so that the spatial autocorrelation structure varies across the study area (Box 5.1), one should consider statistics that can measure the degree of spatial structure at local scales, such as local indicators of spatial association (LISA; Anselin 1995; Sokal et al. 1998). Geographically weighted regression (GWR) (Fotheringham 2002) can detect and address situations where the relationship between **X** and **Y** is non-stationary, so that the slope coefficients **B** vary across the study area. Such statistical analyses have been used in ecology (e.g., Fortin & Melles 2009; Windle et al. 2012) and evolution (e.g., Ochoa-Ochoa et al. 2014) but they are not yet commonly used in landscape genetics. A recent example used GWR to demonstrate that different models of genetic connectivity for Rocky Mountain tailed frogs (*Ascaphus montanus*) were supported in privately and publicly managed forests despite close spatial proximity of the two types of land ownerships (Spear & Storfer 2010).

5.2.9 Spatial eigenvectors

Spatial eigenvector methods (Griffith 2000; Borcard & Legendre 2002; Griffith & Peres-Neto 2006; Dray et al. 2006, 2012; Jombart et al. 2008, 2009; Dray 2011), such as ***Moran's eigenvector map***s (MEM) (Dray et al. 2006; Dray 2011), are based on eigenanalysis (Box 5.4) of the spatial weights matrix **W** (Fig. 5.2D). In the case of MEM, **W** is made symmetric and column and row means are removed before eigenanalysis. If matrix $\mathbf{W}$ is of full rank one of the n eigenvectors will have zero variance (and an eigenvalue close to zero) and will be discarded. Each of the remaining $n-1$ spatial eigenvectors describes a periodic spatial pattern. If sampling locations form a regular transect, the spatial eigenvectors will represent sine-type patterns (similar to a Fourier decomposition), whereas two-dimensional and irregular sampling designs will result in more complex patterns resembling two-dimensional sine waves (Box 5.3). The $(n-1)$ spatial eigenvectors, which form the columns of matrix $\mathbf{V}$, are orthogonal and uncorrelated among themselves. Just as we can describe a set of 2 points by a single line with two parameters, the variation in n observations can be fully described by any set of n variables or by $n-1$ variables if the mean has been removed. This means that the variation in $\mathbf{Y}$ can be completely modeled by a linear combination of the variables in $\mathbf{V}$ (i.e., a regression of $\mathbf{Y}$ on $\mathbf{V}$ will have $\mathbf{U}=0$ and $R^2=1$). While this full model is not of interest in itself, the matrix $\mathbf{R}_{\mathrm{YV}}$ of correlations between the columns in $\mathbf{Y}$ (alleles) and the columns in $\mathbf{V}$ (spatial eigenvectors) provides a spatial decomposition of the genetic variation at all spatial scales that can be used to quantify multivariate spatial autocorrelation with Moran's I or to model significant spatial genetic variation as spatial predictors that can be included in the regression model (spatial filtering, see Section 5.2.11).

5.2.10 Multivariate Moran's I

When using MEM, the spatial eigenvectors in $\mathbf{V}$ are sorted by the spatial scale they represent, so that the first spatial eigenvector ($k=1$) represents a single sine wave spanning the maximum extent of the study area. Each subsequent spatial eigenvector represents a smaller-scale spatial pattern and the spatial scale of each spatial eigenvector can be quantified by Moran's I, a measure of spatial autocorrelation (Box 5.1). Conveniently, the eigenvalue λ_k associated with each spatial eigenvector k is proportional to its Moran's I_{k}, up to a constant a, which corresponds to the inverse of the average sum of weights per observation (Fig. 5.2E). Moreover, a weighted mean of the eigenvalues λ, weighted by the squared correlations $\mathbf{R}^2{}_{\mathrm{yV}}$ of $\mathbf{V}$ with a response variable $\mathbf{y}$ (i.e., frequency of a specific allele), results in Moran's I_{y} of allele $\mathbf{y}$ (Dray 2011). Multivariate Moran's I_{Y} can then be found as the mean of Moran's I_{y} of all alleles, weighted by column weights $\mathbf{c}$ (Fig. 5.2E; Wagner 2013).

When averaging across alleles and loci, the coding of allele frequencies will affect their relative weight (Smouse and Peakall 1999). As each allele is treated as a variable, it makes sense to weigh the contribution of each allele to Moran's I of the locus by the inverse of the number of alleles per locus, so that the total weight of each locus equals one, as recommended by Jombart et al. (2008). Adjusting for differences in allele frequencies within a locus seems less important (Smouse and Peakall 1999). To average across loci, we further divide weights by the number of loci so that the column weights $\mathbf{c}$ sum to one.

While MEM derives the matrix of eigenvectors $\mathbf{V}$ from the sampling design alone, spatial principal component analysis (sPCA) (Jombart et al. 2008) derives $\mathbf{V}$ from a combination of the sampling design and the observed data. This has the advantage that spatial eigenvectors are sorted by their contribution to multivariate Moran's I of the data. However, sPCA does not provide an additive decomposition of I_{Y}. Furthermore, the matrix $\mathbf{V}$ of spatial eigenvectors derived for a response matrix $\mathbf{Y}$ of allele frequencies and for a predictor matrix $\mathbf{X}$ may differ, which will require more prior knowledge about the key spatial scales of interest in the analysis of the relationship between genetic variation and landscape predictors. sPCA is a new explicitly spatial ordination technique that deals with spatial structures multivariate data and its merits still need to be evaluated.

5.2.11 Spatial filtering

Another approach to include space in multivariate models is spatial filtering, which uses spatial eigenvectors to control for significant spatial autocorrelation at any spatial scale. Matrix $\mathbf{V}$ as defined above can be used to decompose the spatial variation in the response $\mathbf{Y}$ in the matrix of fitted values $\hat{\mathbf{Y}}$ or in the matrix of residuals $\mathbf{U}$. We can thus ask what is the overall spatial structure in the genetic data, which spatial patterns are explained by landscape predictors, and which remain unexplained? Significant spatial eigenvectors for fitted values indicate shared patterns, whereas significant spatial eigenvectors for residuals indicate unexplained patterns that may be related either to dispersal processes or to unmeasured landscape factors.

For testing the statistical significance of spatial eigenvectors, the method by Jombart et al. (2009) should be preferred over forward selection (Blanchet et al. 2008), as the former takes into account the relationship between spatial eigenvectors (Wagner 2013). The test statistic is defined as the maximum (multivariate) variance explained by a single spatial eigenvector. For each spatial eigenvector, its observed multivariate R^2 is compared to the distribution of the test statistic obtained from a permutation test, where observations are permuted randomly (Jombart et al. 2009; Wagner 2013).

In a regression framework (Fig. 5.2F), significant spatial eigenvectors for **Y** or for **U** are added as predictors to the regression model to account for significant spatial structure in the genetic data when testing the association between **X** and **Y** (spatial filtering; Legendre & Legendre 2012, Griffith & Peres-Neto 2006; Peres-Neto & Legendre 2010). Spatial filtering with MEM has been used to account for unmeasured environmental variation when identifying loci that are potentially under selection (Manel et al. 2010). However, the statistical validity of such an approach remains to be thoroughly tested, and the existing evidence from non-genetic simulation studies suggests that spatial filtering with MEM and related methods may lead to biased parameter estimates and inflated type I error rates (Dormann et al. 2007; Bini et al. 2009; Beale et al. 2010).

5.3 HOW TO MODEL ISOLATION-BY-DISTANCE

Isolation-by-distance plays a key role in landscape genetics. First, analyzing landscape effects on genetic variation is only warranted if there is spatial structure in the genetic data (i.e., if the population is not panmictic). Second, the effects of matrix resistance (IBR) should not be tested against a null model of panmixis but against a null model of IBD, thus testing for effects of landscape features on rates of gene flow beyond the effect of IBD. Third, if gene flow is spatially restricted (by IBD or IBR), this creates spatial autocorrelation in the genetic data that needs to be accounted for when testing the association between **X** and **Y** (e.g., in the identification of outlier loci that may indicate selection). Figure 5.3 summarizes approaches for testing and accounting for IBD in the frameworks of spatial regression with conditional autoregressive modeling CAR, spatial filtering with MEM, and regression of distance matrices (MRM).

5.3.1 IBD and spatial regression with CAR

In this section, the symbol **Z** will be used to refer to a "space only" model (IBD) and **X** to other landscape predictors. Isolation-by-distance can be modeled as a stationary ***isotropic*** spatial process, thus assuming that the interaction between sampling locations depends on distance alone (Fig. 5.3A). Spatial regression with a conditional autoregressive (CAR) model is appropriate for modeling gene flow as it explicitly models the interaction between allele frequencies of neighboring sampling locations (i.e., it assumes that allele frequency at location *i* depends on the frequency of the same allele at neighboring locations *j*; Fig. 5.3). In essence, $\mathbf{WYB_W}$ calculates a weighted mean of allele frequencies at neighboring locations, where **W** defines which locations are neighbors and what is their relative weight (i.e., the spatial covariance structure). $\mathbf{B_W}$ defines the contribution of the weighted mean $\mathbf{Z} = \mathbf{WY}$ to the fitted values $\hat{\mathbf{Y}}$(i.e., the strength of positive spatial autocorrelation defined by **W**). The spatial weights matrix **W** remains constant, whereas $\mathbf{B_W}$ contains a separate autoregression coefficient for each variable in **Y**.

A simple CAR model without additional terms can be used to test for IBD as a spatial process defined by **W** against a null hypothesis of panmixis. When testing for landscape effects (representing IBR or selection), partial regression (conditioning by the CAR term $\mathbf{WYB_W}$) can be used to incorporate IBD into the null model. Thus, first an IBD model defined by $\mathbf{Z} = \mathbf{WY}$ is fitted separately to **Y** and to **X**, and the regression of **Y** on **X** is carried out on the residuals of both **Y** and **X**. In Fig. 5.3, the conditioning by **Z** is indicated by subscript $\mathbf{X|Z}$.

Residual analysis (see above) should be performed before interpreting the model. If the spatial regression model is correctly specified and the spatial process (gene flow) is stationary and isotropic, then the residuals **U** should not show any further spatial structure. In multivariate regression, this can be tested with multivariate Moran's I (see above). If the test indicates significant spatial autocorrelation in the residuals, it may be useful to perform a PCA of **U**, which is implicitly done in RDA, and plot the scores of the first few axes (unconstrained RDA axes) in space for visual interpretation of unexplained spatial patterns.

5.3.2 IBD and spatial filtering with MEM

If the spatial scale of gene flow is unknown, or if we expect it may not be stationary and isotropic (e.g., IBR leading to variation in rates of gene flow beyond mere distance effects), the assumptions of spatial regression with a CAR model may be too restrictive. Hence we may want to define a flexible model of gene flow using MEM (Fig. 5.3B). In essence, spatial filtering with MEM uses a subset $\mathbf{Z} = \mathbf{V}_K$ of spatial eigenvectors in $\mathbf{V}$ to model significant spatial structure of any shape and at any spatial scale in the response $\mathbf{Y}$. $\mathbf{Z}$ may thus serve as a proxy for unmeasured landscape factors or unspecified spatial biotic processes.

Using MEM to test for a significant spatial structure in $\mathbf{Y}$ at any scale (i.e., against a null hypothesis that none of the spatial eigenvectors are significant – panmixis) is often not informative. Rejecting this null hypothesis does not inform us whether the deviation is due to IBD or IBR. Spatial filtering with MEM may be most applicable when we want to test for evidence of selection (e.g., identifying outlier loci) while accounting for spatial autocorrelation due to gene flow, whether due to IBD or IBR, without using a predefined model.

Once $\mathbf{Z} = \mathbf{V}_K$ has been defined, it may be useful to perform partial regression conditioned by $\mathbf{Z}$ instead of treating both $\mathbf{X}$ and $\mathbf{Z}$ equally as predictors. This makes sense conceptually (space itself is not a meaningful predictor) and avoids problems of interpretation, as we can then distinguish between conditioned variance (variance explained by $\mathbf{Z}$), constrained variance (variance explained by $\mathbf{X}$ after conditioning for $\mathbf{Z}$), and residual variance (Wagner 2013).

As a caveat, Beale et al. (2010) found that spatial filtering with MEM resulted in biased parameter estimates and inflated type I error rates. While spatial eigenvectors are promising as a flexible tool for modeling spatial structure in landscape genetic data, more research is needed on how they can be applied in hypothesis testing, accounting for the special nature of spatial eigenvectors (Gilbert & Bennett 2010; Wagner 2013).

5.3.3 IBD and multiple regression of distance matrices (MRM)

In MRM, a vector $\mathbf{D_Z}$ of pairwise geographic distances is commonly used to model IBD (Fig. 5.3C). The null model of panmixis is rejected if a simple regression of vectors $\mathbf{D_Y}$ on $\mathbf{D_Z}$ is significant, based on a permutation test where the rows and columns of matrix $\mathbf{D_Y}$ are permuted simultaneously (see above). This is equivalent to a Mantel test between vectors $\mathbf{D_Y}$ and $\mathbf{D_Z}$.

Partial regression of distance matrices can be used to test landscape effects against a null model of IBD by testing whether landscape effects defined in vector $\mathbf{D_X}$ (e.g., IBR) are statistically significant after accounting for vector $\mathbf{D_Z}$. If $\mathbf{D_X}$ is a single vector (which may reflect a complex hypothesis about the resistance values of multiple land cover types), this is equivalent to a partial Mantel test between vectors $\mathbf{D_Y}$ and $\mathbf{D_X}$, conditioning for vector $\mathbf{D_Z}$. In essence, the residuals of a regression of vector $\mathbf{D_Y}$ on $\mathbf{D_Z}$ are regressed on the residuals of a regression of vector $\mathbf{D_X}$ on $\mathbf{D_Z}$. If there are multiple vectors $\mathbf{D_X}$, each representing a different landscape element (e.g., length of a single cover type along the transect between two sampling locations), they may be added together in the same model. In that case, after conditioning for vector $\mathbf{D_Z}$, the regression coefficients for the $\mathbf{D_X}$ vectors are tested as if each was added to the model last (i.e., accounting for all other predictors in the model). In contrast, the effect of vector $\mathbf{D_Z}$ is tested without accounting for vectors $\mathbf{D_X}$ because IBD as defined in vector $\mathbf{D_Z}$ is part of the null model for testing vectors $\mathbf{D_X}$. In many landscape genetic studies, IBD will be an appropriate null model, e.g. for testing IBR, whereas IBR would rarely be an appropriate null model for testing IBD. A notable exception may be the simultaneous testing for a barrier effect (IBB) and IBD, as a complete barrier to gene flow may occur with or without IBD on either side. Note that the causal modeling framework (Cushman et al. 2006) involves testing all possible partial Mantel correlations to rule out alternative hypotheses.

There are several issues with MRM, relating mainly to the shape of the relationship, the validity and statistical power of significance tests, and model selection. While regression assumes a linear relationship and constant variance around the regression line (homoscedasticity), these assumptions may not hold for genetic distances, so that a plot of vector $\mathbf{D_Y}$ against vector $\mathbf{D_Z}$ may show a non-linear relationship or an increase of variance with distance. For instance, Rousset (1997) proposed that $F_{ST}/(1 - F_{ST})$ is approximately linearly related to geographic distance in a one-dimensional stepping stone model, but linearly related to the natural logarithm of geographic distance in a two-dimensional stepping stone model. Under equilibrium conditions in a two-dimensional stepping stone model, Hutchison and Templeton (1999) expected a monotonic increase of pairwise F_{ST} values and an increase of their spread with geographic distance (i.e., non-constant variance), due to a shift in the relative importance of the homogenizing effect of gene flow at short distances and the divergent effect of drift at large distances. Lack of regional equilibrium may result in a scatter plot where pairwise F_{ST} increases at short distances and levels off at larger distances, thus introducing non-linearity (Hutchison & Templeton 1999). In a simulation study, Graves et al. (2013) found that after 300 non-overlapping generations, the relationship between pairwise genetic distances D_{ps} and the matrix of cost distances used to simulate gene flow (i.e., the known truth) was asymptotic rather than linear and could not be linearized by common transformations. Furthermore, measures of genetic distance differ in their sensitivity to population genetic processes (e.g., Whitlock 2011; Raeymaekers et al. 2012; see Chapter 3) and population size or divergence time may be more important for explaining population genetic structure than gene flow (Marko & Hart 2011).

Compared to regression of node-based data, tests based on distance matrices (Mantel test, partial Mantel test, and MRM) have considerably lower statistical power (Legendre & Fortin 2010) and thus higher type II error rates, so that existing landscape effects are less likely to be detected in a regression based on distance matrices. On the other hand, if there is positive spatial autocorrelation in both the response and predictor matrices (e.g., due to IBD), type I error rates may be considerably inflated (Guillot & Rousset 2013), so that spurious effects are more likely to become statistically significant. Moreover, recent findings suggest that accounting for IBD by a vector $\mathbf{D_Z}$ of geographic distances does not sufficiently remove spatial autocorrelation in partial analysis (Guillot & Rousset 2013). Goldberg and Waits (2010) proposed a method to identify and remove observations that are non-independent due to spatial autocorrelation, though this does not account for the issue of inflated sample size.

The pairwise nature of distance data impedes the checking of assumptions and conditions with residual analysis. As noted above, the residuals **U** are not independent of each other and information-theoretic indices commonly used for model selection (AIC, AICc, or BIC) are not applicable to distance matrices (Van Strien et al. 2012; Franckoviak et al. submitted). Finally, the identification of influential points is hampered by the problem that one unusual observation will affect $n-1$ pairwise distance values, so that the influential observations may best be identified with leave-one-out jackknife methods.

5.4 FUTURE DIRECTIONS

For landscape genetics of adaptive variation (e.g., detection of outlier loci; see Chapter 9), multivariate regression provides a natural framework for modeling the individual response of alleles to at-site conditions, such as soil, vegetation, or bioclimatic variables related to selection. However, valid inference needs to account for spatial autocorrelation induced by gene flow. Spatial filtering with MEM is useful for partialling out spatial structure without specifying a particular process of IBD or IBR, as long as the majority of alleles reflects the same process of gene flow. However, the statistical validity of spatial filtering needs to be thoroughly tested and alternatives may need to be developed. Future directions may involve studying selection and gene flow at the same time with multivariate spatial regression (e.g., the CAR model). In contrast to MEM, this involves explicitly modeling the spatial process of IBD or IBR at the appropriate spatial scale(s). The spatial process may be averaged over a majority of alleles with similar parameters (i.e., exclude potential outlier loci that may show a different spatial autocorrelation structure) to obtain a quantification of gene flow, which can then be used to account for gene flow when testing the response to at-site variables.

The study of gene flow focuses on between-site characteristics, which remains a challenge for model selection and valid statistical inference. Solutions are emerging along three avenues:

i Remain in link-level analysis and explicitly model the error structure in MRM (Clarke et al. 2002; Van Strien et al. 2012).

ii Adopt a neighborhood-level approach, where between-site characteristics are integrated in a connectivity measure such as an incidence function model to obtain a connectivity value for each sampling location (e.g., Keyghobadi et al. 2005; James et al. 2011; Rico et al. 2014). These connectivity values can then be used as predictors of genetic variation, where the response is a measure of genetic diversity or differentiation, a matrix of allele frequencies, or a set of PCoA scores (dbRDA).

iii In node-level analysis of allele frequencies **Y**, use one or multiple distance matrices $\mathbf{D}_{\mathrm{X}}$ of landscape predictors to model the covariance structure of the errors **U** (but see Guillot et al. 2014 on valid covariance models). While Bradburd et al. (2013) present an implementation for binary SNP data in a Bayesian framework, the approach could be extended to multinomial logistic regression to accommodate codominant markers such as microsatellites and could potentially be implemented in the framework of generalized linear mixed models (GLMM) (Zuur et al. 2009; Gałecki & Burzykowski 2013).

ACKNOWLEDGMENTS

This chapter was supported by NSERC Discovery Grants to HHW and MJF.

REFERENCES

Anselin, L. (1995) Local indicators of spatial association – LISA. *Geographical Analysis* **27**, 93–115.

Balkenhol, N., Waits, L.P., & Dezzani, R.J. (2009a) Statistical approaches in landscape genetics: an evaluation of methods for linking landscape and genetic data. *Ecography* **32**, 818–30.

Balkenhol, N., Gugerli, F., Cushman, S.A., Waits, L.P., Coulon, A., Arntzen, J.W., Holderegger, R., & Wagner, H.H. (2009b) 'Identifying future research needs in landscape genetics: where to from here?', *Landscape Ecology* **24**, 455–63.

Beale, C. M., Lennon, J. J., Yearsley, J. M., Brewer, M. J. and Elston, D. A. (2010), Regression analysis of spatial data. *Ecology Letters* **13**, 246–64.

Bini, L.M., Diniz-Filho, J.A.F., Rangel, T.F.L.V.B., Akre, T.S.B., Albaladejo, R.G., Albuquerque, F.S., Aparicio, A., Araujo, M.B., Baselga, A., Beck, J., Bellocq, M.I., Bohning-Gaese, K., Borges, P.A.V., Castro-Parga, I., Chey, V.K., Chown, S.L., de Marco, P., Dobkin, D.S., Ferrer-Castan, D., Field, R., Filloy, J., Fleishman, E., Gomez, J.F., Hortal, J., Iverson, J.B., Kerr, J.T., Kissling, W.D., Kitching, I.J., Leon-Cortes, J.L., Lobo, J.M., Montoya, D., Morales-Castilla, I., Moreno, J.C., Oberdorff, T., Olalla-Tarraga, M.A., Pausas, J.G., Qian, H., Rahbek, C., Rodriguez, M.A., RUeda, M., Ruggiero, A., Sackmann, P., Sanders, N.J., Terribile, L.C., Vetaas, O.R., & Hawkins, B.A. (2009) Coefficient shifts in geographical ecology: an empirical evaluation of spatial and nonspatial regression. *Ecography* **32**, 193–204.

Blanchet, F.G., Legendre, P., & Borcard, D. (2008) Forward selection of explanatory variables. *Ecology* **89**, 2623–32.

Bolker, B.M. (2008) *Ecological Models and Data in R*. Princeton University Press, Princeton, NJ.

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

Bradburd, G.S., Ralph, P.L., & Coop, G.M. (2013) Disentangling the effects of geographic and ecological isolation on genetic differentiation. *Evolution* **67**, 3258–73.

Clarke, R.T., Rothery, P., & Raybould, A.F. (2002) Confidence limits for regression relationships between distance matrices: estimating gene flow with distance. *Journal of Agricultural, Biological and Environmental Statistics* **7**, 361–72.

Cushman, S.A. & Landguth, E.L. (2010) Spurious correlations and inference in landscape genetics. *Molecular Ecology* **19**, 3592–602.

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**, 486–99.

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

de Jong, P., Sprenger, C., & van Veen, F. (1984) On extreme values of Moran's *I* and Geary's *c*. *Geographical Analysis* **16**, 17–24.

Dormann, C.F., McPherson, M., Araújo, J.B., Bivand, M., Bolliger, R., Carl, J., Davies, G., Hirzel, R., Jetz, A., Kissling, D.W., Kühn, I., Ohlemüller, R., Peres-Neto, P., Reineking, B., Schröder, B., Schurr, M.F., & Wilson, R. (2007) Methods to account for spatial autocorrelation in the analysis of species distributional data: a review. *Ecography* **30**, 609–28.

Dormann, C.F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G., Marquéz, J.R.G., Gruber, B., Lafourcade, B., Leitão, P.J., Münkemüller, T., McClean, C., Osborne, P.E., Reineking, B., Schröder, B., Skidmore, A.K., Zurell, D., & Lautenbach, S. (2013) Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. *Ecography* **36**, 27–46.

Dray, S. (2011) A new perspective about Moran's coefficient: spatial autocorrelation as a linear regression problem. *Geographical Analysis* **43**, 127–41.

Dray, S., Legendre, P., & Peres-Neto, P.R. (2006) Spatial modeling: a comprehensive framework for principal coordinate analysis of neighbour matrices (PCNM). *Ecological Modeling* **196**, 483–93.

Dray, S., et al. (2012) Community ecology in the age of multivariate multiscale spatial analysis. *Ecological Monographs* **82**, 257–75.

Dutilleul, P., Stockwell, J.D., Frigon, D., & Legendre, P. (2000) The Mantel Test versus Pearson's Correlation Analysis: assessment of the differences for biological and environmental studies. *Journal of Agricultural, Biological and Environmental Statistics* **5**, 131–50.

Dyer, R.J., Nason, J.D., & Garrick, R.C. (2010) Landscape modeling of gene flow: improved power using conditional genetic distance derived from the topology of population networks. *Molecular Ecology* **19**, 3746–59.

Epperson, B.K. (2003) *Geographical Genetics*. Princeton University Press, Princeton, NJ.

Florax, R.J.G.M. & Rey, S. (1995) The impacts of misspecified spatial interaction in linear regression models. In: Anselin, L. & Florax, R.J.G.M. (eds.), *New Directions in Spatial Econometrics*. Springer, Berlin, Heidelberg.

Foll, M. & Gaggiotti, O.E. (2006) Identifying the environmental factors that determine the genetic structure of populations. *Genetics* **174**, 875–91.

Fortin, M.-J. & Dale, M.R.T. (2005) *Spatial Analysis: A Guide for Ecologists*. Cambridge University Press, New York.

Fortin, M.-J. & Melles, S.J. (2009) Avian spatial responses to forest spatial heterogeneity at the landscape level: conceptual and statistical challenges. In: Miao, S., Carstenn, S., &

Nungesser, M. (eds.), *Real World Ecology: Large-Scale and Long-Term Case Studies and Methods*. Springer, New York.

Fotheringham, A.S. (2002) In: Brunsdon, C. & Charlton, M. (eds.), *Geographically Weighted Regression: The Analysis of Spatially Varying Relationships*. John Wiley & Sons, Ltd, Chichester.

Franckoviak, R.P., Jarvis, K., Acuna, I., Landguth, E.L., Fortin, M.-J., & Wagner, H.H. (submitted) Model selection with multiple regression on distance matrices leads to incorrect inferences. *Molecular Ecology Resources.*

François, O. & Durand, E. (2010) Spatially explicit Bayesian clustering models in population genetics. *Molecular Ecology Resources* **10**, 773–84.

Gałecki, A.T. and Burzykowski, T. (2013) *Linear Mixed-Effects Models Using R: A Step-by-Step Approach*. Springer, New York.

Gelfand, A.E. & Vounatsou, P. (2003) Proper multivariate conditional autoregressive models for spatial data analysis. *Biostatistics* **4**, 11–15.

Gilbert, B. & Bennett, J.R. (2010) Partitioning variation in ecological communities: do the numbers add up? *Journal of Applied Ecology* **475**, 1071–82.

Goldberg, C.S. & Waits, L.P. (2010) Comparative landscape genetics of two pond-breeding amphibian species in a highly modified agricultural landscape. *Molecular Ecology* **19**, 3650–63.

Graves, T.A., Beier, P., & Royle, J.A. (2013) Current approaches using genetic distances produce poor estimates of landscape resistance to interindividual dispersal. *Molecular Ecology* **22**, 3888–903.

Griffith, D.A. (1996) Some guidelines for specifying the geographic weights matrix contained in spatial statistical models. In Arlinghaus, S.L. (ed.), *Practical Handbook of Spatial Statistics*. CRC Press, Boca Raton, FL.

Griffith, D.A. (2000) A linear regression solution to the spatial autocorrelation problem. *Journal of Geographical Systems* **2**, 141–56.

Griffith, D.A. & Peres-Neto, P.R. (2006) Spatial modeling in ecology: the flexibility of eigenfunction spatial analyses. *Ecology* **87**, 2603–13.

Guillot, G. & Rousset, F. (2013) Dismantling the Mantel tests. *Methods in Ecology and Evolution* **4**, 336–44.

Guillot, G., Mortier, F. & Estoup, A. (2005) Geneland: a computer package for landscape genetics. *Molecular Ecology Notes* **5**, 712–15.

Guillot, G., Leblois, R., Coulon, A., & Frantz, A.C. (2009) Statistical methods in spatial genetics. *Molecular Ecology* **18**, 4734–56.

Guillot, G., Schilling, R.L., Porcu, E. & Bevilacqua, M. (2014) Validity of covariance models for the analysis of geographical variation. *Methods in Ecology and Evolution* **5**, 329–35.

Hanski, I. (1994) A practical model of metapopulation dynamics. *Journal of Animal Ecology* **63**, 151–62.

Hutchison, D.W. & Templeton, A.R. (1999) Correlation of pairwise genetic and geographic distance measures: inferring the relative influences of gene flow and drift on the distribution of genetic variability. *Evolution* **53**, 1898–914.

James, P.M.A., Coltman, D.W., Murray, B.W., Hamelin, R.C., & Sperling, F.A.H. (2011) Spatial genetic structure of a symbiotic beetle-fungal system: toward multi-taxa integrated landscape genetics. *PLoS One* **6**, e25359.

Jay, F. (2011) *PoPS: Prediction of Population Genetic Structure – Program Documentation and Tutorial*. University Joseph Fourier, Grenoble, France.

Johnson, J.B. & Omland, K.S. (2004) Model selection in ecology and evolution. *Trends in Ecology and Evolution* **19**, 101–8.

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

Jombart, T., Dray, S., & Dufour, A.-B. (2009) Finding essential scales of spatial variation in ecological data: a multivariate approach. *Ecography* **32**, 161–8.

Keyghobadi, N., Roland, J., Matter, S.F., & Strobeck, C. (2005) Among- and within-patch components of genetic diversity respond at different rates to habitat fragmentation: an empirical demonstration. *Proceedings of the Royal Society B: Biological Sciences* **272**, 553–60.

Kühn, I. & Dormann, C.F. (2012) Less than eight (and a half) misconceptions of spatial analysis. *Journal of Biogeography* **39**, 995–8.

Legendre, P. (1993) Spatial autocorrelation: trouble or new paradigm? *Ecology* **74**, 1659–73.

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**, 831–44.

Legendre, P. & Legendre, L. (2012) *Numerical Ecology*, 3rd edition. Elsevier Science & Technology Books, San Diego, CA.

Legendre, P. & Trousselier, M. (1988) Aquatic heterotrophic bacteria: modeling in the presence of spatial autocorrelation. *Limnology and Oceanography* **33**, 1055–67.

Lichstein, J.W. (2007) Multiple regression on distance matrices: a multivariate spatial analysis tool. *Plant Ecology* **188**, 117–31.

Manel, S., Joost, S., Epperson, B.K., Holderegger, R., Storfer, A., Rosenberg, M.S., Scribner, K.T., Bonin, A., & Fortin, M.-J. (2010) Perspectives on the use of landscape genetics to detect genetic adaptive variation in the field. *Molecular Ecology* **19**, 3760–72.

Manel, S., Gugerli, F., Thuiller, W., Alvarez, N., Legendre, P., Holderegger, R., Gielly, L., & Taberlet, P. (2012) Broad-scale adaptive genetic variation in alpine plants is driven by temperature and precipitation. *Molecular Ecology* **21**, 3729–38.

Mantel, N. (1967) The detection of disease clustering and a generalized regression approach. *Cancer Research* **27**, 209–20.

McRae, B.H. (2006) Isolation by resistance. *Evolution* **60**, 1551–61.

Moilanen, A. & Nieminen, M. (2002) Simple connectivity measures in spatial ecology. *Ecology* **83**, 1131–45.

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

Muirhead, J.R., Gray, D.K., Kelly, D.W., Ellis, S.M., Heath, D.D., & MacIsaac, H.J. (2008) Identifying the source of species invasions: sampling intensity vs. genetic diversity. *Molecular Ecology* **17**, 1020–35.

Murphy, M.A., Dezzani, R.J., Pilliod, D.S., & Storfer, A. (2010) Landscape genetics of high mountain frog metapopulations. *Molecular Ecology* **19**, 3634–49.

Ochoa-Ochoa, L.M., Campbell, J.A., & Flores-Villela, O.A. (2014) Patterns of richness and endemism of the Mexican herpetofauna, a matter of spatial scale? *Biological Journal of the Linnean Society* **111**, 305–16.

Orelien, J.G. & Edwards, L.J. (2008) Fixed-effect variable selection in linear mixed models using statistics. *Computational Statistics and Data Analysis* **52**, 1896.

Peres-Neto, P.R. (2006) A unified strategy for estimating and controlling spatial, temporal and phyogenetic autocorrelation in ecological models. *Oecologia Brasiliensis* **10**, 105–19.

Peres-Neto, P.R. & Legendre, P. (2010) Estimating and controlling for spatial structure in the study of ecological communities. *Global Ecology and Biogeography* **19**, 174–84.

Raeymaekers, J.A.M., Lens, L., van den Broeck, F., van Dongen, S., & Volckaert, F.A.M. (2012) Quantifying population structure on short timescales. *Molecular Ecology* **21**, 3458–73.

Rico, Y., Boehmer, H.J., & Wagner, H.H. (2014) Effect of rotational shepherding on demographic and genetic connectivity of calcareous grassland plants. *Conservation Biology* **28**, 467–77.

Rousset, F. (1997) Genetic differentiation and estimation of gene flow from F-statistics under isolation by distance. *Genetics* **145**, 1219–28.

Ryman, N., Palm, S. André, C., Carvalho, G.R., Dahlgren, T.G., Jorde, P.E., Laikre, L., Larsson, L.C., Palmé, A., & Ruzzante, D.E. (2006) Power for detecting genetic divergence: differences between statistical methods and marker loci. *Molecular Ecology* **15**, 2031–45.

Safner, T., Miller, M.P., McRae, B.H., Fortin, M.-J., & Manel, S. (2011) Comparison of Bayesian clustering and edge detection methods for inferring boundaries in landscape genetics. *International Journal of Molecular Sciences* **12**, 865–89.

Saura, S. & Rubio, L. (2010) A common currency for the different ways in which patches and links can contribute to habitat availability and connectivity in the landscape. *Ecography* **33**, 523–37.

Schoville, S.D., Bonin, A., François, O., Lobreaux, S., Melodelima, C., & Manel, S. (2012) Adaptive genetic variation on the landscape: methods and cases. *Annual Review of Ecology, Evolution and Systematics* **43**, 23–43.

Selkoe, K.A. & Toonen, R.J. (2006) Microsatellites for ecologists: a practical guide to using and evaluating microsatellite markers. *Ecology Letters* **9**, 615–29.

Smouse, P.E. & Peakall, R. (1999) Spatial autocorrelation analysis of individual multiallele and multilocus genetic structure. *Heredity* **82**, 561–73.

Smouse, P., Long, J.C., & Sokal, R.R. (1986) Multiple regression and correlation extension of the Mantel test of matrix correspondence. *Systematic Zoology* **35**, 627–32.

Sokal, R.R. & Wartenberg, D.E. (1983) A test of spatial autocorrelation analysis using an isolation-by-distance model. *Genetics* **105**, 21–37.

Sokal, R.R., Oden, N.L., & Thomson, B.A. (1998) Local spatial autocorrelation in a biological model. *Geographical Analysis* **30**, 331–54.

Spear, S.F. & Storfer, A. (2010) Anthropogenic and natural disturbance lead to differing patterns of gene flow in the Rocky Mountain tailed frog, *Ascaphus montanus. Biological Conservation* **143**, 778–86.

Spear, S.F., Balkenhol, N., Fortin, M.-J., McRae, B.H., & Scribner, K.T. (2010) Use of resistance surfaces for landscape genetic studies: considerations for parameterization and analysis. *Molecular Ecology* **19**, 3576–91.

Storfer, A., Murphy, M.A., Spear, S.F., Holderegger, R., & Waits, L.P. (2010) Landscape genetics: Where are we now? *Molecular Ecology* **19**, 3496–514.

Van Strien, M.J., Keller, D., & Holderegger, R. (2012) A new analytical approach to landscape genetic modeling: least-cost transect analysis and linear mixed models. *Molecular Ecology* **21**, 4010–23.

Vignieri, S.N. (2005) Streams over mountains: influence of Riparian connectivity on gene flow in the Pacific jumping mouse (*Zapus trinotatus*). *Molecular Ecology* **14**, 1925–37.

Wagner, H.H. (2013) Rethinking the linear regression model for spatial ecological data. *Ecology* **94**, 2381–91.

Wagner, H.H. & Fortin, M.J. (2013) A conceptual framework for the spatial analysis of landscape genetic data. *Conservation Genetics* **14**, 253–61.

Whitlock, M.C. (2011) G'_{ST} and D do not replace $F_{ST.}$ *Molecular Ecology* **20**, 1083–91.

Windle, M.J.S., Rose, G.A., Devillers, R., & Fortin, M.J. (2012) Spatio-temporal variations in invertebrate–cod–environment relationships on the Newfoundland-Labrador Shelf, vol. 1995–2009. *Marine Ecology Progress Series* **469**, 263–78.

Wright, S. (1943) Isolation-by-distance. *Genetics* **28**, 114–38.

Zuur, A.F., Ieno, E.N., Walker, N.J., Saveliev, A.A., & Smith, G.M. (2009) *Mixed Effects Models and Extensions in Ecology with R*. Springer, New York.