Model Based Approaches for Characterizing Environmental Effects on Spatial Genetic Flow

Ephraim M. Hanks* Mevin B. Hooten† Leslie McFarlane‡ Karen E. Mock§

Abstract
Landscape genetics is the study of the effects of landscape on genetic diversity, but current landscape genetic studies often link inference made on spatial genetic boundaries to landscape features in a post-hoc way. We present a general approach that formalizes links between genetic boundaries and landscape features of interest. This approach builds on existing statistical models for landscape genetics and lends itself to optimal sampling methods. We illustrate the approach through a landscape genetic study of mule deer in Utah and Colorado.

Key Words: Landscape genetics, multiple imputation, optimal sampling, mule deer.

1. Introduction

Landscape genetics is the study of the effects of landscape on genetic diversity (Manel 2003, Gaggiotti 2010), but linking genetic information (e.g., microsatellite data) to the landscape in a rigorous way is challenging. Current landscape genetics studies often employ well-developed statistical models for identifying genetic boundaries that are based on the assumption of a homogeneous underlying landscape. Software exists to easily implement many of these models, including Geneland (Guillot et al. 2005a; Guillot et al. 2005b) and TESS (Chen et al. 2007; Durand et al. 2009). These methods allow for rigorous inference about the location of genetic populations and boundaries under both the Bayesian and classical statistical paradigms. However, boundaries resulting from such an analysis are often linked in a post-hoc way to landscape features, such as through comparing the locations of potential barriers (e.g., roads or rivers) to the mode of the posterior distribution through a geographic information system (GIS) overlay (e.g., Sahlsten et al. 2008; Wheeler et al. 2010).

We present a general approach that formalizes such links between posterior predictions of genetic boundary locations and landscape features of interest. This approach builds on existing statistical models for landscape genetics and lends itself to optimal sampling methods.

We first provide a brief introduction to Geneland (Guillot et al. 2005b), an existing model used to make inference about landscape genetic boundaries. We then present our approach for linking the distribution of genetic boundaries resulting from this Bayesian hierarchical model (BHM) to landscape features of interest, and set forth a method of optimal sampling in which we maximize the information we can obtain about the relationship between landscape features and genetic boundaries using limited resources. We then illustrate this approach to make inference on the effects of various landscape features on the genetic differentiation of mule deer (*Odocoileus hemionus*) in Utah and Colorado, USA.

*Department of Statistics, Colorado State University, Fort Collins, CO 80523

†USGS Colorado Cooperative Fish and Wildlife Research Unit, Colorado State University, Fort Collins, CO 80523

‡Utah Division of Wildlife Resources, Salt Lake City, UT 84114

§Department of Wildland Resources, Utah State University, Logan, UT 84321

2. Methods

2.1 Identifying Genetic Boundaries Using Geneland

The Geneland statistical model (Guillot et al. 2005a) is a BHM for making inference about population boundaries given spatially-referenced genetic data, $\mathbf{A}$. This spatial clustering model is based on a Voronoi tesselation of the study area, and results in the distribution of predicted piece-wise linear boundaries between distinct genetic populations. If $\boldsymbol{\theta}$ are the parameters in this model describing the spatial locations of these genetic boundaries, then we will denote their posterior distribution as $[\boldsymbol{\theta}|\mathbf{A}]$. The Geneland model assumes a homogeneous landscape, and we seek a method for making posterior inference about the relationship of hypothesized landscape features to genetic boundaries.

As mentioned previously, there are other statistical models in the literature that also provide inference about spatial genetic boundaries that are based on similar assumptions (e.g., homogeneous landscape). For a discussion of the relative strengths and weaknesses of some of these models, see Francois and Durand (2010) and Manlove et al. (2010). We have chosen to use Geneland to identify spatial genetic boundaries, but the approach we present could utilize any of the existing models.

2.2 Linking Genetic Boundaries to Landscape Features

We are interested in investigating the correlation between landscape features and gene flow in a manner that is statistically rigorous and allows for optimal sampling. Our approach is to model the relationship between landscape features and genetic boundaries, incorporating spatial structure in the data using geostatistical methods. Let $x_{i,m}$ be the shortest euclidean distance from the i-th sample location to the m-th landscape feature of interest, and let y_i be the euclidean distance from the i-th sample location to the nearest genetic boundary, or a monotonic transformation of the same as is common in regression modeling (e.g., Fox 1997). Then, a model for the relationship between the distance to the nearest genetic boundary and the distance to landscape features is:

$$\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon} \ , \ \ \boldsymbol{\epsilon} \sim N(\mathbf{0}, \Sigma) \ . \tag{1}$$

That is, we assume a linear relationship between some monotonic transformation of the distance to a genetic boundary ($\mathbf{y}$) and the distance to the landscape feature of interest ($\mathbf{X}$), with possible spatial correlation between locations (expressed in Σ). In this way, a statistically significant linear relationship between $\mathbf{X}$ and $\mathbf{y}$, as indicated by inference about $\boldsymbol{\beta}$, would indicate that the landscape features are significantly related to genetic boundaries in mule deer populations. The transformation of the response variable could be chosen based on the data in an attempt to satisfy model assumptions of normality and homoscedasticity (e.g., Fox 1997).

Given $\mathbf{y}$, we can use standard geostatistical techniques to make inference about the distributions of $\hat{\boldsymbol{\beta}}$ and $\hat{\Sigma}$. For example, an iterative technique could be used by first specifying an initial $\hat{\Sigma}$, employing generalized least squares (GLS) to estimate the distribution of $\hat{\boldsymbol{\beta}}$:

$$\hat{\boldsymbol{\beta}}|\mathbf{y}, \hat{\Sigma} \sim N\big((X'\hat{\Sigma}^{-1}X)^{-1}X'\hat{\Sigma}^{-1}\mathbf{y}, (X'\hat{\Sigma}^{-1}X)^{-1}\big),$$

then estimating a parameterized version of $\hat{\Sigma}$ (e.g., exponential spatial correlation) through maximum-likelihood fitting of the semi-variogram of the residuals, $\mathbf{e} = \mathbf{y} - X\hat{\boldsymbol{\beta}}$, and repeating the process until convergence. This would allow inference on the parameters in (1) given $\mathbf{y}$.

Note, however, that we do not have a fixed value for $\mathbf{y}$, but rather a posterior predictive distribution $[\mathbf{y}|\mathbf{A}]$ that incorporates our uncertainty about the location of landscape genetic boundaries. In a sense, we are using the Geneland model as a stochastic transformation of the genetic allele data $\mathbf{A}$ to the distance from the spatial locations of these data to the nearest landscape genetic boundary.

We are interested in inference on $[\hat{\boldsymbol{\beta}}, \hat{\Sigma}|\mathbf{A}]$, the distribution of the parameters in (1) conditioned on the genetic allele data, $\mathbf{A}$. We note that

$$[\hat{\boldsymbol{\beta}}, \hat{\Sigma}|\mathbf{A}] = \int_{\mathbf{y}} [\hat{\boldsymbol{\beta}}, \hat{\Sigma}|\mathbf{y}, \mathbf{A}][\mathbf{y}|\mathbf{A}]d\mathbf{y}.$$

If we make the assumption that the influence of landscape covariates on genetic differentiation ($\boldsymbol{\beta}$) in model (1) depends on the observed allele data ($\mathbf{A}$) only conditionally through $\mathbf{y}$, the distance to the nearest genetic boundary, then the first term in the integral could be simplified:

$$[\hat{\boldsymbol{\beta}}, \hat{\Sigma}|\mathbf{A}] = \int_{\mathbf{y}} [\hat{\boldsymbol{\beta}}, \hat{\Sigma}|\mathbf{y}][\mathbf{y}|\mathbf{A}]d\mathbf{y}. \tag{2}$$

We can approximate this integral using composition sampling, which in this case is a form of multiple imputation (e.g., Rubin 1987). It proceeds by iteratively sampling from $[\mathbf{y}|\mathbf{A}]$ and $[\hat{\boldsymbol{\beta}}, \hat{\Sigma}|\mathbf{A}]$ to obtain samples from $\hat{\boldsymbol{\beta}}, \hat{\Sigma}|\mathbf{A}$. Sampling from $[\mathbf{y}|\mathbf{A}]$ is straightforward, as $\mathbf{y}$ is a simple transformation of any realization from $[\boldsymbol{\theta}|\mathbf{A}]$, the posterior distribution of spatial genetic boundaries using the Geneland model. For $[\hat{\boldsymbol{\beta}}, \hat{\Sigma}|\mathbf{y}]$, the geostatistical methods described above allow us to find the distribution of the estimators $[\hat{\boldsymbol{\beta}}, \hat{\Sigma}|\mathbf{y}]$, which we can then sample from. In this way, we can make inference about the effects of hypothesized landscape genetic boundaries, conditioned solely on the microsatellite allele data.

Simulation studies (not included here) indicate that this approach can distinguish between true and extraneous hypothesized boundaries in the case where there are two genetic populations separated by a single boundary. Thus, if a hypothesized boundary's regression coefficient is significantly different from zero we can conclude that it is significantly related to spatial genetic boundaries.

2.2.1 Model Selection

We are interested in making inference about which landscape features are significantly related to genetic population boundaries. Typically we have many hypothesized boundaries, and desire to compare models with different combinations of these proposed landscape genetic boundaries. Model selection in regression models is commonly accomplished through methods that seek to balance model parsimony and goodness-of-fit (see e.g., Burnham and Anderson 2002). Our approach (2) for making inference about model parameters $\hat{\boldsymbol{\beta}}, \hat{\Sigma}$ conditioned on the allele data $\mathbf{A}$ makes standard measures of goodness-of-fit, such as R^2 or the likelihood of the data, given a specific model, difficult to obtain. We propose an approach for obtaining a measure of the goodness-of-fit of a model based on a mixture model interpretation of (2).

Consider three potential models: the first underfits the allele data by not including landscape features that are important for spatial gene flow, the second overfits by including all landscape features that are important for spatial gene flow, plus some extraneous landscape features, and the third includes only landscape features that are important to spatial gene flow. If the underfitting model is missing a covariate that represents a different spatial scale (coarser or finer) of genetic differentiation, then it is reasonable to assume that the distribution of landscape genetic boundaries (e.g., the Geneland posterior $[\boldsymbol{\theta}|\mathbf{A}]$) would be bi-modal (or multi-modal) as the model attempts to account for multiple scales of differentiation. In the case of Geneland, this would manifest itself in realizations from the posterior showing varying degrees of coarseness in the tesselation of the study area. Thus, we might expect bimodal behavior in $[\hat{\boldsymbol{\beta}}|\mathbf{A}]$ with one mode centered at $\mathbf{0}$ representing the lack of model fit at realizations where the model lacks covariates at the correct spatial scale or location. The second mode would represent the distribution of $\hat{\boldsymbol{\beta}}|\mathbf{A}$ at realizations where the model is adequate.

Thus, as a meta-analysis, assume a Gaussian mixture-model approximation to $[\hat{\boldsymbol{\beta}}|\mathbf{A}]$:

$$\hat{\boldsymbol{\beta}}|\mathbf{A} \sim \begin{cases} N(\mathbf{0}, \Sigma_0) & \text{, with probability } p \\ N(\boldsymbol{\mu}, \Sigma) & \text{, with probability } 1-p \end{cases} \tag{3}$$

where p is the probability that the specified model does not adequately explain the spatial genetic boundaries manifested in the data. Thus, large values of p (e.g., greater than 0.05) would indicate an inadequate model.

Similarly, we would expect a model that contains all landscape features that are important to spatial gene flow to have low values of p. Models that are overfit would also have small values of p, but we would expect that removing unnecessary terms from the model would have little effect on p.

To balance parsimony with goodness-of-fit, we could select the most parsimonious model for which $p < 0.05$. In this way we can distinguish hypothesized landscape features that are important for spatial genetic flow from those that are not.

2.3 Optimal Sampling

Having specified an approach for making inference about $\boldsymbol{\beta}$, the effect of landscape features on spatial gene flow, we now present an approach for sampling that allows us to maximize the information we obtain about hypothesized genetic barriers, while minimizing the cost of obtaining and analyzing new genetic samples. We assume that we have potential genetic allele data ($\mathbf{A}$) at a large number M of spatial locations, yet are only able to obtain and analyze a small subset of this potential data. Optimal sampling methods for spatially-referenced data are well-developed (see e.g., Diggle and Lophaven 2006; Lark 2001; Martin 2001; Muller 2007), and include methods for optimal dynamic sampling (e.g., Hooten et al. 2009) in ecological and environmental studies. The general approach to optimal sampling of spatial data is as follows. A prospective sample is chosen in such a way that the process of interest can be studied at various spatial scales. This prospective sample is used to fit the a model that includes spatial structure, such as (1) and (2), and predictions based on this model fit inform the choice of the next set of locations to sample, the first retrospective sample. Successive retrospective samples are chosen using updated model predictions using all samples analyzed to that point.

2.3.1 Prospective Sampling

The main objective of the prospective sample is to obtain a set of genetic samples that cover the spatial range of the data and allow us to estimate the structure of the spatial autocorrelation between samples. Accurately estimating the spatial structure (Σ) is essential to making inference about $\boldsymbol{\beta}$.

We advocate a “lattice and close pairs” approach (Diggle and Lophaven 2006; Lark 2001; Martin 2001; Muller 2007), common in the literature, for a prospective sample of spatial genetic data. Half of the samples in the prospective design are selected in a way that maximizes spatial coverage (the “lattice” samples), and one “close pair” is selected randomly for each lattice sample in such a way that results in a range of spatial scales between sample locations. In this way, the lattice and close pairs approach strikes a balance between the spatial coverage necessary to learn about large-scale trends in the data and the clustering needed to learn about spatial structure at multiple scales.

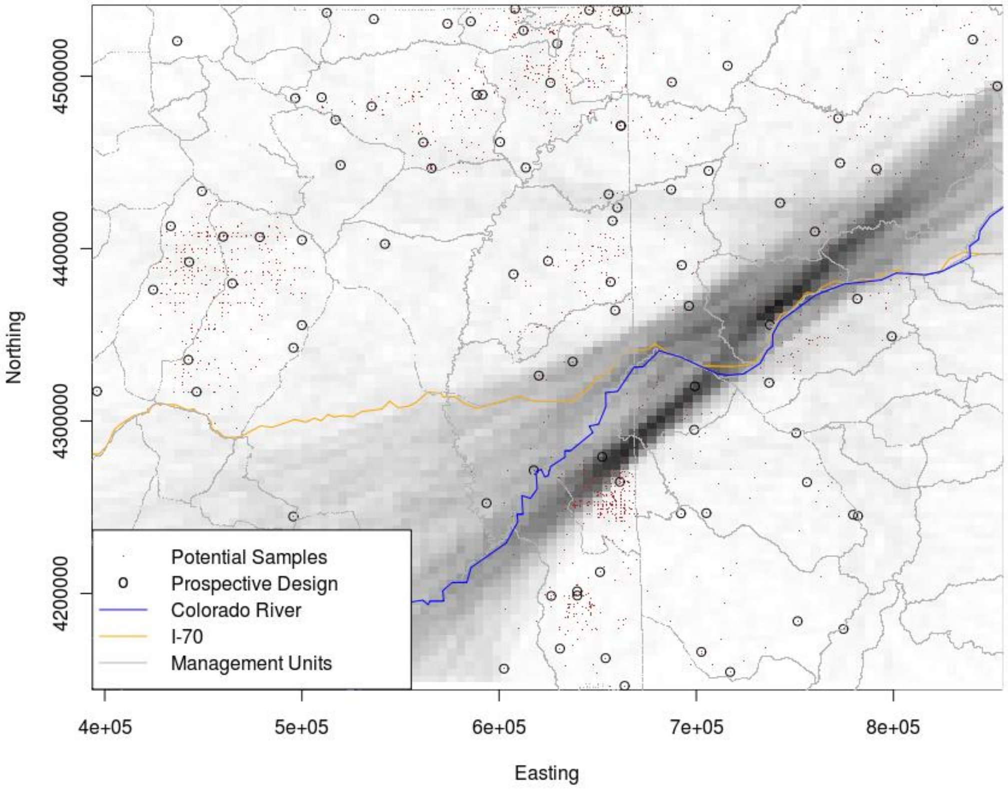

Figure 1: The locations of the 90 samples in the prospective design are shown with the hypothesized boundaries and a representation of the Geneland posterior distribution of genetic boundary locations. Dark cells indicate locations where there is a high probability of a genetic boundary, while lighter cells indicate locations with a low probability.

2.3.2 Retrospective Sampling

Once spatially-referenced genetic allele data have been obtained, we wish to use existing data to pick new locations to sample in a way that maximizes the information gained about $\hat{\boldsymbol{\beta}}$, and thus the relationship between landscape features and spatial genetic flow. Predictive inference about the parameters in (1), based on existing data from a prospective sample, can be used to predict the influence that adding new spatially-referenced data will have on inference about the parameters of interest.

Let $[\mathbf{y}^*|\mathbf{A}]$ be the distribution of the distance to the nearest genetic boundary for all spatial locations where we have potential genetic allele data, and let $\mathbf{X}^*$ be a matrix containing the distance from the same locations to the landscape features of interest. Note that we have only observed A at a subset of the full set of locations represented by $\mathbf{y}^*$. The n observed locations in $\mathbf{y} \subseteq \mathbf{y}^*$ are those in the prospective sample. We seek to choose a retrospective sample: a set of m new locations to obtain and analyze genetic data that would result in $\mathbf{y}_{\text{new}} \subset \mathbf{y}^*$. Thus we could write our combined prospective and retrospective sample as:

$$\tilde{\mathbf{y}} = \begin{bmatrix} y_{\text{new}} \\ y \end{bmatrix}.$$

We could then think about a sampling or design matrix, $\mathbf{K}$, that selects the prospective and retrospective samples from $\mathbf{y}^*$:

$$\tilde{\mathbf{y}} = \mathbf{K}\mathbf{y}^* = \mathbf{K}(\mathbf{X}^*\boldsymbol{\beta} + \boldsymbol{\epsilon}^*).$$

Then equation (1) would become:

$$\tilde{\mathbf{y}} \sim N(\mathbf{K}\mathbf{X}^*\boldsymbol{\beta}, \mathbf{K}^T\Sigma^*\mathbf{K}). \tag{4}$$

We seek to find a set of new samples, and the corresponding $\mathbf{K}$, that maximizes our learning about the relationship of the landscape features to genetic boundaries. Minimizing the design criterion:

$$Q_K = tr(Var[\hat{\boldsymbol{\beta}}|\mathbf{A},\mathbf{K}]) = tr\left(\left((\mathbf{K}\mathbf{X}^*)^T(\mathbf{K}^T\hat{\Sigma}^*\mathbf{K})^{-1}\mathbf{K}\mathbf{X}^*\right)^{-1}\right) \tag{5}$$

minimizes the predicted variance of $\hat{\boldsymbol{\beta}}$ and thus seeks to maximize the information gained about the relationship between landscape features and spatial gene flow.

We seek a quasi-optimal solution through an iterative process in which we change K slightly at each iteration, and then accept the change if it improves Q_K, or reject the change if it does not. Calculation of this design criterion as written in (5) involves a matrix inversion with each change in K, and for each realization of $\mathbf{y}|\mathbf{A}$ and thus can be quite computationally demanding. Appendix A describes an alternate formulation of this design criterion that is equivalent, but is significantly more computationally efficient, allowing us to obtain quasi-optimal designs through iterative switching algorithms.

3. Application: Mule Deer

We apply our approach to a landscape genetic study of mule deer in Utah and Colorado. Mule deer are one of the primary hosts for chronic wasting disease (CWD) in the region (Miller et al., 2006). CWD is caused by a contagious prion-based pathogen that is capable of persistence in the environment (e.g., Miller et al., 2006). Understanding causes of barriers to mule deer movement, and thus gene flow, may be helpful in understanding the spread of this disease. In this study, we chose to investigate the effects of three hypothesized genetic boundaries, the Colorado river, a major highway (Interstate-70), and boundaries of the wildlife management units (WMUs) in Utah and the deer animal units (DAUs) in Colorado. We will refer to the WMU and DAU boundaries collectively as management unit boundaries.

In 1999-2004, spatially-referenced genetic samples were collected from 2,046 mule deer in Utah and Colorado. With resources available to obtain microsatellite allele data from only a fraction of these genetic samples, we applied the optimal sampling methods described above to best learn about the effects of the landscape features described above.

For our prospective sample, we selected 45 “lattice” samples in such a way that the minimum distance between any of these samples was maximized, thus ensuring regular coverage of the geographic extent of the deer in the survey. For each lattice sample, the unexamined samples that were closer to that lattice sample than to any other lattice sample were identified, and a “close pair” was chosen randomly from these unobserved samples. Microsatellite allele data was obtained for 17 loci for each of the animals in the prospective sample.

We utilized the Geneland R-package (Guillot et. al., 2005b) to generate 1,100,000 iterations of the Geneland MCMC algorithm, using the allele data from the prospective sample. We discarded the first 100,000 as burn-in, and thinned 1/1000 of the remaining realizations, resulting in 1000 posterior realizations of the distribution of genetic population boundaries.

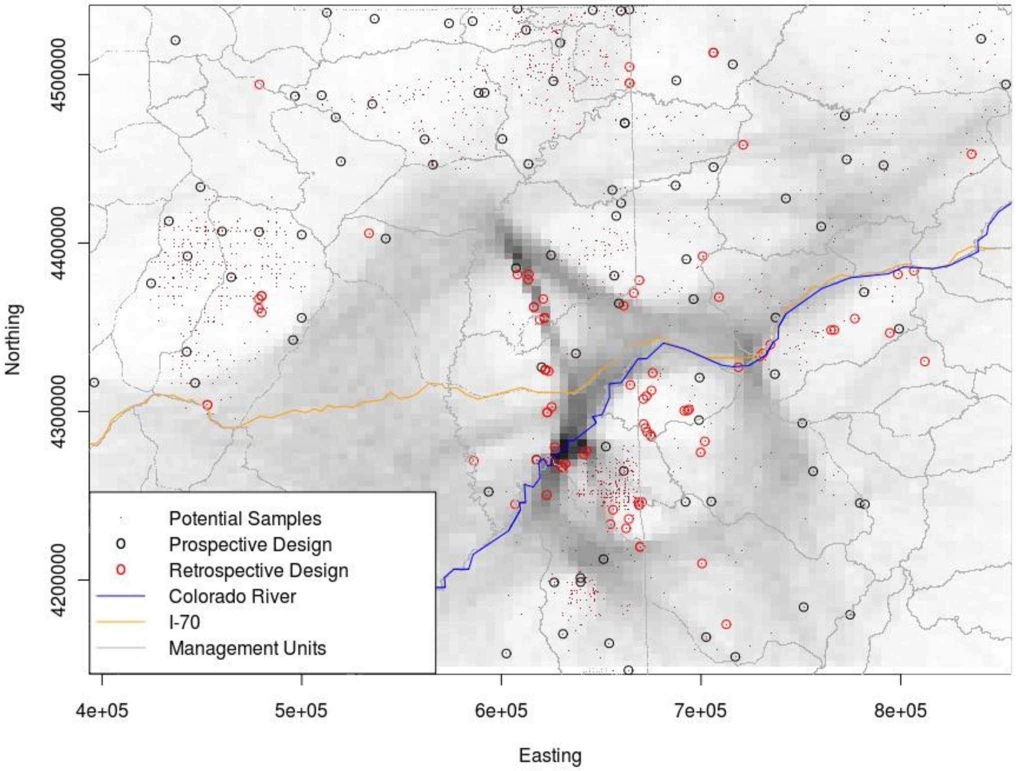

Figure 2: The prospective design, retrospective design, and hypothesized boundaries with a representation of the Geneland posterior distribution of genetic boundary locations. Dark cells indicate locations where there is a high probability of a genetic boundary, while lighter cells indicate locations with a low probability.

These realizations were used in the approach we have described to make inference about $[\hat{\boldsymbol{\beta}}, \hat{\Sigma}|\mathbf{A}]$. When applied to the mule deer microsatellite data, the Geneland posterior suggests two main genetic populations in the spatial domain (Figure 1). As the Colorado river and the interstate I-70 are co-incident for a good portion of the study area (Figure 1), ad-hoc methods such as a GIS overlay would have a difficult time making inference about which of these landscape features are significantly related to gene flow, if any. Using the untransformed distance to closest genetic boundary, results for the prospective sample are shown in Table 1.

We applied the optimal sampling approach described above to select 90 retrospective samples, based on a full model with all three landscape features included. We thus are selecting samples which allow us to maximize our learning about the effect, or lack thereof, of these landscape features on spatial gene flow. The retrospective samples, and the resulting posterior distribution of spatial genetic boundaries, are shown in Figure 2.

The addition of the retrospective sample has allowed us to see genetic boundaries at a finer spatial scale than was previously possible. The posterior distribution of spatial genetic boundaries appears to coincide with the Colorado river through the middle of the study region, though there is significant structure in the posterior that does not visibly coincide with any of our hypothesized boundaries. With this additional data, a log-transform of the distance to nearest genetic boundary variable resulted in better model fits. Log-transforming the response in this way reflects an exponential relationship between the distance to hypothesized landscape features and spatial genetic boundaries.

Table 1: Inference on $\hat{\boldsymbol{\beta}}|\mathbf{A}$ for the prospective sample. In this case the response variable in the regression model is untransformed.

		95% Equal-Tailed CI	
	Mean	2.5%	97.5%
Prospective Sample: all hypothesized boundaries			
Intercept	6.7613E+03	-1.7725E+04	4.9031E+04
Colorado river	6.8052E-01	-4.6307E-01	1.4228E+00
Management unit boundaries	3.7566E-02	-2.5188E-01	6.1877E-01
Interstate-70	2.3046E-01	-3.8445E-01	1.0888E+00
Prospective Sample: only Colorado river			
Intercept	1.5506E+04	-1.3163E+04	6.4713E+04
Colorado river	8.2166E-01	6.4144E-02	1.1533E+00

We apply the model selection approach, as described previously, and show estimates of p for models with various combinations of covariates in Table 2. The best model with one covariate is model 2, with just the Colorado river. The best model with two covariates is model 5, with the intercept and the Colorado river. These models have estimates of p greater than 0.05, indicating that it is still inadequate at characterizing the distribution of spatial genetic boundaries. The best model with three covariates is model 9, which includes the intercept, the river, and the management unit boundaries. The 95% credible intervals for the parameters of models 5, 9, and 12 are shown in Table 3, and histograms of these parameters are shown in Figure 3.

From these results, we can draw some conclusions about the importance of each of the hypothesized landscape boundaries. The Colorado river is significantly linked to spatial genetic boundaries. The best model of every size, as measured by p, contains the Colorado river, and removing this covariate results in a significant increase in p. No model without the Colorado river adequately accounts for the posterior distribution of spatial genetic boundaries. In contrast, there are models that do not include the management unit boundaries (model 8) and the Interstate (model 9), but have p-values that provide no evidence of lack-of-fit. Model 9, which includes the management unit boundaries, better fits the data (as measured by p) than does model 8, which includes the Interstate. Figure 2 shows little correspondence between the posterior distribution of genetic boundaries and the actual management unit boundaries, but the fine spatial scale of the management unit boundaries, relative to the Colorado river and the Interstate, allows this covariate to account for finer-scale genetic differentiation than is accounted for by the Colorado river.

In summary, we might conclude the following about the effects of our hypothesized landscape boundaries. The Colorado river is significantly related to large-scale gene flow in the study region. There is some gene flow not accounted for by the Colorado river, possibly at a finer spatial scale, as evidenced by the significance of the management unit boundaries or Interstate-70 when paired with the Colorado river. This is likely to be at a finer spatial scale, as evidenced by the better model fit obtained when the management unit boundaries are paired with the Colorado river than when the Interstate is paired with the Colorado river. Further work will investigate additional potential barriers to spatial gene flow within this framework.

Table 2: We compare multiple models by contrasting the probability p that the model does not adequately characterize the posterior distribution of spatial genetic boundaries.

	Model Terms				
Model Index	Intercept	Colorado river	Interstate-70	WMU and DAU Boundaries	p
1	X				0.986
2		X			0.209
3			X		0.392
4				X	0.402
5	X	X			0.209
6	X		X		0.400
7	X			X	0.346
8	X	X	X		0.008
9	X	X		X	$< 10^{-3}$
10	X		X	X	0.276
11		X	X	X	$< 10^{-3}$
12	X	X	X	X	$< 10^{-3}$

4. Discussion

We have presented an approach for linking genetic allele data to hypothesized landscape boundaries to spatial gene flow, and illustrated how this approach lends itself to optimal sampling methods. While this approach can identify landscape features that are not significantly linked to gene flow, care must be taken in interpreting the results. The interpretation of $\hat{\boldsymbol{\beta}}|\mathbf{A}$, for example, is unclear. Further work will address interpretability of the model, as well as alternative approaches for model-based characterization of environmental effects on spatial gene flow.

Optimal sampling methods can significantly improve the utility of genetic data, especially in data that are spatially clustered, as were the mule deer samples in our application. Simulation studies (not shown here) show that, in the case where there are two genetic populations separated by one boundary, a set of optimal retrospective samples can result in at least an order of magnitude reduction in the design criterion (trace of the variance matrix) when compared to random samples. This can amount to significant savings in both cost and time as less data, and thus less laboratory work and analysis, can lead to comparable information about spatial gene flow.

Our retrospective optimal design criterion is based on minimizing the predicted trace of the variance matrix of the coefficients. This will maximize the predicted information gained through adding retrospective samples. However, there is an inherent trade-off between reducing the bias and variance of our predicted distribution. The minimization is based on samples already analyzed, and, in an extreme example, samples could be selected at the

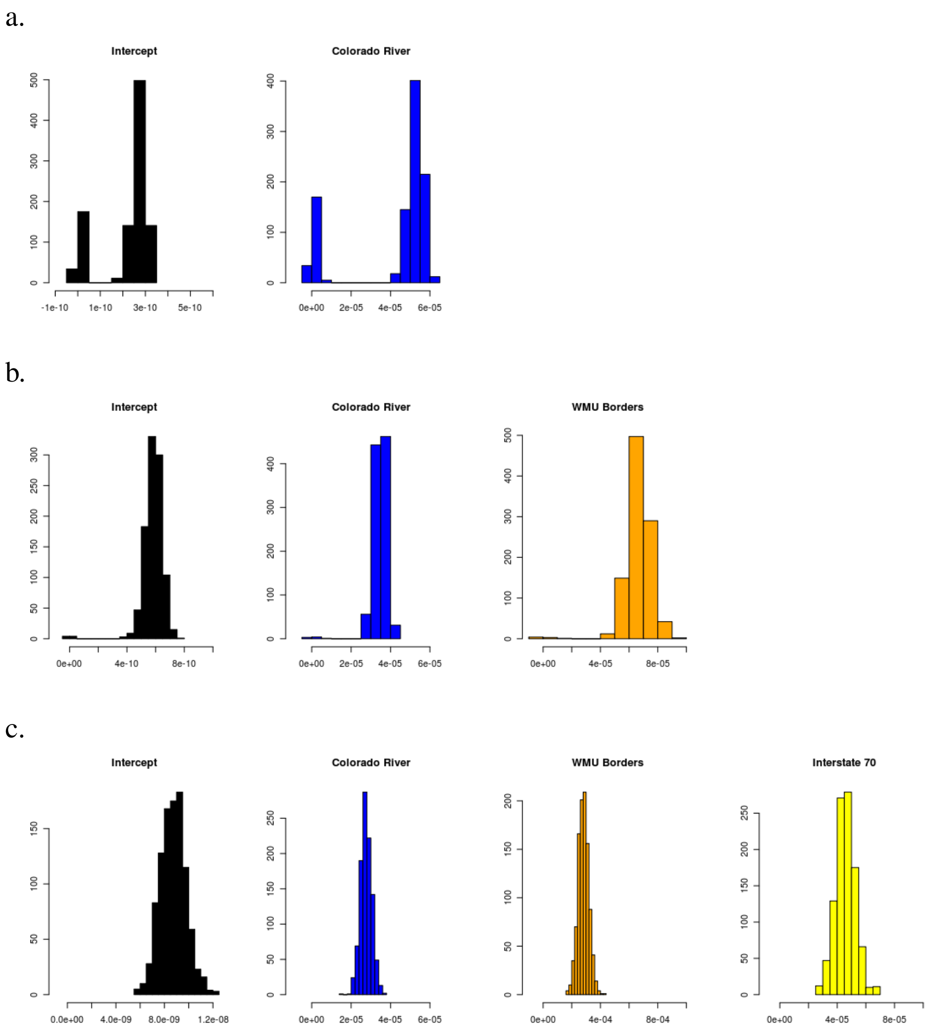

Figure 3: Histograms of $\hat{\boldsymbol{\beta}}|\mathbf{A}$ for the best models with two, three, and four componants. Bi-modality, with one mode centered at $\mathbf{0}$, indicates a model that does not fully account for the posterior distribution of spatial genetic boundaries at all scales.

exact locations where we already have genetic data. The spatial structure in the data would lead us to predict that the new data are similar to existing data, which would lead to reduced variance but very little change in the predicted mean values of the coefficients. Thus, the existing data could bias the optimal sampling procedure, if there is a large discrepancy between the sizes of the prospective and retrospective samples.

Our approach to model selection is based on a two-componant multivariate Gaussian mixture model, and assumes that spatial genetic differentiation is occuring on two different spatial scales. This could be extended to cover the case where genetic differentiation is occuring at 3 or more scales by increasing the number of mixture components. Inference could be made on the number of mixture components under a Bayesian heirarchical modeling framework using reversible-jump or birth-death Markov chain Monte Carlo methods (Green 1995; Stephens 2000).

Acknowledgements

The authors would like to thank the Colorado Division of Wildlife and the Utah Division of Wildlife for providing data for this study, and Carol Rowe for performing the molecular analyses. Funding for this study was provided by the USGS Cooperative Agree-

Table 3: Inference on $\hat{\beta}|\mathbf{A}$ for three models in the mule deer study.

		95% Equal-Tailed CI	
	Mean	2.5%	97.5%
Model 5			
Intercept	2.1636E-10	-2.6663E-15	3.2125E-10
Colorado River	4.2002E-05	-1.0401E-07	5.8859E-05
Model 9			
Intercept	5.8274E-10	4.6719E-10	6.9000E-10
Colorado River	3.4558E-05	2.8181E-05	4.0146E-05
WMU and DAU Borders	6.6617E-05	5.0692E-05	8.1096E-05
Model 12			
Intercept	8.7248E-09	6.6738E-09	1.0967E-08
Colorado River	2.7606E-05	2.1944E-05	3.3259E-05
WMU and DAU Borders	2.8177E-04	2.0777E-04	3.5354E-04
Interstate-70	4.5832E-05	3.2496E-05	5.9516E-05

ment 1434-06HQRU1555.

REFERENCES

Burnham, K.P., and Anderson, D. (2002), “Model selection and multi-model inference”, Springer, New York, New York USA, 496p.

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

Diggle, P., and Lophaven, S. (2006), “Bayesian geostatistical design”, *Scandinavian Journal of Statistics*, 33(1):5364.

Durand, E., Jay, F., Gaggiotti, O.E., and Francois, O. (2009), “Spatial inference of admixture proportions and secondary contact zones”, *Molecular Biology and Evolution*, 26:1963-1973.

Fox, J. (1997), “Applied regression analysis, linear models, and related methods”, Sage Publications, Thousand Oaks, California USA, 624p.

Francois, O., and Durand, E. (2010), “Spatially explicit Bayesian clustering models in population genetics”, *Molecular Ecology Resources*, 10:773-784.

Gaggiotti, O.E. (2010), “Advances in the analyzsis of spatial genetic data”, *Molecular Ecology Resources*, 10(5):757-759.

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

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

Guillot, G., Estoup, A., and Mortier, F. (2005b), “Geneland: a computer package for landscape genetics”, *Molecular Ecology Notes*, 5(3):712-715.

Hooten, M.B., Wikle, C.K., Sheriff, S.L., and Rushin, J.W. (2009), “Optimal spatio-temporal hybrid sampling designs for ecological monitoring”, *Journal of Vegetation Science*, 20(4):639-649.

Lark, R. (2001), “Optimized spatial sampling of soil for estimation of the variogram by maximum likelihood”, *Geoderma*, 105(1-2):4980.

Manel, S., Schwartz, M.K., Luikart, G., and Taberlet, P. (2003), “Landscape genetics: combining landscape ecology and population genetics”, *TRENDS in Ecology and Evolution*, 19(4):189-197.

Manlove, K., and Higgs, M. (2010), “A closer look at models developed for landscape genetics”, presentation at *Joint Statistical Meetings 2010*.

Martin, R.J. (2001), “Comparing and contrasting some environmental and experimental design problems”, *Environmetrics*, 12(3):273287.

Miller, M.W., Hobbs, N.T., and Tavener, S.J. (2006), “Dynamics of Prion Disease Transmission in Mule Deer”, *Ecological Applications*, 16:2208-2214.

Muller, P.M., Berry, D.A., Grieve, A.P., Smith, M., and Krams, M., (2007), “Simulation-based sequential Bayesian design”. *Journal of Statistical Planning and Inference*, 137:3140-3150.

Rubin, D.B. (1987), “Multiple Imputation for Nonresponse in Surveys”, John Wiley, New York, New York USA, 320p.

Sahlsten, J., Thorngren, H., and Hoglund, J., (2008), “Inference of hazel grouse population structure using multilocus data: a landscape genetic approach”, *Heredity*, 101(6):47582.

Stephens, M., (2000), “Bayesian analysis of mixture models with an unknown number of components-an alternative to reversible jump methods”, *Annals of Statistics*, 28(1):40-74.

Wheeler, D.C., Waller, L.A., and Biek, R., (2010), “Spatial analysis of feline immunodeficiency virus infection in cougars” *Spatial and Spatio-temporal Epidemiology*, 1(2-3):151161.

Appendix A: A Computationally-Efficent Design Criterion For Optimal Retrospective Sampling

We derive a computationally efficient form for our retrospective design criterion, $Q_{\mathbf{K}} = tr(Var(\hat{\boldsymbol{\beta}}|\mathbf{A},\mathbf{K}))$, utilizing the following properties (e.g., Harville 2008) of the trace of a matrix:

$$tr(\gamma A + \delta B) = \gamma tr(A) + \delta tr(B) \;\;,\;\; \gamma, \delta \text{ constant} \tag{6}$$

$$tr(ABC) = tr(CAB) = tr(BCA). \tag{7}$$

We also make extensive of Moore-Penrose generalized inverses for non-square matrices in intermediate steps of our derivation.

$$\begin{aligned} Q_{\mathbf{K}} &= tr\left[Var(\hat{\boldsymbol{\beta}}|\mathbf{A},\mathbf{K})\right] \\ &= tr\left[E_{\mathbf{y}^*}\left(Var(\hat{\boldsymbol{\beta}}|\mathbf{y}^*,\mathbf{A},\mathbf{K})\right) + Var_{\mathbf{y}^*}\left(E(\hat{\boldsymbol{\beta}}|\mathbf{y}^*,\mathbf{A},\mathbf{K})\right)\right] \end{aligned}$$

Implicit in our statistical model (1) and (2) is the assumption that $[\hat{\boldsymbol{\beta}}|\mathbf{y}^*,\mathbf{A},\mathbf{K}] = [\hat{\boldsymbol{\beta}}|\mathbf{y}^*,\mathbf{K}]$. That is, the influence of landscape covariates on genetic differentiation ($\hat{\boldsymbol{\beta}}$) depends on the observed allele data ($\mathbf{A}$) only conditionally through $\mathbf{y}$, the distance to the nearest genetic boundary. Thus,

$$\begin{aligned} &= tr\left[E_{\mathbf{y}^*}\left(Var(\hat{\boldsymbol{\beta}}|\mathbf{y}^*,\mathbf{K})\right) + Var_{\mathbf{y}^*}\left(E(\hat{\boldsymbol{\beta}}|\mathbf{y}^*,\mathbf{K})\right)\right] \\ &= tr\left[E_{\mathbf{y}^*}\left(Var(\hat{\boldsymbol{\beta}}|\mathbf{y}^*,\mathbf{K})\right)\right] + tr\left[Var_{\mathbf{y}^*}\left(E(\hat{\boldsymbol{\beta}}|\mathbf{y}^*,\mathbf{K})\right)\right] \\ &= E_{\mathbf{y}^*}\left[tr\left(Var(\hat{\boldsymbol{\beta}}|\mathbf{y}^*,\mathbf{K})\right)\right] + tr\left[Var_{\mathbf{y}^*}\left(\hat{\boldsymbol{\beta}}_{y\mathbf{K}}\right)\right] \\ &\quad \text{where } \hat{\boldsymbol{\beta}}_{y\mathbf{K}} = [(\mathbf{K}X)'(\mathbf{K}\Sigma_{\mathbf{y}^*}\mathbf{K}')^{-1}(\mathbf{K}X)]^{-1}(\mathbf{K}X)'\Sigma^{-1}\mathbf{K}\mathbf{y}^* \\ &= E_{\mathbf{y}^*}\left[tr\left(Var(\hat{\boldsymbol{\beta}}|\mathbf{y}^*,\mathbf{K},A)\right)\right] + tr\left[E_{\mathbf{y}^*}\left(\hat{\boldsymbol{\beta}}_{y\mathbf{K}}\hat{\boldsymbol{\beta}}_{y\mathbf{K}}'\right) - E_{\mathbf{y}^*}(\hat{\boldsymbol{\beta}}_{y\mathbf{K}})E_{\mathbf{y}^*}(\hat{\boldsymbol{\beta}}_{y\mathbf{K}}')\right] \end{aligned}$$

$$= E_{\mathbf{y}^*}\left[tr\left(Var(\hat{\boldsymbol{\beta}}|\mathbf{y}^*,\mathbf{K})\right)\right] \tag{8}$$

$$+\, tr\left[E_{\mathbf{y}^*}\left(\hat{\boldsymbol{\beta}}_{y\mathbf{K}}\hat{\boldsymbol{\beta}}_{y\mathbf{K}}'\right)\right] \tag{9}$$

$$-\, tr\left[E_{\mathbf{y}^*}(\hat{\boldsymbol{\beta}}_{y\mathbf{K}})E_{\mathbf{y}^*}(\hat{\boldsymbol{\beta}}_{y\mathbf{K}}')\right] \tag{10}$$

We consider each of the terms (8)-(10) in turn. First, using generalized inverses, (8) becomes:

$$\begin{aligned} E_{\mathbf{y}^*}\left[tr\left(Var(\hat{\boldsymbol{\beta}}|\mathbf{y}^*,\mathbf{K})\right)\right] &= E_{\mathbf{y}^*}\left[tr\left([(\mathbf{K}X)'(\mathbf{K}\Sigma_{\mathbf{y}^*}\mathbf{K}')^{-1}(\mathbf{K}X)]^{-1}\right)\right] \\ &= E_{\mathbf{y}^*}\left[tr\left((\mathbf{K}X)^{-1}\mathbf{K}\Sigma_{\mathbf{y}^*}\mathbf{K}'(\mathbf{K}X)'^{-1}\right)\right]. \end{aligned}$$

If we have J realiztions of $\mathbf{y}^*|\mathbf{A}$, we could use Monte Carlo integration to approximate (8):

$$E_{\mathbf{y}^*}\left[tr\left(Var(\hat{\boldsymbol{\beta}}|\mathbf{y}^*,\mathbf{K})\right)\right] \approx \frac{1}{J}\sum_{j=1}^{J} tr\left((\mathbf{K}X)^{-1}\mathbf{K}\Sigma_{\mathbf{y}^*}\mathbf{K}'(\mathbf{K}X)'^{-1}\right). \tag{11}$$

Expression (9) can be written:

$$\begin{aligned} tr\left[E_{\mathbf{y}^*}\left(\hat{\boldsymbol{\beta}}_{y\mathbf{K}}\hat{\boldsymbol{\beta}}'_{y\mathbf{K}}\right)\right] &= E_{\mathbf{y}^*}\left[tr\left(\hat{\boldsymbol{\beta}}_{y\mathbf{K}}\hat{\boldsymbol{\beta}}'_{y\mathbf{K}}\right)\right] \\ &= E_{\mathbf{y}^*}\Big[tr\Big([(\mathbf{K}X)'(\mathbf{K}\Sigma_{\mathbf{y}^*}\mathbf{K}')^{-1}(\mathbf{K}X)]^{-1}(\mathbf{K}X)'\Sigma_{\mathbf{y}^*}^{-1}\mathbf{K}\mathbf{y}^* \\ &\quad \times\left([(\mathbf{K}X)'(\mathbf{K}\Sigma_{\mathbf{y}^*}\mathbf{K}')^{-1}(\mathbf{K}X)]^{-1}(\mathbf{K}X)'\Sigma_{\mathbf{y}^*}^{-1}\mathbf{K}\mathbf{y}^*\right)'\Big)\Big]. \end{aligned}$$

Utilizing generalized inverses and cancelling terms yields:

$$= E_{\mathbf{y}^*}\left[tr\left((\mathbf{K}X)^{-1}(\mathbf{K}\mathbf{y}^*)(\mathbf{K}\mathbf{y}^*)'(\mathbf{K}X)'^{-1}\right)\right]$$

which can be approximated by

$$\approx \frac{1}{J}\sum_{j=1}^{J} tr\left((\mathbf{K}X)^{-1}(\mathbf{K}\mathbf{y}^{*(j)})(\mathbf{K}\mathbf{y}^{*(j)})'(\mathbf{K}X)'^{-1}\right), \tag{12}$$

where $\mathbf{y}^{*(j)}$ is the j-th realization from $[\mathbf{y}^*|\mathbf{A}]$.

The third term (10) can be approximated by:

$$\begin{aligned} tr\left[E_{\mathbf{y}^*}(\hat{\boldsymbol{\beta}}_{y\mathbf{K}})E_{\mathbf{y}^*}(\hat{\boldsymbol{\beta}}'_{y\mathbf{K}})\right] &\approx tr\left[\left(\frac{1}{J}\sum_{j=1}^{J}G_{j\mathbf{K}}\mathbf{y}^{*(j)}\right)\left(\frac{1}{J}\sum_{j=1}^{J}G_{j\mathbf{K}}\mathbf{y}^{*(j)}\right)'\right] \\ &\qquad \text{where } G_{j\mathbf{K}} = [(\mathbf{K}X)'(\mathbf{K}\Sigma^{(j)}\mathbf{K}')^{-1}(\mathbf{K}X)]^{-1}(\mathbf{K}X)'\Sigma^{(j)-1}\mathbf{K} \\ &= tr\left[\frac{1}{J^2}\sum_{m=1}^{J}\sum_{n=1}^{J}G_{m\mathbf{K}}\mathbf{y}^{*(m)}\mathbf{y}^{*(n)'}G'_{n\mathbf{K}}\right], \end{aligned}$$

which yields the following, after expanding and cancelling terms:

$$= tr\left[\frac{1}{J^2}\sum_{m=1}^{J}\sum_{n=1}^{J}(\mathbf{K}X)^{-1}(\mathbf{K}\mathbf{y}^{*(m)})(\mathbf{K}\mathbf{y}^{*(n)})'(\mathbf{K}X)'^{-1}\right]. \tag{13}$$

Thus, combining (11), (12), and (13) yields an approximation for $Q_{\mathbf{K}}$.

$$\begin{aligned} Q_{\mathbf{K}} &= tr(Var[\hat{\boldsymbol{\beta}}|\mathbf{A},\mathbf{K}]) \\ &\approx tr\Big[(\mathbf{K}X)^{-1}\Big(\frac{1}{J}\sum_{j=1}^{J}\mathbf{K}\Sigma^{(j)}\mathbf{K}' + \frac{1}{J}\sum_{j=1}^{J}(\mathbf{K}\mathbf{y}^{*(j)})(\mathbf{K}\mathbf{y}^{*(j)})' \\ &\qquad - \frac{1}{J^2}\sum_{m=1}^{J}\sum_{n=1}^{J}(\mathbf{K}\mathbf{y}^{*(m)})(\mathbf{K}\mathbf{y}^{*(n)})'\Big)(\mathbf{K}X)'^{-1}\Big] \\ &= tr\Big[(\mathbf{K}X)^{-1}\mathbf{K}\mathbf{Z}\mathbf{K}'(\mathbf{K}X)'^{-1}\Big], \\ &\quad \text{where } \mathbf{Z} = \frac{1}{J}\sum_{j=1}^{J}\left(\Sigma^{(j)} + \mathbf{y}^{*(j)}\mathbf{y}^{*(j)'}\right) - \frac{1}{J^2}\sum_{m=1}^{J}\sum_{n=1}^{J}\mathbf{y}^{*(m)}\mathbf{y}^{*(n)'}. \end{aligned}$$

This has a distinct advantage in that $\mathbf{Z}$ can be calculated once, and then used at each iteration of the optimal sampling switching algorithm. In practice, we have found that stability is vastly improved by removing the need to calculate the generalized inverses of the rank-deficient matrices $(\mathbf{K}X)$ and $(\mathbf{K}X)'$, leading to an equivalent and more stable formulation:

$$Q_{\mathbf{K}} = tr\Big[\big((\mathbf{K}X)'(\mathbf{K}\mathbf{Z}\mathbf{K}')^{-1}(\mathbf{K}X)\big)^{-1}\Big]. \tag{14}$$

This final formulation is a highly efficient and stable design criterion that facilitates obtaining quasi-optimal samples of genetic data that maximize the information gained about the relationship between spatial gene flow and hypothesized landscape genetic boundaries.

REFERENCES

Harville, D.A. (2008), “Matrix Algebra from a Statistician’s Perspective”, Springer, New York, New York, USA, 634p.