Kezia R. Manlove
Department of Mathematical Sciences
Montana State University

May 15, 2009

A writing project submitted in partial fulfillment
of the requirements for the degree

Master of Science in Statistics

APPROVAL

of a writing project submitted by

Kezia R. Manlove

This writing project has been read by the writing project director and has been found to be satisfactory regarding content, English usage, format, citations, bibliographic style, and consistency, and is ready for submission to the Statistics Faculty.

Date	Mark Greenwood Writing Project Director

Investigation of a Bayesian Hierarchical Model for Landscape Genetics

Kezia R. Manlove

May 15, 2009

Contents

Abstract

Existing genetic data contain a wealth of untapped information of value to ecologists. One major area of ecological interest is the use of spatially point-referenced genetic data to track individuals over a landscape, in an effort to delineate among separate breeding groups. In this paper, we investigate a Bayesian model for inferring population structure and ranges based on point-referenced genetic data. We assess the capacity of the model to correctly assign individuals to populations, examine the model's ability to accurately describe the genetic structure of the original populations, and observe the accuracy of spatial mappings of the model-generated populations. Additionally, we discuss limitations on the model's scope of inference imposed by making (or not making) certain assumptions.

1 Introduction

Genetic data analysis is rapidly emerging as a valuable tool in population and community ecology. By tracking individuals genetically, researchers can better examine how population allele frequency distributions differ with landscape covariates. Identifying such relationships subsequently leads to improved classification of individuals to populations and expanded knowledge about gene flow patterns over space.

While techniques for dealing with classification on a strictly genetic basis are abundant (e.g. Cavalli-Sforza (1971)), techniques for identification of barriers and generation of maps for population epicenters have emerged only recently. A relatively new Bayesian hierarchical model for landscape genetics conditions genetic data on point-referenced spatial location data to infer the number, genetic structure, and spatial organization of populations over a landscape (Guillot, 2005). The Bayesian model allows researchers to address questions concerning how may populations are present, the rate at which these populations are drifting apart from a common ancestral population, the ranges occupied by each population, and from which population an individual arises. In this paper, we investigate this model's performance over three criteria (genetics, spatial distribution of populations, and individual assignment) via simulation.

1.1 Traditional Methods

Historically, genetic methods for classifying individuals to populations included use of the Expectation-Maximization algorithm (e.g. Millar (1987)), multinomial likelihood approaches (e.g. Cavalli-Sforza (1971)), traditional methods of multivariate classification analysis and mixture modeling. Methods for examining the relationship between geographic covariates and populations' home-ranges include Mantel and partial Mantel methods (eg. Arnarud 2003, Banks

et al. 2005, Hitchings and Beebee 1997, Vignieri 2005), implementation of Moran's I (Arnaud 2003), use of correlograms (Banks *et al* 2005), simulated annealing (Banks *et al.* 2005), use of F-statistics (Hitchings and Beebee 1997), Monmonier's algorithm (Liepelt *et al.* 2002), examination of spatial autocorrelation (Pfenninger 2002), PCA and kriging (Piertney *et al.* 1998), neighbor-joining trees (Poissant *et al.* 2005), assignment tests, and information theoretic approaches (Roach *et al.*, 2001) (list from Storfer *et al.*, 2007).

Typical approaches for spatial analysis like those listed above fall short in landscape genetics for a variety of reasons, but mainly because genetic data is typically observed as a multilocus genotype. Therefore, the observations made on a single individual are not a direct measurement of a spatial process, but rather, they are only meaningful when considered in conjunction with other individuals' or populations' genetic structures. The challenge is that multilocus data cannot readily be expressed as point data, so methods typically used for point-process spatial analyses fall short. An additional major drawback to commonly used population genetics statistics (eg. F_{ST}, Nei's D) is the requirement of an *a priori* delineation of populations. Since determining how many populations are present is often a study objective, such *a priori* assumptions can compromise a study's scientific integrity.

Bayesian approaches like those presented by Pritchard (2000) and Falush (2003) allow for relaxation of the delineation requirement. Use of Voronoi polygons to introduce spatial information into population delineation (through an application of Monmonier's algorithm) was first put forth by Manni *et al.*, 2004. Guillot's model extends the work on Manni, Pritchard, and Falush through the implementation of a spatial assignment algorithm, which incorporates Falush's (2003) genetic modeling structure.

1.2 Genetic Data

Genetic data are categorical and multivariate by nature: an individual's genotype consists of a set of genetic markers (loci) located at specific positions on the genome. Locations can be chosen to be independent, either by using loci on different chromosomes or by selecting loci that are located far enough apart on the same chromosome that recombination renders them effectively independent.

The allele observed at each location is the outcome of a phenomenon that is assumed to be random at several levels: Under an assumption of random mating, the parents' genomes are considered random draws from a population, and the particular allele that each parents passes to its offspring is also random.

Because of this doubly-random process, we can think of an observed genotype as being an outcome (a state) of a nested Markov process. The overall consequence is that a genotype is essentially a realization of a string of independent (within a population) categorical random variables.

2 Guillot's Bayesian Hierarchical Model

Spatial models have been extensively developed for quantitative data, on the premise that individuals located in spatial proximity to one another are more likely to be similar than individuals

who are located far apart. When spatial autocorrelation between individuals is not accounted for, impacts of various environmental features on the genetic structure over a landscape are generally overestimated, since spatial proximity acts as a lurking variable. Similar habitats are spatially clustered, as are individuals from the same population. By failing to include spatial proximity between individuals from the same population, the impact of environmental covariates present in those locations may appear more pronounced than it actually is. This is a particularly relevant concern in conservation ecology, where the specification of individuals to a particular habitat is of paramount importance for targeting specific habitat types for preservation.

There are two main reasons for adopting a Bayesian approach for detecting spatial discontinuities in allele frequencies over the landscape:

1. the Bayesian approach allows us to explicitly examine the statistical uncertainty associated with each parameter of interest and

2. prior information (i.e. the spatial clustering of individuals over the landscape) can be incorporated into the model as an additional information source (Falush (2003)).

Markov chain Monte Carlo (MCMC) techniques have been used extensively in genetics research over the last ten years (e.g. Rannala and Mountain, 1997; Pritchard *et al.*, 2000, Falush *et al.*, 2003). The premise of the genetics models is that observed individuals form a representative mixture from a set of populations of potentially unknown size. An individual can come entirely from one population, or arise from two or more populations in a scenario referred to as ad-mixture (Falush *et al.*, 2003). The role of admixture individuals in Guillot's model is discussed later. All populations in the study region can be identified by a unique set of allele frequencies at each examined location on the genome. Within-population allele frequencies are assumed to be in Hardy-Weinberg equilibrium, and linkage equilibrium is also assumed for all examined loci. Hardy-Weinberg equilibrium refers to situations where the following conditions are met (Freeland 2005, pg 68):

1. within-population mating is random.

2. the alleles of interest are not under selection.

3. the effects of migration and mutation are negligible.

4. population size is effectively infinite.

5. the alleles segregate via Mendelian inheritance.

Linkage disequilibrium can occur at two levels, a population level and a within-genome level. At the population level, genetic drift of two current populations from a common ancestral population results in genetic differentiation across all alleles. The allele frequency distributions across all loci within a given population will be correlated with one another even if those loci assort completely independently during meiosis. Population linkage disequilibrium refers to this genome-wide, correlated divergence between several populations arising from a common ancestral group.

At the within-genome level, linkage disequilibrium refers to dependence in assortment. That is, two loci located close together on the same chromosome may occur more commonly in some

pairings (that is, pairings where crossing over didn't occur between the two loci) than pairings where crossing over split the two loci. Consistent recombination of one locus being associated with recombination of another other is within-genome linkage disequilibrium (Freeland 2005, pg 76). Additional background linkage disequilibrium may also exist, but it is not dealt with explicitly in the models examined here.

Under the assumptions outlined above, the likelihood of a given genotype arising in a given population is simply the product of the probability of seeing each observed allele in the given population

$$L(a|j) = \prod_{i=1}^{2l} a_{ij} \tag{1}$$

where a is the genotype being examined, j is the population of interest, and l is the number of loci examine.

The net implication of these assumptions is that the probability of a given genotype, conditional on the genotype originating from the j^{th} population, is the probability of obtaining a given genotype from population j is the product of the probabilities of seeing each two-allele combination at each locus within a certain population. This relationship was thoroughly described by Rannala and Mountain (1997) and is summarized below.

Bayesian analysis of genetic data was first introduced in Rannala and Mountain (1997), who took advantage of the Dirichlet-Multinomial relationship to express observed genotypes as draws from a Dirichlet prior. In their approach, The prior probability density of allele frequencies at the j^{th} locus in the i^{th} population is as follows:

$$Pr(x_{ij}) = \prod_{k=1}^{k_j} \frac{{x_{ij}}^{(1/k_j)-1}}{\Gamma(\frac{1}{k_j})}. \tag{2}$$

After conditioning on the sampled alleles from a given population, the posterior probability density of allele frequencies is

$$Pr(x_{ij}|n_{ij} = \frac{Pr(n_{ij}|x_{ij})Pr(x_{ij})}{Pr(n_{ij})} \tag{3}$$

where

$$Pr(n_{ij}|x_{ij}) = \left(\begin{array}{c} n_{ij} \\ n_{1ij}, ..., n_{k_j ij} \end{array} \right) \prod_{k=1}^{k_j} {x_{hij}}^{n_{hij}}$$

and

$$Pr(n_{ij}) = \prod_{k=1}^{k_j} \frac{\Gamma(n_{hij} + \frac{1}{k_j})}{\Gamma(n_{hij} + 1)\Gamma(\frac{1}{k_j})}.$$

Equation (3) then simplifies to

$$Pr(x_{ij}|n_{ij}) = \Gamma(\theta) \prod_{k=1}^{k_j} \frac{{x_{hij}}^{\theta a_h - 1}}{\Gamma(\theta a_h)} \tag{4}$$

where $\theta = n_{ij} + 1$ (Rannala and Mountain, 1997). This prior-posterior relationship forms the groundwork on which most Bayesian analysis of population genetics is conducted (e.g. Pritchard *et al.* (2000), Falush *et al.* (2003), Corander *et al.* (2003)). This is a modification of the parameterization for common Dirichlet distributions.

It is worth noting that for very small populations, Hardy-Weinberg equilibrium is lost due to increased inbreeding levels. This problem was recognized in Francois *et al.* (2006)'s follow-up to Guillot *et al.*'s model.

In order to extend this genetic formulation to include spatial variables measured on each individual, consider samples to be draws from a joint probability distribution containing allele frequency distributions from the population of origin, conditional on the genotypes of sampled individual. This probability distribution cannot be sampled from directly, so the samples are derived using a Markov chain (Pritchard *et al.*, 2000). Inference using Monte Carlo Markov Chain (MCMC) algorithms is described in section 2.3.1.

2.1 Entry of Information into the Model

Model inputs are genotypes and point-indexed spatial locations for individuals. Spatial clusters ("cells") of individuals with similar genotypes are identified. Depending on how distinct allele frequencies within a cell are from the other cells, a cell may be identified as its own population or may be grouped with other cells with similar allele frequencies elsewhere to form a single population. A prior favoring fewer cells makes it more likely that cells within close spatial proximity to one another are collapsed into a single population. Locations of clusters are driven by location of individuals on the landscape, and within-cluster individual relatedness.

The implication here is that a continuous group of individuals actually composed of two adjacent, but separate, populations, could be identified as two separate cells (based on their distinct allele frequency distributions) as opposed to one cell based on their spatial locations. Using the genetic data alone to identify individuals to populations would ignore the information that the two populations are located side-by-side. Researchers obtain the most complete picture of population distribution over the landscape by using both the spatial and genetic information. Additionally, using the spatial model allows for identification of spatially discontinuous populations.

The degree to which spatial information drives the model is controlled by a Poisson random variable, which defines the number of cells present on the landscape. A Poisson point process over the landscape lays out a set of points which form the centers of the cells. Then, each point within the space is assigned to the cell whose center is nearest. This forms a Voronoi tiling of the sampled region, which will be discussed in the next section. Higher estimated values of the Poisson parameter indicate more cells and greater spatial mixing among populations.

For example, consider a situation where only two clusters are identified and define two distinct populations. In this setting, the model would suggest that only two populations exist, and that spatial mixing between those two populations is very low, so very few cells are needed to describe the spatial organization of populations over the landscape. In a scenario with high spatial mixing, the model incorporates more and more cells, therefore increasing the estimate of the Poisson parameter, until each individual occupies its own cell in a homogenously mixed setting.

2.2 Hierarchical Spatial Model

The model assumes that a representative sample is drawn from panmictic populations separated by geographic borders (i.e. rivers, mountain ranges, etc.). The initial stage of the model is a specification of the spatial organization of the populations. Statistical representation of the genetic properties of the various populations are then specified conditional on the spatial organization. Formally, we sample from joint posterior distribution of the parameterization given space (t) and genetics (z),

$$\pi(\theta|t,z) \tag{5}$$

where $\theta = (K, m, u, c, d, f, f_A, s)$ The likelihood can be written as

$$\pi(t,z|\theta) = \pi(t|\theta)\pi(z|t,\theta) = \pi(t|\theta)\prod_{i=1}^{n}\prod_{l=1}^{L}\pi(z_{i,l}|\theta). \tag{6}$$

Inference about the elements of θ are then made through the posterior distribution, $\pi(\theta|t,z)$. This hierarchy is laid out in Guillot *et al.*, (2005).

Consider K different populations occupying a spatial domain, Δ, such that Δ can be partitioned into K different subdomains, each occupied exclusively by a single population. Let Δ_K be the spatial domain of population K. Each population's subdomain is approximated by a union of convex polygons. This assumption does not limit the shape of the population's subdomain, since any shape can be approximated in this manner as long as an adequate number of polygons is used.

The centers of these polygons are modeled as random variables distributed uniformly over the entire spatial domain. This is analogous to saying that a homogeneous Poisson point process with some realization $(u_1, u_2, \ldots, u_t)$ is present over the landscape. Each point in the process defines a set A_i around it, such that A_i includes all points closer to u_i than to any other u. Each set A_i is then a convex polygon, and the set of all $A_i s$ is a Voronoi tessellation of the region, Δ. Assume that each A_i contains individuals from only one population. This allows us to label each A_i with a population, effectively "coloring" that Voronoi tile.

Now, consider a set of individuals distributed over the partitioned landscape. An individual is assigned to the population label for the set A_i in which it is located.

2.3 Full Bayesian specification

The overriding assumption of this model is that some spatial dependence exists among individuals spread over a landscape. If no spatial dependence is present, Guillot's model should default to Falush's genetic model, since under no spatial dependence, each indidivual should get its own cell in the posterior, λ should then be close to the number of indidivuals sampled, and individual locations should be uniformally distributed over the sampled region. The model relies on *a priori* information on individual locations and spatial organization. Guillot's model generates estimates of the ancestral and current population allele frequency distributions, the drift rate for each current population from the ancestral population, the spatial organization of subdomains over the landscape (as characterized by the intensity of the Poisson point process, and the location and population assignment of each cell), and error in the observed locations of

individuals based on recorded individual locations and genotypes. Priors for each unobserved quantity are tabled below.

Parameter	Description	Prior	Hyper-Prior	Subjective Input(s)
K	Number of Populations	Discrete Uniform	-	Usually, minimum = 1; maximum = number of individuals sampled
m	Number of Tiles in Voronoi tessellation	Poisson	Uniform on *Lambda*	Hyper-prior bounded above by the number of individuals sampled
$(u_1, ..., u_m)$	Location of center of each cell	Uniform over the entire spatial domain	-	
c	Population ("color") to which a tile is assigned	Uniform	-	Takes on values between 1 and K
Δ	Population subdomains	-	-	Union of all Voronoi tiles assigned to a particular population
d	Drift (degree of genetic differentiation among existing populations)	Uniform(0, 1) or *Beta* (2, 20)	-	Selection of the particular prior can be driven by knowledge of genetic differentiation between populations on the landscape
f_A	Frequencies of ancestral alleles	Dirichlet (1, , 1)	-	
f	Frequencies of current alleles	Dirichlet (1-d/d, , 1-dk/dk)	-	Driven by parameterization of d
s	Actual (as opposed to recorded) individual location	Suitable parametric distribution for *epsilon* in t = *s + epsilon*	-	Distributional form for *epsilon*

2.3.1 Markov chain Monte Carlo inference

In order to obtain the joint posterior distribution of the parameters, a Monte Carlo Markov chain (MCMC) algorithm is used. The objective of the MCMC algorithm is to sample from the joint posterior distribution of the parameters given space and genetics,

$$\pi(\theta|t, z) \tag{7}$$

where $\theta = (K, m, u, c, d, f, f_A, s)$, a vector containing all unknown parameters defined in Table 1. Then, the likelihood of the data (t, z) can be expressed as in equation (6) above.

Through the Monte Carlo Markov chain simulation we generate a Markov process in the space of the parameter vector, θ, which converges to the joint posterior distribution for the parameters. Two important assumptions implicit in using the Markov chain simulation are that

1. the stationary distribution of the Monte Carlo Markov chain is specified to be the posterior, and
2. the simulations are run long enough that the distribution of draws is close to the specified stationary distribution.

In Guillot *et al.*'s MCMC algorithm investigated here, starting values are randomly initialized from the prior.

The defining feature of Monte Carlo Markov chain simulation is that samples are drawn sequentially, so that the distribution of the sample draws depends solely on the last value drawn. The method works because the approximate distributions from which samples are drawn are improved with each step of the chain (that is, they converge to the target distribution). This differentiates MCMC algorithms from importance sampling, where the distribution from which draws are made remains the same throughout the entirety of the sampling process (Gelman, 1995) .

2.3.2 Implementation of Gibbs Sampling

Gibbs sampling is used when the complete conditional probability distribution of the variable of interest (conditioning on all other parameters in the model) can be stated explicitly, and can be written in closed form, but the posterior distribution cannot be expressed in a closed form. For example, in Guillot *et al.*'s MCMC, a Gibbs step is appropriate when updating the present allele frequencies from the ancestral allele frequencies because the conditional distribution of present allele frequencies given ancestral allele frequencies and drift parameters can be written out in closed form. That is, conditional on ancestral alleles and drift parameters, we can explicitly express the probabilities associated with the present allele frequency distributions. Because the complete conditional distribution can be stated explicitly (although a closed form of the posterior cannot), this is a case where Gibbs sampling is appropriate.

2.3.3 Implementation of Metropolis-Hastings Updates

In general, the Metropolis-Hastings algorithm is used to get a sample from a distribution. One way it can be implemented is to sample from a complete conditional distribution within a Gibbs sampler. Metropolis-Hastings updates are used when the complete conditional distribution cannot be explicitly stated, but instead must be simulated as a set of draws from the parameter space describing the conditional distribution. For each update, the algorithm proposes a new draw of a parameter, and the density for the new parameter value is compared to the current value evaluated at the last parameter value. If the density at the proposal value exceeds the current one, then the current value is rejected in favor of the proposal value. Then in the next step the old proposal value is treated as the current value. If the density at the current value exceeds the proposal, then the current value is retained and a new proposal in the next update

is compared to the current value (Gelman, 1995). In Guillot *et al.*'s model, Metropolis-Hastings updates are used for the drift parameters and allele frequencies, parameters for which complete conditional distributions cannot be stated in a closed form.

Inference about the elements of θ will be made through an investigation of its posterior distribution, $\pi(\theta|t,z)$. A hybrid Gibbs sampler (a combination of Metropolis-Hastings and Gibbs updates for various parameters in the posterior) based on sequential updates of blocks of parameters is used. All parameters are randomly initialized from the prior, and subsequent moves are made as follows:

1. Update drift parameters, d (Metropolis-Hastings update)
2. Update the ancestral allele frequency distribution, f_A (Metropolis-Hastings update)
3. Update the current allele frequency distribution, f (Gibbs update)
4. Update the coloring of each Voronoi tile, c (Metropolis-Hastings update)
5. Update the location of each tile center, u_j (Metropolis-Hastings random-walk update)
6. Update the error term associated with each individual, s (Metropolis-Hastings random-walk update)
7. Add or discard a tile (randomly choose between a birth or death of a tile, with equal probability. If a birth is chosen, propose a new random point in the current state, whose location is drawn from the uniform prior. The coloring of this tile is drawn from the discrete uniform prior on population labeling of tiles.)
8. Split or merge existing populations

Convergence of the MCMC on the stationary distribution (the posterior) follows from balance, irreducibility, and aperiodicity. Convergence was taken to have been achieved based on the diagnostic plots show below for drift, and others examined for the other parameters.

3 Investigation of Guillot *et al.*'s Model

Ecologists hoping to implement Guillot *et al.*'s model are immediately presented with several dilemmas. They must decide whether they will use a model that assumes the allele frequency distributions for their assorted populations of interest are independent, or they will treat allele frequencies as correlated (since they actually arose from a common ancestral population). Guillot does not allow estimation of K in the correlated version of his model, due to consistent overestimation. Thus, the ecologist is presented with a second dilemma: what is the tradeoff between using the correlated model with K set *a priori* and using the uncorrelated model, which doesn't accurately model the actual biological process, but has the capacity to estimate K more accurately? If they choose to fix K, how do they select an appropriate value for K?

What are the implications of using the model that includes no spatial information, as opposed to including spatial information in the prior, in either the correlated or uncorrelated cases? How does the model deal with admixture individuals who arise from multiple populations of origin?

Figure 1: MCMC Diagnostics

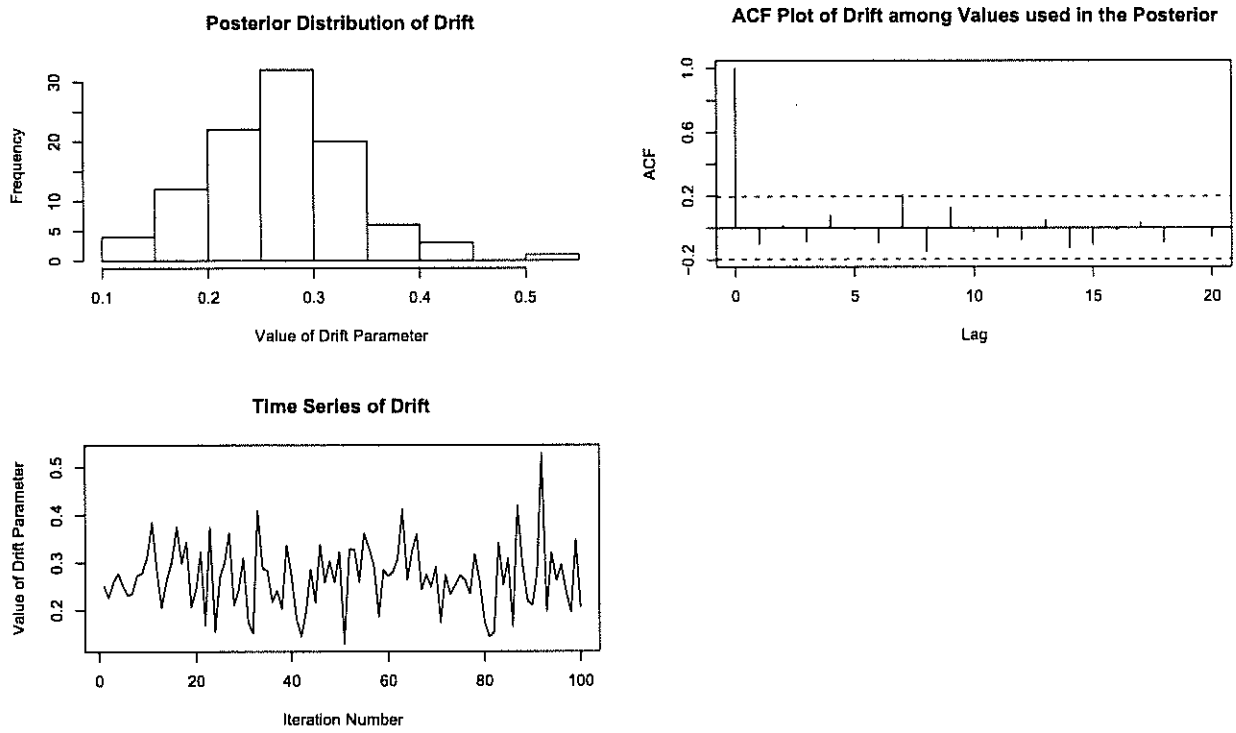

Figure 2: MCMC algorithm diagnostics for Guillot *et al.*'s algorithm.

Finally, what limitations to scope of inference does selection of a certain model impose? The objective of the remainder of this project is to address each of these questions.

3.1 Simulation Study

The impact of using the correlated vs. uncorrelated framework is of interest for several reasons. The uncorrelated model allows for estimation of K from the data, while the correlated model requires a fixed K. However, the uncorrelated model does not represent the process by which genetic data arises as accurately as the correlated model. Of the models we examined, only the correlated model with spatial information uses the spatial hierarchy laid out in Guillot *et al.*'s work. In order to compare modeling capacities for the two models, we conducted a series of simulations. Data were simulated using five models:

1. The correlated structure and a known number of populations (Correlated K)
2. The uncorrelated model with K set *a priori* (Uncorrelated K)
3. The uncorrelated model with K estimated through the model (Uncorrelated, no K)
4. The correlated model relying on genetics alone (Correlated, no location)
5. The correlated model with spatial information included (Correlated with location).

The models were compared by maps they generated of the original populations, through use of a statistic measuring assignment accuracy, and in terms of how closely model and parametric allele frequency distributions overlapped.

3.2 Model Comparison

The objective of the Guillot model is to accurately define genetic populations on a landscape, both in terms of population ranges and population genetics. Ranges can readily be compared via maps of the true population ranges and the ranges inferred by the model using simulated data. Relatedness among the true populations can be compared to relatedness among the model populations via Wright's F_{ST}. A useful model should exhibit between-population relatedness that approximates relatedness in the true populations, produce maps that correspond well to true maps of population distributions over space, and attain a high rate of common assignments between the true populations and the model populations. Failure of the model in each of these criteria (genetics, space, and assignment) will be examined separately.

3.3 Estimating K, the Number of Populations

Estimation of the number of populations present on the landscape is fundamentally linked to how related two groups of organisms are "allowed" to be before they are considered to be members of the same population. In this model, that decision is guided by a combination of information from the observed data (in the form of clusters of indidivuals with similar allele

combinations) and the prior distribution placed on the drift parameters (referred to as F in Falush *et al.*, 2003, to indicate its relationship to Wright's F_{ST}).

If the prior on drift favors low values of F_{ST}, then the cut-off point for distinct populations is lowered, and groups that have similar (but not identical) allele frequency distributions are more likely to be defined as separate populations. Falush *et al.* (2003) emphasize the relationship between the drift parameters and the effective population size of a given population since the time that population diverged with the common ancestral population. Thus large values of the drift parameter correspond to smaller effective population sizes. Falush *et al.*, relying on the advice of Nicholson *et al.* (2002) suggest the use of a truncated gamma as a prior for the drift parameters, whereas Guillot *et al.* (2005) suggest a *Beta* (2, 20). Falush emphasizes that the "harshness" (that is, the weight the prior places on low values of drift) manifests itself as the degree to which the model differentiates between populations. So, a harsh prior on low values of drift corresponds to settings where F_{ST}s are low, which is appropriate when there exists strong information that populations are closely related (Falush *et al.*, 2003).

Guillot *et al.*'s correlated no-location model tends to overestimate the number of populations present on the landscape, to the degree that Guillot *et al.* advise users to fix K from the outset In an effort to avoid relying strictly on expert opinion, we experimented with using the uncorrelated model to estimate the number of populations present, and then applying that estimate as an *a priori* parameter for calculating the correlated spatial model. This approach failed: the number of populations estimated by the model increased steadily with increasing genetic differentiation between the populations.

We modified our initial approach by estimating K using the uncorrelated model, fitting the correlated spatial model, and then combining populations based on minimal pairwise F_{ST} values until the correct number of populations was reached. This approach worked well in some cases, but poorly in others. In some situations, the additional populations are fragments of the actual populations, so combining populations with the smallest pairwise F_{ST}s worked well. However, in some cases the additional populations are genetic intermediaries between the two actual populations. In this case, all pairwise F_{ST}s are very similar, so selecting two populations to combine becomes somewhat arbitrary. Since no acceptable solution was readily available for estimating K, we advocate using expert opinion to guide *a priori* selection of the K parameter.

3.4 Justification for Overlooking the Admixture Model

Guillot *et al.* discourage using Falush *et al.*'s model which allows for admixture (the possibility that a single individual arose from multiple populations). Falush *et al.*'s admixture model did not perform well in Guillot *et al.*'s scenario where genetic data was conditioned on location, an unsurprising result when one considers the objectives of Guillot *et al.*'s model as compared to an admixed model.

Decisions about the inclusion or exclusion of admixed individuals in models speaks to the ongoing discussion of what exactly a biological model is (see, for example, Waples and Gaggiotti, 2006). Guillot *et al.*'s model tries to draw boundaries among discretely breeding groups in order to examine what boundaries exist on the landscape, thus a population is taken to be a breeding group. If admixture is occurring, then by definition the groups determined by the model are not breeding discreetly. Furthermore, inclusion of admixture is difficult to model from a strictly

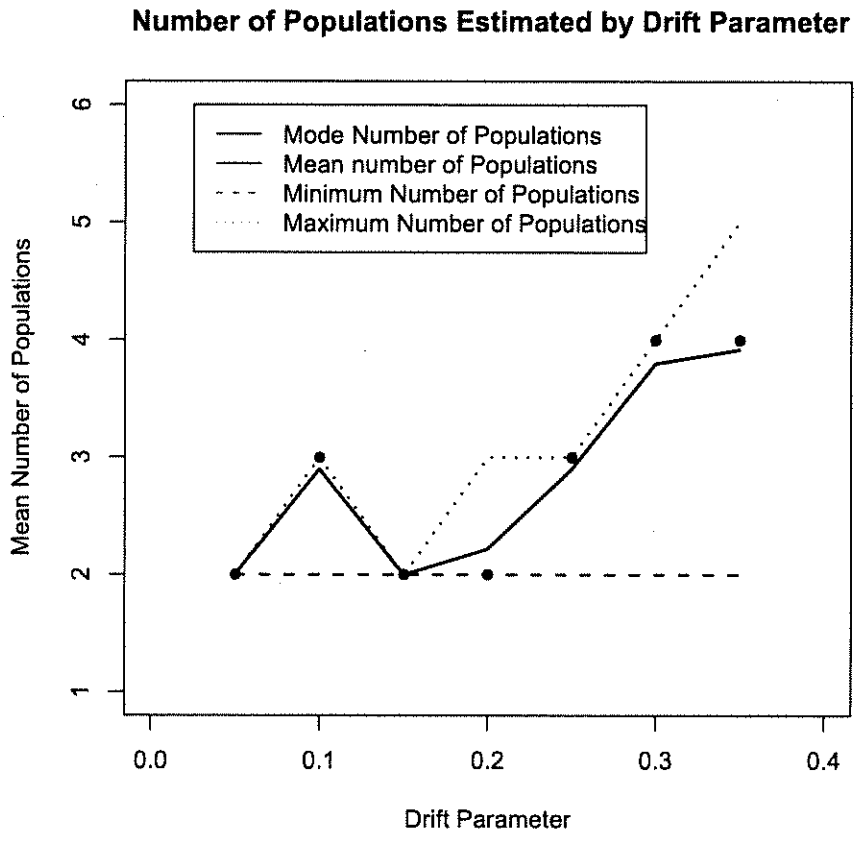

Figure 3: Number of populations estimated by Guillot *et al.*'s model. Note the steady increase in estimated population number as the populations become more and more closely related.

genetic perspective under the parameters defined above because it presents a ridge-type scenario: the same data could be interpreted as two highly diverged populations with lots of admixture or two subtly diverged populations with no admixture. Those two situations would produce (nearly) indistinguishable genetic data, so the model wouldn't be unique. This problem alone is enough to drive researchers away from using the admixed approach to modeling genetic data at all, let alone over space.

3.4.1 Genetics

The major feature separating the uncorrelated model from the two correlated ones is that the uncorrelated model makes no assumption about populations drifting from a common ancestral population. As a result, closely related populations may be virtually indistinguishable from one another in terms of allele frequency distributions at a single locus. It is combinations of alleles across several loci that make one population identifiable from another. In the uncorrelated model, the covariance across loci is taken to be negligible. Although two populations may be a certain distance apart at each locus, there is no multi-locus clustering of alleles, so populations that are measurably different when taken in the multivariate, correlated context appear identical in the uncorrelated model. By contrast, once population differentiation passes a certain threshold value (in Figure 3, that value appears to be roughly .22), the allele frequency distributions within a single locus are distinct enough to distinguish between populations, so the uncorrelated model starts to perform as well as its correlated counterparts. These results are in line with results reported in Falush *et al.*, 2003.

3.4.2 Individual Assignment

The decision about which model to use (correlated with and without space, uncorrelated with and without allowing the number of model populations to vary) will frequently be driven by questions about how well the model assigns individuals to populations. At first glance, this seems like a straight-forward question of assignment rates, but since the populations defined by the model are not necessarily labeled the same as the true populations that exist on the landscape, running a traditional assignment assessment using a classification procedure is inadequate.

An alternative approach which does not rely on assuming that model populations are the same as true populations is to compare matrices describing which individuals are assigned to the same population in the model to block diagonal matrices describing whether individuals were simulated from the same parametric population. The distance between the parametric matrix and the model matrix provides a measure of how well individuals are assigned: very low distance between the two matrices indicates good assignment; high distance indicates poor assignment. A measure using this basic idea structure was defined in Francois *et al.*, 2006. It is the error

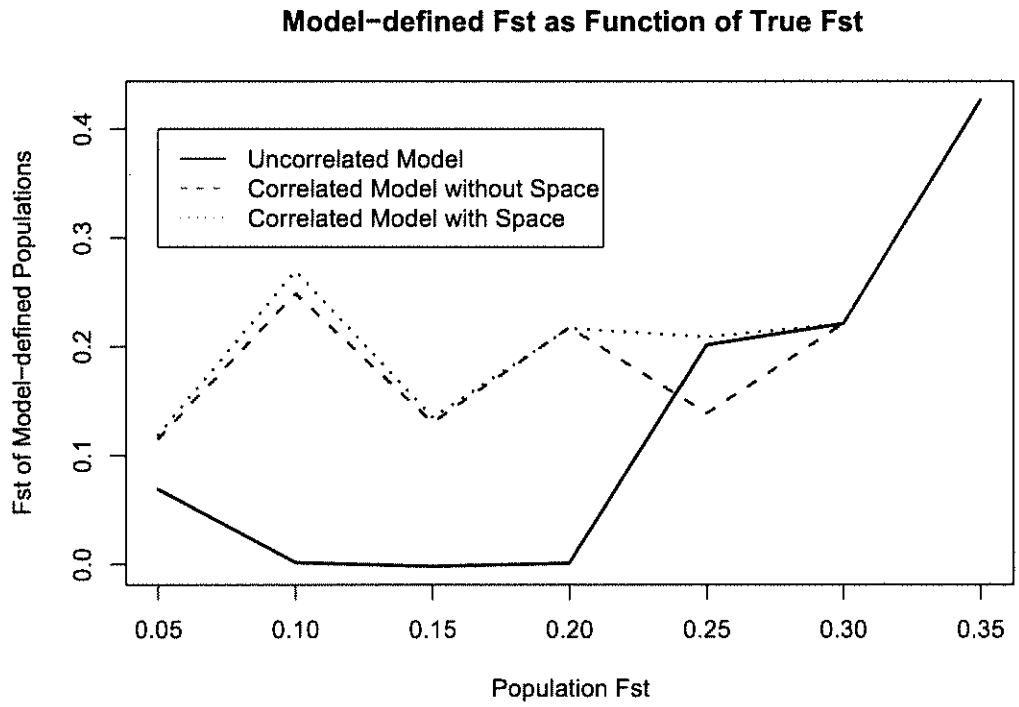

Figure 4: F_{ST} accuracy of model estimated populations as a function of the true population F_{ST}s. Note that a model that perfectly encapsulates the true populations would exhibit a slope of 1.

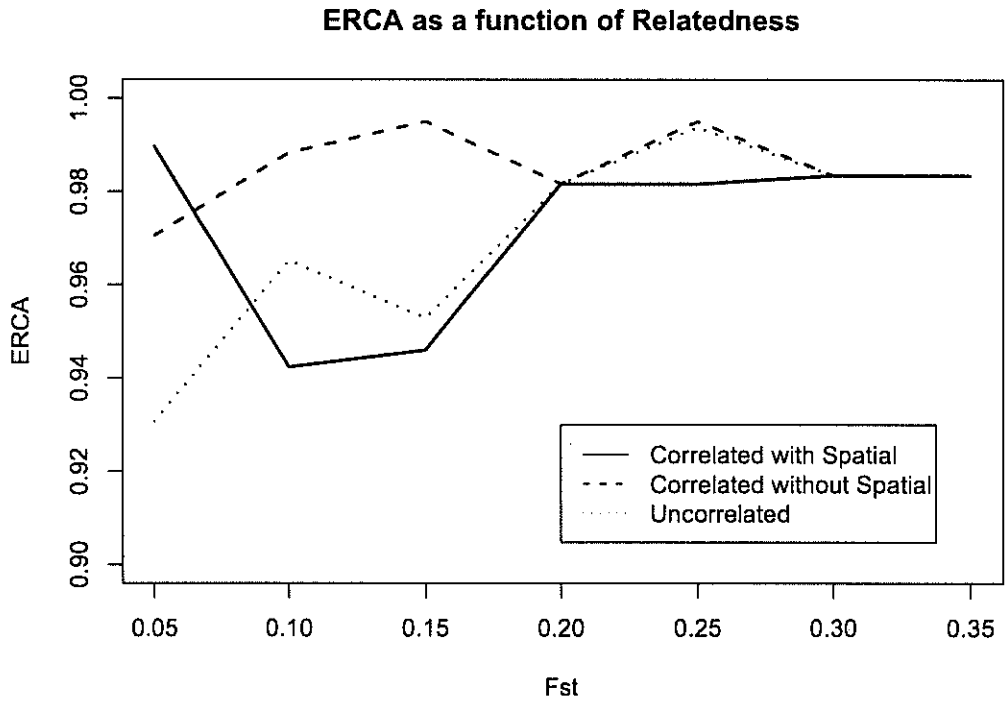

Figure 5: ERCA for the various models as a function of F_{ST}. ERCA is expected to increase with increasing F_{ST} since inviduals from populations that are less closely related should be more genetically distinct (thus easier to assign) than individuals from closely related populations.

rate in coassignment (ERCA), and is

$$ERCA = \frac{2}{n(n-1)} \sum (1 - \delta_{x_{i,j}, x-hat_{i,j}}) \tag{8}$$

where $\delta_{x_{i,j}, x-hat_{i,j}}$ is the Kronecker product of indicator variables for whether the i^{th} and j^{th} individuals were in the same simulated population ($x_{i,j}$) and whether the i^{th} and j^{th} individuals were in the same model population ($x - \hat{x_{i,j}}$). See Figure (4) for ERCA as a function of relatedness for the various models examined here.

3.4.3 Spatial Distribution on the Landscape

Complexity of the maps generated by the different models are best compared by simply examining the maps. Maps of posterior probabilities of belonging to one population are shown for three models: the uncorrelated model with K fixed to be two, the correlated model without spatial information, and then correlated model with spatial information.

3.5 Impact of Model Choice on Utility

Each of the four models examined contains slightly different assumptions. In the uncorrelated model when the number of populations is fixed, the genetic model relies heavily on expert opinion about the number of populations present. Assumptions about K limit the clustering patterns available for the data, so all inference must be made conditional on K being equal to the set value. By contrast, the models that are allowed to estimate K are not limited to inference conditioned on K; their inference extends to all possible values of K (however, these inferences are somewhat limited in their accuracy depending on the prior placed on K and that prior's influence relative to the data). The correlated model generates data in a biologically realistic manner, giving rise to correlation across multiple loci that may not be accounted for in the uncorrelated model, thus it may out-perform the uncorrelated model at low F_{ST}s. However, further consideration of the stochastic processes underlying constant migration vs. recent population divergence is necessary to provide strong biological evidence to corroborate this suggestion.

One great strength of the spatial model is that by conditioning on space, additional information about how individuals are clustered over the landscape is allowed into the model. Traditional genetic methods have not conditioned on space, which poses limitations because apparent clustering of genetic data when space isn't accounted for may actually be simply a model where location, a latent variable when not explicitly conditioned on in the model, drives the measured variables, genotypes.

If mapping population boundaries is the primary goal of a study, the conditional model is probably best, since it does the best job of getting spatial population boundaries correct, and if a study's primary goal is to describe genetic clusters, conditioning on space is appropriate in order to remove the latent influence of location if space isn't conditioned on. In short, if there are biological questions best addressed by a model of genotypes alone, we are currently unaware of them, and feel that the model conditioned on space offers the most advantages.

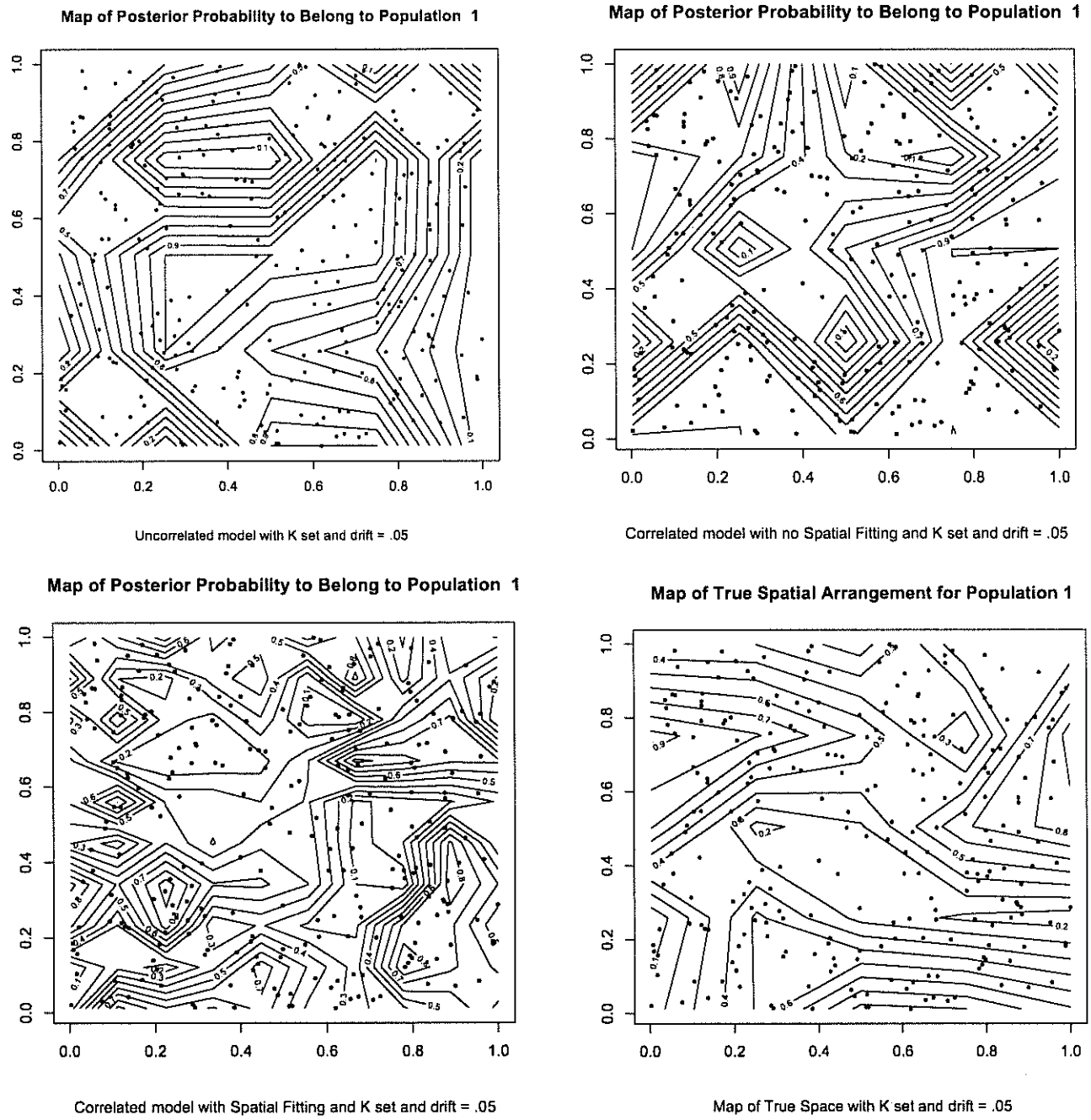

Figure 6: Maps from several model scenarios. Note that the map in the lower right is the true spatial arrangment of population 1. Observe how much closer the map from the correlated model that conditions on space is to the true spatial arrangement than maps from the two models that don't include space. The propensity of the non-spatial models to have deep holes and peaks driven by a few individuals seems to be countered by conditioning on individual locations in the lower left map.

It is critical to recall that these spatial models are intended to represent only the spatial distribution of individuals over the landscape at a single point in time. No inference as to the point of origin or movement patterns for modeled individuals can be inferred, since temporal arrangements are not accounted for in this modeling scenario.

4 Alternative Approaches

Inclusion of spatial data as prior information in delineating among breeding groups appears to have great potential. Under simulation, the models that include correlation among individuals who are located in close spatial proximity to one another do a better job of correctly classifying individuals to populations, accurately reflecting population F_{ST}s, and describing the distribution of populations over the landscape. The exclusion of admixed individuals from spatial models is not limiting, since the whole object in using spatial information is to determine geography barriers between breeding groups, and the presence of admixed individuals allows for the assumption that breeding groups over the landscape are not discreet. The uncorrelated model fails at low F_{ST}s because it does not account from the natural correlation that arises even among independent loci within a population. By contrast, the correlated models that account for that relationship perform well at low F_STs.

The next application of Guillot *et al.*'s model is probably to combine or allign it with Francois *et al.*'s (2006) model using Markov random fields for improved estimation of the K parameter. Francois *et al.* present a method that improves performance of the K-estimate by using regularization, a technique for prevention of overfitting through the introduction of additional information into a model, generally in the form of a penalty for complexity. Common uses of regularization include ridge regression, lasso estimators and feedforward neural networks. In this case, the authors use a Bayesian clustering algorithm that employs hidden Markov random fields as priors for the clustering configurations. Early examinations indicate that the Francois *et al.* model does a better job of estimating K than any of the other models examined herein.

Several additional alternative approaches for delineating between genetic populations over space have emerged in the last several years. Methods include

1. Manel *et al.*'s use of overlaid probability maps for each individual's genotype over space, generated through a moving-window approach. Maps are fit for all individuals, and then a composite map showing regions of maximum change in probabilities is built. Regions of intense change indicate population boundaries.

2. Corander *et al.*'s application of another Bayesian hierarchical model, which is similar to Guillot's, but which doesn't rely on the Voronoi tessellation for delineating among populations in space.

We are unaware of literature at this point in time that uses frequentist approaches for conducting the genetic clustering analysis. While frequentist methods do exist, the literature has moved in a largely Bayesian direction (e.g. Rannala and Mountain (1997), Pritchard *et al.* (2000), Corander *et al.* (2003)).

5 Conclusions

Guillot *et al.*'s model for landscape genetics represents a major advance in the methods through which researchers approach genetic data. The hierarchical structure Guillot *et al.* apply to space could easily be extended to other covariates (e.g. birth location for humans, distinguishing phenotypic characteristics that may impact breeding, etc.), allowing for genetic models conditioned on other measured covariates besides just location. However, even this adjustment does not expand model inference beyond the cross-sectional picture of time mentioned above. Furthermore, it has been shown that the hierarchical model outperforms its non-spatial counterparts in terms of mapping populations over the landscape and in assignment of individuals to populations (Guillot *et al.*, 2005). Due to these strengths, we expect the prevalence of this modeling approach in genetics to increase dramatically over the next few years.

It is essential that researchers recognize that models like the Guillot *et al.* (2005) model described here only address the first of Manel (2003)'s two-step process for landscape genetics. After the genetic populations are accurately mapped, researchers will want to allign genetic maps with maps containing relevant landscape features, to investigate whether genetic boundaries correspond with landscape features. This second step of the landscape genetics process, however, can only be conducted accurately when the location of the various genetic populations is well understood. Though their model doesn't always work ideally, Guillot *et al.* provide a key step toward accurate map construction.

6 References

Arnaud J-F. "Metapopulation genetic structure and migration pathways in the land snail Heliz aspersa: influence of landscape heterogeneity. Landscape Ecology (18): 333-346. 2003.

Banks, S.C., D.B. Lindenmayer, S.J. Ward, A.C. Taylor. "The effects of habitat fragmentation via forestry plantation establishment on spatial genotypic structure in the small marsupial carnivore, Antechinus aglis." Molecular Ecology (14): 1667-1680. 2005.

Corander J., P. Waldmann, M. Sillanpaa. "Bayesian analysis of genetic differentiation between populations". Genetics 163: 367-374. 2003.

Cressie, N. Statistics for Spatial Data. John Wiley and Sons, Inc., New York. 1991. pp 410-421 (Markov random fields).

Falush, D., M. Stephens, and J.K. Pritchard. "Inference of Population Structure Using Multilocus Genotype Data: Linked Loci and Correlated Allele Frequencies." Genetics (164): 1567-1587. 2003.

Francois, O., S. Ancelet, G. Guillot. "Bayesian clustering using hidden Markov random fields in spatial population genetics". Genetics (174): 805-816. 2006.

Freeland, J. Molecular Ecology. Wiley and Sons. 2005

Gelman, A., J.B. Carlin, H.S. Stern, D.B. Rubin. Bayesian Data Analysis. Chapman and Hall. Boca Raton. 1995

Guillot, G, A. Estoup, F. Mortier, J. Cosson. "A Spatial Statistical Model for Landscape Genetics". Genetics (170): 1261- 1280. 2005.

Guillot, G.A., F. Mortier, A. Estoup. "Geneland: a computer package for landscape genetics." Molecular Ecology Notes (5): 712-715. 2005.

Griffith, D. A., and P. R. Peres-Neto. "Spatial Modeling in Ecology: the Flexibility of Eigenfunction Spatial Analyses". Ecology 87(10): 2603-2313. 2006.

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

Manel, S., F. Berthoud, E. Bellemain, M. Gaudeul, G. Luikart, J. Swenson, L. Waits, P. Taberlet, Intrabiodiv Consortium. "A new individual-based spatial approach for identifying genetic discontinuities in natural populations". Molecular Ecology 16: 2031-2043. 2008.

Manni, F., Guerard E., Heyer E. "Geographic patterns of (genetic, morphologic, linguistic) variation: how barriers can be detected using Monmonier's algorithm". Human Biology (76): 173-190. 2004.

Piertney, S., A. MacColl, P. Bacon and J. Dallas. "Local genetic structure in red grouse (*Lagopus lagopus scoticus*: evidence from microsatellite DNA markers. Molecular Ecology 7(12): 1645-1654. 1998.

Pourret, O., P. Naim, B. Marcot. Bayesian Networks: A practical guide to applications. John Wiley and Sons, Inc. Hoboken, NJ. 2008.

Pritchard, J. K., M. Stephens, and P. Donnelly. "Inference of Population Structure Using Multilocus Genotype Data". Genetics (155): 945-959. 2000.

Rencher, A.C. Methods of Multivariate Analysis, Ed. 2. John Wiley and Sons: Hoboken, NJ. 2002.

Schabenberger, O. and A. Gotway. Statistical Methods for Spatial Data Analysis. Boca Raton, FL: Chapman and Hall/CRC, 2005. ISBN 1-58488-322-7.

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

Vounatsou, P., T. Smith and A. Gelfand. "Spatial modelling of multinomial data with latent structure: an application to geographical mapping of human gene and haplotype frequencies". Biostatistics 1(2): 177-189. 2000.

Waples, R. and O. Gaggiotti. "What is a population? An empirical evaluation of some genetic methods for identifying the number of gene pools and their degree of connectivity". Molecular Ecology 15: 1419-1439. 2006.

Wright, B. and C. Cockerham. "Estimating F-statistics for the analysis of population structure". Evolution 38(6): 1358-1370. 1984.