A Spatial Statistical Model for Landscape Genetics

Gilles Guillot,* 1 Arnaud Estoup,1 Frédéric Mortier1 and Jean François Cosson§

*Unité de Mathématiques et Informatique Appliquées, INRA-INAPG-ENGREF, Paris, France 75231, 1Centre de Biologie et de Gestion des Populations, INRA-ENSAM-IRD-CIRAD, Montpellier, France F-34988, 2CIRAD, Département Forêt, Montpellier, France F-34398 and §Centre de Biologie et de Gestion des Populations, INRA-ENSAM-IRD-CIRAD, Montpellier, France F-34988

Manuscript received July 21, 2004
Accepted for publication October 31, 2004

ABSTRACT

Landscape genetics is a new discipline that aims to provide information on how landscape and environmental features influence population genetic structure. The first key step of landscape genetics is the spatial detection and location of genetic discontinuities between populations. However, efficient methods for achieving this task are lacking. In this article, we first clarify what is conceptually involved in the spatial modeling of genetic data. Then we describe a Bayesian model implemented in a Markov chain Monte Carlo scheme that allows inference of the location of such genetic discontinuities from individual georeferenced multilocus genotypes, without a priori knowledge on populational units and limits. In this method, the global set of sampled individuals is modeled as a spatial mixture of panmictic populations, and the spatial organization of populations is modeled through the colored Voronoi tessellation. In addition to spatially locating genetic discontinuities, the method quantifies the amount of spatial dependence in the data set, estimates the number of populations in the studied area, assigns individuals to their population of origin, and detects individual migrants between populations, while taking into account uncertainty on the location of sampled individuals. The performance of the method is evaluated through the analysis of simulated data sets. Results show good performances for standard data sets (e.g., 100 individuals genotyped at 10 loci with 10 alleles per locus), with high but also low levels of population differentiation (e.g., \( F_{ST} \) < 0.05). The method is then applied to a set of 88 individuals of wolverines (Gulo gulo) sampled in the northwestern United States and genotyped at 10 microsatellites.

RECENT developments in molecular markers and statistical tools, combined with powerful computers have led to the emergence of a new scientific field, landscape genetics, which is an amalgamation of population genetics and landscape ecology (Manel et al. 2003). This discipline aims to provide information on how landscape and environmental features influence gene flow, population structure, and local adaptation. It also aids in identifying cryptic genetic discontinuities, which are breaks in gene flow without any obvious cause, or secondary contact among previously isolated populations. The spatial delineation of genetic discontinuities within a species allows that of operational units, an important issue for species management (i.e., for pest control, as well as the monitoring of game or threatened species). Moreover, identifying the abiotic and biotic factors involved in evolutionary processes is essential for modeling and predicting the evolution of genetic diversity under different scenarios, especially those related to environmental changes due to human activity (e.g., habitat fragmentation). Landscape genetics usually makes use of data obtained at highly variable mark-
coding the assignment to a population is *a priori* independent and identically distributed among individuals. As a consequence, they do not make use of spatial coordinates of sampled individuals, except in some *ad hoc* postprocessing schemes like those consisting of drawing by hand the spatial convex hull of each inferred population. Hence these methods cannot objectively identify the spatial location of genetic discontinuities between populations.

Although a large body of statistical literature is available on the clustering of spatially explicit data (Lawson and Denison 2002), the models available are devoted to the analysis of quantitative and univariate data, whereas genetic data are categorical and strongly multivariate by nature. The work of Vounatsou et al. (2000) should, however, be mentioned, since it aims to relate haplotype frequencies to spatial coordinates and environmental covariates. Nevertheless, this method starts from a known clustering of data, which is injected as a prior information on the spatial dependence through a conditional autoregressive model. In a different spirit, Dupanloup et al. (2002) proposed to cluster populations by maximization of a differentiation criterion between population clusters, the criterion being the proportion of total genetic variance due to differences between clusters. As the effective maximization is numerically prohibitive, Dupanloup et al. (2002) suggest using a simulated annealing approach in which the random search strategy makes use of the spatial coordinates of the population samples (note that the unit of treatment could be an individual rather than a population). This can lead to certain local maxima easy to reach from a given spatial sampling configuration and to the identification of relevant genetic discontinuities between groups of populations. However, their criterion does not rely on the coordinates in its definition; therefore the global maximum searched does not depend itself on the coordinates, and the method turns out to be spatial mainly through the heuristic optimization strategy, whose convergence properties still have to be assessed.

In this article, we describe a new statistical model that aims at inferring and locating genetic discontinuities between populations in space from individual multilocus genetic data. Our central assumption throughout this work is that some spatial dependence is often present among individuals. On the basis of this sensible assumption, we developed a hierarchical spatial model in which we formally inject *a priori* information on how the individuals are spatially organized. In addition to the detection of genetic discontinuities between populations, our method also addresses the following points: (i) denoising blurred coordinates of sampled individuals, (ii) estimating the number of populations in the studied area, (iii) quantifying the amount of spatial dependence in the data, (iv) assigning individuals to their population of origin, and (v) detecting individual migrants between populations.

**Hierarchical Spatial Model**

The global set of sampled individuals is viewed as representative of one or several panmictic populations separated by geographic borders across space. Our modeling strategy is hierarchical in the sense that we first specify how the populations are spatially organized and then we specify the statistical genetic properties of each population conditionally on this spatial organization.

**Hidden model of spatial organization through Voronoi tessellation:** Let $\Delta$ be the geographical region under study. We denote by $z = (z_1, \ldots, z_n)$ the vector of genotypes of the $n$ diploid individuals (although other ploidy could be considered) observed at $L$ loci, $z_i$ being a collection of pairs of alleles $z_{i,l} = (a_{i,l}, b_{i,l})$ with $l = 1, \ldots, L$, and we denote by $t = (t_1, \ldots, t_s)$ the vector of the two-dimensional spatial coordinates of these individuals.

We consider that there are $K$ different populations present in the spatial domain under study and that those populations occupy some subdomains $\Delta_1, \ldots, \Delta_K$. The $\Delta_i$ form a partition of $\Delta$, namely $\Delta = \Delta_1 \cup \ldots \cup \Delta_K$ with $\Delta_i \cap \Delta_j = \emptyset$, for $k \neq l$. We consider the setting where we have no knowledge about the shape and location of these subdomains, and part of this article is devoted to their estimation from spatial and genetic data $t$ and $z$. A geographical subdomain being possibly extremely complex, we need to make a few assumptions on the shape and locations of the $\Delta_i$, to reduce the complexity, namely the number of parameters, of the problem. For several reasons discussed later, we assume that each subdomain $\Delta_i$ can be approximated by a union of convex polygons. This assumption is not restrictive as any complex domain can be arbitrarily well approximated by such union of convex polygons provided enough polygons are considered. As we do not know where such polygons should be placed to approximate some true spatial organization of the population under study, we model the locations of these polygons as random variables with uniform distribution over the whole spatial domain.

More formally, we consider that there is a point process (a set of random points in the spatial domain, the number of points itself being random) that has realization denoted by $(u_1, \ldots, u_m)$. Each $u_k$ (referred to hereafter as nucleus) defines a set $A_k$ around it, defined as $A_k = \{s, \text{dist}(s, u_k) \leq \text{dist}(s, u_j), \forall j = 1, \ldots, m\}$. Namely, $A_k$ is the set of geographical sites closer to $u_k$ than to any other points among $(u_1, \ldots, u_m)$. Each $A_k$ is a convex polygon and the set $A_1, \ldots, A_m$ is known as the Voronoi tessellation of $\Delta$, as it splits $\Delta$ in $m$ non-overlapping subdomains. We now assume that each of these $A_k$ contains individuals of one subpopulation only. Hence, each $A_k$ can be labeled by a number between 1 and $K$ that we denote by $c(u_k)$.

In mathematical terms, we parameterized these subdomains assuming that they are unions of some underly-
Figure 1.—Random tessellation of a unit square into two spatial domains through a colored Voronoi tiling. Left, realization of a Poisson point process with Voronoi tessellation induced. Right, partition obtained after union of tiles belonging to the same population (coded as two colors).

ing Voronoi cells induced by a homogeneous Poisson point process. For any arbitrary point \( x \) in \( \Delta \), \( c(x) \) is defined as the population of its closest nucleus. The domain finally covered by population \( k \) is the union of cells of the same population, namely \( \Delta_k = \bigcup_{i: c(x)=i} A_i \). The belonging of any point of the domain to the populations can be thought of as a coloring; hence this model is sometimes referred to as colored Voronoi tiling. (See Figure 1 for an illustration with \( K = 2 \) and \( m = 5 \).) It is widely used in earth sciences to model spatial organization of categorical variables such as geological formations, or soil occupation (Lantuéjoul 2002), and has also been used in genetics in a different framework by Dupanloup et al. (2002).

We assume that all populations have \textit{a priori} equal probability; therefore, we assume that each tile belongs to a population with probability \( 1/K \), independently. Although reasonably simple, our model allows departure from the so-called independent identically distributed (i.i.d.) mixture model commonly used in nonspatial cluster models (Pritchard et al. 2000; Falush et al. 2003). In the latter method, the joint prior probability that individual \( i \) belongs to population \( k \) and individual \( i' \) belongs to population \( k' \) is equal to \( 1/K^2 \) (the product of the marginals) wherever the geographical distance between the individuals may be.

In the colored Voronoi tiling model that we propose to use, the marginal is uniform, but the joint probability does not factorize. More specifically, the joint probability that any two individuals belong to the same population decreases with the geographical distance between them. In other words, i.i.d. mixture models such as those of Pritchard et al. (2000) or Falush et al. (2003) put equal prior weights to all partitions, whereas our model tends to favor partitions that are spatially organized. To a certain extent, our model is very similar in spirit to those used in image analysis, where the purpose is to retrieve a true scene blurred by a certain noise. In this context, it is widely admitted that even when very little is known about the true scene, it is useful to use a prior assuming some spatial organization; see Besag (1986) or Hurn et al. (2003) for a recent review. The advantages of our spatial model are (1) the possibility for better classification with limited data due to a more informative prior and (2) the direct inference about range boundaries. However, with large amounts of data (and a given \( K \)), the posterior assignments of individuals should be the same under the spatial model as for the nonspatial model.

The amount of spatial dependence prescribed by the colored Voronoi tessellation depends on how the domains \( \Delta_i \) are themselves fragmented in smaller polygons \( A_i \). Let us denote by \( \lambda \) the rate of the Poisson process \( u_1, \ldots, u_m \). This parameter controls the number of polygons in \( \Delta \) and hence the amount of spatial dependence in the hidden clustering. Low values of \( \lambda \) correspond to weakly fragmented partitions of \( \Delta \) and thus to strong dependence of the hidden spatial organization of populations, whereas large values of \( \lambda \) correspond to high fragmentation and weak spatial dependence. When the number of points \( m \) is very large, each tile contains only one sampled individual and our tessellation model behaves like an i.i.d. mixture model similar to the prior on the clustering used by Pritchard et al. (2000), Corander et al. (2003), or Falush et al. (2003).

It is worth noting that the model allows the definition of complex spatial domains including situations for which those domains appear as unconnected pieces in
domain of origin. The amount of differentiation depends on many factors.

Moreover, as is shown later, this feature makes it possible to visually detect migrants and identify their spatial distribution at a given time, their allele frequencies will differ. The populations have split from an ancestral population and are potentially frequent cases. This equation accounts for the fact that if a set of populations have diverged according to drift factors $d_1, \ldots, d_K$. In this second model (referred to as F-model and spatial F-model when embedded in our spatial scheme), the frequencies at each locus in the ancestral population follow a Dirichlet$(1, \ldots, 1)$ distribution. Falush et al. (2003) relate frequencies in the present populations to those of the ancestral population and to the drift factors through

$$f_{kl} \sim \text{Dirichlet}\left(f_{\alpha l} \frac{1 - d_k}{d_k}, \ldots, f_{A_{kl}} \frac{1 - d_k}{d_k}\right),$$

where $f_{kl}$ denotes the vector whose entries are $f_{kl}$ (frequency of allele $j$ at locus $l$ in population $k$), $f_{\alpha l}$, $f_{A_{kl}}$ are Dirichlet parameters.

This equation accounts for the fact that if a set of populations have split from an ancestral population at a given time, their allele frequencies will differ. The amount of differentiation depends on many factors that are quantitatively unknown. The vector $d_1, \ldots, d_K$ parameterizes our uncertainty about the level of differentiation.

Conditional model for genotypes: Given the partition and the allele frequencies $f_{kl}$, we assume that the genotypes in each population are independent draws from the discrete multivariate distribution specified by the $f_{kl}$, which is equivalent to the assumption of Hardy-Weinberg equilibrium within, and linkage equilibrium, between loci.

FULL BAYESIAN SPECIFICATION

All nonobserved quantities involved are treated as unknown. For full Bayesian inference we place priors on them, and the model can be summarized as follows.

Number of populations: $K \sim \text{Uniform}(\{K_{\min}, \ldots, K_{\max}\})$. This is a weakly informative prior for which the only subjective inputs are $K_{\min}$, usually set to 1, and $K_{\max}$, set to a large value as compared to the maximum number of populations that can be reasonably expected, so that the choice is numerically inconsequential.
Spatial Model for Genetic Data

Number and location of tiles: We consider that \( m \sim \text{Poisson}(\lambda) \) and that \((u_1, \ldots, u_m) \sim \text{Uniform}(D)\). This is the usual homogeneous Poisson process. The complete randomness of the locations of the points makes this prior noninformative about the locations of the borders between populations, and it is numerically convenient. The unknown amount of spatial organization is controlled by \( \lambda \), on which we place a flat hyper-prior, assuming that \( \lambda \sim \text{Uniform}(0, \lambda_{\text{max}}) \). As previously mentioned, for large \( \lambda \)-values, the tessellation model behaves like an i.i.d. mixture model. As \( \lambda_{\text{max}} \) should be taken large enough to cover a large range of spatial organization from strong (\( \lambda \approx 0 \)) to weak spatial organization, we suggest taking \( \lambda_{\text{max}} \) equal to the number of sampled individuals.

Colors of tiles: \( \epsilon_1, \ldots, \epsilon_m \sim \text{Uniform}[1, \ldots, K] \).

Subdomains of populations: \( \Delta_k = \bigcup_{j=1}^k A_j \) for \( k = 1, \ldots, K \), where \((A_1, \ldots, A_m)\) are the Voronoi tiles induced by \((u_1, \ldots, u_m)\).

Drift: The drift parameters prescribe the amount of genetic differentiation between the present-time populations and the ancestral population and, hence, between the present-time populations themselves. Although it may be sometimes possible to have a rough idea of the amount of differentiation between populations under study (e.g., in terms of classical measures, such as \( F_{ST} \)), it seems, however, difficult to express this prior knowledge in terms of a prior on the drift parameters. To assess the relation between the drift factors and the differentiation we carried out a small set of simulations. We simulated 1000 data sets, each made of two populations of 50 individuals genotyped at 10 loci with 10 alleles per locus. All frequencies were preliminarily sampled from the F-model with drift \( d_1 = d_2 \) uniform on \([0, 1]\). The level of differentiation between population samples was measured using the parameter \( F_{ST} \) estimated following \textit{Weir and Cockerham} (1984). Results show a close linear relation between \( d \) and \( F_{ST} \). It therefore sounds natural to place a prior on \( d \), putting most weight on the left side of \([0, 1]\) with a negligible weight for values \( >0.3 \), instead of a prior uniform on \([0, 1]\). An independent Beta(2, 20) prior complies well with these requirements (see Figure 3). Note that a limited number of test simulations indicated that the choice of prior on \( d \) did not affect the results of the overall algorithm (results not shown).

Frequencies in the ancestral population: \( f_{u_1}, \ldots, f_{u_m} \sim \text{Dirichlet}(1, \ldots, 1) \), \( l = 1, \ldots, L \).

Frequencies in the present population: \( f_i \sim \text{Dirichlet}(f_{u_1}(1-d_1)/d_1), \ldots, f_{u_l}(1-d_2)/d_2) \), \( k = 1, \ldots, K \), \( l = 1, \ldots, L \).

Locations of individuals: The recorded location \( t_i \) of sampled individual \( i \) may not be representative of the true location of the individual. Consider, for instance, the case of sedentary animals (where the true location makes sense) that have moved during capture, or the case of units localized at a coarse administrative level only, or simply the case of measurement errors. To account for this uncertainty, each \( t_i \) stands for the observed location whereas the true location is denoted by \( s_i \). These unobserved \( s_i \) are treated as unknown hereafter and are part of the parameters to be estimated. The true coordinates \( s_i \) are naturally related to observed coordinates \( t_i \) through

\[
t_i = s_i + \epsilon_i,
\]

where \( \epsilon_i \) is an i.i.d. additive noise chosen in a suitable parametric distribution. Even when the coordinates are recorded with a good precision with respect to the size of the domain under study, it is useful to introduce such an additive noise to allow individuals with the same coordinates to belong to different populations.

MARKOV CHAIN MONTE CARLO INFERENCE

We denote by \( \theta = (K, m, u, c, d, f, f_s, s) \) the vector of unknown parameters to be estimated. The likelihood of the data \((t, \epsilon)\) is

\[
\pi(t, \epsilon | \theta) = \pi(t | \theta) \pi(\epsilon | t, \theta) = \pi(t | \theta) \prod_{i=1}^{n} \pi(\epsilon_i | t_i, \theta). \tag{4}
\]

The terms of the product are given by the allelic frequencies:

\[
\pi(\epsilon_i | t_i, \theta, (\alpha, \beta)) = \begin{cases} 
\frac{f_{\alpha}}{f_{\beta}} & \text{if } \alpha \neq \beta \\
\frac{f_{\beta}}{f_{\alpha}} & \text{if } \alpha = \beta.
\end{cases}
\]

The inference of \( \theta \) will be made through the investigation of its posterior distribution \( \pi(\theta | t, \epsilon) \). We consider a hybrid algorithm based on sequential updates of the various blocks of parameters. All parameters are randomly initialized from the prior. Then the moves are proposed in a deterministic order as follows: (1) update \( d \); (2) update \( f_s \) (these two steps are skipped in the D-model; see APPENDIX, Details of MCMC computations); (3) update \( f_{u_l} \); (4) update \( c_i \) for \( j = 1, \ldots, m \); (5) update \( u_j \) for \( j = 1, \ldots, m \); (6) update \( s_i \) for \( i = 1, \ldots, n \); (7) discard or add a tile (increase or decrease \( m \) by 1); and (8) split one existing population into two or merge two into one \((i.e., \text{increase or decrease } K \text{ by 1})\). Convergence follows from detailed balance, irreducibility, and aperiodicity (see APPENDIX for computational details).

In practice the number of populations is first estimated by computing the mode of the posterior distribution: \( \hat{K} = \text{Argmax}_K \pi[K | t, \epsilon] \) from a first run \((\theta_i)_{i=1,\ldots,T}\). Then, because it is meaningless to compute empirical means on values of \( \theta \) corresponding to different values of \( K \), one can either work on subsamples of \((\theta_i)_{i=1,\ldots,T}\) obtained by restricting to the states corresponding to \( K = \hat{K} \) or simply rerun the Markov chain Monte Carlo (MCMC) algorithm with fixed \( K \) set to \( \hat{K} \).

As in all mixture problems, this model is not identifiable as the likelihood is invariant under any relabeling of the populations \((\text{Celeux} 1997; \text{Stephens} 1997; \text{Robert} \ldots))
and Mengersen 1999; Celeux et al. 2000; Stephens 2000). Consequently, it may happen that along an MCMC run (especially if the numbers of loci and alleles are small), a move from say state $c_i = k$ to $c_i = l$ does not correspond to a reassignment of the $i$th cell but to a relabeling of the populations (switch between $k$ and $l$). Although more theoretically rooted methods are available (Celeux et al. 2000; Stephens 2000) we used the expedient (when label switching was suspected), which consists of imposing after the run the following identifying constraints on the frequency of the first locus,

$$f_{ii1} < f_{i'1} \quad \text{for all } k, k', k < k',$$

(6)

where the loci are sorted by an increasing number of alleles. This rule enabled us to fix the label-switching issue only when the frequencies did not overlap in the chains, and we found that it is advisable to check visually the trace of frequencies in the MCMC run to detect a possible switch.

We observed that label switching could be frequent when working with small data sets ($\leq$100 individuals, 1 or 2 loci, few polymorphisms) and becomes rare with larger and more traditional microsatellite data sets (e.g., at least 100 individuals genotyped at 10 loci with 10 alleles per locus). When such a relabeled sample is available, the conditional posterior $\pi(c(s)|K = \hat{K}, z, t)$ is estimated by the corresponding empirical mean. Similarly, $d$, $f_x$, and $f$ are estimated by their means over the corresponding populations.

The whole algorithm has been programmed in Fortran 77, making use of the numerical library for random number generation, Randlib1.3. The machine time required for $10^5$ iterations on a data set of 1000 individuals genotyped at 10 loci with 10 alleles per locus with 10 loci is typically of 1 hr on a PC equipped with a 2-GHz chip set.

**RESULTS FROM SIMULATED DATA SETS**

**Number of populations:** A first interesting feature of this model is its ability to deal with an unknown number of populations. Because the effective value of the model depends on the precision in the estimation of $K$, this key point has been investigated by analyzing data sets simulated using the prior of the previously described spatial F-model. To retrieve the known $K$ parameters that served to build the data set, we ran the MCMC scheme using the spatial F-model or the spatial D-model as a prior in the inference and computed $\hat{K}$ after 50,000 iterations.

We first started by building 50 data sets of $n = 50$ individuals with $K = 1$ (only one population). For each data set, we get estimated $K$, and thus we end up with 50 estimated values corresponding to the 50 simulated data sets. This procedure has been repeated for various values of $K$ with data sets of 50 individuals. Everywhere the number of loci was $L = 10$ and the number of alleles

**Figure 3.**—Relationship between drift factor and $F_{ST}$. Results are from simulated data. The drifts $d_1$ and $d_2$ were equal and sampled from a Beta(2, 20) prior (x-axis).
The empirical relationship between $F_{ST}$ and the number of populations $K^\hat{}$ estimated using the spatial $F$-model (circle) and the spatial $D$-model (triangles) as a prior in the MCMC inference. The 50 simulated data sets are made of $K = 2$ populations, with $L = f_{l=1...L} = 10$.

**Figure 5.**—Empirical relationship between $F_{ST}$ and the number of populations $K$ estimated using the spatial $F$-model (circle) and the spatial $D$-model (triangles) as a prior in the MCMC inference. The 50 simulated data sets are made of $K = 2$ populations, with $L = f_{l=1...L} = 10$.

Each data set has been analyzed in parallel by the four methods, from which we can derive the false classification rate defined as

$$FCR = \frac{\text{number of wrongly assigned individuals}}{\text{total number of individuals}}.$$ (7)

Results on the whole set of simulations and on various subsets with different levels of genetic and spatial struc-

Assignment to population of origin: One other feature on which our model can be evaluated is its ability to assign individuals to their population of origin. We have several goals in mind here: (i) assessing the ability of the model to correctly classify individuals; (ii) comparing our model to nonspatial approaches suggested in the literature; and (iii) as the model is specifically tailored for populations displaying structure, we want to describe how the model behaves with respect to the amount of spatial and genetic structure present in the data. We simulated 1000 data sets using the spatial $F$-model, with 100 individuals organized in two populations (50 individuals per population), with 10 loci and 10 alleles per locus and then with only 3 loci and 10 alleles per locus. Each of these data sets has been first analyzed using the spatial $F$-model as a prior in the MCMC scheme. However, since we know from the results of the previous section that using the $F$-model can be misleading about the number of populations, we also analyzed our simulated data sets with the spatial $D$-model as a prior in the MCMC scheme.

An inference of all parameters (including the number of populations) could have been made. However, we wished to compare our method to methods that do not handle the inference of $K$. Therefore we chose to consider $K$ as a known parameter and to compare the different methods when $K$ is known and equal to 2. As already mentioned, most of the existing clustering models are based on i.i.d. mixtures. Therefore we have compared our spatial model to nonspatial clustering models obtained by replacing our spatial prior on the clustering variable $c$ by an i.i.d. prior, giving the nonspatial $F$-model and nonspatial $D$-model. The nonspatial $D$-model turns out to be the model described by Pritchard et al. (2000) in the no-admixture case, whereas the nonspatial $F$-model can be viewed as a simplified version of the model of Falush et al. (2003), where linkage equilibrium is assumed.

Each data set has been analyzed in parallel by the four methods, from which we can derive the false classification rate defined as

$$FCR = \frac{\text{number of wrongly assigned individuals}}{\text{total number of individuals}}.$$ (7)
TABLE 1
Average false classification rates (in percentage) for all simulated data sets and subsamples with various levels of genetic and spatial structure

<table>
<thead>
<tr>
<th>Structure</th>
<th>Spatial</th>
<th>F-model</th>
<th>D-model</th>
<th>Nonspatial</th>
<th>F-model</th>
<th>D-model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Genetic</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>All</td>
<td>All</td>
<td>1.8</td>
<td>2.6</td>
<td>3.8</td>
<td>3.3</td>
<td></td>
</tr>
<tr>
<td>$F_{ST} &lt; 0.04$</td>
<td>All</td>
<td>7.8</td>
<td>14.2</td>
<td>15</td>
<td>13.5</td>
<td></td>
</tr>
<tr>
<td>$F_{ST} &lt; 0.06$</td>
<td>All</td>
<td>4.7</td>
<td>7.6</td>
<td>9</td>
<td>8.5</td>
<td></td>
</tr>
<tr>
<td>$F_{ST} &gt; 0.11$</td>
<td>All</td>
<td>0.3</td>
<td>0.3</td>
<td>0.2</td>
<td>0.2</td>
<td></td>
</tr>
<tr>
<td>All</td>
<td>$m &lt; 12$</td>
<td>2.3</td>
<td>1.9</td>
<td>11.4</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>All</td>
<td>$m &lt; 25$</td>
<td>1.7</td>
<td>1.8</td>
<td>6.8</td>
<td>4.4</td>
<td></td>
</tr>
<tr>
<td>All</td>
<td>$m &gt; 80$</td>
<td>2.2</td>
<td>3</td>
<td>2.8</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>$F_{ST} &lt; 0.06$</td>
<td>$m &lt; 25$</td>
<td>2.7</td>
<td>5.3</td>
<td>11.8</td>
<td>9.5</td>
<td></td>
</tr>
<tr>
<td>$F_{ST} &lt; 0.04$</td>
<td>$m &lt; 12$</td>
<td>3.5</td>
<td>1</td>
<td>24</td>
<td>16.7</td>
<td></td>
</tr>
</tbody>
</table>

Results with 10 loci

Results with 3 loci

All | All | 11.3 | 12.5 | 17.5 | 17.5 |

The level of genetic and spatial structure increases with $F_{ST}$ and decreases with $m$, respectively. Results are shown from 1000 simulated data sets of 100 individuals in two populations, with $L = f_{1,...,L} = 10$ and $L = 3, f_{1,...,L} = 10$.

Individual genotypes were obtained with the software EasyPop (Balloux 2001) by simulating two populations exchanging different numbers of migrants per generation. In a first data set (referred to hereafter as set A), the two populations are considerably differentiated ($F_{ST} = 0.16$), in a second set (B) the populations are less differentiated ($F_{ST} = 0.06$), and in a third data set (C), they are very weakly differentiated ($F_{ST} = 0.01$). In each set we have $n = 200$ individuals (with 100 individuals per population) and $L = 10$ loci with $f_{1,...,L} = 10$ alleles per locus. Individuals of each population were randomly located on each part of an oscillating curve on the unit square.

The ability of the model to find the actual partition of space is illustrated in Figure 7, set A, which displays the posterior probability $\pi(c(s) = k|t, z)$ for any pixel to belong to the two populations. We observe that the true partition is very well detected. We obtained similar results on the two other data sets, although the boundary between the two domains was detected with less precision with data sets B and C (results not shown). We then quantitatively assessed how the level of differentiation influences the precision in the estimation of the border. This can be viewed from the map of $\delta(s)$, where $\delta(s)$ is any suitable measure of the dispersion of $\pi[c(s)|t, z, K]$. As $\pi[c(s) = j|t, z, K]$ is a probability measure on $[1, \ldots, K]$ whose weights can be zero, the entropy is not defined, and therefore we used instead

\[
\delta(s) = \sum_{j=1}^{K} \pi^2[c(s) = k|t, z, K]
\]

as an index of the dispersion of the distribution $\pi[c(s) = k|t, z, K]$. The right-hand side of Equation 8 is minimal when all the colors have equal probability, which corresponds to a flat posterior. Thus, the points $s$ for which $\delta(s)$ is high are those confidently classified, whereas those where $\delta(s)$ is low correspond to poorly classified points that might correspond to transition regions or to sparsely sampled regions. The maps of $\delta(s)$ given in Figure 8 for the three data sets show that the lower values of $\delta(s)$ are along the true line of discontinuity (darker color), and that the accuracy in the estimation increases with the level of differentiation between populations.

Detection of migrants: Since populations exchange migrants, it is sensible to assess whether the presence of such migrants would affect the spatial detection of genetic discontinuities and whether migrants could be detected and spatially located by our method, for different levels of differentiation between populations.

To address these points, we mimicked the presence of first-generation migrants in our previous data sets A, B, and C by moving one individual from the upper population to the lower population and another one in the opposite direction. Figure 9 shows that the presence of these first-generation migrants did not affect the accuracy of the method to detect the two populations and
Figure 6.—Examples of simulated spatial organization of 100 individuals (black dots) into two populations (coded as two colors) with various levels of spatial dependence. This level is controlled by parameter $m$ (number of Voronoi tiles). The nuclei of the tiles are not depicted for clarity.

$\text{Figure 6.}$

Effect of errors on the locations of individuals: We show here how errors on the locations of individuals may lead to poor results and how accounting for errors in the positioning of individuals allows us to retrieve most of the underlying signal. This question was addressed through the analysis of a new set of simulations.

Three independent populations separated by straight lines were positioned on the unit square. The number of individuals was $n = 150$ (50 individuals per population), $L = J_{i=1}^L = 10$, and the level of differentiation between populations was relatively low ($F_{ST} = 0.08$). The true positions $s_i$ were blurred by an additive noise $\varepsilon$, uniform on $[-0.15, 0.15]^2$, so that the observed positions were $t_i = s_i + \varepsilon$. The model was first run considering that the given positions were true ($\varepsilon = 0$) and in a second run the uncertainty was accounted for by injecting the information that $\varepsilon \sim \text{Uniform}[-0.15, 0.15]^2$. We also considered the case where the true coordinates were used whereas wrong coordinates were assumed, and finally we give the results for true coordinates considered as true coordinates. The results are summarized in Figure 10. It can be clearly observed from the figure that accounting for uncertainty in the positioning of individuals substantially increases the pre-

to spatially locate the genetic discontinuity. The figure also shows that the two migrants are easily detected (i.e., visualized) for data sets A, B, and C. The population of origin of each migrant could be easily deduced from its coloring pattern, including when more than two populations shared the domain (results not shown).
cision in the detection of the true borders when some errors on the position of individuals exist (Figure 10, line 1 vs. line 2). Moreover, adding position noise in the model does not alter the results for data sets without position errors (Figure 10, line 3 vs. line 4).

APPLICATION TO MONTANA WOLVERINES

We now analyze a previously published data set on wolverines (*Gulo gulo*), a medium-sized carnivore wildly distributed in North America. Wolverines are highly mobile, with the ability to disperse up to 300 km within a year, but are also highly sensitive to habitat disturbance by humans. Eighty-nine individuals were sampled in Montana and genotyped at 10 microsatellite loci Cegelski et al. (2003). Samples are nearly evenly distributed over an area that corresponds to a landscape highly fragmented by human development and disturbance. Using nonspatial Bayesian clustering procedures and assignment tests implemented in the programs *Structure* (Pritchard et al. 2000) and *Geneclass* (Cornuet et al. 1999), Cegelski et al. (2003) provided some evidence for the existence of three populations of wolverines in Montana, with *F*$_{\text{ST}}$ values ranging from 0.038 to 0.110. Hardy-Weinberg equilibrium could not be rejected in two of the four inferred populations (Fisher’s exact test, *P* > 0.05; Raymond and Rousset 1995). Hence, our spatial method gives strong evidence for the presence of (at least) one more population than those previously determined by Cegelski et al. (2003).
The main added value of our spatial approach is thus the delineation of a fourth population located north of population 3. Populations 3 and 6 were previously confounded using nonspatial methods. Pairwise $F_{ST}$ between the four populations inferred by our method are $F_{ST,3-1} = 0.151$, $F_{ST,4-1} = 0.13$, $F_{ST,6-1} = 0.174$, $F_{ST,4-3} = 0.108$, $F_{ST,6-3} = 0.079$, and $F_{ST,4-6} = 0.176$. The $F_{ST}$-value between the previously confounded populations is thus the lowest of all pairwise $F_{ST}$.

Interestingly, a close examination of the land cover map in Figure 2 of Cegelski et al. (2003) shows that the spatial domains of our populations 3 and 6 are separated by a narrow extent of man-made habitats (i.e., grasslands and grazing lands) that may have reduced wolverine movements.

Our spatial approach also allows detection of five individuals that genetically differ considerably from their spatial neighbors (Figure 13). These individuals can be interpreted as first-generation migrants as suggested by our previous analysis of simulated data sets. Some of
these putative individual migrants have been previously detected by Cegelski et al. (2003).

DISCUSSION

The two key steps of landscape genetics are the detection and location of genetic discontinuities and the correlation of these discontinuities with landscape and environmental features. Efficient methods to achieve the first step have been lacking so far; our method provides the first efficient tool for locating genetic discontinuities within a landscape from individual geo-referenced multi-locus genotype data, without any a priori knowledge on the populational units and limits. Once genetic discontinuities have been detected and spatially located using the observed genetic data, accurate landscape descriptors implemented, for example, in a geographic information system (GIS) can be used to associate the inferred genetic discontinuities with landscape features and hence generate hypotheses about the cause of genetic boundaries; see Piertney et al. (1998) for an attempt in this direction.

Our spatial method appears well suited for revealing cryptic spatial genetic structure and also provides a suit-
Figure 10.—Maps of the posterior probability $\pi(c(s) = k | t, z)$ when coordinates $s$ are blurred by a uniform noise. First row, wrong coordinates, assumed true; second row, wrong coordinates, assumed wrong; third row, true coordinates, assumed wrong; fourth row, true coordinates, assumed true. $F_{ST} = 0.08$, $L = f_{d=1..6} = 10$. Dashed black lines are the true borders.

A possible approach for the detection of migrants (i.e., individuals poorly genetically related to their spatial neighbors) and their assignment to their population of origin. This has been illustrated by the analysis of a set of wolverine individuals sampled in the northwestern United States and genotyped at microsatellite markers (Cegelski et al. 2003). In addition to the populations previously identified by Cegelski et al. (2003) using nonspatial methods,
our spatial approach allowed us to delineate a fourth population separated from others by a narrow extent of human-made habitats. Our analysis hence strengthens the conclusion that, even for a highly mobile species, habitat disturbance by humans may considerably limit movements and creates spatial genetic structure. Our approach also allowed detection of migrants between populations that were not previously distinguished, using nonspatial approaches (e.g., Figure 13, two migrants originating from population 6). The overall larger number of potential migrants detected by Cegelski et al. (2003) using assignment methods implemented in the package Geneclass (Cornuet et al. 1999) may be due to: (i) a lower power of our spatial method for detecting migrants and offspring of migrants, although this remains to be assessed; (ii) some Wahlund effect in at least one population identified by Cegelski et al. (2003); and (iii) an excess of the first type of error (i.e., resident individuals identified as migrants) produced by most assignment methods, as previously shown from simulated and empirical data sets (Berry et al. 2004; Paetkau et al. 2004). The exact behavior of our spatial method with respect to old migrants (e.g., F₁, F₂, and backcross migrants) still has to be assessed. Note that these cases are explicitly treated in the admixture case of the Structure approach (Falush et al. 2003).

Although theoretically applicable to any type of qualitative variable, our method was more specifically designed for genetic codominant markers (e.g., allozymes, microsatellites, single nucleotide polymorphisms). It assumes that the marker loci are unlinked and at linkage equilibrium with one another within populations (the effect of deviations from the second assumption is discussed further in this section). The accuracy of our method increases with the sampling effort (number of individuals and loci) as well as the strength of the genetic discontinuity between populations. However, our analysis of both simulated and real data sets showed good performances of the method for data sets of standard size (e.g., 100 individuals genotyped at 10 loci with 10 alleles per locus), with mild to low levels of population differentiation (e.g., $F_{ST} < 0.1$). Regarding individual sampling strategy, efficient inference in landscape genetics implies random sampling across the entire study area and not just sampling some individuals in each of several a priori defined populations (Manel et al. 2003). This also holds for our spatial method. Because methods to treat spatially approximately evenly distributed individual genetic data sets have not been available so far, such a sampling design has been rarely applied. We anticipate that our spatial method will stimulate population geneticists and ecologists interested in determining landscape and environmental factors influencing population genetic structure to modify their sampling strategy. The effect of a traditional sampling scheme (i.e., several groups of individuals collected on a limited area) on the performance of our method still needs to be assessed using simulated data sets. However, preliminary tests suggest that this effect may be small if enough of such sampling groups have been collected.

An interesting feature of our method is its propensity to infer that several spatial domains that may be apparently unconnected within the sampling window can belong to the same population unit. This represents a significant advantage in comparison to previous methods that aimed to define population (or individual) groups and hence genetic discontinuities. Hence our method is capable of treating complex spatial situations. In the meantime, we have demonstrated that our model does not artificially enforce a spatial substructure when it does not exist (see especially case $K = 1$ in Figure 4).

Another major advantage of our approach as a whole compared to earlier methods (Pritchard et al. 2000; Dupanloup et al. 2002; Falush et al. 2003) is that the number of population units is treated here as an unknown parameter. Although Corander et al. (2003) also did so, their method is not spatially explicit and aims first at grouping populations rather than individuals. Regarding inference of the number of population units, the spatial D-model was found to perform better than the spatial F-model. The latter model tends to largely overestimate the number of population units, especially for low levels of population differentiation. To capture subtle genetic structures, the F-model gives a rather loose definition of what a population is. Embedding it in a fully Bayesian model where the number of these loosely defined populations is itself unknown seems to place too much flexibility in the algorithm. These results are in
agreement with those of Falush et al. (2003), who found that, in a nonspatial context, the F-model was in general more permissive than the D-model (additional populations being fitted to a data set), as it permits the existence of two or more populations with very similar allele frequencies (particularly if the prior on the drift factor is chosen to favor small values).

We proposed a full Bayesian inference of the parameters of our spatial model through MCMC simulations. Estimation of parameters, uncertainty about the estimated values, and graphical outputs about all parameters involved can be derived from one single MCMC run. However, from a user point of view, it is more convenient to launch a first run processing all parameters in $\theta$, from which only $K$ is derived, and then to launch a second run with $K = \hat{K}$, and from which all remaining parameters will be investigated. All computations described in this article used the previous rule. It is worth mentioning that additional test computations performed on some other data sets for which some of the basic...
assumptions of our model were somewhat violated have shown that results from one single run could be misleading, as the Markov chain could get stuck to some local modes of the posterior. In this case we found, in agreement with Corander et al. (2003, 2004), that processing several independent runs and ranking them according to the posterior density could be an aid in the interpretation of results. We believe, however, that in such situations, more could be gained from model improvement than from inference algorithm improvement.

Postprocessing issues were also encountered when computing modal populations on the wolverine data as their number was smaller than the number of inferred populations. Such an issue has not been encountered with simulated data. The reason why this has been observed in the wolverine data set has to be further assessed. However, we may speculate that, as often in the MCMC algorithm for mixture models (Falush et al. 2003), some rarely visited states lead to inferred populations that do not appear to be the modal population for any individuals. Such populations may be interpreted as spurious populations that have not been successfully removed by the MCMC algorithm (and, hence, as a convergence flaw of the algorithm) or as populations standing for complex multimodality in the posterior. From a user point of view, these ghost populations have just to be ignored, and focus can be restricted to modal populations.

Some other forms of spatial dependence may occur in addition to that due to the presence of genetic discontinuities. These include isolation by distance between individuals, i.e., a regular increase of differentiation between individuals with geographic distance due to limited dispersal (Rousset 2000; Leblois et al. 2003); kin clustering, i.e., spatial clustering of highly related individuals at least before dispersal (e.g., Kelly 1994); and a certain rate of selfing reproduction for some species (Wolf and Takebayashi 2004). The problem of isolation by distance has been already mentioned by Dupanloup et al. (2002), who showed that under a stepping-

---

**Figure 12.—Continued.**

**Figure 13.—** Map of the mode of the posterior probability to belong to each class for the wolverine data. Large character numbers indicate population labels. Arrows indicate putative migrants.
stone migration model, their method found significant clustering of populations in the absence of any genetic boundary. It is expected that isolation by distance, kin-clustering, or selfing decrease the performance of our spatial model since the latter assumes Hardy-Weinberg and linkage equilibrium among loci within populations separated by genetic discontinuities. In particular, it is expected that deviations from random assortment that are not caused by genetic discontinuity will tend to overestimate the number of population units (Pritchard et al. 2000; Falush et al. 2005). However, to what extent additional forms of spatial dependence would affect the performance of our model to detect and locate genetic discontinuities between populations still needs to be assessed, analyzing simulated data sets. Given the complexity of genetic data sets at highly variable markers, our spatial model had to remain as simple as possible. A simple way to achieve spatial dependence was to assume a dependence on the class variable and an independence of $z(s)$ within populations. However, this basic model might be improved in several ways. For instance, Allard and Guillot (2000) proposed in a different context a model where the population label variable is not spatially dependent (unlike here where the population variable $c$ is spatially organized) and the data $z$ are spatially dependent within populations. This model tended to cluster samples in a way very different from that of the current model. Although this model is not straightforwardly transposable to genetic data, the introduction of within-population dependence while keeping the spatial dependence of the class variable is certainly a promising way to improve spatial models for genetic data and, more specifically, to take into account other forms of spatial dependence such as isolation by distance, kin clustering, or selfing.

The whole computer code (including Fortran MCMC, postprocessing subroutines, and R data handling and graphical interfaces) used to carry out computations under our spatial model will be made available soon on the web page of G. Guillot and Comprehensive R Archive Network (IHaka and Gentleman 1996; R Development Core Team 2004) as an R package called GeneLand.

The wolverine data set was kindly provided by Chris Cegelski and Lisette Waits from the Department of Fish and Wildlife Resource, University of Idaho. We thank Stuart Baird for constructive comments on the manuscript and Mark Hewison for comments on an earlier draft. G. Guillot thanks Joel Chadeau, Christian Robert, Eric Parent, Olivier François, Jesper Møller, and Rasmus Waagepetersen for stimulating comments at various stages of this work. We also thank two reviewers and Laurent Excoffier as associate editor for constructive remarks. This work was partly supported by a grant to the authors from Bureau des Ressources Génétiques.

LITERATURE CITED


Paetkau, D., R. Slade, M. Burdens and A. Estoup, 2004 Genetic assignment methods for the direct, real-time estimation of migra-
tion rate: a simulation-based exploration of accuracy and power.


Communicating editor: L. Excoffier

APPENDIX

Description of simulation study: The exact algorithm described in Number of populations is detailed in pseudo-code below:

do K in (1, 2, 5, 10)
do isim in 1:50

draw (m, u, c, d, f, fλ, s) from πK\text{model} (m, u, c, d, f, fλ, s|K)
draw z from π(z|θ)
simulate sample (θ)\text{isim} from π\theta|K\text{model}(θ|z)
compute \tilde{K}_\text{model}[isim] = posterior mode of K in (θ)\text{isim}\text{model}
simulate sample (θ)\text{isim}\text{model} from π\theta|\tilde{K}_\text{model}(θ|z)
compute \tilde{K}_\text{model}[isim] = posterior mode of K in (θ)\text{isim}\text{model}
endo
plot histogram of \tilde{K}_0
plot histogram of \tilde{K}_F
endo

In words, we simulated spatialized data sets with various sets of parameters and tried to retrieve these parameters, considered as unknown, via our algorithm with a special emphasis on the number of populations.

Details of MCMC computations: Six block updates (namely, those of u, c, p, f, fλ, and s) do not change the dimensionality of θ, and two updates (K and m) are of jump type, increasing or decreasing the length of θ.

Update of drifts: This is done through Metropolis-Hastings updates, as described by Falush et al. (2003).

Update of frequencies in the ancestral population: This is done through Metropolis-Hastings updates, as described by Falush et al. (2003).

Update of frequencies in the present-time population: This is done through Gibbs updates, as described by Falush et al. (2003).

Update of the colors of tile: This is a Metropolis-Hastings update. We make componentwise updates of c: sequentially for each tile with color c_j, we propose a new value c_j* from the prior, namely π(c_j* = 1|K = k) = 1/k.

This proposal is accepted according to the usual M-H acceptance probability

$$
\alpha(\theta, \theta^*) = 1 \wedge \frac{\pi(t, z|K, m, u, c^*, d, f, fλ, s)}{\pi(t, z|K, m, u, c, d, f, fλ, s)},
$$

where c^* = (c_{-j}, c_j^*), c_{-j}, denoting the vector c deprived from its jth entry (the one currently updated).
**Update of the locations of tiles:** This is a Metropolis-Hastings random-walk update. We make a componentwise update of $u$: sequentially for each tile centered in $u_j$, we propose a new value $u^*_{j}$ obtained by a small random perturbation of the current position.

This proposal is accepted according to the usual M-H acceptance probability

$$\alpha(\theta, \theta^*) = 1 \wedge \frac{\pi(t, z | K, m, u^*, c, d, f, f_\delta, s)}{\pi(t, z | K, m, u, c, d, f, f_\delta, s)}, \quad (A2)$$

where $u^* = (u_{-j}, u^*_{j})$, $u_{-j}$ denoting the vector $u$ deprived from its $j$th entry (the one currently updated).

**Update of the true unknown locations of individuals:** This is a Metropolis-Hastings random-walk update. We make a componentwise update of $s$: sequentially for each tile individual $s_j$, we propose a new value $s^*_{j}$ obtained by a small random perturbation of the current position.

This proposal is accepted according to the usual M-H acceptance probability

$$\alpha(\theta, \theta^*) = 1 \wedge \frac{\pi(t, z | K, m, u, c, d, f, f_\delta, s^*)}{\pi(t, z | K, m, u, c, d, f, f_\delta, s)}, \quad (A3)$$

where $s^* = (s_{-j}, s^*_{j})$, $s_{-j}$ denoting the vector $s$ deprived from its $i$th entry (the one currently updated).

**Birth or death of a tile:** In step (7) we first randomly choose between a birth or a death of a tile with equal probability.

In the case where a birth is proposed, we have to propose a new random point to the current state $u_1, \ldots, u_m$, which we denote by $u^*_{m+1}$. This new proposed tile will also receive a color denoted by $c^*_{m+1}$. $u^*_{m+1}$ and $c^*_{m+1}$ are drawn from the prior. Namely, $u^*_{m+1}$ is uniform on the spatial domain and $c^*_{m+1}$ is uniform on the current set of colors $\{1, \ldots, K\}$.

We denote by $\theta$ the current state and by $\theta^* = (K, m^*, u^*, c^*, d, f, f_\delta, s)$ the proposed value. Since we consider the case of adding a tile, $m^* = m + 1$, $u^* = (u, u^*_{m+1})$, $c^* = (c, c^*_{m+1})$, and the reversible jump acceptance probability (see, e.g., van Lieshout 2000 or Roberts and Dellaportas 2003) takes the following form:

$$\alpha(\theta, \theta^*) = 1 \wedge \frac{\pi(\theta^* | z) q(\theta | \theta^*)}{\pi(\theta | z) q(\theta^* | \theta)} = 1 \wedge \frac{\pi(z | \theta^*)}{\pi(z | \theta)} \lambda \frac{1}{m + 1}. \quad (A4)$$

In the case where a death is proposed, a tile is uniformly selected at random over $(u_1, \ldots, u_m)$, and the acceptance probability takes a similar form (reverse expression; precisely; see Byers and Raftery 2002).

**Birth or death of a population:** The strategy used here to draw a new population or to discard an existing one is similar to the one described by Richardson and Green (1997), with three differences: we have to maintain a spatial consistency within the geometry of the current state and the one of the proposed state (hence, not all proposals will make sense), the data are categorical (there is no straightforward order between alleles), and they are also multivariate (which makes it impossible even to sort quantitative mean parameters, like frequencies).

The strategy to draw a new population consists of splitting an existing population into two distinct populations. Reversibly, a death of an existing population will be obtained by merging two existing ones. Thus, the first step consists of choosing between splitting and merging. This is done with equal probability.

Then a step of, say, split type involves the following:

Propose a new allocation of individuals (over $K + 1$ populations instead of $K$) through a new coloring $c^*$, namely propose the population $k$ to be split with equal probability over the $K$ existing populations (the split population will give birth to two new populations labeled as $k$ and $K + 1$),

count the number of tiles $v_i$ belonging to population $k$ (remember that their union forms the set $\Delta_k$),

choose a uniform number of tiles $v$ between 0 and $v_i$,

select randomly $v$ of those tiles in $\Delta_k$ that will be given a new color,

propose a new allocation of individuals $c^*$ by changing deterministically colors of the $v$ selected tiles of $\Delta_k$ from $k$ to $K + 1$.

Then propose frequencies $f^*_{k}, f^*_{K+1}$ and drifts $d^*_{k}, d^*_{K+1}$.

The proposed state is denoted by $\theta^* = (K^*, m, u, c^*, d^*, f^*, f_\delta, s)$ with $K^* = K + 1$, $f^* = (f_{-1}, f^*_k, f^*_{K+1})$, $d^* = \ldots$. 

where

\[ n_k \equiv (d_{-1}, d_{1}^*, d_{k+1}^*) \text{, where } f_{-1} \text{ (resp. } d_{-1} \text{) denotes the set } f \text{ (resp. } d \text{) deprived of parameters corresponding to population } k. \]

The drifts \( d_{-1}^* \) and \( d_{k+1}^* \) are sampled from the prior [Beta(2, 20) in the present case]. And to have a fairly high acceptance probability, the frequencies \( f_{-1}^* \) and \( f_{k+1}^* \) are sampled from the conditional distributions \( \pi(f_{-1}^* | d_{-1}^*, f_{-1}, m, c^*, u, z, t) \) and \( \pi(f_{k+1}^* | d_{k+1}^*, m, c^*, u, z, t) \) (in the spirit of a Gibbs sampler).

The acceptance ratio for the birth of a population is as usual

\[ \alpha(\theta, \theta^*) = 1 \land \mathcal{R}(\theta, \theta^*), \quad (A6) \]

where

\[ \mathcal{R}(\theta, \theta^*) = \frac{\pi(z | \theta^*) \pi(\theta^*) q(\theta | \theta^*)}{\pi(z | \theta) \pi(\theta) q(\theta^* | \theta)}. \quad (A7) \]

Most of the terms in the ratio of the prior probabilities vanish, and we keep only

\[ \frac{\pi(\theta^*)}{\pi(\theta)} = \left( \frac{K}{K + 1} \right)^{\alpha} \frac{\pi(f_{-1}^* | f_{-1}, d_{-1}) \pi(f_{k+1}^* | f_{k+1}, d_{k+1})}{\pi(f_{-1} | f_{-1}, d_{-1})}. \quad (A8) \]

The acceptance ratio in the ratio of proposals and we get

\[ \frac{q(\theta | \theta^*)}{q(\theta^* | \theta)} = 2^{(v + 1)} C_v \frac{\pi(f_{-1} | d_{-1}, c, u, z, t)}{\pi(f_{-1} | d_{-1}, c, u, z, t)} / \pi(f_{k+1}^* | f_{k+1}, d_{k+1}, f_{k+1}, m, u, c^*, z, t) \quad (A9) \]

while the last term in Equation A9 factorizes as

\[ \pi(f_{-1} | f_{-1}, d_{-1}, m, u, c^*, z, t) = \pi(f_{-1} | d_{-1}, m, u, c^*, z, t) \pi(f_{k+1}^* | f_{k+1}, d_{k+1}, m, u, c^*, z, t). \quad (A10) \]

The likelihood in Equation A7 has the expression given in Equation 4.

We have just to give more explicit expressions for the terms involving frequencies in Equations A8 and A9. These conditional distributions are Dirichlet by standard conjugacy properties. Denoting by \( n_{k} \) (resp. \( n_{k+1}^* \)) the number of copies for locus \( l \) observed in population \( k \) for the current state (resp. the proposed state), we get

\[ \frac{\pi(f_{-1} | d_{-1}, c, u, z, t)}{\pi(f_{-1} | d_{-1}, c, u, z, t)} = \prod_{j=1}^{l} \frac{\Gamma(q_{jk} + n_{k})}{\Gamma(q_{jk})} \frac{\Gamma(q_{k+1}^* + n_{k}^*)}{\Gamma(q_{k+1}^*)}. \quad (A11) \]

where \( n_{k} = \sum_{j=1}^{l} n_{jk} \) and \( q_{jk} = d_{jk}/(1 - d_{jk}) \) (and similar obvious notation for \( n_{k+1}^* \) and \( q_{k+1}^* \)).

Combined with Equations A7–A9 we end up with

\[ \mathcal{R}(\theta, \theta^*) = 2^{(v + 1) C_v} \left( \frac{K}{K + 1} \right)^{\alpha} \prod_{j=1}^{l} \frac{\Gamma(q_{jk} + n_{jk})}{\Gamma(q_{jk})} \frac{\Gamma(q_{k+1}^* + n_{k+1}^*)}{\Gamma(q_{k+1}^*)} \times \prod_{j=1}^{l} \frac{\Gamma(f_{jk} q_{jk}^* + n_{k+1}^*)}{\Gamma(f_{jk} q_{jk}^*)} f_{jk}^{n_{jk} - n_{jk}^*} \frac{\Gamma(f_{jk} q_{jk} + n_{k})}{\Gamma(f_{jk} q_{jk})} f_{jk}^{n_{jk} - n_{jk}^*}. \quad (A12) \]

Model nesting: The same computer code can be used to make inference in spatial and nonspatial models (i.e., with i.i.d. class variable \( c \)) as well as in the F-model and D-model. The more general algorithm performs simulations of the posterior in the spatial F-model. The nonspatial version can be obtained with \( \lambda \) set to a very large value, but, more efficiently, by initializing the number of tiles to the number of individuals, with the center of the tiles at the position of the individuals (\( m = n \) and \( u = t \)), with \( \epsilon_i \) to 0 (\( i = 1, \ldots, n \)), and then by skipping steps (5)–(7).

Simulations of the posterior in the D-model instead of the F-model can be made by initializing \( d_i = 0.5, f_{ij} = 1 \) and skipping steps (1) and (2).