MBE Advance Access published May 21, 2009

Spatial inference of admixture proportions and secondary contact zones

Eric Durand[1] Flora Jay[1] Oscar E. Gaggiotti[2]
Olivier François[1]

May 14, 2009

Address 1: Faculty of Medicine, TIMC-IMAG, University Joseph Fourier, Grenoble IT, Group of Mathematical Biology, F38706 La Tronche, France.

Address 2: Laboratoire d'Ecologie Alpine, LECA, UMR CNRS 5553, University Joseph Fourier, BP 53, 2233 Rue de la Piscine, 38041 Grenoble Cedex 9, France.

Running head: Spatial admixture and contact zones.

Keywords: Admixture, Bayesian inference, spatial trends, spatial autocorrelation, secondary contact zones.

Corresponding author: Olivier François, TIMC-IMAG, University Joseph Fourier, Grenoble IT, Group of Mathematical Biology, F38706 La Tronche, France.
Phone: +33 4 56520025
E-mail: olivier.francois@imag.fr

Genetic admixture of distinct gene pools is the consequence of complex spatio-temporal processes that could have involved massive migration and local mating during the history of a species. However current methods for estimating individual admixture proportions lack the incorporation of such a piece of information. Here, we extend Bayesian clustering algorithms by including global trend surfaces and spatial autocorrelation in the prior distribution on individual admixture coefficients. We test our algorithm by using spatially explicit and realistic coalescent simulations of colonization followed by secondary contact. By coupling our multiscale spatial analyses with a Bayesian evaluation of model complexity and fit, we show that the algorithm provides a correct description of smooth clinal variation, while still detecting zones of sharp variation when they are present in the data. We also apply our approach to understanding the population structure of the killifish, *Fundulus heteroclitus*, for which the algorithm uncovers a presumed contact zone in the Altantic coast of North-America.

Biological data based on geographic surveys often display global trends and spatial autocorrelation (Sokal and Oden 1978). Spatial autocorrelation is the correlation of a geographic variable with itself but at a certain distance apart. This phenomenon complicates the analysis of spatial patterns by creating departure from the standard independence hypothesis (Slatkin and Arter 1991; Epperson and Li 1996). This pattern may be driven by endogeneous factors like dispersal limitation, or by exogenous factors like an important environmental determinant that is spatially structured and that implies spatial structuring in the observed variable. It is widely acknowledged that underestimating autocorrelation in ecological data can bias inference from statistical models (Lichstein et al 2002; Dormann 2007).

Traditional spatial statistical analyses take these points into account by decomposing the spatial variation of a response variable, z, into global and local effects

$$z = m(x) + y,$$

where x are the two-dimensional spatial coordinates. The first term, $m(x)$, is a trend surface – often defined as a first order polynomial, $m(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$ – capturing regional or long-range variation. The second term, y, is a spatially autocorrelated residual which represents short-range variation. This approach is sometimes called *universal kriging* (Ripley 1988) or spatial trend analysis (Bocquet-Appel and Sokal 1989).

An important question that could greatly benefit from a more precise modeling of spatial patterns is the study of genetic admixture. The demography of natural populations is the result of phases of expansion, contraction and migration, or local mating that can produce shifting patches of genotypes. In such conditions, populations isolated for a long time may be brought into contact in a certain area,

leading to the genetic admixture of different gene pools (Chakraborty 1986). Admixture is particularly pervasive in humans because migratory movements have brought together peoples from different origins (Cavalli-Sforza et al 1994), and its precise assessment is important for association studies that are susceptible to biases due to population structure (Pritchard et al 2000; Yu et al 2006). In addition, admixture between populations originating in different continents can be exploited to detect disease susceptibility loci at which risk alleles are distributed differentially between these populations (Chakraborty and Weiss 1988; Reich and Patterson 2005; Smith and O'Brien 2005).

Under natural conditions, admixture is known to happen in secondary contact zones, and it may generate Hardy-Weinberg and linkage desequilibria at unlinked loci (Barton and Hewitt 1985; Durrett et al 2000). These zones are places where the hybrid offspring of the interbreeding populations are present, and where their allele frequencies form a cline (Barton and Gale 1993; Endler 1977). Secondary contact or hybrid zones have often been described as the consequence of post-Pleistocene re-colonization of landmasses after the ice retreat (Taberlet et al 1998). Detecting and identifying the relative contributions of these refugia to current populations is of paramount interest to the reconstruction of the demographic history of many organisms (Hewitt 2000).

Many admixture models compute population coefficients, considering hybrid genes as proportionally inherited from two or more populations that are thought of as being the relicts of some parental populations. The quantities being estimated, the admixture coefficients, are the respective contributions of the parental populations to the hybrid gene pools. Several approaches to estimating these proportions in populations have been proposed during the last few decades, including least-square regression (Roberts and Hiorns 1965), maximum likelihood (Long

1991), estimation of coalescence times (Bertorelle and Excoffier 1998), Markov chain Monte Carlo (MCMC) algorithms or likelihood-based methods (Chikhi et al 2001; Wang 2003), and approximate Bayesian computation (Excoffier et al 2005). Regarding the estimation of admixture proportions in individuals, current methods are based on computer-intensive programs like STRUCTURE (Pritchard et al 2000; Falush et al 2003), ADMIXMAP (Hoggart et al 2004), INSTRUCT (Gao et al 2007), LAMP (Sankararaman et al 2008). Spatial models have been implemented in TESS (Chen et al 2007) and BAPS (Corander et al 2008). Recent examples of the use of individual-based Bayesian clustering algorithms are for the genetic analysis of hybridization between two species of lemurs (Pastorini et al 2009), the inference of a strong subdivision between two subpopulations of the lepidopteran *Chilo suppressalis* in China (Meng et al 2008), the demographic history of European population of the model plant *Arabidopsis thaliana* (François et al 2008) or the recolonization of the Swiss Alps by the Valais shrew *Sorex antinorii* (Yannic et al 2008). Principal component analysis (PCA) may provide concurrent means to estimate admixture proportions, and spatial versions of PCA might also be relevant to this framework (Patterson et al 2006; Jombart et al 2008).

In this study we extended the hierarchical Bayesian algorithm implemented in TESS in order to include spatial prior distributions on the individual admixture proportions, and we assessed the abilities of this approach to detect admixture in secondary contact zones. The proposed approach adopts a formulation similar to universal kriging in which a response variable – here admixture proportion – can be modelled as the sum of two components: a trend surface plus a Gaussian autoregressive residual term (Besag 1975; Ripley 1981; Cressie 1993). The trend surface and the residual terms attempt to capture the broad-scale and fine-scale patterns that may be expected under migration or local isolation-by-distance pro-

cesses (Bocquet-Appel and Sokal 1989).

The objective of the proposed algorithm is to improve inference of admixture proportions when admixture proportions are variable across space. The inference method is tested on synthetic data obtained from simple models and from spatially explicit scenarios simulating secondary contact and mimicking realistic migration routes for a species that colonized Europe from two glacial refugia. We measure the relative fit of spatial and non-spatial models in terms of statistical information criteria, and we display their posterior spatial predictions using a two-dimensional graphical method. The approach is applied to analyzing an hypothesized contact zone in the marine species *Fundulus heteroclitus*, with individuals genotyped at 8 microsatellite loci in 15 samples along the east coast of North-America (Adams et al 2006).

1 Materials and Methods

A spatial prior for admixture proportions. We consider N individuals genotyped at L loci, and we assume that their geographic coordinates were recorded at the sampling locations. Individuals can be either diploid or haploid. As in the algorithm underlying STRUCTURE (Pritchard et al 2000), we assume that the individuals represent a mixture from at most $K_{\max}$ unobserved clusters, and a matrix denotes the admixture proportions for all the individuals. Each element of the matrix, q_{ik}, is the proportion of individual i's genome that originated from cluster k.

We perform inference of population structure in a Bayesian framework by incorporating individual geographic covariates in the prior distributions on the admixture coefficients. More specifically, we assume a Dirichlet distribution on

the q_{ik}'s for each individual i,

$$q_{i.} \sim \mathscr{D}(\alpha_{i1}, \ldots, \alpha_{iK_{\max}}), \tag{1}$$

where α_{ik} is proportional to the average admixture coefficient, $\mathrm{E}(q_{ik})$. The novelty is that we consider a log-normal model for the α_{ik}, viewed as unobserved response variables

$$\log(\alpha_{i.}) = f(x_i)^T \beta_{.} + y_{i.}, \tag{2}$$

where x_i represents a two-dimensional vector of spatial covariates for i, for example latitude and longitude. Log-linear regressions of the average admixture levels on the spatial covariates are performed in each of the $K_{\max}$ clusters. The definition of the two terms appearing in the right-hand-side of equation (2) is given hereafter.

The hidden regression model described in equation (2) is similar to *universal kriging* (Ripley 1981; Cressie 1993), and it can be separated into two components. The first component, $m = f(x_i)^T \beta$, represents the mean response, and it is modelled as a (possibly) non-linear trend surface. Although this was not stated explicitly, latent regression models that may incorporate trend surfaces were previously considered by Gaggiotti et al (2004), Foll and Gaggiotti (2006) and Faubet and Gaggiotti (2008) who studied population divergence measures and recent migration rates. We limited our further analyses to linear trend surfaces, but the proposed method is valid for arbitrary polynomial shapes, and our computer program allows the use of quadratic or cubic models. The second component, y_i, represents a zero-mean spatially autocorrelated random variable. This term is a conditional autogressive Gaussian model (CAR, Besag 1975; Vounatsou et al 2000). In the CAR model, the conditional expectation of y_i, given the response at all other locations, is a weighted sum of the mean-centered coefficients at neigh-

boring locations

$$E(y_i \mid y_j \text{ at other locations}) = \rho \sum_{j \text{ neighbor of } i} w_{ij} y_j \tag{3}$$

where ρ is a parameter that determines the magnitude of the spatial neighborhood effect, and w_{ij} are weights that determine the relative influence of location j on location i. The CAR model is mathematically defined as a Gaussian random field, and it may represent the locally structured part of the variation. To better account for local mating, we defined neighbors from the Dirichlet tessellation (François et al 2006), and we used an exponential covariance matrix to model the decay of correlation with geographic distance

$$w_{ij} = \exp(-d_{ij}/\theta) \tag{4}$$

where d_{ij} is the great-circle distance between the sites i and j, and θ is a scale parameter that may be related to the intensity of gene dispersal. More specifically, the expression (3) for y_i implies the covariance matrix $\Lambda = \sigma^2(\mathrm{Id} - \rho W)^{-1}$ where W is an $N \times N$ matrix with zeros on the diagonal and the neighbor weights (w_{ij}) in the off-diagonal positions, Id is the identity matrix, and σ^2 is the variance of the CAR. Equation (3) underlines that ρ and θ are not simultaneously identifiable parameters, and that estimates should focus on the product ρW. In practice, we set θ equal to the mean value of great-circle distances between the individuals locations, and ρ is estimated from the data. We further refer to the model defined in equation (2) as the full regression model. A model without the CAR component is termed a trend model.

To give correct interpretations of linear trend surfaces, one should keep in mind that the assumption is not that the admixture proportions vary linearly in space. In fact the model assumes that the q_{ik}'s have sigmoidal shapes across space,

mirroring theoretical predictions for allele frequency curves in hybrid zones (Barton and Hewitt 1985). To give an illustration of the shape of the admixture proportions under a linear trend model, we simulated realizations of the prior model using two clusters. Assuming dependence on the longitude, x, we parameterized the trend surface as $m_1 = a - bx$ in cluster 1 and as $m_2 = -a + bx$ in cluster 2 ($a = 25, b = 5$), and we sampled individuals along the longitudinal gradient ($x \in [4,6]$). A rough approximation of the average admixture proportion in cluster 1 at longitude x can then be given by $q_{x,1} = 1/(1+\exp(-2(a+bx)))$, which can be represented by a sigmoid curve. Figure 1 shows that the curve of the expected admixture proportions indeed varies spatially with a sigmoidal shape, staying almost constant in each cluster and decreasing sharply at the boundary between the two clusters. Simulations with 3 adjacent clusters displayed similar patterns, with admixture coefficients showing stable values over large regions and varying substantially at their boundaries.

Implementation details. Our Bayesian model was implemented as a hybrid MCMC algorithm, following Gelman et al (2004) for the priors on regression models and Metropolis-Hastings rules for the CAR model (Supporting Text ST1). For the parameter ρ, we used a non-informative prior over the interval $(0, 1/\lambda_{\max})$, where $\lambda_{\max}$ is the largest eigenvalues of W, and we implemented Metropolis-Hastings updates. An important feature of the hidden regression approach was the possibility to display posterior predictive maps of admixture coefficients. These maps can show the predictions of admixture proportions for an individual at an arbitrary geographic location, adding useful information to the standard unidimensional bar chart representations.

Since the model specified in equation (2) is not the unique way to define a spatially explicit prior for admixture, we implemented variants of the above Bayesian

approach. One alternative is to use a multinomial logit regression model for the admixture proportions, q_{ik}, instead of the log-normal model for their average-proportional values, α_{ik}. Another alternative is to use a convolution Gaussian prior with two variance parameters, τ^2 and σ^2, as defined by Besag et al (1991) and used by Mollié (1996) in an epidemiological context (BYM model; Supporting Text ST1). The CAR and the BYM models are close to each other. Both of them were implemented in the TESS computer program, and were used in the subsequent data analyses. They generally led to similar results. For the BYM model we used non-informative priors on variance parameters, and updates of these parameters were performed according to a Gibbs sampling algorithm.

Model choice. Following Pritchard et al (2000), we suggest performing analyses of population structure for a range of values of $K_{\max}$. When choosing the number of population, we need to account for the fact that including a trend surface implies a regularization of the number of observed clusters, so that the actual number of clusters, K, may be less than the number specified by the mixture model (François et al 2006). To decide which $K_{\max}$ (and K) may provide the best fit to the genetic data, we used the Deviance Information Criterion (DIC, Spiegelhalter et al 2002). The DIC was computed along MCMC runs as the average model deviance plus a penalty term, p_D, that counts the effective number of parameters in a model. To select the number of clusters, the program was run for a range of values of $K_{\max}$, and we considered the values for which the DIC first reached a plateau, like it is usually done for STRUCTURE with the logarithm of evidence (Evanno et al 2005). The DIC was also useful for selecting among a non-spatial prior, a trend-only prior or the full model (trend plus CAR prior). It allowed us to assess the presence of clines or clusters and to measure the relative importance of large-scale and local effects. In this case the focus of model selection shifted to

choosing the best regression model, and we utilized a conditional version of the DIC based on the average residuals of the hidden regression model (Celeux et al 2006).

Simulations of recent admixture of two parental populations. In a first series of experiments, we simulated spatial genetic data mimicking the instantaneous admixture of two weakly differentiated parental populations. The parental populations were assumed to be in migration/drift equilibrium, and genotypes for $n = 400$ diploid individuals were obtained from structured coalescent simulations with two islands, constant levels of gene flow and constant mutation rate (infinite allele model, $4\mu N_e = 1$). We controlled the simulations by varying the effective migration rate $M = 4mN_e$ between 4 and 12 so that the F_{ST} of the parental gene pool varied in the range $[.02, .05]$ (estimated with HIERFSTAT (Goudet 2005)). To create a spatial framework, the individual locations were randomly generated with Gaussian distributions around two centroids put at distance 2 on a longitudinal axis (SD = 1). The genotype of each individual at each of L loci was built as follows. For each individual and each locus, we computed the distance d_1 (d_2) to the left (right) centroid, and we assumed that each allele originated in the first (second) parental populations with probability $d_2/(d_1+d_2)$ ($d_1/(d_1+d_2)$). We used $L = 100$ loci. This simulation of individual levels of admixture was similar to the ones classically used in studies of population samples (Griebeler et al 2006; Chikhi et al. 2001). The simulation imposed a longitudinal trend to the genetic data, with individuals at lower longitude sharing more alleles with the first parental population than with the second one. Spatial autocorrelation was neglected in this simulation process.

Simulations of contact zones in Europe. In a second series of simulations, we used spatially explicit simulations to generate synthetic population genetic

data following secondary contact. Simulations were performed using SPLATCHE (Currat et al 2004), a computer program that allows incorporation of geographic and environmental information in the migration scenario. The simulation of the demographic phase occurred in a two-dimensional non-equilibium stepping-stone model defined on a lattice of ∼25,000 cells (or demes) covering Europe. Each deme represented a surface of ∼450 km^2, and exchanged migrants with its four neighbors at rate m. Topographic information was imported from a geographical information system, and it was encoded into distinct friction values for each cell. In these simulations, measures of genetic differentiation at neutral loci increased with geographic distance. Population sizes grew logistically at rate r in each deme, and saturated at their carrying capacity, C. The three parameters r, m, C determined the speed of the wave-of-advance. In our study, the growth rate was set to $r = .6$, the migration rate ranged between [.2, .9] and carring capacities were set either to $C = 100$ or to $C = 1,000$ in each deme. With the tested parameter settings, Europe was colonized in less than 600 generations.

The dynamics were started from an ancestral population of effective size $N_e = 1,000$ individuals. After an initial divergence phase of about $300 - 500$ generations, populations started to colonize Europe from two distant southern foci, one in the Iberian peninsula and the other one in Turkey. Secondary contact occurred in Central Europe, in an area close to Germany. We used a friction map that made migration toward mountainous areas more difficult, and water-masses were impossible to cross. We added two isthmi that connected the British Isles to France and Scandinavia to Denmark. We used two values for the total number of generations, $T = 1,000$ and $T = 2,500$. The genetic data were simulated as short tandem repeats at either $L = 10$ or $L = 100$ neutral loci according to the stepwise mutation model. We used a mutation rate of 5×10^{-4} per locus and generation, and we

sampled 60 populations at random locations in Europe containing either 3 or 20 individuals per sample. Combining all the simulation parameters we generated a total of 16 data sets.

Simulations of equilibrium stepping-stone models. In a third series of experiments, we used EASYPOP (Balloux 2001) to generate spatial genetic data sets under an equilibium model of isolation-by-distance. Under this scenario, theory shows that measures of genetic differentiation at neutral loci increase with geographic distance, due to the well-known process of accumulation of local genetic differences under geographically restricted dispersal (Wright 1943). Allele frequencies vary across the region, but they do not exhibit regional shapes. Equilibrium stepping-stone simulations are examples of data that do not correspond well to Bayesian clustering model assumptions. In absence of a reasonable number of source populations, the inferred value of the number of clusters and the corresponding allele frequencies in each cluster can be rather arbitrary (Pritchard et al 2003).

The simulation took place in a two-dimensional stepping-stone model defined on a 10 by 10 lattice. We generated data sets for 60 populations of diploid inviduals genotyped at 10 microsatellite loci. The mutation rate was set to $\mu = 5 \times 10^{-4}$, and the migration rate, m, was varied in the interval $[.3, .9]$. Then we created two data sets by randomly resampling 3 individuals in each population. The presence of long-range isolation-by-distance was assessed by regressing the pairwise differentiation measures $F_{\mathrm{ST}}/(1 - F_{\mathrm{ST}})$ on the geographic distances.

Application to Fundulus heteroclitus data. The mummichog *Fundulus heteroclitus* is a small killifish. Its habitat ranges from northern Florida to the Gulf of St Lawrence along the eastern coast of North America. It has been shown that *F. heteroclitus* exhibited a steep latitudinal cline using allozymes, mtDNA and mi-

crosatellite markers (Power 1991; Adams et al 2006). Several hypotheses for this clinal variation have been proposed, including secondary contact between two divergent populations or a northward expansion from a southern refugium after the last glacial age. Using 731 diploid individuals genotyped at 8 microsatellite loci, Adams et al (2006) showed that a pure northward expansion might not explain the observed nuclear pattern of variation, and they suggested an alternative model of post-glacial colonization.

MCMC runs. We studied a total of 22 simulated data sets plus one biological example. The scale parameter θ was set to 1 in the first four data sets (recent admixture) and to $\theta = 1,000$ in the other ones (contact zones). In the scenarios of recent admixture and the equilibrium isolation-by-distance simulations, we present results for the CAR model (similar results were obtained with the BYM model). In secondary contact simulations and for the killifish, we used the CAR and BYM models. Results were almost identical for both models, and we reported results for the second one.

For each data set we investigated which of a non-spatial, a linear trend or a full model provided the best fit. These analyses were performed for values of $K_{\max}$ ranging from 2 to 7. MCMC algorithms were run for a length of 50,000 sweeps with burn-in periods of 40,000 sweeps. For each data set and for each model, we ran the algorithm 100 times, retained the 10 runs with the best DICs, and averaged admixture estimates using CLUMPP (Jakobsson and Rosenberg 2007). As the full analysis required 41,400 runs, we put restriction on some computations when the results were obvious (scenarios 1-6). Runs were performed using an upgraded version of the program TESS (Chen et al. 2007) on a cluster of computers.

2 Results

Recent admixture of two parental populations. For $K_{\max} = 2$ and $F_{ST} \geq .04$, the smooth longitudinal cline created in the simulated data was uncovered by the spatial algorithms (Figure S1A). Note that the F_{ST} values were computed before creating admixture, and that these numbers were likely to overestimate the true levels of differentiation in the data. Using a conditional version of the DIC for the hidden regression, we evaluated the fit of the non-spatial (trend of degree 0), longitudinal trend (trend of degree 1), and both longitudinal and latitudinal trends (trend of degree 1) models in Table 1 (no autocorrelation term). Minimum values were computed over 100 runs (Min) and averages over the ten best runs (Mean and Standard Deviations). The best values are bolded and marked with a star. Values for $F_{ST} = .02$ were similar to those reported for $F_{ST} = .03$. The non-spatial algorithm was unable to obtain correct estimates of the admixture proportions when $F_{ST} = .04$. The clustering algorithms failed to uncover the cline at $F_{ST} \leq .03$. There was a steep decrease of DICs when the cline was detected, shifting from values around 420 to values around 370. In the latter case, the DIC analysis selected the longitudinal trend model (DIC = 362-366) in agreement with the synthetic data generation process. The correlation between the estimated admixture proportions and their true values was also highest for the longitudinal trend model ($r = .97$, $p < 10^{-10}$), indicating that the cline was almost perfectly reconstructed by the algorithm. Similar results were obtained for $K_{\max} = 3-4$ for which $K = 2$ effective clusters were actually detected when $F_{ST} \geq .04$. We also obtained slightly better performances for these data sets when we used a multinomial logit regression model for the admixture proportions, uncovering the cline at $F_{ST} = 0.03$ (Figure S1B).

Table 1: Conditional DIC for data sets simulating recent admixture of two populations.

F_{ST}	no covariate			longitudinal trend			linear trend surface		
	Min	Mean	sd	Min	Mean	sd	Min	Mean	sd
0.03	414.7	419.8	3.87	417.4	422.7	4.31	419.4	424.8	3.41
0.04	416.0	422.9	4.98	**362.0***	369.1	4.94	367.5	398.4	5.72
0.05	387.7	397.7	4.77	**366.4***	379.5	3.72	373.6	383.4	3.22

The strength of the spatial effect was measured by the regression coefficients. Table 2 presents these coefficients for the trend model and for the scenario with $F_{ST} = .05$. As expected, there was a clear effect of longitude on the admixture proportions. Latitude, on the other hand, had no detectable influence since the credibility interval of its regression coefficient included zero (Figure S2). sFinally, the symmetric role of the two parental populations was reflected by regression estimates that were approximately symmetric for each cluster.

Table 2: Regression table for a data set simulating recent admixture of two populations.

	CLUSTER 1		CLUSTER 2	
	Estimate	95% C.I.	Estimate	95% C.I.
Intercept	-25.13	$[-33.57, -16.44]$	23.86	$[14.07, 34.02]$
latitude	1.03	$[-0.56, 2.71]$	0.34	$[-1.41, 1.97]$
longitude	4.58	$[3.18, 5.92]$	-4.51	$[-6.01, -3.13]$

Contact zones in Europe. The levels of differentiation in the 16 simulated data sets ranged from .02 to .28. The highest F_{ST}'s were observed for the smaller mi-

gration rate, number of generation and carrying capacities. In accordance with classical models, the F_{ST} decreased when one of these parameters increased. In all data sets, longitudinal clines separating the western and eastern part of the continent were inferred as soon as we set $K_{max} \geq 2$. These patterns clearly exhibited a contact zone localized in central Europe.

Separate DIC analyses were ran for the BYM model and for the small (180 individuals, 10 loci) and the large (1,200 individuals, 100 loci) data sets (Figure S3 and S4). When the small data sets were used to compare the non-spatial, trend and full models using $K_{max} = 3$, the relative differences in DIC were in favor of the inclusion of spatial covariates. The DIC selected the full regression model $7/8$ times and the trend model for $1/8$ data set (Figure 2A). For $K_{max} = 5$, the spatial models outperformed the non-spatial models, except for one data set (Figure 2B). The full model was also selected more often (5/8) than the trend model (2/8). For these data, the effective number of cluster varied between 2 and 4, with the lowest K's found in data sets with small F_{ST}'s. Figure 2C-D details the DIC analysis for two data sets (labels 1 and 8). Similar conclusions were reached for the big data sets, but the trend model was selected more often (3/8) than the full model (1/8) as more loci and larger samples were used.

The main features observed in the spatial population structure analyses are illustrated in Figure 3A-B, considering one particular data set with demographic parameters $T = 1,000$, $C = 100$, $m = .3$, total sample size $n = 1200$ and $L = 100$ loci. For these data, the DIC selected the linear trend model (DIC = 181,456) and a value of $K_{max} \approx 5$ (label 6, Figure S4). For $K_{max} = 2$, the admixture estimates exhibited a clinal pattern in central Europe. For $K_{max} = 3$, a clear separation was identified in Scandinavia, a pattern that was observed in a majority of the simulations. For $K_{max} = 4$, a small cluster – particular to the studied data – was found in

the North-East of Europe (blue cluster). Setting $K_{\max} \geq 5$ did not modify the estimates of admixture proportions significantly. Figure 3B displays a posterior map of predicted admixture levels in Europe. The hidden regression model predicted a long and narrow contact zone (in red pixels) consistent with the shape of hybrid zones observed in many species (Barton and Hewitt 1985).

Equilibrium stepping-stone simulations. For the data set with the largest migration rate, the pairwise F_{ST}'s ranged from 0.0004 to 0.11, and the mean differentiation was equal to 0.042 (SD= 0.019). The extent of long-range isolation-by-distance was assessed in Figure S5. With the smallest value of the migration rate, the levels of differentiation ranged from 0.0004 to 0.0037. Varying $K_{\max}$ between 2 and 9, we used the DIC to compare the non-spatial models, CAR models (trend of degree 0, $\rho > 0$), trend models (trend of degree 1, $\rho = 0$), and full models (trend of degree 1, $\rho > 0$). For the largest value of the migration rate m, the DIC analysis revealed that the value $K_{\max} = 4 - 5$ received the highest support, and that no model performed better than the non-spatial models (Figure S6). No cluster was effectively discovered. The results for the smallest value of m were similar to those obtained for the largest value. With more extensive sampling (20 individuals in each population) and more genetic data (100 microsatellite neutral markers), again no model performed better than the non-spatial models. We obtained 4 clusters, located in the corners of the study area that were not subsets of those obtained with $K_{\max} = 3$, suggesting that they might correspond to mathematical artifacts.

Application to Fundulus heteroclitus. Using the same scheme as for the simulated data, we fitted non-spatial, linear trend and full hidden regression models to the *Fundulus* data (BYM model). The linear trend model obtained the best DICs for values of $K_{\max}$ in $[2,7]$ (Figure 4). Increasing $K_{\max}$ above 3 did not lead to a significant decrease in DICs, and the clustering results remained unchanged

suggesting that the effective number of cluster could be estimated as $K = 3$ (Figure 3C-D). The best models detected a cline separating the northern and southern populations, and grouped two isolated samples to the south to the study area. The posterior predictive map localized the cline to the east to New Jersey (in red pixels, Figure 3D) and agreed with the findings of Adams et al (2006).

3 Discussion

We proposed a Bayesian algorithm to estimate individual admixture proportions by incorporating spatial trends and spatial autoregressive processes in the prior distribution on these coefficients. The priors were defined as hidden regression models with autocorrelated residuals including spatial effects at multiple scales. Although spatial autoregressive models have been known for a long time in the statistical literature, they have been considered in population genetics only recently (Vounatsou et al 2000; Wasser et al 2004). The new algorithms extend a previous work by François et al (2006) who implemented a hidden Markov random field in a model without admixture. The results of our simulation study indicate that our method can outperform those that ignore spatial information, especially when genetic information is not extensive. For example, this is the case in non-model species for which extensive genomic data sets are not yet available.

Regression of admixture proportions. Regression of admixture coefficients has received much attention in population genetics in recent years. For example, regression was previously utilized to examine the relationships between admixture and geographic distance in Europeans. This was done in order to support the hypothesis of a large contribution of the Neolithic farmers to the current European gene pool. Surveys of admixture clines in this context uncovered an approximate linear relationship between admixture proportions and distance to a putative

eastern origin (Chikhi et al 2001; Dupanloup et al 2004; Belle et al 2006), or a true eastern origin when simulations were used (Currat and Excoffier 2005). Because they assumed statistically uncorrelated residuals, regressions of posterior estimates might differ from those obtained by our approach in a drastic way. In our approach, the regression is part of the modeling process. Polynomial trend surfaces may account for clines in all directions, and autocorrelated residuals may account for isolation-by-distance. Including spatial information in the prior distribution on the admixture proportions can also provide posterior estimates that have been corrected for genealogical correlation between individuals. This is achieved in a rather natural fashion using the hierarchical Bayesian approach (Gelman et al 2003).

Model selection and DIC. An important intrinsic feature of imposing spatially structured priors was the possibility for the MCMC algorithm to eliminate a number of spurious clusters automatically. When we input a maximum of $K_{\max}$ clusters to the model, the effective number of cluster in the data may be a smaller value, K. In this case, the DIC sometimes selects models in which $K_{\max}$ is greater than K. An explanation may be the variability in estimated DICs. Theory predicts that errors in information criterion comparisons are of order $\sqrt{n}$, where n is the number of observations (Ripley 1996). We suspect that the constant term in this large-sample approximation could be rather big, especially in complex hierarchical models as implemented in this study. In the killifish example, models with $K_{\max} = 4-5$ clusters were given smaller values of the DIC than models with $K_{\max} = 3$ clusters. Nevertheless it was obvious from the direct inspection of the posterior estimates that the effective number of population was equal to 3 in the selected models. It is possible that the DIC decrease – around 100 units – may not be large enough to justify a choice of a model with a larger number of clusters.

Note that although the DIC is widely acknowledged to be a useful measure, it does not always lead to choosing the best model (Brooks 2002).

Simulation analyses. In the simulations of recent admixture, a given level of admixture was assigned to each individual according to a pure longitudinal trend model. These simulations were an approximation of more complex spatially explicit processes, for which we neglected spatial autocorrelation. The DIC analysis selected the correct covariate, and the observed number of cluster in the data agreed with two parental populations. The posterior estimates of the admixture coefficients exhibited a longitudinal clinal shape, as we expected. In secondary contact simulations, the best models were obtained when we included both the trend and the autocorrelation terms in the statistical model. The estimated trends were appearent in the prediction maps, and they were oriented along a longitutinal axis. They were visible for $K_{\max} \geq 2$, and they captured the signature of the simultaneous range expansion from the two refugia. The inclusion of autocorrelation in the best model was not a surprising result as sampling was dense enough to observe the short range effects that are inherent to the stepping-stone simulation (the average distance between nearest samples was around 300 kilometers). The prediction maps for the admixture proportion described and highlighted the areas where the hybrids resided. These hybrid zones conformed to their theoretical predictions (Barton and Hewitt 1985). In some runs, more than three clusters were actually found, especially when we used the larger number of loci and the larger sample sizes. Only the continental cline and the northern cluster were consistently present in all runs. The additional clusters were often located in the North-East or in the British Isles, and might have resulted from drift or localized founder effects within the main cline. Such founder effects were more frequent when the Baltic sea was crossed, leading us to observe a Scandinavian cluster more frequently.

One potential source of misleading interpretations is with data sets arising from homogeneous short-range migration process across time and space. Such data clearly violate the spatial admixture model assumptions. The formulation of the admixture model accounts for short-range isolation-by-distance effects by way of the autocorrelated residuals, and for regional effects by means of the latent regression model and the trend surface. Under an equilibrium stepping-stone model we expect a long-range isolation-by-distance pattern. Because there are no regional effects, the trend surface is not useful, and genetic variation is partitioned over artificial clusters like for other Bayesian clustering algorithms. In addition, we observe that the estimated clusters are inconsistent over increasing values of $K_{\max}$. In contrast, the reason why it works well in the case of a secondary contact zone is that, in this case, variation is more structured and exhibits regional trends. Regional effects are well taken into account by the latent regression, which makes clusters easier to identify than in pure equilibrium situations. The residual autoregressive term can improve the admixture model by taking care of short-range isolation-by-distance. Note that the goal of the proposed algorithm differs from detecting isolation-by-distance. For an approach able to separate the effects of isolation-by-distance from migration and to give an estimate of the scale at which each process operate see (Bocquet-Appel and Sokal 1989).

Secondary contact hypothesis for the killifish. The killifish *Fundulus heteroclitus* has served as a model for understanding local adaptation to variable environments (Avise 2004). This species is known to exhibit latitudinal clinal variation in a number of physiological traits, and patterns at mitochondrial and nuclear DNA loci have suggested a complex history of spatially variable selection and secondary contact, with an abrupt genetic transition between northern and southern populations (Adams et al 2006). The spatial population structure analysis

inferred a cline that separated the northern and southern populations. Adams et al (2006) suggested that this cline was the result of recolonization of the whole current habitat from unfrozen water at the end of the last glacial age, creating a secondary contact zone between northern and southern populations. The best model did not include spatial autocorrelation effects. An explanation may be the use of population samples, which perhaps removed some local aspects of variation. We think that including spatial autocorrelation would have been more useful if individual sampling had been performed uniformly within the study area. A third cluster corresponded to the two southernmost samples of killifish. Because these two samples were geographically isolated from the rest, it was difficult to decide whether the smooth variation observed to the south of the area could be attributed to isolation-by-distance, i.e. an artificial cluster, or to historical patterns of migration. In any case, coupling Bayesian clustering methods with additional demographic analyses seems always necessary, as secondary contact and isolation-by-distance in an irregular sampling design might produce confounding signals.

Clines and clusters. The methods presented in this study have the potential to detect coexisting clines and clusters through the inferred variation of admixture proportions (see Rosenberg et al (2005) for a related discussion on clustering algorithms). This was emphasized by the analysis of simulations of range expansion from two refugia. In these spatially explicit simulations, the algorithm detected a contact zone at the same time as it found clusters in the north of Europe and elsewhere. In general it might be difficult to distinguish between clines and clusters without a good spatial coverage of the study area. In this case, a DIC analysis will provide an assessment of the relative contribution of clines and clusters to the posterior estimates of admixture coefficients. For example, a non-spatial analysis

for the killifish data suggested the existence of four clusters partitioning the southern cline, but a spatial analysis coupled to a DIC evaluation indicated that a cline merging two clusters better explained the data.

Comparisons with simpler methods. Relationships between Bayesian clustering algorithms and PCA have been emphasized by Patterson et al (2006) who considered a model of genetic structure in which populations have diverged from an ancestral population recently. If the model assumes K populations, PCA is then expected to have $K-1$ significant components under the Tracy-Widom theory (Patterson et al 2006). Applying PCA to the killifish, the cline and the southern genetic cluster were visible in the first and in the third eigenvectors (PC1 and PC3; Figure S7). In this example, the patterns found in PC1 and PC3 match those computed by the Bayesian clustering program. In simulations of recent admixture, the tests were significant for PC1 only, and this axis of variation clearly captured clinal variation at the contact zone. This was to be expected, because the informative panel F_{ST} was low and the theory could be expected to perform very well. In contrast, the Tracy-Widom theory yielded more than fifteen significant axes of variation ($p < .01$) in some simulations of contact zones in Europe (Figure S8). For these components, the genetic meaning was hard to interpret. This happened in situations where the informative panel F_{ST} was high (>.10), and the Tracy-Widom theory less valid. In this case, the Bayesian algorithm was more robust as it always detected no more than five clusters, and provided interpretable values for the admixture proportions. Nevertheless the first PCs always included the cline and clusters found by the Bayesian clustering algorithm, and we believe that the two methods are useful complementary exploratory tools.

Concluding remarks. Bayesian algorithms for inference of population structure have traditionally focussed on finding clusters, whereas less efforts have been

devoted to detecting clinal variation. To provide a better description of the relative contribution of clines and clusters, we coupled a multiscale spatial admixture analysis with a Bayesian assessment of model complexity and fit. This approach reduces the number of spurious cluster when the underlying variation is mainly clinal, while still detecting zones of small genetic discontinuities. Our new algorithm provides more accurate estimates of admixture proportions compared to standard non-spatial methods, and this suggests its use when studying spatial population structure, secondary contact zones, and when correcting for population structure in phenotype-genotype association studies.

Acknowledgments

We are grateful to Stephanie Adams for communicating the *Fundulus* data, and to Nicolas Ray for providing us a recent version of SPLATCHE. We also thank Michael GB Blum, Nick Patterson, Jonathan K Pritchard and an anonymous referee for their comments. Simulations were run on the the UJF-CIMENT cluster of computers (http://healthphy.grenoble.cnrs.fr/). OF was supported by grant ANR BLAN06-3-146282 MAEV.

4 Cited Literature

Adams SM, Lindmeier JB, Duvernell, DD (2006) Microsatellite analysis of the phylogeography, Pleistocene history and secondary contact hypotheses for the killifish, *Fundulus heteroclitus*. Mol. Ecol. 15:1109-1123.

Avise JC (2004) Molecular markers, natural history, and evolution, 2nd edn. Sinauer Associates Sunderland, Massachusetts.

Balloux F (2001) EASYPOP (version 1.7): A computer program for the simulation of population genetics. J. Heredity 92:301-302.

Barton NH, Gale KS (1993) Genetic analysis of hybrid zones. In: Harrison RG, editor. Hybrid zones and the evolutionary process. Oxford University Press, Oxford. p. 13-45.

Barton NH, Hewitt GM (1985) Analysis of hybrid zones. Ann. Rev. Ecol. Syst. 16:113-148.

Belle EMS, Landry P-A, Barbujani G (2006) Origins and evolution of the Europeans'genome: evidence from multiple microsatellite loci. Proc. R. Soc. B 273:1595-1602.

Bertorelle G, Excoffier L (1998) Inferring admixture proportions from molecular data. Mol. Biol. Evol. 15:1298-1311.

Besag, J (1975) Statistical analysis of non-lattice data. The Statistician 24:179-195.

Besag J, York J, Mollié A (1991) Bayesian image restoration with two applications in spatial statistics (with discussion). Ann. I. Stat. Math. 43:1-59.

Bocquet-Appel JP, Sokal RR (1989) Spatial autocorrelation analysis of trend residuals in biological data. Syst. Zool. 38(4):333-341.

Brooks SP (2002) Discussion on the paper by Spiegelhalter, Best, Carlin and van der Linde. J. Roy. Stat. Soc. B 64:616-639.

Cavalli-Sforza LL, Menozzi P, Piazza A (1994) The history and geography of human genes. Princeton University Press, Princeton.

Celeux G, Forbes F, Robert CP, Titterington DM (2006) Deviance Information Criteria for Missing Data Models. Bayesian Analysis 1:651-674.

Chakraborty R (1986) Gene admixture in human populations: models and predictions. Yearb. Phys. Anthropol. 29:1-43.

Chakraborty R, Weiss KM (1988) Admixture as a tool for finding linked genes and detecting that difference from allelic association between loci. Proc. Natl. Acad. Sci. USA 85:9119-9123.

Chen C, Durand E, Forbes F, François O (2007) Bayesian clustering algorithms ascertaining spatial population structure: A new computer program and a comparison study. Mol. Ecol. Notes 7:747-756.

Chikhi L, Bruford MW, Beaumont MA (2001) Estimation of admixture proportions: a likelihood-based approach using Markov chain Monte Carlo. Genetics 158:1347-1362.

Corander J, Sirén J, Arjas E (2008) Bayesian spatial modeling of genetic population structure. Computation. Stat. 23:111-129.

Cressie NAC (1993) Statistics for spatial data. Wiley, New York.

Currat M, Ray N, Excoffier L (2004) SPLATCHE: a program to simulate genetic diversity taking into account environmental heterogeneity. Mol. Ecol. Notes 4:139-142.

Currat M, Excoffier L (2005) The effect of the Neolithic expansion on European molecular diversity. Proc. R. Soc. B 272:679-688.

Dormann CF (2007) Effects of incorporating spatial autocorrelation into the analysis of species distribution data. Global Ecol. Biogeogr. 16:129-138

Dupanloup I, Bertorelle G, Chikhi L, Barbujani G (2004) Estimating the impact of prehistoric admixture on the Europeans' genome. Mol. Biol. Evol. 21:1361-1372.

Durrett R, Buttel L, Harrison R (2000) Spatial models for hybrid zones. Heredity 84:9-19.

Endler JA (1977) Geographic variation, speciation, and clines. Princeton University Press, Princeton N.J.

Epperson BK, Li T (1996) Measurement of genetic structure within populations using Moran's spatial autocorrelation statistics. Proc. Natl. Acad. Sci. USA 93:10528-10532

Excoffier L, Estoup A, Cornuet JM (2005) Bayesian analysis of an admixture model with mutations and arbitrary linked markers. Genetics 169:1727-1738.

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.

Faubet P, Gaggiotti OE (2008) A new Bayesian method to identify the environmental factors that influence recent migration. Genetics 178:1491-1504.

Foll M, Gaggiotti OE (2006) Identifying the environmental factors that determine the genetic structure of populations. Genetics 174:875-891.

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

François O, Blum MGB, Jakobsson M, Rosenberg NA (2008) Demographic history of European populations of *Arabidopsis thaliana*. PLoS Genet. 4(5): e1000075.

Gaggiotti OE, Brooks SP, Amos W, Harwoods J (2004) Combining demographic, environmental and genetic data to test hypotheses about colonization events in metapopulations. Mol. Ecol. 13:811-825.

Gao HS, Williamson S, Bustamante CD (2007) A Markov chain Monte Carlo approach for joint inference of population structure and inbreeding rates from multilocus genotype data. Genetics 176:1635-1651.

Gelman A, Carlin JB, Stern HS, Rubin DB (2004) Bayesian data analysis. Chapman and Hall/CRC Press, Boca Raton, Florida.

Goudet J (2005) HIERFSTAT, a package for R to compute and test hierarchical

F-statistics. Mol. Ecol. Notes 5:184-186.

Griebeler EM, Müller JC, Seitz A (2006) Spatial genetic patterns generated by two admixing genetic lineages: a simulation study. Conserv. Genet. 7:753-766.

Hewitt G (2000) The genetic legacy of the quaternary ice ages. Nature 405:907-913.

Hoggart CJ, Shriver MD, Kittles RA, Clayton DG, McKeigue PM (2004) Design and analysis of admixture mapping studies. Am. J. Hum. Genet. 74:965-978.

Jakobsson M, Rosenberg NA (2007) CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. Bioinformatics 23:1801-1806.

Jombart T, Devillard S, Dufour A-B, Pontier D (2008) Revealing cryptic spatial patterns in genetic variability by a new multivariate method. Heredity 101:92-103.

Lichstein JW, Simons TR, Shriner SA, Franzreb KE (2002) Spatial autocorrelation and autoregressive models in ecology. Ecol. Monogr. 72:445-463.

Long JC (1991) The genetic structure of admixed populations. Genetics 127: 417-428.

Meng XF, Shi M, Chen XX (2008) Population genetic structure of *Chilo suppressalis* (Walker) (Lepidoptera: Crambidae): strong subdivision in China inferred from microsatellite markers and mtDNA gene sequences. Mol. Ecol. 17: 2880-2897.

Mollié A (1996) Bayesian mapping of disease. In: Markov chain Monte Carlo in practice, Gilks WR, Richardson S, Spiegelhalter DJ, Chapman & Hall, London.

Pastorini J, Zaramody A, Curtis DJ, Nievergelt CM, Mundy NI (2009) Genetic analysis of hybridization and introgression between wild mongoose and brown lemurs. BMC Evol. Biol. 9(1):32.

Patterson N, Price AL, Reich D (2006) Population structure and eigenanalysis. PLoS Genet. 2:e190.

Power DA, Lauerman T, Crawford DL, DiMichele L (1991) Genetic mechanisms for adapting to a changing environment. Annu. Rev. Genet. 25:629-659.

Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. Genetics 155:945-959.

Pritchard JK, Wen X, Falush D (2003) Documentation for structure software: Version 2.3. Department of Human Genetics, University of Chicago.

Reich D, Patterson N (2005) Will admixture mapping work to find disease genes? Phil. Trans. R. Soc. B 360:1605-1607.

Ripley BD (1981) Spatial statistics. Wiley, New York.

Ripley BD (1996) Pattern recognition and neural networks. Cambridge University Press.

Roberts DF, Hiorns RW (1965) Methods of analysis of the genetic composition of a hybrid population. Hum. Biol 37:38-43.

Rosenberg NA, Saurabh S, Ramachandran S, Zhao C, Pritchard JK, Feldman MW (2005) Clines, clusters, and the effect of study design on the influence of human population structure. PLoS Genet. 1:660-671.

Sankararaman S, Kimmel G, Halperin E, Jordan MI (2008) On the inference of ancestries in admixed populations. Genome Res. 18:668-675.

Slatkin M, Arter HE (1991) Spatial autocorrelation methods in population genetics. Am. Nat. 138:499.

Smith MW, O'Brien SJ (2005) Mapping by admixture linkage disequilibrium: advances, limitations and guidelines. Nat. Rev. Genet. 6:623-632.

Sokal RR, Oden NL (1978) Spatial autocorrelation in biology. I. Methodology. Biol. J. Linn. Soc. 10:199-228.

Spiegelhalter DJ, Best NG, Carlin BP, van der Linde A (2002) Bayesian measures of model complexity and fit (with discussion). J. Roy. Stat. Soc. B 64:583-639.

Taberlet P, Fumafalli L, Wust-Saucy AG, Cosson J-F (1998) Comparative phylogeography and postglacial colonization routes in Europe. Mol. Ecol. 7:453-464.

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

Wang J (2003) Maximum-likelihood estimation of admixture proportions from genetic data. Genetics 164:747-765.

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.

Yannic G, Basset P, Hausser J (2008) Phylogeography and recolonization of the Swiss Alps by the Valais shrew (*Sorex antinorii*), inferred with autosomal and sex-specific markers. Mol. Ecol. 17:4118-4133.

Yu J, Pressoir G, Briggs WH, Vroh Bi I, Yamasaki M, Doebley JF, McMullen MD, Gaut BS, Nielsen DM, Holland JB, Kresovich S, Buckler ES (2006) A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nat. Genet. 38:203-208.

Figure 1. Expected admixture proportions q_x as a function of space under a longitudinal linear model, $\log(\alpha_x) = 25 - 5x$, in a simulation of the spatial prior distribution using 2 clusters. The results for the second cluster are symmetric with respect to the middle of the area.

Figure 2. Bayesian model choice for secondary contact scenarios. (A-B) All simulations (180 individuals, 10 loci). Data sets are labelled 1-8. The plots represent relative differences in DIC for 2 hidden regression models. The reference model is the model without covariate (non-spatial model, dashed lines). (C) DIC as a function of $K_{\max}$ for one simulated data set ($T = 2,000$, $m = .9$, $C = 1000$) for 3 models. (D) DIC as a function of $K_{\max}$ for one simulated data set ($T = 1,000$, $m = .3$, $C = 1000$) for 3 models.

Figure 3. Posterior estimates of admixture proportions and predictive maps for selected models. (A-B) Range expansion from 2 refugia ($T = 1,000$, $m = .3$, $C = 100$, 1,200 individuals, $L = 100$ loci). These results are representative of a majority of the data sets. The contact zone is highlighted in red pixels. (C-D) *Fundulus heteroclitus.* In (C), the individuals are sorted by latitude. The cline at latitude $40°41.2'$ (red pixels, black line) corresponds to the observation of Adams et al (2006).

Figure 4. DIC as a function of $K_{\max}$ for *F. heteroclitus*. The vertical dashed line corresponds to estimated effective number clusters $K = 3$ obtained from the linear trend model.

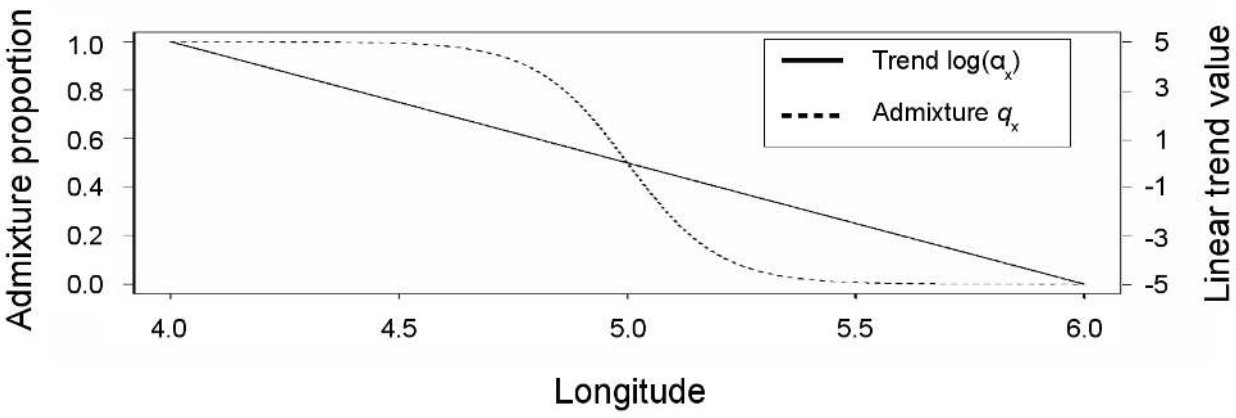
Admixture proportion
1.0
0.8
0.6
0.4
0.2
0.0
4.0
4.5
5.0
5.5
6.0
Longitude
Trend log(α_x)
Admixture q_x
5
3
1
-1
-3
-5
Linear trend value

Figure 1.

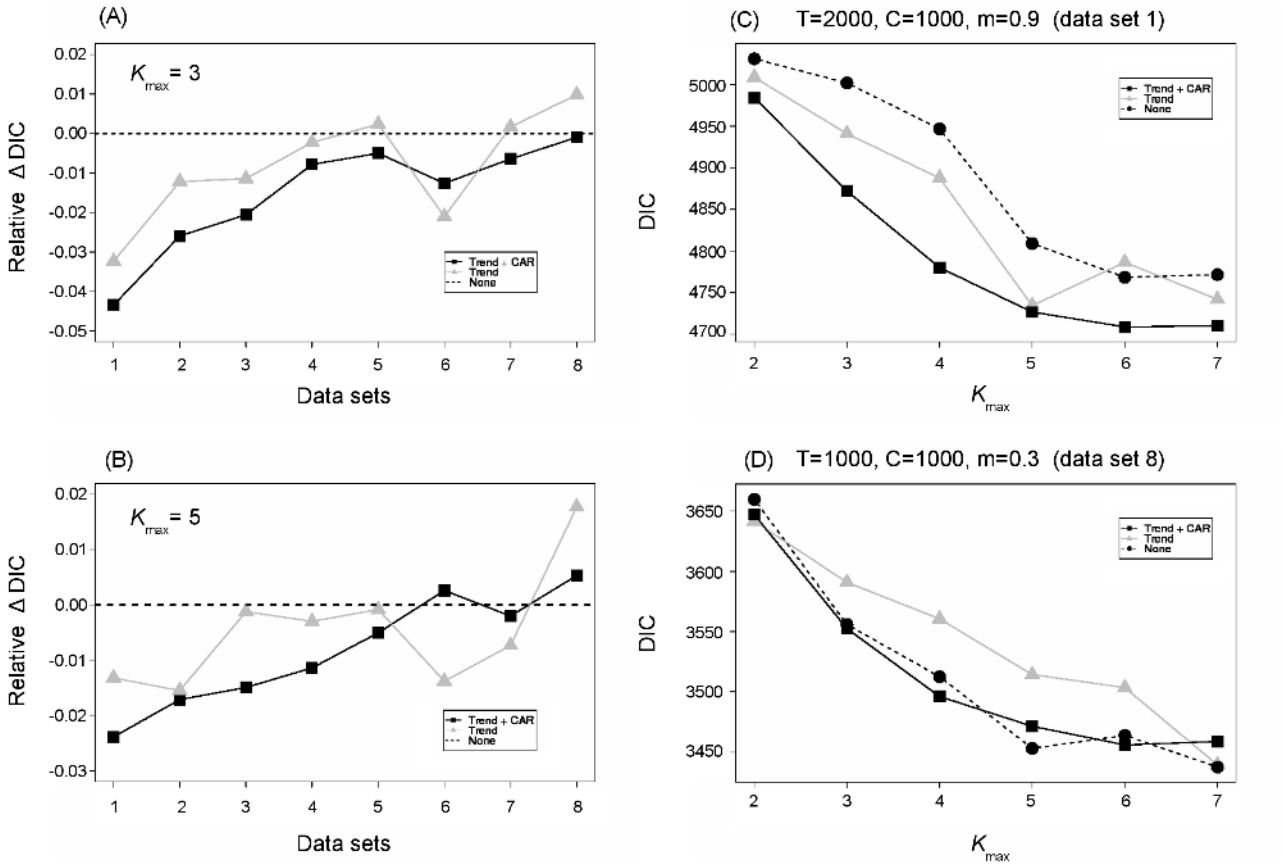
(A)
K_{max} = 3
Relative Δ DIC
Data sets
Trend + CAR
Trend
None
(B)
K_{max} = 5
Relative Δ DIC
Data sets
Trend + CAR
Trend
None
(C) T=2000, C=1000, m=0.9 (data set 1)
DIC
K_{max}
Trend + CAR
Trend
None
(D) T=1000, C=1000, m=0.3 (data set 8)
DIC
K_{max}
Trend + CAR
Trend
None

Figure 2.

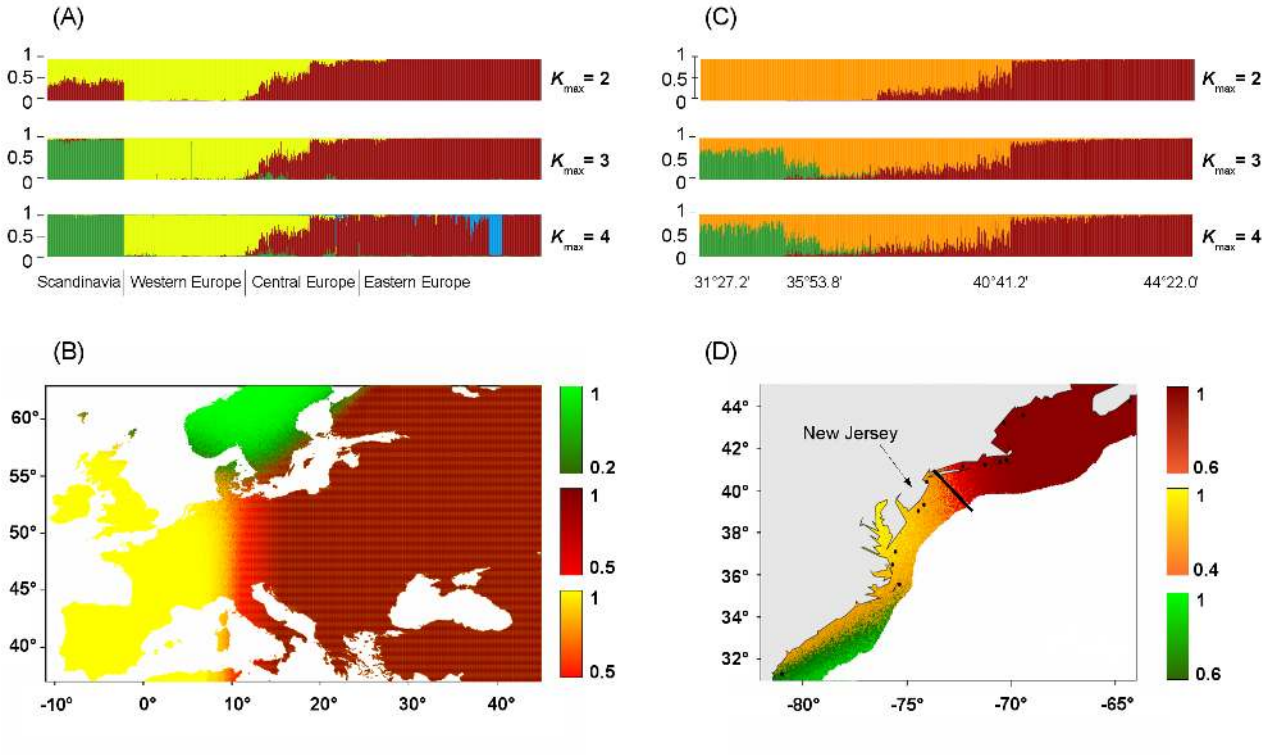
(A)
1
0.5
0
K_{max} = 2
K_{max} = 3
K_{max} = 4
Scandinavia
Western Europe
Central Europe
Eastern Europe
(C)
31°27.2'
35°53.8'
40°41.2'
44°22.0'
(B)
60°
55°
50°
45°
40°
-10°
0°
10°
20°
30°
40°
1
0.2
0.5
(D)
New Jersey
44°
42°
40°
38°
36°
34°
32°
-80°
-75°
-70°
-65°
0.6
0.4

Figure 3.

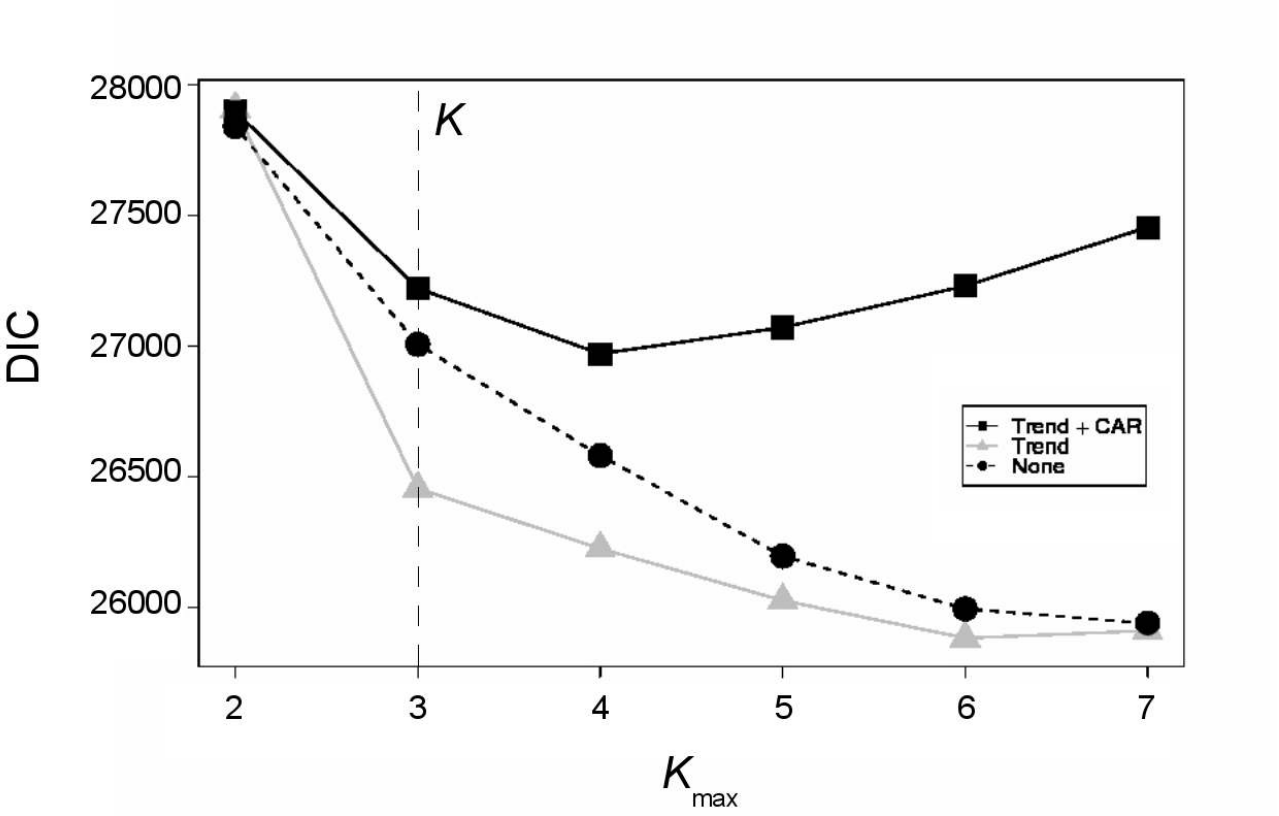
28000
27500
27000
26500
26000
DIC
K
Trend + CAR
Trend
None
2
3
4
5
6
7
K max

Figure 4.