HIDDEN MARKOV RANDOM FIELDS AND THE GENETIC STRUCTURE OF THE SCANDINAVIAN BROWN BEAR POPULATION

Sophie ANCELET [1], Gilles GUILLOT [1], Olivier FRANÇOIS [2*]

ABSTRACT

Spatial Bayesian clustering algorithms can provide correct inference of population genetic structure when applied to populations for which continuous variation of allele frequencies is disrupted by small discontinuities. Here we review works which used Bayesian clustering algorithms for studying the Scandinavian brown bears, with particular attention to a recent method based on hidden Markov random field. We provide a summary of current knowledge about the genetic structure of this endangered population potentially useful for its conservation.

Keywords : Population genetic structure, Spatial Bayesian analysis, Clustering analysis, Scandinavian brown bear.

RÉSUMÉ

Les algorithmes de classification bayésienne spatiale sont utiles afin d'étudier la structure génétique de populations pour lesquelles on observe une variation des fréquences d'allèles généralement continue en espace, mais localement interrompue par de petites discontinuités. Dans cet article, nous présentons une synthèse de travaux récents appliquant ces algorithmes à l'étude de l'ours brun de Scandinavie et nous résumons les connaissances actuelles sur la structure de cette population potentiellement utiles pour sa conservation.

Mots-clés : Structure génétique des populations, Analyse bayésienne spatiale, Analyse par méthodes d'agrégation, Ours brun de Scandinavie.

1. Unité de Mathématiques et Informatique Appliquées, INRA-INAPG-ENGREF, 75732 Paris Cedex 15, France.

2. TIMC Equipe TIMB, Faculté de Médecine, F38706 La Tronche, France.

* Corresponding author: Olivier.Francois@imag.fr

1. Introduction

The improvements of molecular tools in population genetics and ecology have led to an increasing use of Bayesian clustering algorithms in studies of population structure. The aim of conservation biologists and managers is to determine what constitutes a natural break in populations. But the ability to delineate evolutionary significant or conservation units strongly depends on detecting population subdivision (Manel *et al.*, 2003). In some situations, it is easy to define subpopulations on the basis of spatial clustering of individuals. However, individuals are not always arranged in clearly identified clusters, but they may be uniformly distributed across space.

The detection of genetic discontinuities and the correlation of these discontinuities with environmental or spatial features is a typical objective of the users of the Bayesian clustering algorithms developed by Pritchard *et al.*(2000), Dawson and Belkhir (2001), Corander *et al.*(2003), which achieve this goal without assuming predefined populations. Nevertheless, in these algorithms the spatial data are not part of the modelling. In addition, it is still a matter of debate to decide whether (or not) clusters identified by these algorithms are artificially detected structures emerging from uneven sampling along geographical clines, i.e. directions along which allele frequencies vary continuously (Serre and Pääbo, 2004).

We recently argued that Bayesian models offer a natural and appropriate framework for including spatial prior information when assigning an individual to a fixed number of clusters (François *et al.*, 2006). We presented a hierarchical Bayes algorithm that incorporated models for the variation of allele frequencies across space. This was achieved by using *Hidden Markov Random Fields* (HMRF) as prior distributions on cluster membership. Markov Random Fields are indeed 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 the book by Preston (1974)). The term *hidden* indicates that the cluster configuration is unobserved, and that it should be inferred from observations, often using Monte Carlo sampling. In spatial genetics, continuous population usually refers to Wright's famous concept of isolation by distance (Wright, 1943), which can in turn be understood in terms of the stepping stone model (Malécot, 1948), (Kimura and Weiss, 1964). Because it considers interacting demographic units on a lattice, the stepping stone model exhibits the same type of spatial Markov property as does the HMRF model. However using the stepping stone model within a Bayesian framework poses conceptual difficulties, whereas HMRF can capture conditional independence in an efficient way.

In this study, we illustrate the application of HMRFs with the study of the genetic structure of the Scandinavian brown bear population. Brown bears are an example of a wild population with presumed continuous variation in allele frequencies. We showed that HMRFs were powerful at detecting geographical discontinuities in allele frequencies and regulating the number of clusters. Here we briefly discuss the implication of these findings for the conservation of Scandinavian brown bears. Most of the material presented in this review

can be found in recent articles by Blum *et al.* (2004), Manel *et al.* (2004) and François *et al.* (2006).

2. Hierarchical Bayes model

We devised a model-based clustering algorithm that identifies subgroups that have distinctive allele frequencies, and which accounts for the fact that nearby individuals are likely to share similar membership to the subgroups. To achieve this goal, we used a hierarchical Bayesian model based on a HMRF, extending a procedure implemented in the computer program STRUCTURE (Pritchard *et al.* 2000) which places individuals into K clusters, where K is chosen in advance but can be varied across independent runs of the algorithm.

The input data $z = (z_i)$ consist of multilocus genotypes obtained from n diploid individuals located at fixed sampling sites which usually correspond to the habitat. A genotype z_i records paired alleles at L loci ($z_{i\ell}^a$ and $z_{i\ell}^b$, $\ell = 1, \ldots, L$). Each individual originates from a geographical cluster which may span several sampling sites. The cluster to which individual i belongs is labelled as c_i, and the set of all labels $c = (c_i)$ is called the *cluster configuration*.

In the model *with no admixture*, Pritchard *et al.* (2000) made a number of simplifying biological assumptions. First, recombination events have eliminated the potential correlation between the genetic markers (linkage equilibrium). This is a reasonable assumption when the markers are separated by large physical distances. The second assumption was Hardy-Weinberg equilibrium within clusters, which implicates that genes evolve under selective neutrality and local random mating. The program STRUCTURE can actually achieve the statistical inference of $\theta = (c, f)$ where $f = (f_{k\ell j})$ are the unknown allele frequencies, $k = 1, \ldots, K$, $j = 1, \ldots, J_\ell$, and J_ℓ is the number of distinct alleles observed at locus ℓ. The probability of observing the n genotypes given the parameter θ was computed as follows

$$\pi(z|\theta) = \prod_{i=1}^{n} \prod_{\ell=1}^{L} \pi(z_i^\ell | c_i, f_{c_i,\ell,.}) = \prod_{i=1}^{n} \prod_{\ell=1}^{L} \mathcal{L}_k(f_{c_i \ell z_{i\ell}^a}, f_{c_i \ell z_{i\ell}^b}) \tag{1}$$

where $\mathcal{L}_k(f, f) = f^2$ and $\mathcal{L}_k(f, g) = 2fg$ for $f \neq g$.

In the Bayesian approach we compute the posterior density function for (c, f) by combining the likelihood function in (1) with a prior density for (c, f), which we represent in general terms as $\pi(c, f) = \pi(f|c)\pi(c)$.

$$\pi(\theta|z) \propto \pi(z|f, c)\pi(f|c)\pi(c). \tag{2}$$

Conditional on the cluster label $c_i = k$, the priors on allele frequencies $f_{k,\ell,.}$ were Dirichlet distributions $\mathcal{D}(\alpha_k, \ldots, \alpha_k)$. In practice we set $\alpha_k = 1$ for all k.

In the HMRF model, spatial information was modelled through the prior distribution $\pi(c)$. HMRF can account for the conditional independence of individual cluster labels given the neighbours' labels

$$\pi(c_i|(c_j)\,,\, j \neq i) = \pi(c_i|(c_j)\,, j \text{ neighbours of } i)$$

for all i in $1, \ldots, n$. For this reason, this concept is particularly useful for population genetics, because it can model the fact that individuals are more likely to share cluster membership with their close neighbours than with distant representatives. HMRF were also successfully applied in several domains such as computer image analysis (Destrempes *et al.*, 2005) or spatial epidemiology (Green and Richardson 2002). More specifically we defined

$$\pi(c) = \frac{\exp(\psi U(c))}{Z}, \quad c \in \{1, \ldots, K_{\max}\}^n, \tag{3}$$

where ψ is a nonnegative number called the interaction parameter, $U(c)$ is the number of neighbouring pairs that share the same labels in c, and Z is a normalizing constant called the partition function. While the definition of neighbourhood is immediate in the case of grid observations, it is less obvious in the case of irregular sampling. In this study, we used the neighbourhood structure obtained from the so-called *Delaunay graph*. Denoting by (s_i), $i = 1, \ldots, n$, the set of observation sites, each s_i is surrounded by regions made of 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 were neighbours if their cells shared a common edge.

Because computing partition functions is an highly difficult problem, inferences on θ were carried out by simulating the posterior distribution $\pi(\theta|z)$ through an MCMC algorithm. With ψ equal to 0, the model assumed a non-informative spatial prior, and then matched the Bayesian clustering model of Pritchard *et al.*

Note that Eq. 3 assumed the existence of at most $K_{\max}$ clusters, i.e., $c_i \in \{1, \ldots, K_{\max}\}$. In practice the constant $K_{\max}$ may be considered larger than the true (or presumed true) number of clusters, K. In order to estimate K we used the approach proposed by François *et al.* (2006). This approach may be viewed as a *regularisation* method that, loosely speaking, let the algorithm decide which number of clusters can achieve the best trade-off between the influences of genetic and spatial data on the inference of θ.

3. Scandinavian brown bears

As in many other places in Europe, brown bears *Ursus arctos* were almost exterminated in Scandinavia by the end of the nineteenth century. But bounties elimination in 1893 and making killed bears State property in 1927, were efforts that contributed to protect bears in Sweden. The near extinction and recovery of bears in Scandinavia has been well documented and thus provides an excellent record of a population bottleneck and subsequent population expansion (Swenson *et al.*, 1994), (Swenson *et al.*, 1995), (Swenson *et al.*, 1998). After the protection efforts in Sweden, the bear population has recovered from four female concentration areas. These areas were mainly identified from hunting data during the years 1981-1993 as North North (NN), North South (NS), Middle (M) and South (S) (see Fig. 1). Until recently these

areas were believed to represent the surviving relict subpopulations after the 1930's bottleneck maintained separately because of the strong philopatry of females (see e.g. (Waits *et al.*, 2000)). Using a coalescent approach, Blum *et al.* (2004) computed a female spatial dispersal rate and found an estimate of 9 km per generation, which was consistent with field observations.

The structure of the Scandinavian brown bear population into subpopulations was studied both from mtDNA data (Taberlet *et al.*, 1995) and nuclear DNA markers, which give further characterization of the population genetic status (Waits *et al.*, 2000). Waits *et al.* used 19 microsatellite markers collected from 380 bears in this population, and assignment tests to quantify and compare the levels of nuclear DNA diversity for the total population and for each of the four predefined subpopulations. They also estimated the degree of genetic differentiation and the level of gene flow among these four subpopulations. Using F-statistics, they were unable to confirm the existence of a contact zone S/M identified from mtDNA by Taberlet *et al.* (1995).

Manel *et al.* (2004) investigated the persistence of the four relict geographical areas using the multilocus genotypes without predefining populations. From two independent methods (neighbour-joining trees and the Bayesian clustering algorithm structure), a new subdivision of the population was identified. They found four genetic clusters which also matched with geographical clusters, but two of them were distinct from the original female concentration areas.

Because of the low dispersal rate, continuity can be considered as a reasonable assumption to be included in a Bayesian model for Scandinavian brown bear genetic diversity. We analysed the same data set as did the two previous studies. We first used a full-Bayes approach where the prior on ψ was uniform over $(0, 1)$. Values of ψ in this range allowed the prior coexistence of several clusters (simulations not reported), and we ran the algorithm with $K_{\max} = 4 - 7$. After 30,000 cycles, the runs with $K_{\max} = 4$ led to the same clusters as described by Manel *et al.* (2004). 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 and the fourth (S) was splitted into two subsets with random shapes. The spatial interaction parameter ψ had posterior mode in the range $(0.6, 0.8)$ (95% credible interval). However, the irregular shapes of the two S subclusters indicated that the MCMC might have not been run long enough for warranting convergence, perhaps due to the large amount of computational resource spent into the estimation of ψ. Therefore we performed 10 additional runs of the algorithm for two values of interaction parameter $\psi = 0.7 - 0.8$. The runs that reached the highest likelihood resulted in the same four clusters as previously observed (see Fig. 1).

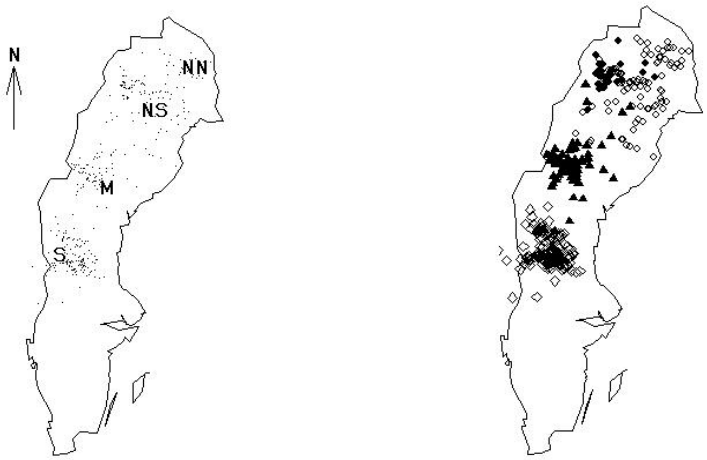

FIG 1. — Left: The spatial distribution of brown bears in Scandinavia. The four subpopulations (NN, NS, M, S) were defined as areas of female concentration. Right: Estimated cluster configuration using the HMRF model. $\psi = 0.7$ and $K_{\max} = 6$. Two clusters (diamonds, triangles) coincided with predefined the populations S and M. Two clusters (black and white circles) differred from the predefined populations NN and NS.

4. Discussion

Detecting population subdivision is a subject of great interest to population geneticists, and a large body of approaches have been developed to this aim. In this study, we presented a Bayesian clustering algorithm that incorporated HMRFs as prior distributions on cluster configurations. The Scandinavian brown bear was an example for which local genetic similarities can be explained by the fact that female disperse at a very low rate. Because of the low dispersal rate in this population, MRF can be considered as an appropriate prior distribution to be included in a Bayesian model. The results provided a reasonable estimate of the number of clusters (four clusters). They confirmed that the genetic structure of the Scandinavian brown bear matches with the four relict clusters only partially, because two of the identified clusters were distinct from the four female concentration areas inferred from female bears killed by hunters.

A potential issue of non-spatial Bayesian algorithms is that they may produce spurious clustering due to irregular sampling design. Inferences were carried out using a large fixed value of the interaction parameter. This large value favored cluster configurations made of few large clusters. The fact that we obtained the same clusters as the non-spatial algorithm provided evidence that the 4 clusters were robust to the inclusion of a continuity prior. In fact,

this study gave support to the hypothesis of 4 clusters resulting from genetic discontinuities within the population rather than artificial clusters created by sampling artifacts.

A long shared genealogical history is one criterion (among others) for biologists to define a significant evolutionary unit of conservation. A closer look at the NWN cluster showed that this cluster actually consisted of few individuals (about 34). A parentage analysis was conducted by Manel *et al.* (2004). This analysis concluded that bears were closely related within the group. Actually, one male was responsible for 88% of the descendants (the male was the father of 70% of them, grandfather of 12% and great-grandfather for 6% of them, and probably the uncle for 9% of them). The cluster might then be explained by matriarchal structure which is known to occur in bears (Rogers, 1987) or by a recent founder effect caused by the expansion of the population. These results suggested to aggregate the NWN and NN clusters into a single evolutionary unit, because the NWN cluster is probably too recent to meet the significance criterion. The overall results confirmed that there was no particular reason for distinguishing the NS and NN bear subpopulations, and we recommended that the Scandinavian brown bear population be viewed as three subpopulations connected by male-mediated gene flow and separated by small relict genetic discontinuities.

Acknowlegments: We are grateful to two reviewers for their useful suggestions and their constructive comments. Olivier François thanks Avner Bar-Hen and Eric Parent for their invitation to present this article at the "Statistical Modelling in the Environment with Special Reference to Biodiversity and Spatio-temporal Approaches" meeting in Paris, May 2006. This work was supported by grants from the ACI IMPBio (Interface Mathématiques, Physique, Biologie) and the ANR Project MAEV (Modèles Aléatoires pour l'Évolution du Vivant).

References

Blum M., C. Damerval, S. Manel, and O. François (2004). Brownian models and coalescent structures. Theoretical Population Biology **65:** 249-261.

Corander J., P. Waldmann, and M. Sillanpää (2003). Bayesian analysis of genetic differentiation between populations. Genetics **163:** 367-374.

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.

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

GREEN P. and S. RICHARDSON (2002). Hidden Markov models and disease mapping. Journal of the American Statistical Association *97* (460): 1055-1070.

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.

MANEL S., M. SCHWARTZ, G. LUIKART, and P. TABERLET (2003) Landscape genetics: combining landscape ecology and population genetics. Trends in Ecology and Evolution *18*(4): 189-197.

PRESTON C. (1974). *Gibbs States on Countable Sets*. Cambridge: Cambridge University Press.

PRITCHARD J., M. STEPHENS, and P. DONNELLY (2000). Inference of population structure using multilocus genotype data. Genetics **155:** 945-959.

ROGERS L. (1987). Effects of food supply and kinship on social behavior, movements, and population dynamics of black bears in northeastern. Minnesota Wildlife **97:** 1-72.

SERRE D. and S. PÄÄBO (2004). Evidence for gradients of human genetic diversity within and among continents. Genome Research **14:** 16791685.

SWENSON J., F. SANDEGREN, A. BJARVALL, A. SODERBERG, M. WABAKKEN, and M. FRANZEN (1994). Size, trend, distribution and conservation of the brown bear, Ursus arctos, population in Sweden. Biological Conservation **70:** 9-17.

SWENSON J. E., F. SANDEGREN, A. BJARVALL, M. FRANZEN, and A. SODERBERG (1995). The near extinction and recovery of brown bears in Scandinavia in relation to the bear management policies of Norway and Sweden. Wildlife Biology **1:** 11-25.

SWENSON J. E., F. SANDEGREN, and A. SODERBERG (1998). Geographic expansion of an increasing brown bear population: evidence for presaturation dispersal. Journal of Animal Ecology **67:** 819-826.

TABERLET P., J. SWENSON, F. SANDEGREN, and A. BJARVALL (1995). Localization of a contact zone between two highly divergent mitochondrial DNA lineages of the brown bear Ursos arctos in Scandinavia. Conservation Biology **9:** 1255-1261.

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.

WRIGHT S. (1943). Isolation by distance. Genetics **28:** 114-138.