Computational Statistics (2008) 23:111–129
DOI 10.1007/s00180-007-0072-x

ORIGINAL PAPER

Bayesian spatial modeling of genetic population structure

Jukka Corander · Jukka Sirén · Elja Arjas

Accepted: 14 September 2006 / Published online: 13 July 2007

Abstract Natural populations of living organisms often have complex histories consisting of phases of expansion and decline, and the migratory patterns within them may fluctuate over space and time. When parts of a population become relatively isolated, e.g., due to geographical barriers, stochastic forces reshape certain DNA characteristics of the individuals over generations such that they reflect the restricted migration and mating/reproduction patterns. Such populations are typically termed as genetically structured and they may be statistically represented in terms of several clusters between which DNA variations differ clearly from each other. When detailed knowledge of the ancestry of a natural population is lacking, the DNA characteristics of a sample of current generation individuals often provide a wealth of information in this respect. Several statistical approaches to model-based clustering of such data have been introduced, and in particular, the Bayesian approach to modeling the genetic structure of a population has attained a vivid interest among biologists. However, the possibility of utilizing spatial information from sampled individuals in the inference about genetic clusters has been incorporated into such analyses only very recently. While the standard Bayesian hierarchical modeling techniques through Markov chain Monte Carlo simulation provide flexible means for describing even subtle patterns in data, they may also result in computationally challenging procedures in practical data analysis. Here we develop a method for modeling the spatial genetic structure using a combination of analytical and stochastic methods. We achieve this by extending a novel theory of Bayesian predictive classification with the spatial information available, described here in terms of a colored Voronoi tessellation over the sample domain. Our results for real and simulated data sets illustrate well the benefits of incorporating spatial information to such an analysis.

J. Corander (✉) · J. Sirén · E. Arjas
Department of Mathematics and Statistics, University of Helsinki,
P. O. Box 68, 00014 Helsinki, Finland
e-mail: corander@mappi.helsinki.fi

Keywords Bayesian inference · Genetic structure · Spatial modeling · Statistical learning theory · Unsupervised classification

1 Introduction

The study of ancestral histories of natural populations is a common theme within a broad range of fields of biological sciences. For instance, in conservation biology it is often of importance to understand how a specific population of endangered species living in a fragile and discontinous habitat has been founded, and whether it is supported by migration from other geographical areas. In general, natural populations often have complex histories consisting of phases of expansion and decline, and the migratory patterns within them may fluctuate over space and time. When parts of a population become relatively isolated from each other, e.g., due to changing environmental conditions or geographical barriers, stochastic forces reshape certain DNA characteristics of the individuals over generations such that they reflect the restricted migration and reproduction patterns. Thus, the study of the DNA characteristics of a sample of individuals from a population can reveal detailed information about the ancestry of the population. For an excellent introduction to the principles of population genetics underlying any such analyses, see Hartl and Clark (1997).

A population consisting of several parts having a certain set of their DNA characteristics divergent from each other, is typically termed as genetically structured. The statistical analysis of populations structured in this sense was pioneered by Wright (1951, 1965), who derived various summary statistics for molecular data that could be used to judge the degree of genetic differentiation within a population. An analogous approach has been used almost exclusively until relatively recently, when several statistical methods utilizing model-based clustering for the analysis of such data were introduced. In the model-based approach, the genetically divergent parts of a population are represented as latent clusters which are inferred from the DNA characteristics of a sample of individuals. In particular, the Bayesian statistical approach has in this context attained a vivid interest among biologists. Several models and special-purpose software packages have been introduced, e.g., Pritchard et al. (2000), Dawson and Belkhir (2001), Falush et al. (2003), Corander et al. (2003, 2004, 2007). Also, the versatility of Bayesian models for the assignment of individuals into their geographical origin using molecular information has been recognized, see Rannala and Mountain (1997), Pella and Masuda (2001), Corander et al. (2006). However, only very recently, the possibility of incorporating the spatial information in a sample to the model of population structure has been considered by Guillot et al. (2005) and Wasser et al. (2004). Related earlier Bayesian spatial methods for analyzing molecular data were developed in Vounatsou et al. (2000), and Gelfand and Vounatsou (2003).

The traditional and still widely applied approach to examining the genetic structure of a population from a spatial perspective was developed very early, see e.g., Wright (1943), Kimura and Weiss (1964) and Sawyer (1977). Generally, in this framework, a simple regression of pairwise genetic distances between the sampled units on their geographical distance is considered. Since this necessitates a pre-assignment of the sampled individuals into biologically relevant units, the recent, more flexible Bayesian

models are expected to soon gain popularity among biologists. A modern, although not model-based approach to estimating geographical positions of genetic barriers in a population was introduced in Manni et al. (2004).

The main focus of Wasser et al. (2004) was on a continuous assignment of sampled individuals from unknown origins to a geographical map, given their DNA characteristics and a priori reference information about variation in DNA characteristics over different regions of the map, the latter provided in the form of individuals with known map coordinates. Guillot et al. (2005), in turn, assumed the spatial coordinates for the sample individuals to be known, and showed that such information can be utilized effectively to complement relatively sparse molecular data. The methodology introduced here has similar aims as Guillot et al. (2005), however, our model and computational techniques are different.

One of the main practical obstacles in using Markov chain Monte Carlo (MCMC) simulations in the context of complex hierarchical Bayesian models, is the need to manually tailor and monitor the algorithms to avoid misleading inferences due to convergence problems and label switching. In particular, for large data sets with a complex structure, the efforts needed to practically control such analysis may be prohibitive for a scientist outside the statistical community. Therefore, our aim is to develop a computationally efficient model for Bayesian spatial modeling of genetic population structure, which requires a minimum of statistical expertise on the part of a user. The methods introduced here are freely available in the BAPS software downloadable at http://www.rni.helsinki.fi/jic/bapspage.html.

To derive the present Bayesian spatial model for the genetic structure, we utilize recently introduced statistical learning theory for unsupervised classification, see Corander et al. (2004, 2007). The structure of the paper is as follows. The statistical learning task connected with the unsupervised classification relating to genetic population structure is formulated in Sect. 2. Thereafter, we consider incorporation of the spatial information to the learning problem. Algorithms for inference are described in Sect. 4, and the method is illustrated by analyses of both real and simulated data sets in Sect. 5. Some remarks are given in the final section.

2 Statistical learning about genetically structured populations

Genetic population structure is primarily investigated using so called molecular markers, which are also commonly utilized for forensics purposes (see, e.g., Balding and Nichols 1997). Markers can be expressed DNA regions (genes) or DNA segments that have no known coding function but whose inheritance pattern can be followed. DNA sequence differences are especially useful markers because they are abundant and easy to characterize precisely. Markers must be polymorphic to be useful, that is, alternative forms (called alleles) must exist among individuals so that they are detectable among different members of the investigated population.

The percentage of an allele at a specific marker site in the genome (called locus) in a given time in a population is typically referred to as the allele frequency, which can be statistically estimated by sampling a number of individuals and analyzing their DNA in a laboratory. Although allele frequencies tend to remain fairly constant

in a population over generations, however, a number of distinct forces affect them. The most basic neutral force (i.e. not related to the natural selection) is the genetic drift, which corresponds to random fluctuations in allele frequencies over generations due to occurring mating patterns in the population. In cases where migration (even indirectly) between two parts of a population is negligible, the genetic drift causes the allele frequencies of these parts to become increasingly distinct over generations if they had earlier been equal. This effect is further intensified by decreasing sizes of the population parts. A dramatic form of such changes is called a genetic bottleneck, where a sudden and extensive decrease has taken place in the number of individuals able to reproduce. A similar scenario is caused by the founder effect, where a small group of individuals act as founders of a new local population. Rearrangements in the DNA composition due to mutations also affect the allele frequencies over generations, although such changes typically take place in a quite slow pace for a given marker set.

Statistically, the possibility to distinguish samples from genetically differentiated parts of a population stems from the fact that the underlying allele distributions are different over the considered marker loci. This renders the joint observation of certain allele combinations over the loci much more probable in one part of the population than in another part. By scoring sampled individuals at a multitude of independent molecular marker loci, it is possible to increase considerably the statistical power to detect even quite subtle patterns in the genetic structure and to identify potential migrants or close descendants of such individuals. However, the statistical power also depends on a number of factors such as the sample size, degree of genetic differentiation between different parts of the investigated population and the level of polymorphism of the used markers (i.e. cardinalities of the existing allelic forms).

Our method will be suitable for considering common molecular marker types, such as microsatellites, single-nucleotide polymorphisms (SNPs), and amplified fragment length polymorphisms (AFLPs). However, for dominant markers, such as AFLPs, our model targets the genotype (a pair of observed alleles at a marker locus) frequencies in a population, instead of the allele frequencies which are considered for co-dominant marker types. The formulae given in the sequel refer to the co-dominant case, and should be re-interpreted accordingly when genotype frequencies are considered.

The marker loci are assumed to be unlinked (conditionally independent), which is generally a valid assumption in biological studies of the kind considered here. In most studies some of the alleles remain unresolved for a subset of the individuals due to problems with DNA amplification, which leads to missing data. When the occurrence pattern of the missing data is assumed to be random, it is appropriately taken into account by our model. More explicitly, the sufficient statistics emerging for our classification model reflect the potential differences in the amount of information present in the data over the loci.

Consider a sample of n individuals indexed by i, and suppose that DNA samples are available from them. For notational simplicity, we assume that the individuals represent a diploid species (i.e. having two copies of the basic number of chromosomes), however, our model can be applied to haploid (one copy) or tetraploid (four copies) data as well. Let (y_{i1}, y_{i2}) denote coordinates on a plane, representing the geographic origin of individual i. The spatial information will be jointly abbreviated as $\mathbf{y}$. Genotypes over N_L molecular marker loci can be specified as an $N_L \times 2$ genetic profile matrix

$\mathbf{x}^{(i)}$, where the elements (alleles) are coded by integers, such that for each row (locus) j there are r_j, $j = 1, \ldots, N_L$, possible distinct values. In the sequel we use extensively a set notation where $N = \{1, \ldots, n\}$ denotes the sample individuals. The observed genetic data will be jointly referred to as $\mathbf{x}^{(N)}$, whereas for an arbitrary subset $s \subseteq N$ of the individuals, $\mathbf{x}^{(s)}$ is used.

As stated earlier, the biological question of interest here, is the putative existence of genetically distinct parts in the studied population, that is areas where alleles in at least some of the marker loci are systematically appearing in distinct proportions compared to the other areas. This question can be investigated using the traditional latent class model (Duda et al. 2003) with a fixed number k of classes, as was done, e.g., in Pritchard et al. (2000), and Falush et al. (2003). The latent class model specifies the genetic structure of a population in terms of a mixture of product multinomial distributions for the observed allele patterns.

A different approach was taken by Corander et al. (2003, 2004), who represented the genetic structure in terms of a partition of the sampled individuals, and showed the advantages of this modeling strategy, though without the spatial information considered here. Later, the partition based approach to unsupervised classification was more formally derived in a general context in Corander et al. (2007). Here we utilize in particular the results of Corander et al. (2006, 2007) to specify the model and to make inferences from $\mathbf{x}^{(N)}$ and $\mathbf{y}$.

Let $S = (s_1, \ldots, s_k)$ $(1 \leq k \leq n)$ be a partition of N, i.e., a collection of classes of N, such that $\bigcup_{c=1}^{k} s_c = N$, $s_c \cap s_{c^*} = \varnothing$, for all $c, c^* = 1, \ldots, k, c \neq c^*$. Let $\mathcal{S}$ denote the space of all such partitions for a given n. The partition will be given a biological relevance by stating that the individuals allocated in the same class represent a sample from a so called random mating unit of the investigated population, whereas any collection of classes does not form a random mating unit, due to differences in the underlying frequencies the alleles are present in these units (for further details about random mating populations, see, e.g., Hartl and Clark 1997). Thus, S can be interpreted as a description of the genetic structure of a population in terms of the sample N.

By incorporating the spatial information contained in $\mathbf{y}$ to the Bayesian predictive classification framework utilized in Corander et al. (2004), we can write the prior predictive distribution for the molecular data as

$$p(\mathbf{x}^{(N)}|\mathbf{y}) = \sum_{S \in \mathcal{S}} p(\mathbf{x}^{(N)}|S)p(S|\mathbf{y}). \tag{1}$$

Here $p(S|\mathbf{y})$ describes the a priori uncertainty about S, given the spatial information, and $p(\mathbf{x}^{(N)}|S)$ is the prior predictive distribution of the molecular data given the structure. The posterior distribution of S is now determined by Bayes' rule as

$$p(S|\mathbf{x}^{(N)}, \mathbf{y}) = \frac{p(\mathbf{x}^{(N)}|S)p(S|\mathbf{y})}{\sum_{S \in \mathcal{S}} p(\mathbf{x}^{(N)}|S)p(S|\mathbf{y})}. \tag{2}$$

The uniform prior for S utilized in the earlier applications of this model is not always a biologically motivated choice in the presence of the spatial information. Therefore,

in particular when the amount of molecular information is only modest, we expect to strengthen the inferences about S by introducing a biologically meaningful spatial dependence through $p(S|\mathbf{y})$. However, in situations where the molecular information is overwhelming (i.e. N_L is large), the prior has a negligible impact on the posterior inferences, and the two approaches will tend to yield similar conclusions.

It was shown by Corander et al. (2007) in a general context for item classification using discrete-valued observed features that under an extension of de Finetti's exchangeability condition, the prior predictive distribution given a structure S has the unique product form

$$p(\mathbf{x}^{(N)}|S) = \prod_{c=1}^{k} p\left(\mathbf{x}^{(s_c)}\right) = \prod_{c=1}^{k}\prod_{j=1}^{N_L}\left[\frac{\Gamma\left(\sum_{l=1}^{r_j}\lambda_{cjl}\right)}{\Gamma\left(\sum_{l=1}^{r_j}\left(\lambda_{cjl}+n_{cjl}\right)\right)}\prod_{l=1}^{r_j}\frac{\Gamma\left(\lambda_{cjk}+n_{cjl}\right)}{\Gamma\left(\lambda_{cjl}\right)}\right], \tag{3}$$

where $\Gamma(\cdot)$ is the gamma function, and here the collection of counts n_{cjl} over alleles and loci is identified as a sufficient statistic for the data from the individuals allocated in the class s_c. Furthermore, the hyperparameter $\lambda_{cjl} > 0$, for all index values. This hyperparameter has the operational meaning of representing the a priori beliefs about the allele frequencies underlying the classes in the population structure. As in the earlier applications of this classification model, we use the reference value

$$\lambda_{cjl} = r_j^{-1}, \quad l = 1, \ldots, r_j; \;\; j = 1, \ldots, N_L; \;\; c = 1, \ldots, k, \tag{4}$$

which was derived originally by Perks (1947). In the case of dominant markers (such as AFLPs), the quantities n_{cjl} and λ_{cjl} in the above formulae refer to genotype counts and genotype frequencies, respectively.

3 Incorporating the spatial information to the structure model

In the earlier methods for modeling genetic population structure spatial information has been accounted for in a number of different ways. For instance, Gelfand and Vounatsou (2003) used a fixed Voronoi tessellation to specify a neighborhood structure for the spatial domain of the sample, whereas Guillot et al. (2005) modelled the domain by using a tessellation in which the number, shapes and locations of the cells were all random. The genetic structure can in such a framework be conveniently expressed in terms of a coloring of the cells, such that each genetically distinct unit is represented by a unique color. In general the Voronoi tessellation is considered a very versatile concept (see e.g., http://www.voronoi.com), and Bayesian random tessellation models have been earlier considered in a wide variety of spatial contexts, see e.g., Heikkinen and Arjas (1998, 1999), and Denison and Holmes (2001).

A Voronoi tessellation is a decomposition of a metric space (here a bounded 2-dimensional domain) determined by distances to a specified discrete set of points

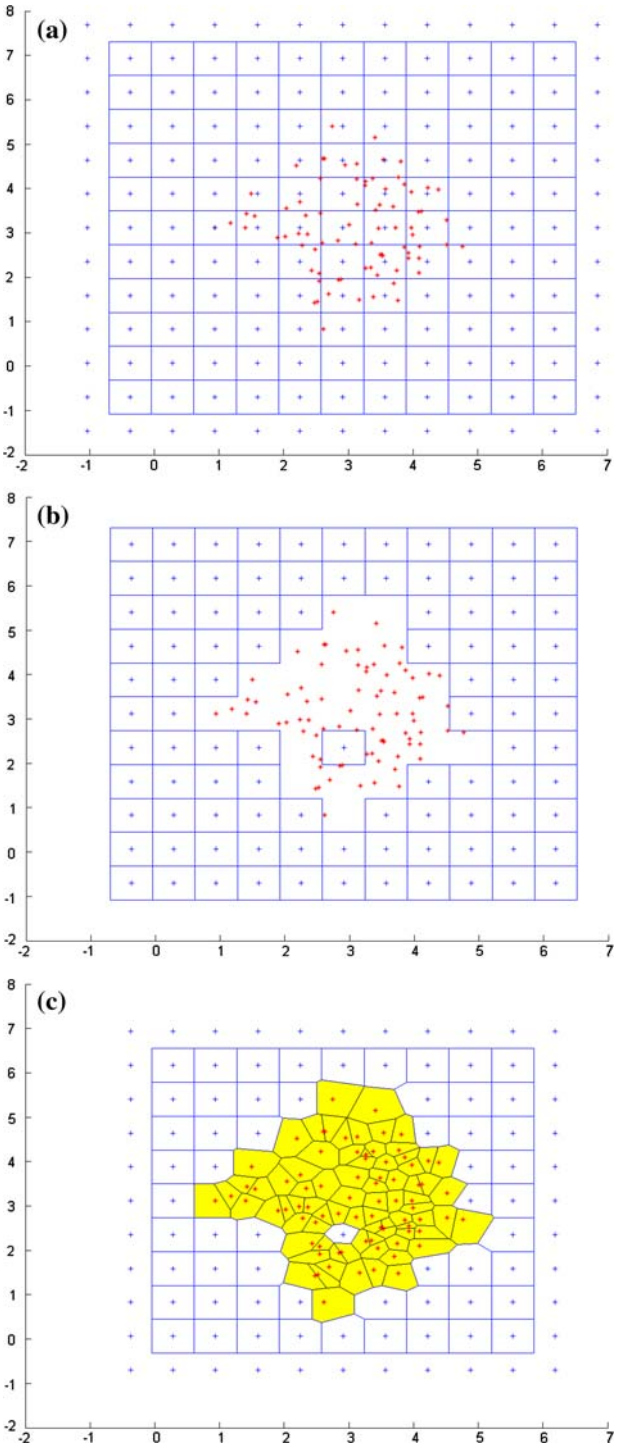
(here the coordinates of the n individuals in N). In general, for a given point (y_{i1}, y_{i2}) in the discrete set of coordinates, the set of all points closer to (y_{i1}, y_{i2}) than to any other coordinate point of N is the interior of a convex polytope called the Voronoi cell for (y_{i1}, y_{i2}). It is important to notice that the dual for a Voronoi tessellation is the Delaunay triangulation for the same set of points. The Delaunay triangulation for a set of points in the plane is the triangulation such that none of the points is inside the circumcircle of any triangle. The usefulness of the Voronoi tessellation in the current modeling context stems from two fundamental aspects; the numerical tractability for even quite extensive coordinate sets, and the possibility to represent efficiently joint prior probabilities over local neighborhoods in a spatial domain.

Let Δ be a spatial domain including the points in $\mathbf{y}$. As in Gelfand and Vounatsou (2003), we create a neighborhood structure in Δ by a Delaunay triangulation of the points in $\mathbf{y}$. However, here the triangulation is subject to certain restrictions imposed by a regular grid over Δ. Let $\min \mathbf{y}_l$ and $\max \mathbf{y}_l$ denote the minimum and maximum values of the coordinate l, $l = 1, 2$, in $\mathbf{y}$, respectively. Further, let $\beta_1 > 0$, $\beta_2 > 0$, be specified constants, which are used to determine a regular grid over Δ. For simplicity, assume that no two individuals have exactly equal coordinates (the relaxation of this assumption is discussed in the final section). The number of points on a row of the grid is defined as $n_{\text{row}} = \lceil n^{\beta_1} \rceil + \beta_2$, leading to the total of n_{row}^2 points in the whole grid. The purpose of β_2 is to determine the number of grid points allocated between the boundaries of Δ and the outermost data points. Here $\lceil n^{\beta_1} \rceil$ denotes the smallest integer greater than or equal to n^{β_1}. The distances between the grid points are defined for the two axes as:

$$(\max \mathbf{y}_l - \min \mathbf{y}_l)/(n_{\text{row}} - \beta_2), \quad l = 1, 2. \tag{5}$$

Given the grid, a Voronoi tessellation is constructed from its points, leading to a single square for each of these. Thereafter, the data points $\mathbf{y}$ are examined, to determine whether an arbitrary square includes any points. For those squares lacking any points, the coordinates of the corresponding grid point are added to a support vector $\mathbf{z}$. Finally, a Delaunay triangulation is formed from the set $(\mathbf{y}, \mathbf{z})$ of points. To numerically construct the tessellations and the Delaunay triangulation we use the Qhull algorithm, see Barber et al. (1996). The used triangulation effectively avoids including potential gap areas in the domain Δ to the natural neighborhood of the original data points. We have experimented with various choices of β_1 and β_2, concluding that values in the intervals $\beta_1 \in [1/4, 1/2]$, $\beta_2 \in [5, 10]$, seem to work appropriately for numerous different data configurations. In Fig. 1, the constrained triangulation is illustrated for a set of random coordinates.

The resulting Voronoi tessellation can be connected with the genetic structure S by coloring each cell according to the class membership of the particular individual. Thus, with k classes in S, there will also be k colors in the tessellation. It is clear that the distribution implied on the coloring by the uniform prior on S, is not in general biologically sensible since the colors would not be expected to be randomly scattered over the spatial domain. A non-uniform prior on S expressing such beliefs may then be derived by utilizing the theory of graphical models, see e.g., Lauritzen (1996). Our


(a)
(b)
(c)

◀ **Fig. 1** An example of a constrained Delaunay triangulation (**a,b**) and the resulting tessellation (**c**) for a random set of points, based on the regular grid approach. Four population kernels ($k = 4$) were simulated from a uniform distribution over the unit circle, and final coordinates for 20 individuals for each particular population were then randomly scattered around the corresponding kernel. The random coordinates were generated by adding to the kernel coordinates the sum of $Z_1 \sin(\pi/k)\frac{4}{5}$ and $Z_2 \sin(\pi/k)\frac{2}{5}$, where Z_1 is a random number from the standard normal distribution, and Z_2 a random number from the uniform distribution over [0, 1]. This simulation framework allows a straightforward generation of spatially ordered populations with a certain degree of overlap at the edges. The values $\beta_1 = 0.4, \beta_2 = 7$ were used for creating the grid in the figure

derivation uses several central concepts of graph theory and graphical models, the exact definitions of which can be found in Lauritzen (1996).

Given the constrained Delaunay triangulation, a neighborhood structure for the domain is here introduced by considering an undirected graph $G = G(N, E)$ with nodes corresponding to the points marked with the individual labels. The edge set E of the graph G is defined by setting $\{i, j\} \in E$ if and only if i and j belong to the same triangle, for all $i, j \in N, i \neq j$.

A graph G is called *complete* when all possible pairs of nodes are connected with an edge. Let $a \subset N$ be a subset of the nodes. Such a subset is called a *clique*, if (1) the subgraph G_a of G on the nodes in a is complete, and (2) any addition of arbitrary nodes in $N \backslash a$ to a results in a non-complete subgraph of G. Hence, a clique corresponds to a maximally complete subgraph. A *chordless cycle* of length l of a graph G is a sequence of l nodes, such that only successive nodes in the sequence are adjacent in the graph. When a graph does not contain any chordless cycles of length $l \geq 4$, it is called *chordal*. For a chordal graph there exists a *perfect ordering*, according to which the cliques can be arranged successively into a *tree*. The *separators* of the cliques correspond to the intersections of successive cliques in such a tree. Both the sets of cliques and separators, say $cl(G)$ and $sep(G)$, respectively, are uniquely determined for a chordal graph G.

In the sequel it is assumed that G is chordal. In cases where the original graph resulting from the Delaunay triangulation is non-chordal, we replace it by a chordal version obtained by adding a minimum number of edges. For this generally considered NP-complete problem, we use an algorithm analogous to that introduced in Berry (1999).

Using the graph G, we now derive a prior expressing biological beliefs about a fairly smooth spatial distribution of the points representing the individuals in each class s_c. Let $\gamma_i \in \{1, \ldots, k\}$ denote the color index of tessellation cell $i, i = 1, \ldots, n$, with respect to a genetic structure having k classes. Assume for a moment that, a priori, these colors appear independently of each other, which ignores any spatial dependence of the individuals. Then we can specify the joint distribution of the color indices by

$$p(\gamma_1, \ldots, \gamma_n) = \prod_{i=1}^{n} p(\gamma_i) = \prod_{c=1}^{k} p(\gamma(s_c))^{|s_c|}, \tag{6}$$

where $\gamma(s_c)$ indicates the color of class s_c. The operational meaning of the probabilities in the above formula is to represent the relative abundance of the colors in the tessellation.

To introduce spatial dependence among the tessellation cells, we consider a model for the dependence of the occurrence of a color in a particular cell, on the colors appearing in the neighborhood of the cell. The rationale behind this is to, via such a dependence structure, arrive at a smaller effective number of parameters. Using the standard results for graphical models given in Lauritzen (1996), we can write the joint probability of the coloring under the dependence structure defined according to graph G as

$$\frac{\prod_{a \in cl(G^*)} p(\gamma_a)}{\prod_{d \in sep(G^*)} p(\gamma_d)}, \tag{7}$$

where $p(\gamma_a)$ is the probability of the colors appearing in any clique $a \subset N$, and $p(\gamma_d)$ is the corresponding probability for the separator $d \subset N$.

Let $m(c, a)$ be the count of color c appearing in clique a of G. The maximum number of colors appearing in clique a equals $\min(k, |a|)$. Using a Dirichlet(α,…,α) prior with $\alpha = 1/\min(k, |a|)$ for the probabilities of the colors leads to the following marginal expression for the observed color counts

$$t(a) = \frac{\Gamma\left(\sum_{c=1}^{\min(k,|a|)} \alpha\right)}{\Gamma\left(\sum_{c=1}^{\min(k,|a|)} \alpha + m(c, a)\right)} \prod_{c=1}^{\min(k,|a|)} \frac{\Gamma\left(\alpha + m(c, a)\right)}{\Gamma(\alpha)}. \tag{8}$$

For separators, an analogous definition is used. The prior for the tessellation coloring, and consequently for the structure parameter S, is now specified by

$$p(S|\mathbf{y}) \propto \frac{\prod_{a \in cl(G)} t(a)}{\prod_{d \in sep(G)} t(d)}. \tag{9}$$

In the sequel we refer to this as the spatial prior for the genetic structure, on the contrary to the non-spatial, uniform prior used in Corander et al. (2003, 2004, 2006, 2007).

4 Inference about the genetic structure

Overall, the possibilities for practical implementation of Bayesian hierarchical models have been steadily improving with the developments in the MCMC algorithm theory (for a recent review see Andrieu et al. 2004). However, it is well-known that mixture models, in particular with a variable number of components, pose often a considerable challenge for the algorithms to work in practice. A major issue for mixture models specified in terms of latent classes is consensus inference about the number of mixture components that are supported by the data. Also, label-switching in the MCMC inference for mixtures with a varying number of components is a well-known problem, as was discussed in the current context by Guillot et al. (2005).

In the present partition-based classification framework the labels attached to the classes carry no meaning whatsoever since the classes are identified through their contents. Furthermore, exclusion of the prevalence parameters representing the relative sizes of the latent classes in the sample enables the derivation of the predictive likelihood in an analytical form, given a particular value of the structure. This is important for samples representing a complex underlying population structure, since the inferences can then be made on a numerically more stable basis.

Corander et al. (2007) introduced a non-reversible parallel MCMC algorithm to resolve problems with the standard Metropolis-Hastings algorithm for variable-dimensional classification models. Furthermore, Corander et al. (2006) developed a less computationally intensive stochastic search method utilizing intelligent moves in the partition space $\mathcal{S}$. Here we have applied the latter algorithm to learning the genetic structure under the spatial prior.

An important reason for introducing the spatial dependence aspect, as was done in the previous section, is to arrive at a model specification in which the posterior of the genetic structure S can be obtained in an analytic form, albeit with an unknown normalizing constant. This enhances considerably the possibilities to apply the classification model to the challenging data sets that are likely to arise in practice. It should be noticed that under the standard latent class mixture approach, fitted using the basic Gibbs sampler algorithm, the model may easily contain up to several thousands of parameters within a single iteration.

A straightforward Bayesian estimate for the genetic structure under the current framework is the partition associated with maximum posterior probability,

$$\hat{S} = \arg\max_{S \in \mathcal{S}} p(S|\mathbf{x}^{(N)}, \mathbf{y}). \tag{10}$$

Clearly, the mode cannot be identified through enumerative methods, except for unrealistically small data sets ($n \leq 12$). To solve this problem, Corander et al. (2007) introduced a non-reversible parallel Metropolis-Hastings algorithm, which utilizes a number of dependent stochastic processes. Although their parallel solution to identifying the posterior mode has been shown to produce excellent results, it is computationally demanding in particular when implemented under a single-machine architecture. Therefore, for complex data sets the convergence towards representative areas of $\mathcal{S}$ may take a long time. To improve the posterior estimation in this respect, we have implemented the spatial prior under the stochastic search algorithm developed in Corander et al. (2006). Brief details of the search operators are given below.

The maximum a posteriori estimate of the genetic structure is sought by using three distinct operations:

1. Re-allocation of single individuals to other classes (each individual is considered in a random order).
2. Combination of classes (each pair of classes is considered in a random order).
3. Split of classes into at least two subclasses guided by Hamming distances of individuals to identify maximally homogeneous subclasses (distances are calculated from marker data).

The Hamming distance is here defined as the number of differing alleles any two individuals carry over the loci, such that at any particular locus the genotypic difference is between zero and two alleles.

Each proposal operation is accepted when it leads to an improvement in the posterior $p(S|\mathbf{x}^{(N)}, \mathbf{y})$, except for the split step, where even inferior states are temporarily accepted. As opposed to MCMC, where random proposals are used, the intelligent search makes the computation considerably more efficient as it is able to identify better candidate structures both in the close vicinity of the current structure (re-allocation of single individuals) as well as at quite large distances (split/combine steps). In particular the random split operator used in Corander et al. (2004, 2007) is very inefficient compared to the splits guided by the Hamming distance.

The time complexity of our method is reasonable even for large data sets, as compared to the computation times required by MCMC methods. For a typical small data set with 100 individuals and 10–30 marker loci, 30 replicate runs of the posterior optimum search take only 1–2 min to perform on a standard PC (2.8 GHz Pentium 4 processor). For moderate sized data with, say 300–400 individuals, a corresponding analysis has taken approximately 1–2 h, and for large data sets with up to 1,000 individuals the computations have taken approximately 24 h. These figures are quite favorable to the computation times for standard MCMC algorithms, which in the current context are at least 10–20 fold. Also, our method has considerably low memory intensivity compared to MCMC, as the unknown allele frequencies for each of the classes are analytically treated and need not be stored in the Markov chains.

Representing the peakedness of the posterior distribution with respect to an estimated mode $\hat{S}$, is in general a complicated task. Since the dimensionality of the structure parameter is high, manual monitoring of the values in the estimated posterior distribution is tedious, if not practically impossible with standard computational facilities. Thus, it is important to characterize the uncertainty about the spatial structure by some reasonably manageable means. For this purpose, both graphical and numerical summaries can be used. To display the local uncertainty connected with the estimate $\hat{S}$, Corander et al. (2006) used conditional posterior probabilities over potential alternative classifications for each individual in the sample.

Let $S^*(i, c)$ denote the structure where individual i has been moved from its class in $\hat{S}$ to class c. The conditional posterior probabilities over the range of putative classifications of i are then defined by

$$\frac{p(\mathbf{x}^{(N)}|S^*(i,c))p(S^*(i,c)|\mathbf{y})}{\sum_{c=1}^{k} p(\mathbf{x}^{(N)}|S^*(i,c))p(S^*(i,c)|\mathbf{y})}, \quad c = 1, \ldots, k. \tag{11}$$

Overall, the probabilities (11) can be investigated to see to which extent the data supports alternative classifications of the individuals. In particular, when class c corresponds to the optimal classification determined by $\hat{S}$, the probability (11) measures the local support for keeping the color of the particular cell unchanged in the tessellation. A 3D graphical representation can then be provided, where the height of the surface of each tessellation cell is given by one minus the probability yielded by formula (11). As for typical data sets most values of (11) are expected to be fairly

high for the optimal classification, the reverse dependence on the posterior probility prevents the peaks of the 3D graph from obscuring the other parts. Furthermore, in our software implementation of the method, we have included the possibility to rotate the 3D graphics to an optimal viewing angle by the user.

5 Example analyses

To illustrate our spatial model, we use both simulated and real data sets. Allele frequencies for the simulated data have been generated such that for each population c, $c = 1, \ldots, k$, a random number u_{cjl} has first been drawn from the uniform distribution on [0,1], independently at each locus j, $j = 1, \ldots, N_L$, and allele l, $l = 1, \ldots, r_j$. Then, the allele frequencies of the populations p_{cjl} have been specified as $u_{cjl}^{(r_j-1)^{1/2}}$, to avoid them being "too uniform" at any particular locus. Given the generated allele frequencies, alleles were simulated independently for each individual and population. Both diploid and haploid types of data sets were generated.

To introduce spatially structured populations, we simulated k population kernels from a uniform distribution over the unit circle. The individuals of any particular population were then randomly scattered around the corresponding kernel by adding to the kernel coordinates a random number from the uniform distribution over $[0, 2\sin(\pi/k)]$. The presented results are based on the stochastic greedy search, modified to incorporate the spatial priors. In most cases repeated runs of the estimation algorithm for the same data set produced nearly identical results.

In Fig. 2 the estimated spatial genetic structure for the point configuration given in Fig. 1 is displayed. To generate the corresponding marker data, 16 loci each with four possible alleles were used for diploid individuals. It can be seen that the true underlying structure is resolved exactly, despite of the slight spatial discontinuity of the populations at some of the boundaries.

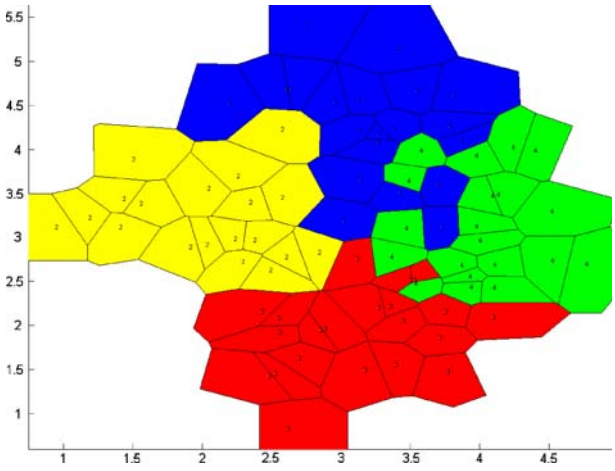


Fig. 2 The estimated spatial genetic structure for the point configuration given in Fig. 1

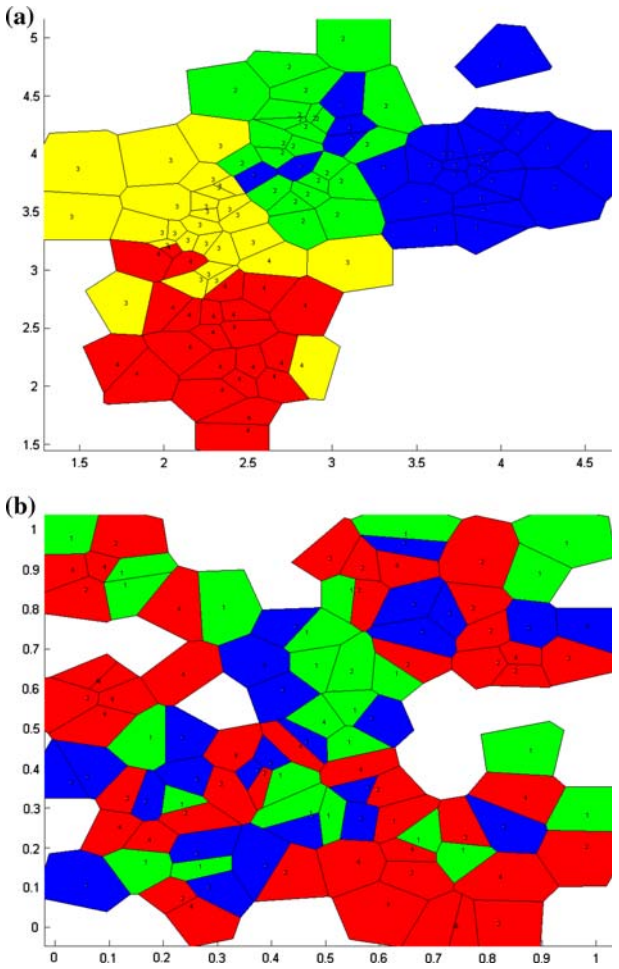


Fig. 3 The estimated genetic structure under spatially smooth (**a**) and random (**b**) coordinates for four distinct populations. The same molecular data is used for both estimates

An aspect of concern is the performance of the spatial prior in cases where a population is genetically structured in a non-spatial manner. It is clear that the prior may yield an over-smooth description of the molecular data under such circumstances. To compare the evidence displayed from the same molecular data, we generated the coordinates for the individuals under two distinct scenarios. One of these corresponded to the above spatial simulation scheme and the other allocated the individuals randomly over a domain. Molecular data was generated from four distinct populations for diploid individuals, over 14 loci each having four possible alleles. The posterior estimates of the genetic structure are displayed in Fig. 3a and b, respectively, and a summary is given in Table 1. It is seen that while the spatial prior discovers the true population structure almost perfectly (4 misclassified individuals in total) for the spatially ordered data (referred to as the data set 5 in Table 1), a slightly over-smooth

Table 1 Classification results under various spatially ordered sampling configurations and two distinct priors

Data 1: $n = 80, k = 3, N_L = 16$	MCI	$\hat{k}$
Spatial prior	2	4
Uniform prior	3	17
Data 2: $n = 90, k = 5, N_L = 15$	MCI	$\hat{k}$
Spatial prior	3	5
Uniform prior	19	17
Data 3: $n = 100, k = 2, N_L = 10$	MCI	$\hat{k}$
Spatial prior	2	2
Uniform prior	2	20
Data 4: $n = 90, k = 5, N_L = 15$	MCI	$\hat{k}$
Spatial prior	3	5
Uniform prior	19	17
Data 5: $n = 100, k = 4, N_L = 14$	MCI	$\hat{k}$
Spatial prior	4	4
Uniform prior	16	13
Data 6: $n = 100, k = 4, N_L = 14$	MCI	$\hat{k}$
Spatial prior	29	3
Uniform prior	16	13

The uniform prior refers to a uniform distribution over $\mathcal{S}$. MCI refers to the number of misclassified individuals with respect to the spatial reference populations, i.e. an individual is counted as misclassified when placed in a class where the majority of individuals represents another spatial region. For data set 2, the underlying true population sizes were 35, 30, 10, 10, 5. In other data sets all populations have an equal number of individuals. Individuals in data sets 1–4 are haploid and the number of possible alleles per locus was five. Data sets 5 and 6 are those corresponding to the results presented in Figs. 3a and b, respectively. Since the molecular data is the same for these two data sets, the uniform prior yields the same result

estimate is obtained for the randomly scattered data (referred to as the data set 6 in Table 1). Two of the underlying populations are still well discovered, whereas the other two are combined. However, this simulation scenario represents in a sense an extreme test case, as genetically structured natural populations are in reality seldom expected to completely lack a spatial order with respect to the genetic structure. While the uniform prior allocated a majority (84/100) of the individuals correctly in the randomly scattered data, it also allowed the emergence of nine false small outlier classes (containing one or two individuals) in addition to the classes corresponding to the underlying populations.

Table 1 presents also results from other comparisons between the spatial and uniform priors, when the underlying structure is spatially relatively smooth (data sets 1–4). The spatial prior performed consistently well, whereas the uniform prior showed tendency to over-estimate k. However, in simulations with more informative genetic data (exact results not shown here), the two different methods tended to yield highly similar structures, which were all well in agreement with the underlying generating model. It is entirely reasonable that the importance of the prior diminishes with increasing

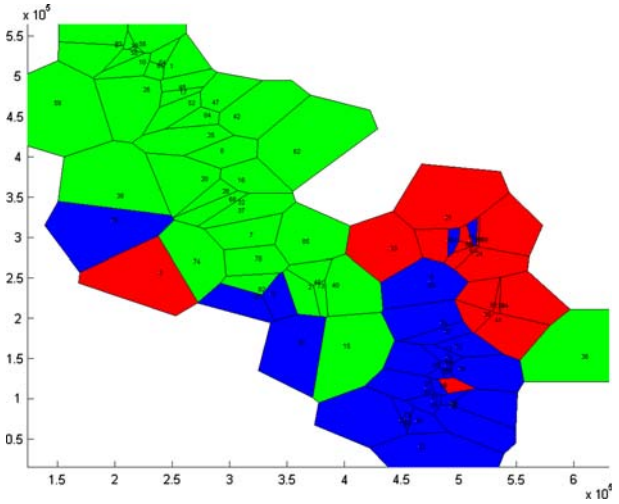


Fig. 4 Posterior mode tessellation for the wolverine data

number of informative marker loci. Surely, we do not wish to on the basis of these simulation results over-interpret the gain from using the spatial prior, but there is a clear tendency towards better predictive ability for small or moderate amounts of marker data. Generally, the earlier model based on a uniform prior has yielded biologically reasonable results in the numerous applications reported in the literature.

To illustrate the spatial model for real molecular data we consider a microsatellite data set, which has been investigated also by Guillot et al. (2005) using an alternative spatial Bayesian model. The data set introduced by Cegelski et al. (2003) contains 88 individuals of wolverines (*Gulo gulo*) with known coordinates from the Montana region in the Unites States, all of which were genotyped for ten microsatellite loci. Originally, Cegelski et al. (2003) found evidence for three genetically distinct groups using the non-spatial model of Pritchard et al. (2000). In Fig. 4, we show the estimated spatial genetic structure, and also, in Fig. 5, the local stability of the estimates is illustrated using the 3D graphics described in the previous section. This latter figure illustrates that the classification is very stable for a majority of the sample, whereas for some individuals there is substantial uncertainty about the classification.

Qualitatively, our results in Fig. 4 are well in agreement with those of Guillot et al. (2005). However, as discussed in their paper, empty "ghost" classes emerged in the MCMC simulation, and needed to be manually removed. The resulting spatial structure has four genetically distinct regions, of which three are similarly represented in our results.

6 Discussion

In the previous section we illustrated the potential of the spatial approach to inferring the genetic population structure in populations. Since our approach is computationally efficient even for large datasets, and can handle missing molecular observations, it is

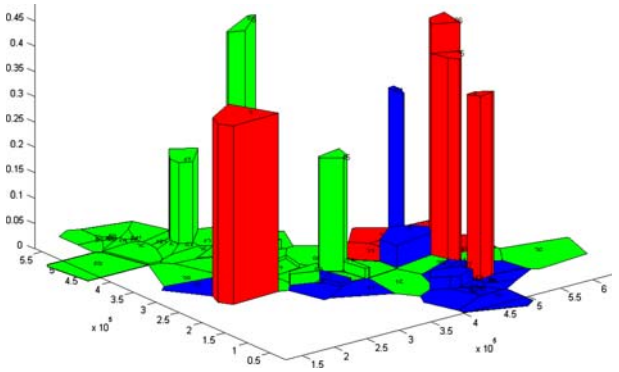


Fig. 5 Local uncertainty in the posterior mode tessellation for the wolverine data. The height of the surface of each tessellation cell is given by one minus the conditional posterior probability yielded by formula (11). Thus, for very flat cells the molecular data is decisively allocating the cell to a particular class as the conditional posterior probability is near unity

applicable to a wide range of real biological settings. However, it is also necessary to consider the meaningfulness of the assumed relative smoothness and the representativeness of the observed coordinates for each potential application. Often biologists have extensive a priori knowledge about the investigated species, which can help to judge the reasonability of the model assumptions and the results. Since we formulated our method without flexible hyperparameters, it may result in over-smoothing when the investigated population completely lacks a spatial structure. Fortunately, such instances are quite rare in the study of natural populations. A considerable advantage of our model formulation is the resulting computational framework which does not require manual monitoring of MCMC runs and specification of hyperprior settings. To arrive at the particular prior presented here, we originally experimented with a variety of distinct but related ideas based on the neighborhood representation in terms of the Delaunay triangulation. Among these alternatives, the presented prior showed most consistent behavior in analytical and simulation-based comparisons.

An important aspect of our spatial prior is its general tendency to express the idea that all genetic structures are not a priori equally likely, which has been used in Corander et al. (2003, 2004, 2007). However, the specified spatial prior is quite weak compared to the total amount of information contained in a moderately sized molecular data set, and therefore, deviations from spatial smoothness are still fairly easily allowed in the posterior inference (as is seen, e.g., in Fig. 3b). One purpose of the prior is to prevent weak stochastic fluctuations in the allele frequencies to emerge as evidence of genetic structure which would allow the number of clusters to increase without firm support. This is especially relevant for highly polymorphic microsatellite data from weakly structured populations, for which the uniform prior may yield small spurious outlier clusters in some situations. Under such circumstances, we feel that it is more optimal to have a conservative prior about the amount of genetic differentiation, which requires more molecular evidence to support an increase in the number of classes. In any case, it is highly recommendable to analyze data both with the spatial and non-spatial method, to investigate the congruence between the results. When these

disagree, genetic distances between individuals and groups of individuals may provide a useful insight concerning the biological plausibility of the inferences.

A central issue in the utilization of the spatial information is to consider how much the findings at a particular geographic location should impact the beliefs about genetic configuration of the population in areas where no observations have been made. Here we have restricted the use of spatial information to the immediate neighborhood of the observed points, as opposed to the model of Guillot et al. (2005), where the whole domain Δ is modeled through a colored tessellation. As is well visible in their analysis of the wolverine data, this may have the consequence that, in some areas of Δ, the posterior inference is merely reflecting the prior opinion, since there are no spatial data to be learned from. Under circumstances where the observed points are not well covering the domain Δ, one should be cautious in inferring generally the shapes and extents of the genetically distinct regions.

In Sect. 2 it was assumed that all individuals are observed with unique coordinates. In cases where this is not true, some random variation can be added to the equal coordinates before the analysis. However, it should be noted that if a large number of individuals are given nearly the same coordinates, the prior will give more support for them being genetically homogeneous. Such situations are primarily expected to arise when the sampling scheme is group-based, such that several individuals are collected at a restricted geographical site which can be intrepreted as a single unit from the evolutionary perspective. An example of such a sampling strategy is the nest-based sampling often used for the investigation of ants, see e.g., Seppä et al. (2004). A biologically more sensible approach to the spatial analysis of such data can then be obtained by a pre-grouping of the data in the same fashion as in Corander et al. (2003), since the model for the molecular data remains unchanged. When the samples consist of groups of individuals with coordinates given for each geographical area representing a group, the groups can be spatially clustered by our model. For many species, such as plants, it may indeed be more relevant to cluster the habitat patches of the individuals rather than the individuals themselves. In our software implementation of the method we have also included this additional possibility for the spatial analysis.

Acknowledgments This research has been supported by the Centre of Population Genetic Analyses, University of Oulu, Finland (Academy of Finland, grant no. 53297). The authors would like to thank Christine Cegelski for kindly providing access to the wolverine data.

References

Andrieu C, Doucet A, Robert CP (2004) Computational advances for and from Bayesian Analysis. Stat Sci 19:120–129

Balding DJ, Nichols RA (1997) Significant genetic correlations among Caucasians at forensic DNA loci. Heredity 78:583–589

Barber CB, Dobkin DP, Huhdanpaa HT (1996) The Quickhull algorithm for convex hulls. ACM Trans Math Software 22:469–483

Berry A (1999) A wide-range efficient algorithm for minimal triangulation. Proceedings of the tenth annual ACM-SIAM symposium on Discrete algorithms, Philadelphia, SIAM, pp 860–861

Cegelski CC, Waits LP, Anderson NJ (2003) Assessing population structure and gene flow in Montana wolverines (*Gulo gulo*) using assignment-based approaches. Mol Ecol 12:2907–2918

Corander J, Waldmann P, Sillanpää MJ (2003) Bayesian analysis of genetic differentiation between populations. Genetics 163:367–374
Corander J, Waldmann P, Marttinen P, Sillanpää MJ (2004) BAPS 2: enhanced possibilities for the analysis of genetic population structure. Bioinformatics 20:2363–2369
Corander J, Marttinen P, Mäntyniemi S (2006) Bayesian identification of stock mixtures from molecular marker data. Fish Bull 104:550–558
Corander J, Gyllenberg M, Koski T (2007) Bayesian unsupervised classification framework based on stochastic partitions of data and a parallel search strategy. Adv Data Analysis Classification, under review
Denison DGT, Holmes CC (2001) Bayesian partitioning for estimating disease risk. Biometrics 57:143–149
Duda RO, Hart PE, Stork DG (2000) Pattern classification, 2nd edn. Wiley, New York
Falush D, Stephens M, Pritchard JK (2003) Inference of population structure using multilocus genotype data: Linked loci and correlated allele frequencies. Genetics 164:1567–1587
Gelfand AE, Vounatsou P (2003) Proper multivariate conditional autoregressive models for spatial data analysis. Biostatistics 4:11–25
Guillot G, Estoup A, Mortier F, Cosson JF (2005) A spatial statistical model for landscape genetics. Genetics 170:1261–1280
Hartl DL, Clark AG (1997) Principles of population genetics, 3rd edn. Sinauer Associates, Sunderland
Heikkinen J, Arjas E (1998) Non-parametric Bayesian estimation of a spatial Poisson intensity. Scand J Statist 25:435–450
Heikkinen J, Arjas E (1999) Modeling a poisson forest in variable elevations: a nonparametric Bayesian approach. Biometrics 55:738–745
Kimura M, Weiss GH (1964) The stepping-stone model of population structure and the decrease of genetic correlation with distance. Genetics 49:561–576
Lauritzen SL (1996) Graphical models. Oxford University Press, Oxford
Manni F, Guérard E, Heyer E (2004) Geographic patterns of (genetic, morphologic, linguistic) variation: how barriers can be detected by "Monmonier's algorithm". Hum Biol 76:173–190
Pella J, Masuda M (2001) Bayesian methods for analysis of stock mixtures from genetic characters. Fish Bull 99:151–167
Perks W (1947) Some observations on inverse probability including a new indifference rule. J Inst Actuaries 73:285–334
Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. Genetics 155:945–959
Rannala B, Mountain JL (1997) Detecting immigration by using multilocus genotypes. PNAS 94:9197–9201
Seppä P, Gyllenstrand M, Corander J, Pamilo P (2004) Coexistence of the social types: Genetic population structure in the ant Formica exsecta. Evolution 58:2462–2471
Sawyer S (1977) Asymptotic properties of the equilibrium probability of identity in a geographically structured population. Adv Appl Prob 9:268–282
Vounatsou P, Smith T, Gelfand AE (2000) Spatial modeling of multinomial data with latent structure; an application to geographical mapping of human gene and haplotype frequencies. Biostatistics 1:177–189
Wasser SK, Shedlock AM, Comstock K, Ostrander EA, Mutayoba B, Stephens M (2004) Assigning African elephant DNA to geographic region of origin: Applications to the ivory trade. PNAS 101:14847–14852
Wright S (1943) Isolation by distance. Genetics 28:139–156
Wright S (1951) The genetical structure of populations. Ann Eugen 15:323–354
Wright S (1965) The interpretation of population structure by F-statistics with special regard to systems of mating. Evolution 52:950–956