

# Bayesian Clustering using Hidden Markov Random Fields in Spatial Population Genetics

Olivier François\*

Sophie Ancelet†

Gilles Guillot †

July 6, 2006

\*TIMC, TIMB (Department of Mathematical Biology), Faculty of Medicine, F38706  
La Tronche, France

†Unité de Mathématiques et Informatique Appliquées, ENGREF, 19 avenue du Maine,  
75732 Paris Cedex 15, France

Running Head: Bayesian clustering in spatial genetics

Key Words: Bayesian Clustering, Spatial Genetics, Continuous Populations, Inbreeding, Hidden Markov Random Fields, Assignment, Estimating the number of populations

Corresponding Author:

Olivier François

`olivier.francois@imag.fr`

## Abstract

We introduce a new Bayesian clustering algorithm for studying population structure using individually geo-referenced multilocus data sets. The algorithm is based on the concept of Hidden Markov Random Field which models the spatial dependencies at the cluster membership level. We argue that (i) a Markov Chain Monte-Carlo procedure can implement the algorithm efficiently, (ii) it can detect significant geographical discontinuities in allele frequencies, and regulate the number of clusters, (iii) it can check whether the clusters obtained without the use of spatial priors are robust to the hypothesis of discontinuous geographical variation in allele frequencies, (iv) it can reduce the number of loci required to obtain accurate assignments. We illustrate and discuss the implementation issues with the Scandinavian brown bear and the human CEPH diversity panel data set.

## INTRODUCTION

It has been a recent matter of debate to decide whether clusters identified by **Bayesian algorithms were artificially detected structures** emerging from uneven sampling along clines or were actually well-differentiated groups (SERRE and PÄÄBO 2004), (ROSENBERG *et al.* 2005). It has indeed been suggested that uneven sampling during the experimental design might influence clustering patterns, and that the degree of clustering might be diminished by use of samples with greater spatial homogeneity. This dilemma has even introduced doubt about whether Bayesian clustering algorithms are appropriate tools for studying genetic structure in populations with continuous variation of allele frequencies.

Such issues have been reported after a study of genetic structure of human populations by ROSENBERG *et al.* (2002). Without the use of predefined populations, this study inferred the geographical ancestries of individuals from 52 world-wide samples with individuals genotyped at 377 microsatellite loci. **Using the Bayesian clustering program STRUCTURE** (PRITCHARD *et al.* 2000) and increasing the number of loci from 377 to 993, ROSENBERG *et al.* (2005) have shown that the six clusters found in their previous study are robust, and, at the noticeable exception of the genetic isolate Kalash, that they match with the major geographic regions in the world. These clusters were interpreted as arising from small discontinuities in allele frequencies when geographical barriers are crossed.

In the latter and other applications of clustering algorithms, the spatial data are actually treated off-line and are not part of the modeling. **Bayesian models such as those developed by** PRITCHARD *et al.* (2000), DAWSON and BELKHIR (2001), or CORANDER *et al.* (2003), nevertheless offer a natural and appropriate framework for including spatial prior information when assigning an individual to a fixed number of clusters. For example, a recent study by (GUILLOT *et al.* 2005) has used spatial explicit priors in a full-Bayes perspective, and has successfully identified genetic barriers in a wolverine population. An assignment method was also used by WASSER *et al.* (2004) to infer the spatial origin of African elephants. Here we argue that modified Bayesian algorithms can provide addi-

tional evidence to solve cline/cluster dilemmas such as those discussed in ROSENBERG *et al.* (2005). A natural way to proceed is to include priors on continuous variation of genetic diversity in the Bayesian model used by STRUCTURE, and check whether the previously discussed clusters are robust or not.

In this study, we present a new hierarchical Bayes algorithm that incorporates models for geographical continuity of allele frequencies. This is achieved by using *Hidden Markov Random Fields* (HMRF) as prior distributions on cluster membership. **An informal definition of HMRFs states that allele frequencies at a specific geographical site are more likely to be close to the allele frequencies at neighboring sites than distant sites.** The problem of local differentiation may also be studied in terms of change in correlation with distance as considered by MALÉCOT (1948) where “individuals living nearby tend to be more alike than those living far apart” (KIMURA and WEISS 1964). The HMRF is basically another formulation of the same idea with statistical correlation hidden at the cluster membership level.

We illustrate some applications of HMRFs in a Bayesian context. First, in populations with presumed continuous variation in allele frequencies, we argue that **HMRFs are powerful when detecting geographical discontinuities in allele frequencies and regulating the number of clusters.** Then, we address the cline/cluster dilemma with HMRFs using a subsample of the CEPH human polymorphism data set, and check that the main clusters obtained with STRUCTURE are robust to the inclusion of continuous variation in allele frequencies through space. In addition, we show that an accuracy similar to the one obtained with non-spatial methods can be achieved while using a smaller number of genetic markers.

#### THE POTTS-DIRICHLET MODEL

In this study, **we borrow from the toolbox of statistical physics the concept of *Markov Random Field* (MRF), also called the *Potts model* (POTTS 1952), (WU 1982),**

(PRESTON 1974). The model has been coined to handle stochastic networks where particles in identical states evolve in patches larger than expected under an absence of interactions. GUTTORP (1995) gives a recent review of the Potts model at a fairly introductory level. Since the 1970s, MRFs have a long tradition in image analysis, where the color of pixels is correlated to the color of neighboring pixels (see e.g. (Geman and Geman 1984), (Besag 1986), (Ripley 1988)). In this context MRFs account for the property that adjacent pixels are more likely to be of the same color than non-adjacent pixels. HMRFs are relatively recent, but they have been successfully applied in several domains (ZHANG *et al.* 2001), (GREEN and RICHARDSON 2002), (DESTREMPES *et al.* 2005). **Ideas from Bayesian spatial genetics were also used in association studies** (THOMAS *et al.* 2003). In analogy with image analysis, MRF can model the fact that individuals from spatially continuous populations are more likely to share cluster membership with their close neighbors than with distant representatives. They seem therefore relevant to study populations for which continuous variation of allele frequencies may be used as a postulate.

Devising MRF models raises a difficulty when the study design is irregular. While the definition of neighborhood is **immediate in the case of lattice observations**, it is less obvious in the case of irregular sampling, because many choices are available. In this study, we use the natural neighborhood structure obtained from the so-called *Dirichlet tiling*. Denoting by  $(s_i)$ ,  $i = 1, \dots, n$ , the set of observation sites for  $n$  individuals, each  $s_i$  is surrounded by points which are closer to  $s_i$  than to any other sampling site. This set of points is known as the *Dirichlet cell* (or tile). Two sampling sites are neighbors if their cells share a common edge. **The use of the sampling locations to define cells is natural unless the sampling locations are unrepresentative of the individual spatial distribution. However, the method works in principle for any fixed tiling, as soon as the user can define a neighborhood structure to incorporate in the Potts model.** In the sequel, we refer to the Potts model build on the Dirichlet tiling generated by sampling sites to as the *Potts-Dirichlet* model.

We denote by  $c_i$  the cluster from which the individual  $i$  originates, and we assume the existence of at most  $K_{\max}$  clusters. As we shall see later, the constant  $K_{\max}$  should indeed be considered to be larger than the true (or presumed true) number of clusters,  $K$ . We let  $c = (c_i)$  denote the cluster configuration, i.e., a map that takes all cells and specifies the clusters to which they belong. In addition, **we let  $U(c)$  denote the number of neighboring pairs** with the same labels in  $c$ . Formally, we have

$$U(c) = \sum_{i \sim j} \delta_{c_i, c_j} \quad (1)$$

where  $i \sim j$  indicates that  $i$  and  $j$  are neighbors, and the Kronecker symbol  $\delta_{c_i, c_j}$  takes the value 1 iff  $c_i = c_j$ , otherwise 0. Large values of  $U(c)$  correspond to spatial patterns with large patches of individuals belonging to the same cluster. Small values of  $U(c)$  (maybe equal to 0) correspond to patterns that do not display any sort of spatial organization.

The Potts model is a probability distribution on the set of cluster configurations. Given  $n$  observation sites, the probability of configuration  $c$  is written as

$$\pi(c) \propto \exp(\psi U(c)), \quad c \in \{1, \dots, K_{\max}\}^n, \quad (2)$$

where  $\psi$  is a nonnegative parameter called the *interaction parameter*. The value  $\psi = 0$  corresponds to the uniform distribution on the configuration space. Large values of  $\psi$  make more likely the observation of largely clustered configurations corresponding to large  $U(c)$ . Two simulations of the Potts-Dirichlet model are displayed in Figure 1 for  $K_{\max} = 3$ ,  $\psi = 0.1$ ,  $\psi = 0.9$ , where the **sites were generated from the uniform distribution on a square domain**. For  $K_{\max} = 3 - 6$ , simulations (not reported) showed that the value  $\psi = 1.0$  can be considered a high level of spatial interaction, for which the probability that pairs of neighbors are in the same cluster is close to one. In contrast, **values of  $\psi \leq 0.4$  correspond to weak interactions**. In this case **the probability that pairs of neighbors are in the same cluster is less than 0.3**. Values of  $\psi$  around  $\psi \approx 0.6 - 0.7$  are suitable for observing the coexistence of several clusters, while for larger values the model has a

tendency to form a single cluster. **We also note that the Potts model** does not assume connected clusters, and the number  $K$  of observed clusters may be lower than  $K_{\max}$ .

In order to work with a well-defined probability distribution, the requirement that **probabilities sum to one must be fulfilled**. This is achieved by taking

$$\pi(c) = \frac{e^{\psi U(c)}}{Z(\psi, K_{\max})} \quad (3)$$

where  $Z(\psi, K_{\max})$  is a normalizing constant called the *partition function*

$$Z(\psi, K_{\max}) = \sum_c e^{\psi U(c)}. \quad (4)$$

Computing the partition function of the Potts model and performing perfect sampling for an arbitrary graph is feasible if there are only a few sampling sites, otherwise it is an highly difficult problem. Historically the METROPOLIS algorithm got round the issue by using an ingenious cancellation of this constant term (METROPOLIS *et al.* 1953).

Besides providing a flexible way to model **a spatially organized population**, the Potts model satisfies a spatial Markov property which states that the conditional probability for membership in  $c_i$  given the configuration at all other sites  $c_{-i} = (c_j)_{j \neq i}$  is equal to the conditional probability **given the state of its neighbors**  $c_{\partial i} = (c_j)_{j \sim i}$ . Mathematically, this property can be written as

$$\pi(c_i | c_{-i}) = \pi(c_i | c_{\partial i}). \quad (5)$$

More specifically, we have

$$\pi(c_i | c_{\partial i}) \propto \exp\left(\psi \sum_{j \sim i} \delta_{c_i, c_j}\right) \quad (6)$$

The above conditional probabilities involve local computations only, and the sum  $\sum_{j \sim i} \delta_{c_i, c_j}$  can be interpreted as the sum of influences of all neighbors of  $i$ . The Markov property is a basis for implementing fast simulation and inference algorithms.

## HIERARCHICAL BAYES

**Model** In this section, we present the hierarchical Bayes model based on an HMRF. With  $\psi$  equal to 0, the HMRF model assumes a non-informative spatial prior, and then

encompasses the classical Bayesian clustering models (PRITCHARD *et al.* 2000), (DAWSON and BELKHIR 2001) and (CORANDER *et al.* 2003) which can be seen as particular cases. In addition to a spatial prior, a second modification of the standard Bayesian clustering model includes departures from the HW equilibrium caused by inbreeding. Inbreeding coefficients represent the probability that two homologous genes are identical by descent. To implement the modification, inbreeding coefficients can be considered as additional statistical parameters  $\phi_k$ . We use notations similar to those used in the previous works:  $L$  is the number of loci,  $J_\ell$  is the number of alleles at locus  $\ell$ , and  $z$  is the collection of all genotypes (the data). Given that the individual  $i$  originates from the cluster  $c_i = k$  and given the allele frequencies  $f_{k..}$  in this cluster, **the conditional probability of observing the genotype  $z_i^\ell = (a_i^\ell, b_i^\ell)$  at locus  $\ell$  is equal to**

$$\pi(z_i^\ell | k, f_{k..}, \phi_k) = \mathcal{L}_k(f_{k\ell a_i^\ell}, f_{k\ell b_i^\ell}) \quad (7)$$

where  $\mathcal{L}_k(f, f) = f^2 + \phi_k f$  and  $\mathcal{L}_k(f, g) = 2fg(1 - \phi_k)$  for  $f \neq g$  (see e.g., (Hartl and Clark 1997)). Diploidy is also assumed.

We write the set of all parameters as  $\theta = (\psi, c, f, \phi)$  with  $\psi$  the interaction parameter,  $c$  the cluster configuration,  $f = (f_{k\ell j})$ ,  $k = 1, \dots, K_{\max}$ ,  $\ell = 1, \dots, L$ ,  $j = 1, \dots, J_\ell$ , the allele frequencies, and  $\phi = (\phi_1, \dots, \phi_{K_{\max}})$  the inbreeding coefficients in each subpopulation. As in STRUCTURE, the priors on allele frequencies are Dirichlet distributions  $\mathcal{D}(\alpha, \dots, \alpha)$ . The prior distributions on the  $\phi_k$ 's are Beta  $\mathcal{B}(\lambda, \mu)$  distributions. Although we have included  $\psi$  in the parameter list in order to implement a full-Bayes approach, the estimation of  $\psi$  nevertheless generates specific **computational difficulties due to the exponential number of terms involved in the partition function  $Z$**  (Gelman and Meng 1998). **For this reason, we often consider fixed values for this parameter with typical values within the range (0.1, 1.0).** This can be formulated with prior distributions on the rescaled interaction parameters  $\psi/\psi_{\max}$  being either Beta distributions or constant (Dirac) distributions. The prior distribution on  $\theta$  reflects the hierarchy of the model, and

takes the following form

$$\begin{aligned}
\pi(\theta) = \pi(\psi, \phi, c, f) &= \pi(\phi)\pi(\psi|\phi)\pi(c|\phi, \psi)\pi(f|c, \psi, \phi) \\
&= \pi(\phi)\pi(\psi)\pi(c|\psi)\pi(f|c)
\end{aligned}
\tag{8}$$

**Assuming linkage equilibrium between loci, the likelihood is defined as follows**

$$\pi(z|\theta) = \prod_{i=1}^n \prod_{\ell=1}^L \pi(z_i^\ell | c_i, f_{c_i, \ell}, \phi_{c_i}) = \prod_{i=1}^n \prod_{\ell=1}^L \mathcal{L}_{c_i}(f_{c_i a_i^\ell}, f_{c_i b_i^\ell})
\tag{9}$$

where  $\mathcal{L}_k$  is defined in equation (7).

**Inference using Markov Chain Monte Carlo** Inferences on  $\theta$  are carried out by simulating the posterior distribution  $\pi(\theta|z)$  through a Markov Chain Monte Carlo (MCMC) sampling algorithm. In this algorithm, we combine sequential updates of blocks of parameters, each block of parameters being either fully or partially updated. The description of the MCMC steps is detailed in the appendix section **DETAILS OF MARKOV CHAIN MONTE CARLO COMPUTATIONS**. A complete update of all blocks of parameters is referred to as a *cycle*.

**Estimating the number of clusters** As other Bayesian clustering methods do, the HMRF model refers implicitly to an unknown number of clusters  $K$ . In practice this number  $K$  has to be estimated. **Previous approaches typically fall into two categories:** 1) Maximizing the likelihood modified with a penalty that decreases with model complexity (e.g., BIC, DIC information criteria), 2) Choosing a prior distribution on  $K$  and maximizing the posterior distribution using trans-dimensional MCMC computations (which are usually time-consuming to develop and to run). Although these methods have proved effective in many cases, we use an alternative approach known as *regularization* in statistics. For this terminology, we refer to the book by Ripley (1996), Chapter 4.3,

p.136. The rationale for regularization and the relationship with the algorithm implemented in STRUCTURE can be explained as follows. Let  $L_s(z, f, c)$  denote the log-likelihood for the complete data (observed + unobserved) in the original approach of (Pritchard *et al.* 2000). When we refer to this approach, we mean the no-admixture model with uncorrelated allele frequencies. Assuming absence of inbreeding, the log-likelihood of the HMRF model can be expressed as

$$L(z, f, c) = L_s(z, f, c) + \psi U(c) + C_\psi \quad (10)$$

where the term  $U(c)$  represents the contribution from the spatial prior, and  $C_\psi$  is a constant that depends on  $\psi$ . For the value  $\psi = 0$ , the model implemented in STRUCTURE is then recovered. In fact equation (10) corresponds to the Lagrangian formulation of an optimization problem where  $\psi$  can be viewed as the Lagrange multiplier. With the data in hands, the optimization problem seeks the most likely cluster assignments under the constraint that a maximal number of neighboring pairs should fall in the same clusters. For small  $\psi$ 's ( $\psi < 0.3$ ), the constraint is weak, and the results are expected to be close to those produced by STRUCTURE. For larger values the results are generally expected to differ.

In the regularization approach,  $K_{\max}$  is a value presumed larger than the true number of clusters  $K$ . When the algorithm is started, the cluster configuration  $c$  spans arbitrary values between 1 and  $K_{\max}$ . As the chain runs, the program attempts to reduce the number of non-empty clusters which is finally considered as an estimate of  $K$ . In practice, one starts with runs with small values of  $K_{\max}$ , and increases  $K_{\max}$  unless the estimated  $K$  is strictly lower than  $K_{\max}$ . Then, one checks that the result remains identical when higher values of  $K_{\max}$  are used. Practice also shows that repeating shorter runs and performing estimation from the runs with the highest likelihood is a reasonable strategy.

The connections between model selection and regularization have been emphasized several times in the statistical literature. Indeed, regularization is a key argument in statistical procedures such as *ridge regression* (HOERL and KENNARD 1970), *lasso estimators* (TIBSHIRANI 1996), feedforward neural networks *weight decay* (BISHOP 1995). **Such methods were successful in various areas such as text mining or gene selection from large transcriptomic data sets. Nevertheless, we are not aware of any published statistical methods that have used regularization in a hidden context as is done here.** The relevance of the regularization principle is carefully assessed in the section SIMULATION STUDY.

### SIMULATION STUDY

In this section we report results from an intensive simulation study. **The goals of our experiments are (i) to give evidence that the MCMC implementation is correct, (ii) to assess the value of predictions obtained from the HMRF model with particular attention paid to estimation of the unknown number of populations  $K$  and the cluster configuration  $c$ , and (iii) to compare the HMRF model with a non-spatial approach and to a lesser extent with the Bayesian clustering algorithm GENELAND developed by GUILLOT *et al.* (2005).**

**Estimating the number of clusters** In order to check the validity of the HMRF model, we performed inferences for 300 simulated data sets obtained as replicates from the model prior distributions. Individual geographical coordinates were generated from a two-dimensional uniform distribution on a square domain. Genotypes with 10 loci and 10 alleles per locus were simulated using multinomial sampling from the Dirichlet  $\mathcal{D}(1, \dots, 1)$  distribution. The interaction parameter  $\psi$  was simulated according to a uniform distribution on  $\{0, 0.1, \dots, 1\}$ . The inbreeding coefficients were simulated according to a Beta  $\mathcal{B}(4, 40)$  distribution. The hidden cluster configurations  $c$  were generated from the Potts-Dirichlet model with  $K = K_{\max} = 1, 2, 3$  classes. Replicates with  $K = 1, 2, 3$  classes were

simulated for  $n = 50, 100, 150$  individuals respectively.

In the full-Bayes inference method (inference of  $\psi$ ), the computation of the partition function  $Z(\psi, K_{\max})$  involved preliminary off-line runs. They were carried out with 20,000 cycles of a Gibbs sampler with a thinning period of 10 cycles. The maximal number of clusters was fixed to  $K_{\max} = 5$ , and 30,000 cycles, a burn-in period of 20,000, a thinning period of 10 cycles were used. The parameter  $\psi$  was kept equal to 0 during the 5,000 first cycles (see the appendix section UPDATING THE INTERACTION PARAMETER  $\psi$  for more details).

The estimation errors are summarized in Figure 2. This Figure displays histograms for the three types of data sets  $K = 1, 2, 3$ . For data sets made of a single population, the HMRF model estimated  $\hat{K} = 1$  in almost all replicates. Data sets made of  $K = 2$  clusters were also identified as being so for more than 80 replicates (out of 100), and in the datasets for which we had  $\hat{K} = 3$  instead of  $\hat{K} = 2$ , the third cluster consisted of less than two individuals. For data sets made of  $K = 3$  populations, perfect estimation dropped to 55%, but a closer look at the results for which we had  $\hat{K} = 4$  instead of  $\hat{K} = 3$  revealed that the third cluster consisted of less than 4 individuals. In these cases, a longer run might empty the spurious cluster (but we did not evaluate how long this might take). In all simulations, each extra cluster consisted of at most 6 individuals. Furthermore,  $K$  was never underestimated. These results are summarized in TABLE 1.

**Estimating cluster membership probabilities** We now turn to the accuracy of inference in terms of correct assignments. We denote by  $(x_{ij})$  the  $n \times n$  matrix whose entries are  $x_{ij} = 1$  if  $c_i = c_j$ , and 0 otherwise. Similarly we denote by  $(\hat{x}_{ij})$  the corresponding matrix obtained from the estimated cluster configuration  $\hat{c}$ . We assessed the accuracy of cluster assignment through the error rate in co-assignment (ERCA) defined as

$$\text{ERCA} = \frac{2}{n(n+1)} \sum_{i,j=1}^n 1 - \delta_{x_{i,j}, \hat{x}_{i,j}}$$

This pair-based measure has the advantage over individual-based indices to be insensitive to the issue of (cluster) label-switching.

To assess the benefit of our approach as compared to models accounting neither for inbreeding nor for spatial structure, we carried out additional experiments from the HMRF model at  $\psi = 0$  and  $\phi = 0$  (HWE assumed). The assumptions of this simpler model (referred to as the NON-SPATIAL MODEL) were similar to those made in the programs STRUCTURE (PRITCHARD *et al.* 2000), PARTITION (DAWSON and BELKHIR 2001) and BAPS (CORANDER *et al.* 2003). The HMRF model with fixed parameters  $\psi = 0$  and  $\phi = 0$  was used instead of these programs in order to avoid potential biases due to specific computer implementations. Typical cluster configurations at low and high  $\psi$ 's are portrayed in Figure 1 for  $K = 3$  (see the Figure again). They correspond to low and high levels of spatial organization ( $\psi = 0.1$  and  $0.9$ ). In this section similar situations were reproduced with  $K = 2$ .

**We simulated 200 datasets from the HMRF model prior distributions with  $K_{\max} = 2$  using simulations from the MCMC program without data (1000 cycles). Running the program for a fixed number of cycles did not warrant the convergence of the MCMC sampler. As the aim of the simulation study was the retrieval of previously stored allele frequencies and cluster memberships, this shortcoming did not affect the performance study. In the sampled data, individuals were occasionally grouped in a single cluster (for values of  $\psi$  greater than 0.8). The clusters had no predefined size, and might consist of very few (less than 10) individuals. The ERCA rates were reported in TABLE 2. In this table, the rates were either averaged over all data sets or over subsets of data that corresponded to different levels of pairwise  $F_{ST}$ , interaction parameter  $\psi$ , and inbreeding coefficients ( $\phi_1$ ,  $\phi_2$ ).**

The results **provided evidence** that the HMRF model increased the number of correct assignments compared to the non-spatial model. A more detailed look at subsets of simu-

lated data revealed that the HMRF model always performed better than the other models whatever the levels of spatial interaction or inbreeding. The highest improvements were obtained at low levels of differentiation ( $F_{ST} \leq 0.08$ ) and high levels of spatial structure ( $\psi > 0.6$ ). The HMRF model achieved the smallest improvements over the other models for high levels of inbreeding, although it still gave very accurate results. In this cases, the inbreeding coefficients were correctly estimated (Results not shown).

The error rates of the non-spatial model were in some cases very high. This was indeed the case for large values of  $\psi$ . These results may be explained as data sets generated from large  $\psi$  sometimes contained a single cluster. Thanks to the regularization procedure, this cluster was successfully detected by the HMRF model (and also by GENELAND) but not by the non-spatial model **which split the unique population** in two arbitrary parts.

These results carried information about the performances of the HMRF model when the initial number of cluster was close to the true number ( $K_{\max} = 2$ ,  $K = 1$  or  $2$ ). We repeated the inference study on the same 200 datasets with  $K_{\max} = 5$ . The global ERCA was around 10 percent, which was still a low misclassification rate.

## REAL DATA ANALYSIS

**Scandinavian brown bears** The Scandinavian brown bear (*Ursus arctos*) is an **example** of a wild population with strong female phylopatry and male mediated gene flows. We analyzed the same data set as two previous studies (WAITS *et al.* 2000), (MANEL *et al.* 2004) from 366 geo-referenced individuals genotyped as 19 microsatellite loci. We first used the full-Bayes HMRF model implemented with the same prior distributions as in the simulation study, and ran the algorithm with  $K_{\max} = 4 - 7$ . After 30,000 cycles, the HMRF model with  $K_{\max} = 4$  converged to the same clusters as described in the previous study. We referred to these clusters as the S (South), M (Middle), NWN (North West North) and NN (North North) areas. With  $K_{\max} = 5 - 7$ , the HMRF model yielded 5 clusters, three of which coincided with the  $K_{\max} = 4$  run while the fourth (S) **was split in**

**two subsets with random shapes.** The spatial interaction parameter  $\psi$  had **posterior mode within the range** (0.6, 0.8) (95% credible interval). However, the random shapes of the two S subclusters were an indicator that the MCMC runs might have not converged, perhaps due to the large amount of computational resource spent into the estimation of  $\psi$ . Therefore we performed 10 additional runs of the algorithm for two values of the interaction parameter  $\psi = 0.7 - 0.8$ . The runs that reached the highest likelihood resulted in the four same clusters as previously observed (see Figure 3). Inferences carried out under a fixed large value of  $\psi$  usually favor cluster configurations made of few large clusters. **The fact that the HRMF model obtained the same clusters** as STRUCTURE gave evidence that these original clusters were robust to the inclusion of a spatial prior. A by-product of the HRMF model is its ability to infer inbreeding coefficients. The inbreeding coefficients posterior estimates were computed as  $\phi_{NN} = 0.022$ ,  $\phi_{NWN} = 0.006$ ,  $\phi_M = 0.013$ ,  $\phi_S = 0.007$ . These small values were consistent with the observation that STRUCTURE worked well for this data set. The HRMF model with fixed parameter setting converged faster than the full-Bayes version. (We used 1,000 cycles for  $K_{\max} = 4$ , and 20,000 cycles for  $K_{\max} = 7$ .) **GENELAND runs at fixed  $K = 4 - 6$  produced the same assignment results as the HRMF model (5,000 cycles).** Using reversible jumps, the posterior distribution of  $K$  exhibited a mode at  $K = 5$  and a 95% credible interval  $K \in (4, 8)$  (50,000 cycles).

**Human data** We used the HGPH-CEPH Human Genome Diversity Cell Line Panel (CANN *et al.* 2002) to further assess the influence and the benefit of including spatial continuity prior hypotheses in the analysis of multilocus genotypes. The HGPH-CEPH diversity panel dataset contains 1056 individuals genotyped **at 377 autosomal microsatellite loci**. It was first studied with the software STRUCTURE by ROSENBERG *et al.* (2002). Without using predefined populations, six main genetic clusters were identified, five of which corresponded to major geographic regions. Here we restricted the study to the Eurasian and East-Asian populations including samples with distinct origins, 8 from Pakistan, 16

from China, 1 from Siberia, Japan and Cambodia (451 individuals). **Two reasons could be given** for limiting the study to Eurasian and East-Asian populations. First, these populations contained **two of the five main clusters** as well as the sixth cluster found by ROSENBERG *et al.* (2002). Second, the 27 populations live on a same mainland, which justified using the Dirichlet tiling without modifying the neighborhood structure (**although our computer program makes it possible**). Coordinates of individuals in each sample were not known explicitly. Instead they were available as sample intervals from CANN *et al.* (2002). For instance, the Kalash from Pakistan have longitudes in the range 35-37 deg.E and latitudes in the range 71-72 deg.N. **Individual coordinates were generated randomly within the specified intervals. We checked that the results presented here were rather independent of the individual coordinates within each sample (not reported).**

To evaluate the inclusion of geographic continuity prior, subsets of data containing 20, 10, and 5 random loci were extracted from the original dataset (20 subsamples for each number of loci). The HMRF model was initialized with  $K_{\max} = 3$  clusters, and then run for 50,000 cycles, with a burn-in period of 500 cycles and a thinning interval of 5 cycles. The interaction parameter  $\psi$  was either estimated from the same prior distributions as in the simulation study (full Bayes) or fixed to  $\psi = 0.6$ . With 20 loci, all outputs contained 2 clusters (Pakistan including Kalash (8 samples), against the other Asian populations) **regardless of the estimation strategy of the interaction parameter  $\psi$** . With 10 loci the HMRF model identified the two main clusters in 18 of the 20 runs. With 5 loci no successful run was observed. The non-spatial version ( $\psi = 0$ ) led to the same outputs when the number of clusters was set to  $K_{\max} = 2$ .

To further highlight the potential of the HMRF model, we focused on the Pakistan dataset and the retrieval of the Kalash cluster. The Kalash sample contains 25 of the 200 individuals from the 8 Pakistan samples. Ranges for sample spatial coordinates are reported in TABLE 3 (CANN *et al.* 2002), and a representation of the resampled individual

locations is displayed in Figure 4. In this study, data sets with 40, 30, 20 randomly chosen loci were extracted from the Pakistan dataset. **The idea here is to use the results from a large number of loci as the “correct” answer, and then see which methods are able to get this correct answer with fewer loci.** Because all the extracted data sets did not contain the same amount of information about genetic structure, we distinguished three distinct levels of potential difficulty (SC, WC and NC) according to the following classification. For each subset, we preliminarily computed a neighbor-joining (NJ) tree using the shared allele distance, see (NEI and KUMAR 2000), which separated the Pakistan samples in two sister clades. Data sets for which one clade contained more than 20 Kalash grouped against the remaining Pakistan representatives were classified as SC (Strong Clustering). Such data sets were expected to be easy for Bayesian clustering algorithms, because a more basic analysis gives a correct answer. **As well, there were data sets for which no obvious clusters could be directly inferred from the NJ tree. These data sets were classified as NC (No Cluster), and they were expected to be difficult for Bayesian clustering algorithms.** We added an intermediate class, named WC (Weak Clustering) for which the Kalash sample generally formed a cluster in the NJ tree, but this was done in association with other samples such as Pathan or Balochi/Brahui. With 40 randomly chosen loci, about 38% of all data sets were in the SC category, 24% were classified as WC, and the remaining 38% were NC. One NJ tree clustered the Balochi/Brahui against the rest of Pakistan. With 30 loci, **these numbers changed to SC + WC = 42% and NC = 58%, and NC increased to 76% in the 20 loci data sets.** These ratios were obtained from 300 distinct data sets.

We performed 10 runs of the HMRF model for 42 subsets (21 subsets with 40 loci and 21 subsets with 30 loci). **The HMRF model was first run for 200 cycles** at  $\psi = 0.4$ , and **these cycles were followed by a further 500-1,000 cycles** at  $\psi = 0.6$ . For the CEPH diversity panel data set, this strategy appeared more efficient than the full-Bayes approach, which was statistically unable to identify the Kalash (we attributed this failure

to the algorithmic complications and the approximations made in estimating  $\psi$ ). The run with the highest likelihood was saved as the final result. The same strategy was also used at  $\psi = 0$  with a larger total number of cycles (up to 2,000). Small burn-in (10 cycles) and thinning (1 cycle) period were implemented. We first used  $K_{\max} = 2$  in both the HMRF and non-spatial versions. In order to compare with published results, we also assumed absence of inbreeding.

For SC datasets with 40 loci, the HMRF model and the non-spatial versions performed similarly and retrieved the Kalash sample. Similar results were reported for STRUCTURE in the literature (BAMSHAD *et al.* 2003), (RAMACHANDRAN *et al.* 2004). The HMRF model failed to identifying the Kalash in a single WC subset whereas the non-spatial version failed twice in this category. The HMRF model identified the Kalash successfully in 75% NC samples whereas the non-spatial version failed in the same ratio (75%). The divergence between the spatial and non-spatial version increased as we reproduced the study with 30 loci. **The HMRF algorithm failed to identify** the Kalash in 37% of the NC cases. The global success rate of the HMRF model was however greater than 85% (including SC, WC and NC cases) whereas this global rate dropped to 47% in the non-spatial algorithm. With 20 loci, both algorithms **failed in a majority of the NC cases**. For all loci, the  $K_{\max} = 3$  results were in strict concordance with the  $K_{\max} = 2$  results for the spatial version although more than 10 runs were sometimes necessary in the NC cases.

## DISCUSSION

Detecting population subdivision is a subject of great interest to population geneticists, and **a large body of approaches have been developed for this**. In this study, we have presented a Bayesian clustering algorithm that incorporates Hidden Markov Random Fields as prior distributions on cluster configurations. Markov Random Fields are mathematical models that account for the “continuity” of discrete random variables on a graph or a network (for a rigorous definition of continuity in this context, refer to (PRESTON 1974)).

**The term Hidden means that the cluster configuration is unobserved, and is instead reconstructed from an MCMC algorithm.** In spatial genetics, **the term *continuous population* usually refers to S. WRIGHT's famous concept of isolation by distance (WRIGHT 1943), which can in turn be understood in terms of the stepping-stone model (KIMURA and WEISS 1964), (ROUSSET 2004).** Because it considers interacting demes on a lattice, the stepping stone model exhibits the same type of spatial Markov property as does the Potts model. **Inserting the stepping-stone model in a Bayesian framework generates conceptual difficulties because its stationary distribution has no known formulation. However the HMRF model may capture its essential properties.**

While STRUCTURE has recently become prominent among clustering algorithms, another recent approach includes spatially explicit priors in a highly structured statistical framework (GUILLOT *et al.* 2005). The approach developed by GUILLOT *et al.* (2005) nevertheless differs from the HMRF model significantly. In GUILLOT *et al.* (2005), population territories are viewed as unions of polygons. A full-Bayes algorithm estimates the number of populations using the reversible-jump MCMC machinery. The simulation study carried out by GUILLOT *et al.* (2005) suggests that their model performs well when genetic discontinuities occur as very simple polygonal lines are crossed (eg, straight lines). A field study and a subsequent analysis by COULON *et al.* (2006) also support these observations. Although simple shaped territories are likely to be quite common, there are also important cases where these assumptions do not hold (for example, limited gene flows in areas with complex geography, mountain ranges, world-wide studies). In the HMRF model, spatial dependencies are prescribed at the individual level directly. The advantage of the HRMF approach **is that it can assign** individuals when the hidden cluster configurations are too complex to be summarized by simple polygonal regions.

The HMRF model involves an interaction parameter  $\psi$  **which corresponds to the intensity with which two neighbors belong to the same cluster.** Estimates of  $\psi$  may

be interpreted as local measures of spatial *clusteredness* for the studied sample. The higher  $\psi$  the more likely the population may consist of a unique cluster with a high level of genetic continuity (e.g. slow clinal variation). Estimates of  $\psi$  found in the studied (real) data sets were generally greater than 0.5 *which indicated the presence of continuous organization*. Nevertheless, interpretations of such parameters would lead us far beyond the scope of this study, because the connection to statistical physics is not so direct in this context. In addition, we have also claimed that  $\psi$  may play a more important role as **a Lagrange multiplier in a constrained optimization problem** where the non-spatial likelihood is optimized while the algorithm attempts to assign a maximal number of neighbor pairs to a same cluster. We have indeed argued that the HMRF algorithm then contains an implicit way for deciding the number of clusters, a major issue in such statistical mixtures algorithms. From this perspective, maintaining fixed values of the interaction parameter  $\psi$  may be preferable to estimating this parameter, and has the additional advantage of avoiding difficult computational issues (GELMAN and MENG 1998). The simulation study evaluated the use of the full-Bayes HRMF algorithm (estimation of  $\psi$ ) only. **This was done because simulations and inferences** with fixed  $\psi$  would have biased the results toward very low ERCAs and very optimistic conclusions. During the analysis of real data, versions of the HMRF model at fixed values of  $\psi$  (around 0.5 – 0.7) nevertheless achieved better performances and were considerably faster than the full-Bayes version.

**The use of HMRF model has been illustrated on two previously published data sets.** The Scandinavian brown bear is an example of a population with a strong female phylopatry. Scandinavian bears were almost exterminated at the beginning of the 20th century. After efforts to protect the species in Sweden, the bear population has recovered from four female concentration areas. Until recently these areas were believed to represent the surviving relict subpopulations after the 1930's bottleneck (see e.g. (WAITS *et al.* 2000)). Using two independent methods (neighbor-joining trees and the Bayesian clustering algorithm STRUCTURE), MANEL *et al.* (2004) found four genetic clusters which

matched with geographical clusters, but two of them were distinct from the original female concentration areas. Using a coalescent approach, BLUM *et al.* (2004) computed the female dispersal rate and found an estimate of 9km per generation. Because of the low dispersal rate in this population, **local genetic similarities can be considered as a reasonable assumption** to be included in a Bayesian model for brown bear genetic diversity. The HMRF model has been used for detecting geographical discontinuities in allele frequencies. The results confirmed previously published results, and provided reasonable estimates for the number of clusters.

Using the Human CEPH diversity panel data set, we checked whether the clusters obtained without spatial priors were robust to the hypothesis of continuous geographical variation in allele frequencies. The results presented here reconciled the two apparently divergent perspectives of SERRE and PÄÄBO (2004) and (ROSENBERG *et al.* 2002), (ROSENBERG *et al.* 2005) which brought into conflict clines and clusters regarding variation of human diversity. Restricting to Eurasian and Asian populations and working with a prior on continuous variation ( $\psi \approx 0.6$ ), **we recovered the 3 main clusters found by the algorithm STRUCTURE**. Some important facts must be mentioned at this stage: 1) The two main clusters (Pakistan/non-Pakistan) were identified with less than twenty randomly chosen loci. The Kalash cluster was identified using less than fifty loci. 2) More importantly, the algorithm was unable to confirm the presence of other clusters in the Pakistan and East-Asia areas, perhaps due to the simultaneous effects of reducing the number of loci (<120 loci) and imposing the continuity prior. **The combination of these effects may have led the neglect of some very small discontinuities** which were previously detected when STRUCTURE was used with large values of  $K$  and a larger number of loci. We performed ten additional runs of the HMRF model using the full set of loci. Regarding the Pakistan data, we were not able to retrieve other clusters either. Regarding the East-Asia data set, we identified one additional cluster in the north-eastern that matched with the Yakut-Japanese samples. This cluster was also apparent in the

NJ tree. 3) The weight given to the prior distribution was a moderate value that also corresponded to the posterior mean estimated from the full-Bayes algorithm when it converged ( $\psi \approx 0.6$ , 95% credible interval (0.5, 0.9)). 4) Stronger level of prior interaction (e.g.,  $\psi \approx 1$ ) led to a unique cluster and gave strong support to SERRE and PÄÄBO's hypothesis of clinal variation within a unique cluster. 5) Weaker levels of prior interaction (e.g.,  $\psi \approx 0.2$ ) led to the same results as STRUCTURE, and supported ROSENBERG's small discontinuities hypothesis. Here we supported the intermediate view of clinal variation of allele frequencies with a number of discontinuities lower than estimated by (ROSENBERG *et al.* 2002). See Figure 5 for a picture of the reconciliation.

**In conclusion we have shown that the HMRF model can achieve accuracy similar to the one obtained with non-spatial methods while using a smaller number of genetic markers. Consequently the use of HMRF algorithms could be advocated in cases where the number of polymorphic loci available to the study is limited, and a prior knowledge about continuous spatial structure could be incorporated with certainty.**

*The source codes used in this study are available as an R package that also provides additional visual displays and the data sets used during this study. The R package was mainly developed by SA, and a version supporting Linux OS and R 3.1.1. can be downloaded from SA's or OF's website. A multiple-platform software will be made available within a few months.*

#### ACKNOWLEDGMENTS

We are grateful to Noah Rosenberg for his suggestions on an early version of this manuscript. We wish to thank Stephanie Manel, Oscar Gaggiotti, Chibiao Chen for fruitful discussions, and Mathieu Emily for his help with simulations of the Potts model on a Dirichlet tiling. We are also grateful to two anonymous reviewers for their constructive comments. OF was supported by grants from the AlpB IMAG project and the French

ministry of research ACI ImpBIO project.

#### LITERATURE CITED

- BAMSHAD, M., S. WOODING, W. WATKINS, C. OSTLER, and M. E. A. BATZER, 2003 Human population genetic structure and inference of group membership. *American Journal of Human Genetics* **72**: 578–589.
- BESAG, J., 1986 On the statistical analysis of dirty pictures. *Journal of the Royal Statistical Society, series B* *48*(3): 259–302.
- BISHOP, C., 1995 *Neural Networks for Pattern Recognition*. Oxford: Clarendon Press.
- BLUM, M., C. DAMERVAL, S. MANEL, and O. FRANÇOIS, 2004 Brownian models and coalescent structures. *Theoretical Population Biology* **65**: 249–261.
- CANN, H., C. TOMA, L. CAZES, M. LEGRAND, V. MOREL, L. PIOUFFRE, J. BODMER, W. BODMER, B. BONNE-TAMIR, A. CAMBON-THOMSEN, Z. CHEN, J. CHU, C. CARCASSI, L. CONTU, R. DU, L. EXCOFFIER, G. FERRARA, J. FRIEDLAENDER, H. GROOT, D. GURWITZ, T. JENKINS, R. HERRERA, X. HUANG, J. KIDD, K. KIDD, A. LANGANEY, A. LIN, S. MEHDI, P. PARHAM, A. PIAZZA, M. PISTILLO, Y. QIAN, Q. SHU, J. XU, S. ZHU, J. WEBER, H. GREELY, M. FELDMAN, G. THOMAS, J. DAUSSET, and L. CAVALLI-SFORZA, 2002 A human genome diversity cell line panel. *Science* **296**: 261–262.
- CORANDER, J., P. WALDMANN, and M. SILLANPÄÄ, 2003 Bayesian analysis of genetic differentiation between populations. *Genetics* **163**: 367–374.
- COULON, A., G. GUILLOT, J. COSSON, J. ANGIBAULT, S. AULAGNIER, B. CARGNELUTTI, M. GALAN, and A. HEWISON, 2006 Genetics structure is influenced by landscape features. Empirical evidence from a roe deer population. To appear in *Molecular Ecology*.

- DAWSON, K. and K. BELKHIR, 2001 A Bayesian approach to the identification of panmictic populations and the assignment of individuals. *Genetical Research* **78**: 59–77.
- DESTREMPES, F., M. MIGNOTTE, and J.-F. ANGERS, 2005 A stochastic method for Bayesian estimation of hidden Markov random field models with application to a color model. *IEEE Transactions on Image Processing* **14**: 1097–1108.
- GELMAN, A. and X. MENG, 1998 Simulating normalizing constants: from importance sampling to bridge sampling to path sampling. *Statistical Science* **13**: 163–185.
- GEMAN, S. and D. GEMAN, 1984 Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **6**: 721–741.
- GREEN, P. and S. RICHARDSON, 2002 Hidden Markov models and disease mapping. *Journal of the American Statistical Association* *97*(460): 1055–1070.
- GUILLOT, G., A. ESTOUP, F. MORTIER, and J. COSSON, 2005 A spatial statistical model for landscape genetics. *Genetics* *170*(3): 1261–1280.
- GUILLOT, G., F. MORTIER, and A. ESTOUP, 2005 Geneland: A computer package for landscape genetics. *Molecular Ecology Notes* *5*(3): 708–711.
- GUTTORP, P., 1995 *Stochastic Modelling of Scientific Data*. Chapman & Hall.
- HARTL, D. and G. CLARK, 1997 *Principles of Population Genetics*. Sunderland MA: Sinauer Associates Inc.
- HOERL, A. and R. KENNARD, 1970 Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* **12**: 55–67.
- HURN, M., O. HUSBY, and H. RUE, 2003 *Spatial Statistics and Computational Methods*, Chapter A tutorial in image analysis, pp. 87–141. *Lecture Notes in Statistics*. Springer.
- KIMURA, N. and G. WEISS, 1964 The stepping stone model of population structure and the decrease of genetic correlation with distance. *Genetics* **49**: 561–575.

- MALÉCOT, G., 1948 *Les Mathématiques de l'Hérédité*. Paris: Masson.
- MANEL, S., E. BELLEMAIN, J. SWENSON, and O. FRANÇOIS, 2004 Assumed and inferred spatial structure of populations: the Scandinavian brown bears revisited. *Molecular Ecology* **13**: 1327–1331.
- METROPOLIS, N., A. ROSENBLUTH, M. ROSENBLUTH, A. TELLER, and E. TELLER, 1953 Equation of state calculations by fast computing machines. *Journal of Chemical Physics* **21**: 1087–1092.
- NEI, M. and S. KUMAR, 2000 *Molecular Evolution and Phylogenetics*. Oxford University Press.
- POTTS, R., 1952 Some generalized order-disorder transformations. *Proceedings of the Cambridge Philosophical Society* **48**: 106–118.
- PRESTON, C., 1974 *Gibbs States on Countable State Space*. Cambridge: Cambridge University Press.
- PRITCHARD, J., M. STEPHENS, and P. DONNELLY, 2000 Inference of population structure using multilocus genotype data. *Genetics* **155**: 945–959.
- RAMACHANDRAN, S., N. ROSENBERG, L. ZHIVOTOVSKY, and M. FELDMAN, 2004 Robustness of the inference of human population structure: A comparison of X-chromosomal and autosomal microsatellites. *Human Genomics* **1**: 87–97.
- RIPLEY, B., 1988 *Statistical Inference for Spatial Processes*. Cambridge: Cambridge University Press.
- RIPLEY, B., 1996 *Pattern Recognition and Neural Networks*. Oxford: Oxford University Press.
- ROSENBERG, N., J. PRITCHARD, J. WEBER, H. CANN, K. KIDD, L. ZHIVOTOVSKY, and M. FELDMAN, 2002 Genetic structure of human populations. *Science* **298**: 2981–2985.
- ROSENBERG, N., S. SAURABH, S. RAMACHANDRAN, C. ZHAO, J. PRITCHARD, and

- M. FELDMAN, 2005 Clines, clusters, and the effect of study design on the influence of human population structure. *PLoS Genetics* **1**(6): 660–671.
- ROUSSET, F., 2004 *Genetic structure and selection in subdivided populations*. Princeton University Press.
- SERRE, D. and S. PÄÄBO, 2004 Evidence for gradients of human genetic diversity within and among continents. *Genome Research* **14**: 16791685.
- THOMAS, D., D. STRAM, D. CONTI, J. MOLITOR, and P. MARJORAM, 2003 Bayesian spatial modeling of haplotype associations. *Human Heredity* **56**: 32–40.
- TIBSHIRANI, R., 1996 Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society (Series B)* **58**: 267–288.
- WAITS, L., P. TABERLET, J. SWENSON, F. SANDEGREN, and R. FRANZEN, 2000 Nuclear DNA microsatellite analysis of genetic diversity and gene flow in the Scandinavian brown bear *Ursus arctos*. *Molecular Ecology* **9**: 610–621.
- WASSER, S., A. SHEDLOCK, K. COMSTOCK, E. OSTRANDER, B. MUTAYOBA, and M. STEPHENS, 2004 Assigning African elephants DNA to geographic region of origin: applications to the ivory trade. *Proceedings of the National Academy of Sciences* **101**(41): 14847–14852.
- WRIGHT, S., 1943 Isolation by distance. *Genetics* **28**: 114–138.
- WU, F., 1982 The Potts model. *Reviews of Modern Physics* **54**: 235–268.
- ZHANG, Y., M. BRADY, and S. SMITH, 2001 Segmentation of brain MR Images through a hidden Markov random field model and the Expectation-Maximization algorithm. *IEEE Transaction on Medical Imaging* **20**: 45–57.

#### DETAILS OF MARKOV CHAIN MONTE CARLO COMPUTATIONS

We iterated updates of blocks of parameters where the basic update was as follows.

**Updating allele frequencies  $f_{k\ell j}$**  We used a componentwise Metropolis-Hastings Markov chain simulation algorithm. For the cluster labelled  $k$  and locus labelled  $\ell$ , an update of  $(f_{k\ell 1}, \dots, f_{k\ell J_\ell})$  selected two alleles at random with indices  $j$  and  $j'$ , and proposed to change their frequencies  $f_{k\ell j}$  and  $f_{k\ell j'}$  as follows. Denoting  $a = 1 - \sum_{\substack{m \neq j \\ m \neq j'}} f_{k\ell m}$ , new frequencies  $f_{k\ell j}^*$  and  $f_{k\ell j'}^*$  are proposed as  $f_{k\ell j}^* = aB_f$  and  $f_{k\ell j'}^* = a - f_{k\ell j}^*$  where  $B_f$  is sampled from a Beta  $\mathcal{B}(\alpha, \alpha)$  distribution (often  $\alpha = 1$ ). This move was accepted with probability

$$1 \wedge \frac{\pi(z|\theta^*) f_{k\ell j}(1 - f_{k\ell j})}{\pi(z|\theta) f_{k\ell j}^*(1 - f_{k\ell j}^*)} \quad (\text{A1})$$

The update was based on the conditional distribution of the Dirichlet distribution (Gibbs sampler). The complete update of allele frequencies replicated this basic step for each loci and in all clusters. Typical values of  $\alpha$  were  $\alpha = 1$  or  $2$ .

**Updating inbreeding coefficients  $\phi_k$**  We implemented a componentwise independent Metropolis-Hastings sampler. For each population we iterated the following basic update. A new inbreeding coefficient  $\phi_k^*$  was sampled from a  $\mathcal{U}[0, 1]$  distribution. We assumed a Beta  $\mathcal{B}(4, 40)$  prior distribution on each  $\phi_k$ , hence  $\phi_k^*$  was accepted with probability

$$1 \wedge \frac{\pi(z|\theta^*) \phi_k^{*3} (1 - \phi_k^*)^{39}}{\pi(z|\theta) \phi_k^3 (1 - \phi_k)^{39}} \quad (\text{A2})$$

as we assumed a uniform prior on  $\phi_k$  and made a symmetric proposal.

**Updating the cluster configuration  $c$**  We used sequential updates for all  $i \in \{1, \dots, n\}$ , where all sites were visited in order. At the  $i$ th step, a new value  $c_i^*$  was drawn from a uniform distribution over all possible cluster labels  $\{1, \dots, K_{\max}\}$ . This new state was accepted with probability

$$1 \wedge \frac{\pi(z|\theta^*) \pi(c^*)}{\pi(z|\theta) \pi(c)} \quad (\text{A3})$$

and then it replaced the current cluster label  $c_i$ . The ratio  $\pi(c^*)/\pi(c)$  can be calculated from a local variation of the function  $U(c)$  very easily

$$\frac{\pi(c^*)}{\pi(c)} = e^{\psi \Delta U_i(c)},$$

where

$$\Delta U_i(c) = \sum_{j \sim i} \delta_{c_j, c_i^*} - \delta_{c_j, c_i}$$

Although this has not received much space in this article, we also conducted numerical checks on the correctness of the MCMC sampler. In particular we checked that the results were consistent with those obtained with STRUCTURE at  $\psi = 0$ , and we checked that prior distributions were well recovered when the algorithm was implemented without data.

**Updating the interaction parameter  $\psi$  (Full-Bayes only).** Metropolis-Hastings updates of  $\psi$  required evaluating ratios of distributions of the form  $\pi(c|\psi^*)/\pi(c|\psi)$  for  $\psi^*$  the new value. From equation (3), this computation involved the ratio  $Z_\psi/Z_{\psi^*}$  which was computationally intractable. To avoid this difficulty, we implemented a statistical physics approach known as thermodynamic integration (Gelman and Meng 1998) previously used by Green and Richardson (2002) in the context of spatial epidemiology studies and also described in details in (Hurn *et al.* 2003). The method consisted of approximating the continuous interval  $(0, \psi_{\max})$  by a discrete set of values  $\{\delta, 2\delta, \dots, \psi_{\max}\}$ , and evaluating  $Z(\psi, K_{\max})$  for each  $\psi$  using importance sampling. Here, we used  $\delta = 0.1$  and the maximal value of the interaction parameter was  $\psi_{\max} = 1$ . The importance sampling method used MCMC computations based on the simulation of the Potts model with 50,000 cycles (thinning period of 100 cycles).

The values  $Z(\psi, K_{\max})$  were stored in a look-up table, and were used in all further computations with the same graph topology. Updates of  $\psi$  were then carried out by a standard Metropolis-Hastings Markov chain.

	$\widehat{K} = 1$	$\widehat{K} = 2$	$\widehat{K} = 3$	$\widehat{K} = 4$	$\widehat{K} = 5$
True $K = 1$	0	0.02	0	0	0
True $K = 2$	–	0	0.0136	–	0.03
True $K = 3$	–	–	0	0.0096	0.0267

Table 1: *Proportions of individuals assigned to extra clusters given the number of estimated clusters  $\widehat{K}$  and their true number  $K$ . The symbol – indicates cases that never occurred during the simulation study.*

Genet. structure	Spatial structure	Inbreeding	NON-SPATIAL	HMRP	GENELAND
$F_{ST}$	$\psi$	$(\phi_1, \phi_2)$	MODEL	MODEL	
all	all	all	16.1	0.7	3.2
$F_{ST} \leq 0.08$	all	all	26.3	1.6	6.6
$0.08 < F_{ST} \leq 0.09$	all	all	7.6	0.6	1.4
$0.09 < F_{ST} \leq 0.1$	all	all	8	0.6	1.4
$F_{ST} > 0.1$	all	all	8.3	0.2	1.1
all	$\psi \leq 0.2$	all	1.1	1	1.1
all	$0.2 < \psi \leq 0.4$	all	1	0.8	1.6
all	$0.4 < \psi \leq 0.6$	all	2.7	0.7	0.9
all	$0.6 < \psi \leq 0.8$	all	28.2	0.4	4.7
all	$\psi > 0.8$	all	42.4	0.5	6.9
all	all	$(<0.06, <0.06)$	17.2	0.3	0.7
all	all	$(<0.06, >0.1)$ or $(>0.1, <0.06)$	10	0.5	1.9
all	all	$(>0.1, >0.1)$	12.3	1	1.5
$F_{ST} \leq 0.08$	$\psi \leq 0.4$	all	2.7	2.1	2.8
$F_{ST} \leq 0.08$	$0.6 < \psi \leq 1$	all	41.8	0.9	9.4
$F_{ST} > 0.1$	$\psi \leq 0.4$	all	0.2	0.1	0.4
$F_{ST} > 0.1$	$0.6 < \psi \leq 1$	all	23.7	0.3	2.4

Table 2: Error rate in co-assignments (ERCA) for 200 simulated datasets ( $n = 100, L = 10, J_\ell = 10$ ) with  $K_{\max} = 2$ . The three models were initialized at  $K_{\max} = 2$ .

Sample name	latitude	longitude	sample size
Brahui	30 – 31N	66-67E	25
Balochi	30 – 31N	66-67E	25
Hazara	33 – 34N	70E	25
Makrani	26N	62 – 66E	25
Shindi	24 – 27N	68 – 70E	25
Pathan	32 – 35N	69 – 72E	25
Kalash	35 – 37N	71 – 72E	25
Burusho	36 – 37N	73 – 75E	25

Table 3: *Latitudes and longitudes for the 8 Pakistan samples (from Cann et al. 2002).*

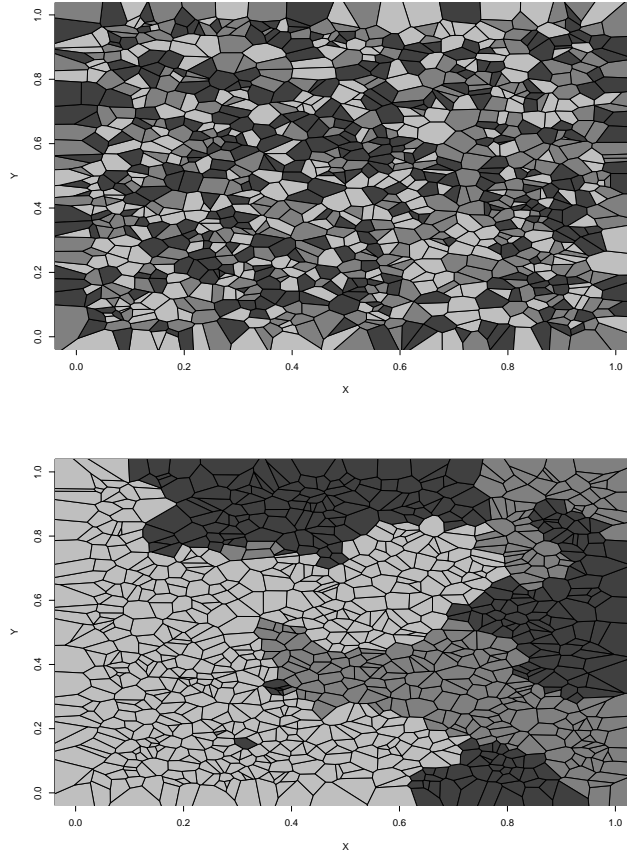


Figure 1: *Two cluster configurations from the 3-states Potts-Dirichlet model. For  $\psi = 0.1$ , no spatial structure can be observed (the situation is close to the non-informative prior used by STRUCTURE). For  $\psi = 0.9$ , a number of non necessarily connected random clusters can be observed.*

**(Low resolution image for review only)**

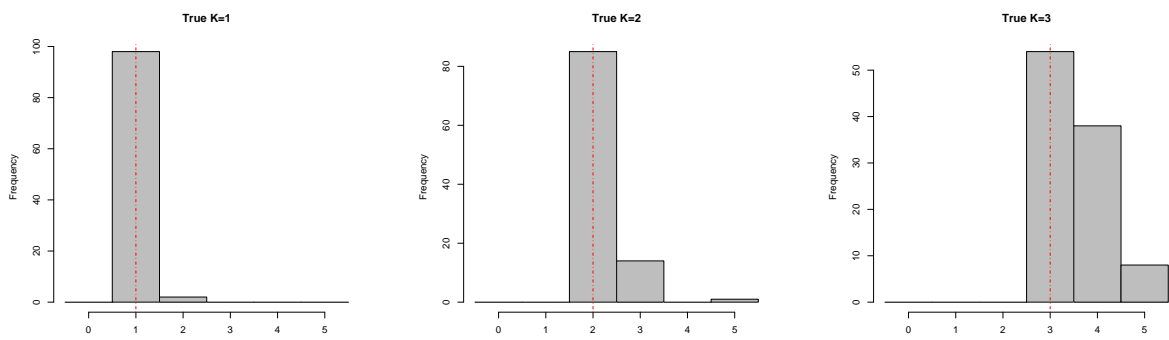


Figure 2: *Distributions of the number of clusters estimated by the HRMF model. Data sets were simulated from the prior distributions of the HRMF model. The vertical lines indicate the true number of populations.*

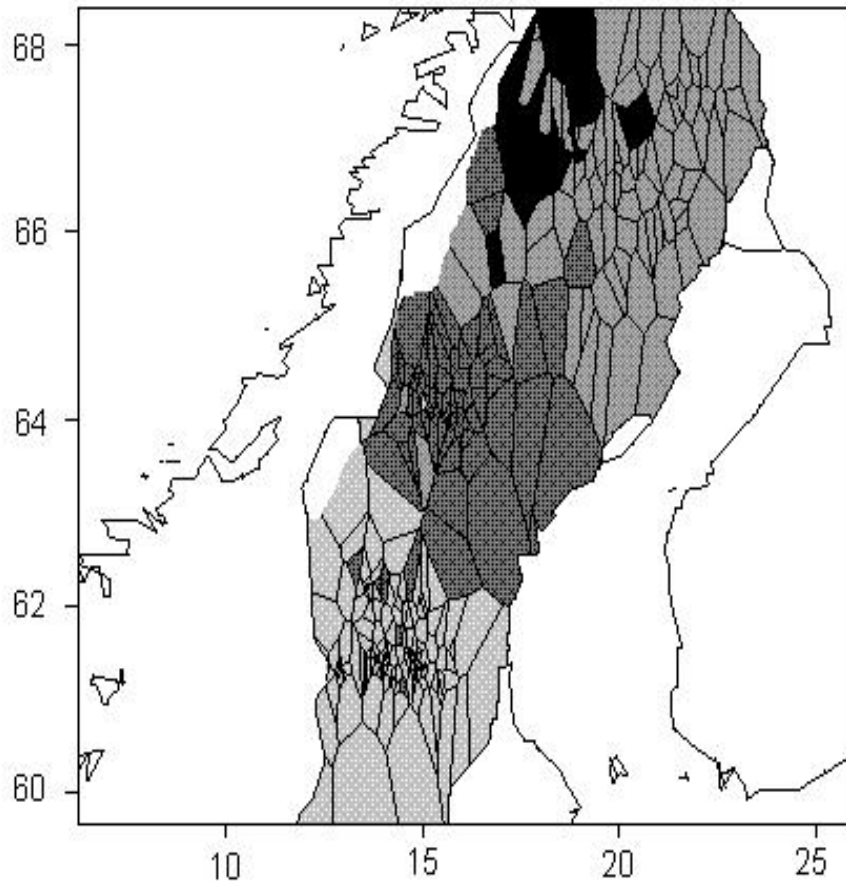


Figure 3: *Estimated cluster configuration for the Scandinavian brown bear data set in North Sweden using the HMRF model (4 clusters).*

**(Low resolution image for review only)**

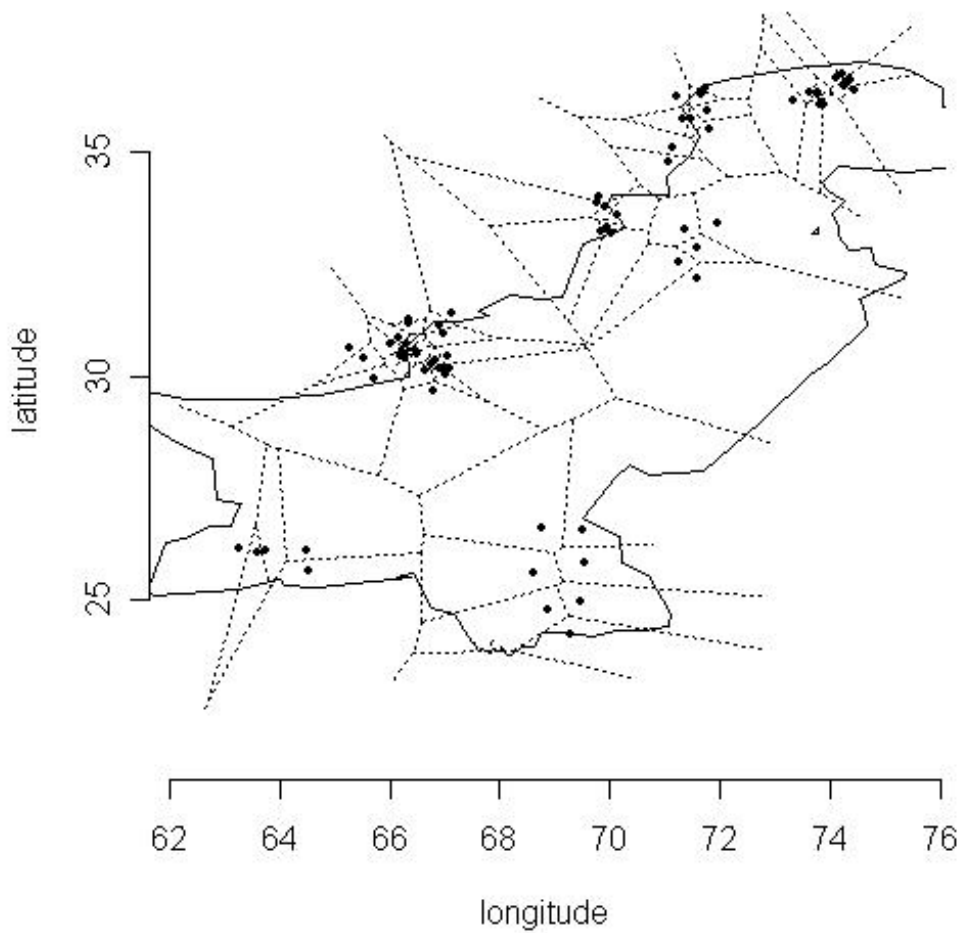


Figure 4: *Sampled geographical coordinates of 70 individuals from the Pakistan data set and the associated Dirichlet tiling. (The full sample was not shown but a similar spatial distribution was assumed for the 200 individuals.)*

**(Low resolution image for review only)**

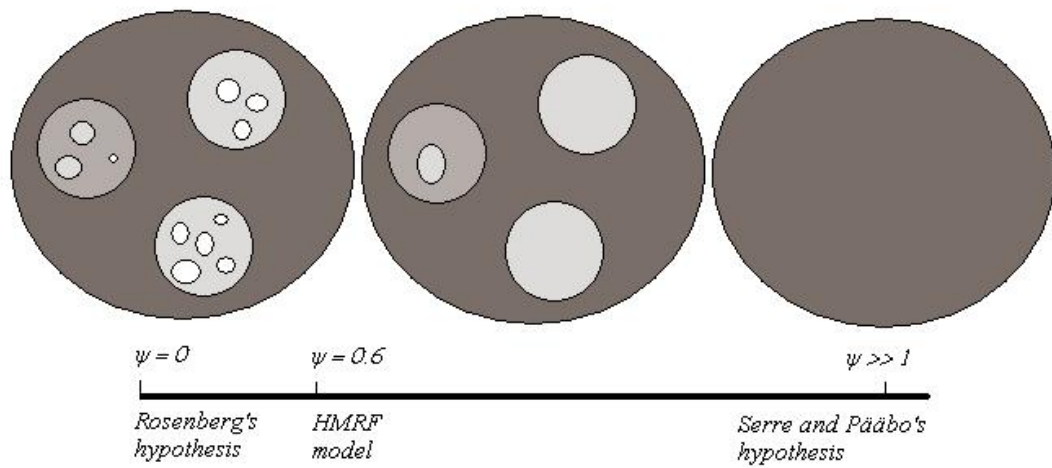


Figure 5: *The reconciliation illustrated. At the left of  $\psi$  axis, a clustering analysis does not account for the spatial continuity of allele frequencies, and may detect more clusters than actually existing. At the right, the pure continuity hypothesis assumes no cluster. Here the vision is intermediate, with the main discontinuities confirmed, but some small clusters may be considered non-significant.*

**(Low resolution image for review only)**