

1 A computer program to simulate multilocus genotype data with
2 spatially auto-correlated allele frequencies

3 Gilles Guillot*and Filipe Santos†

4 October 23, 2008

Abstract

5 Many models for inference of population genetics parameters are based on the assumption
6 that the data set at hand consists of groups displaying within-group Hardy-Weinberg equilibrium
7 at individual loci and linkage equilibrium between loci. This assumption is commonly violated
8 by the presence of within-group spatial structure arising from non-random mating of individuals
9 due to isolation by distance (IBD).

10
11 This paper proposes a model and simulation method implemented in a computer program
12 to flexibly simulate data displaying such patterns. The program permits displaying of smooth
13 spatial variations of allele frequencies due to IBD and more abrupt variations due to presence of
14 strong barriers to gene flow. It is useful in assessing performance of various statistical inference
15 methods and in designing spatial sampling schemes. This is shown by a simulation study aimed
16 at assessing the extent to which IBD patterns affects accuracy of cluster inferences performed in

*Centre for Ecological and Evolutionary Synthesis Department of Biology, University of Oslo, P.O. Box 1066 Blindern, N-0316 Oslo Norway. To whom correspondence should be addressed.

†Centre de Biologie et de Gestion des Populations, INRA / IRD / CIRAD / Montpellier SupAgro, Campus international de Baillarguet, CS 30016, F-34988 Montferrier-sur-Lez cedex, France

17 models assuming panmixia. The program is also used to study the effects of spatial sampling
18 scheme (e.g. sampling individuals in clumps or uniformly across the spatial domain). The
19 accuracy of such inferences is assessed in terms of number of inferred populations, assignment
20 of individuals to populations and location of borders between populations. The effect of spatial
21 sampling was weak while the effect of IBD may be substantial, leading to the inference of spurious
22 populations, especially when IBD was strong with respect to the size of the sampling domain.
23 The model and program are new and have been embedded in the R package Geneland, for user
24 convenience and compliance with existing data formats.
25 See: folk.uio.no/gillesg/Geneland/Geneland.html.
26 More on simulation study: folk.uio.no/gillesg/MER

27 1 Background

28 Panmictic reproduction is a necessary condition for Hardy-Weinberg equilibrium and linkage
29 equilibrium (HWLE) to arise when studying organisms over a large spatial area. When the sam-
30 pling scale within a continuous population is large compared to the dispersal capability of the
31 species, it is less likely that mixing will occur between individuals that are situated further apart
32 than among those that are separated by short distances (Wright, 1943; Rousset, 1997, 2000).
33 This leads to a departure from HWE but also from LE (Epperson, 1995; Ostrowski et al., 2006).
34 Hence, even for organisms where the assumption of random mating is reasonable on a local scale,
35 these assumptions become more questionable as the spatial scale of the study domain increases.
36 If a data set displaying this type of pattern is analyzed with a model assuming HWE and LE
37 within groups, it can be expected that the departure from these assumptions will result in various
38 biases. For instance, when using clustering software such as Structure (Pritchard et al., 2000)
39 or Geneland (Guillot et al., 2005), the presence of IBD may lead to the algorithms splitting the
40 data set into several groups in such a way that HWE and LE will be approximately fulfilled,
41 even though none of the inferred groups correspond to genuine HWLE groups. This bias might
42 be expected to be stronger particularly in the case of spatially irregular sampled data set (i.e.
43 spatially-clumped sampling sites), a common practice in molecular ecology studies. Potential
44 problems due to IBD are not specific to models based on spatial data and are likely to happen
45 with any clustering model assuming HWLE within populations, see (Pritchard et al., 2000; Falush
46 et al., 2003; Serre and Pääbo, 2004; Rosenberg et al., 2005).

47

48 The aim of this paper is first to describe a statistical model for simulation of population
49 genetics data displaying spatial auto-correlation of allele frequencies as it arises in case of IBD.
50 This will be followed by an illustration of the usefulness of this program by assessing the effect of
51 IBD patterns on the accuracy of two clustering algorithms (spatial and non-spatial) that assume
52 within-group HWLE.

53 **2 A statistical model for simulation of spatially structured IBD** 54 **populations**

55 **2.1 Overview of proposed model**

56 The goal of this model is to mimic the features of allele frequencies and genotypes of data sets
57 displaying a structure arising from the presence of IBD and barriers to gene flow. This model
58 brings together the notions of cline of allele frequencies and of spatially-structured population
59 domains following Wasser et al. (2004) and Guillot et al. (2005) respectively.

60

61 It is assumed that the overall population under study consists of several groups separated by
62 barriers to gene flow. These barriers can be known, visible, landscape features such as rivers,
63 roads or urban areas which have quantifiable effects, see (Coulon et al., 2006), or less obvious
64 barriers such as those induced by climate conditions (Stenseth et al., 2004). In any case, it makes
65 sense to assume that these barriers can be represented by rather simple shapes. This has been
66 accounted for by assuming that each group belongs to a spatial domain with boundaries that can
67 be represented by a small number of polygonal-shaped areas. See (Guillot et al., 2005) for details

68 and graphical examples.

69

70 Now, in contrast with a common assumption made by most population genetic clustering
71 software, it is assumed here that, because of isolation by distance, there is a departure from
72 HWLE within each group defined by such boundaries. Consider two individuals belonging to
73 the same group and sampled at sites s and s' respectively. Under HWLE, it is assumed that the
74 genotypes at a given bi-allelic locus for these individuals have been sampled from the same allele
75 frequency f . Our model assumes that this allele frequency varies across space and hence, that the
76 genotypes have been sampled from frequencies $f(s)$ and $f(s')$ respectively, are different, though
77 correlated. Frequencies are therefore modeled as correlated random variables varying across space.

78

79 In statistical terms, this IBD model assumes that allele frequencies are spatially auto-correlated
80 random functions. Their spatial variability can be described through the so-called covariance
81 function defined as a function of two spatial coordinates $C(s, s') = Cov[f(s), f(s')]$. The pro-
82 posed program offers a wide range of possibilities for the form of this covariance function.

83 **2.2 Formal statistical modeling**

84 Throughout this paper, the spatial domain under study is considered to be occupied by K pop-
85 ulations. Defining what is meant by populations and their spatial boundaries is one of the goals
86 of this section. It is assumed that there have been observations of n individuals, located at sites
87 s_1, \dots, s_n with genotypes at L loci z_1, \dots, z_n , where $z_i = (z_i^{l,1}, z_i^{l,2})$, $l = 1, \dots, L$, the exponents $l, 1$
88 and $l, 2$ referring to the allele 1 and 2 harbored by a diploid individual at locus l .

89

90 To begin with, it is assumed that the number of populations present in the domain is one
91 only ($K = 1$). A model has been introduced parameterizing spatial patterns of genotypes. For
92 the sake of clarity, start with genotypes at a single bi-allelic locus and denote possibly observed
93 genotypes by $\{a, a\}$, $\{a, A\}$ and $\{A, A\}$. The kind of spatial patterns observed under IBD is
94 typically that of Fig. 1, where an obvious trend occurs from left to right both in term of allele
95 frequencies and proportion of homozygous $\{a, a\}$.

96

97 If one computes local empirical allele frequencies (e.g. by computing frequencies of allele a ,
98 denoted by $f(s)$ within the circle of a small radius r centered at site s for various s), there is
99 a tendency to observe smooth increases of $f(s)$ from left to right (Fig. 1 top). One could also
100 check (by any appropriate statistical test) that the allele a around site s has been drawn with
101 probability $f(s)$ and allele A with probability $1 - f(s)$. Further, one could also check that geno-
102 types $\{a, a\}$, $\{a, A\}$ and $\{A, A\}$ occur with probabilities $f^2(s)$, $2f(s)(1 - f(s))$ and $(1 - f(s))^2$
103 respectively. In other words, in the simulated data set presented in Fig. 1, around any site s ,
104 there is a 'local' population whose genes a and A are present in proportions $f(s)$ and $1 - f(s)$
105 respectively, and this local population is at HWE.

106

107 The previous remarks actually describe how the data set in Fig. 1 has been simulated: two
108 surfaces $f_A(s)$ and $f_a(s)$ where built in such a way that they were random, add up to one at any
109 site and display some spatial auto-correlation in such a way that they mimic variations in space

110 of local allele frequencies which could be observed under IBD. The surface $f_a(s)$ is displayed in
111 Fig. 1 (top). Then at 200 random sites, the genotypes of individuals were drawn according to
112 the local frequencies $f_A(s)$ and $f_a(s)$, and obtaining the genotypes given at Fig. 1.

113

114 This model of spatial variations is now defined more formally for the field of allele frequencies.
115 Work continues with only one population but now in a more general setting with L independent
116 loci with any arbitrary number of alleles J_l at each locus l . It is assumed that at any site s ,
117 allele j at locus l is present in proportion $f_{lj}(s)$. It is also assumed that for each locus l , the
118 set of J_l surfaces $f_{l1}(s), \dots, f_{lJ_l}(s)$ was obtained by transformation of some hypothetical random
119 surfaces $g_{l1}(s), \dots, g_{lJ_l}(s)$ such that (i) the random variables $g_{lj}(s)$ have Normal(0, 1) distribution,
120 (ii) at any site s , and any $l = 1, \dots, L$, the variables $(g_{lj})_{j=1, \dots, J_l}(s)$, are mutually independent,
121 (iii) at any pair of sites (s, s') the random variables $(g_{lj}(s), g_{lj}(s'))$ have a correlation equal to one
122 whenever $s = s'$ and decreasing with the distance between s and s' . These surfaces are random
123 but they mimic variations in space of local allele frequencies which could be observed under IBD.

124

125 More specifically, it is assumed that the set of surfaces $g_{lj}(s)$, are mutually independent Gaus-
126 sian random fields (Adler, 1984). The $g_{lj}(s)$ surfaces are not necessarily above zero, nor do they
127 fulfill $\sum_j^{J_l} g_{lj}(s) = 1$. Thus, with a positive real parameter α , the g_{lj} values are transformed
128 into point-wise Gamma(1, α) distributed random surfaces and then re-normalized so as to acquire
129 values summing up to one at each site and for each locus. The resulting surfaces $f_{lj}(s)$ are now
130 positive and sum up to one at any site s and could then be interpreted as local allele frequencies

131 for some locus.

132

133 Note that re-normalization of the independent Gamma distributed random surfaces results in
134 surfaces with Dirichlet distribution at any site. Because it is assumed that no statistical linkage
135 exists between loci, the variations of allele frequencies for each locus across space can be described
136 by a set of such random fields $(f_{l1}(s), \dots, f_{lJ}(s))$ for any locus l with independence across loci.

137

138 The interest of Gaussian random fields lies in the fact that their spatial variations can be flexi-
139 bly and parsimoniously parametrized through their covariance function (Chilès and Delfiner, 1999;
140 Lantuéjoul, 2002). This function for a random field $g(s)$ is defined as $C(s, s') = Cov[g(s), g(s')]$.
141 If the quality of being stationary is assumed (namely invariance under spatial translation of
142 statistical distributions), then $C(s, s')$ actually depends only on the lag $s - s'$ (length and di-
143 rection). If isotropy is assumed (no directional effect), $C(s, s')$ eventually depends only on the
144 length $|s - s'|$. Many suitable parametric functions C have been introduced in the geostatistical
145 literature to describe observed variations of environmental variables (and also to comply with a
146 mathematical requirement known as positive-definiteness). Detailed lists of such function can be
147 found in (Wackernagel, 1995) or (Schlather, 1999).

148

149 In the present work, it is assumed that all Gaussian fields involved have the covariance function
150 sometimes referred to as the stable model and defined by

$$C(s, s') = \exp(-(|s - s'|/\beta)^\gamma) \tag{1}$$

151 The parameter β prescribes the rate at which $C(s, s')$ decreases as $|s - s'|$ increases, while
 152 γ describes the behavior of C at small lags, hence the smoothness of the underlying process.
 153 See (Lantuéjoul, 2002) for illuminating graphical examples. This model shares some features
 154 with that of Wasser et al. (2004). However, our model departs from the latter in the following
 155 way: the Gaussian fields are transformed into Gamma distributed random variables, resulting in
 156 Dirichlet distributed vectors of frequencies. This model thus generalizes the classical non-spatial
 157 Dirichlet assumption for allele frequencies into a spatial context. Thus, this model is locally in
 158 good agreement with the Wright-Fisher theory, see (Tavaré and Zeitouni, 2001), and potentially
 159 so with real data.

160

161 The parameter γ prescribes the regularity properties of the realizations of the random field.
 162 Although a detailed analysis of the meaning and implications of this parameter has yet to be
 163 carried out, it corresponds to subtle local properties of the variations of allele frequencies across
 164 space and can be set to a fixed value as an initial benchmark.

165

166 The fixed value $\gamma=1.5$ is set everywhere in the simulations reported below. The value 1.5
 167 is a trade-off between the classical exponential model ($\gamma = 1$) with very irregular realizations
 168 and the so-called Gaussian model ($\gamma = 2$) with excessively regular realizations. In this case,
 169 the transformation of g values into Gamma($1, \alpha$) distributed values generates Dirichlet(α, \dots, α)

170 distribution for allele frequencies. Although simulations with any arbitrary values for α are
171 possible, for the sake of simplicity we set it to 1 in our simulations . These parameter values
172 approximate real data where something close to HWE is expected locally (on a small scale) and
173 smooth variations of allele frequencies are observed on a larger scale.

174 [Figure 1 about here.]

175 In addition to IBD leading to smooth spatial variation of allele frequencies, groups of individ-
176 uals may also be genetically isolated by barriers to gene flow (i.e. environmental features such as
177 mountains, rivers, roads, deforested areas), leading to abrupt change in allele frequencies. These
178 will be hereafter considered as spatial borders between populations. Because one major aim of
179 landscape genetic analysis is to identify and locate spatially such borders, this paper will now
180 consider a model accounting for both IBD and barriers to gene flow.

181

182 For the modeling of the spatial repartition of IBD populations, a model is used that was
183 previously introduced in population genetics by Guillot et al. (2005) who parameterized barriers
184 to dispersal as borders of polygonal regions. More precisely, Guillot et al. (2005) it is assumed
185 that there is a hidden Poisson process generating a set of Voronoi polygons with mutually in-
186 dependent population membership. The union of polygons with the same color forms a spatial
187 domain standing for the domain of one population. See Guillot et al. (2005) for more details and
188 illustrating examples.

189

190 We now embed the smooth variation in allele frequencies typical of IBD continuous popula-
191 tions within Guillot et al.'s (2005) model of polygonal territories. It is assumed that K IBD
192 populations are spread over the spatial domain. The respective territories of these populations
193 are unions of some Voronoi polygons induced by a hidden Poisson process, while each of popu-
194 lation k has allele frequencies varying smoothly as described previously. Since the independence
195 of allele frequencies between different populations is assumed, the surface of allele frequencies
196 displays discontinuities along their borders as shown on Fig. 2. See supplementary material and
197 program documentation for further details.

198 [Figure 2 about here.]

199 **3 Simulation study**

200 **3.1 Overview of simulation study**

201 The model presented above is used to assess the extent to which the presence of IBD patterns
202 affects inference of boundaries between groups in clustering models assuming HWLE. The accu-
203 racy of inferences is assessed in terms of number of groups and individual memberships to these
204 groups. Inferences were carried out with Geneland under the uncorrelated frequency model and
205 using the spatial option and the non-spatial option. Although the goal is to report results that
206 are related to models rather than to software, it is worth mentioning that the non-spatial model
207 of Geneland assumes a non-structured prior probability distribution for the individual popula-
208 tion memberships. Hence, it is similar to the model implemented in all non-spatial Bayesian
209 programs such as Structure (Pritchard et al., 2000), Partition (Dawson and Belkhir, 2001) or the

210 non-spatial version of BAPS (Corander et al., 2003).

211

212 In the assessment of performance of the spatial algorithm, inferences were also carried out
213 on data sets where the individuals are spatially collected at a small number of bunches across
214 space. In this case, the center of bunches are sampled uniformly and independently on $[0, 1] \times$
215 $[0, 1]$ and the individuals within each bunch are sampled from a centered bi-variate Gaussian
216 distribution with a standard deviation of 0.01 (see supplementary material and figure 3 therein for
217 details and examples). For each simulation condition, 500 independent simulation replicates were
218 performed with spatially-regular sampling along with 500 simulation replicates with spatially-
219 irregular sampling.

220 **3.2 Detail on simulated data-sets.**

221 All the simulations described hereafter, except those for graphic examples, are located on a
222 squared domain of dimension $[0, 1] \times [0, 1]$. Unless otherwise specified, the genotypes are simulated
223 for 200 individuals at 10 loci with 10 alleles per locus. This corresponds to most commonly-
224 published empirical microsatellite datasets. The individuals are located either regularly (sampled
225 uniformly and independently on $[0, 1] \times [0, 1]$) or as 10 bunches of 20 individuals. In this case, the
226 center of bunches are sampled uniformly and independently on $[0, 1] \times [0, 1]$ and the individuals
227 within each bunch are sampled from a centred bi-variate Gaussian distribution with a standard
228 deviation of 0.01 (Fig. 3). For each simulation condition, a set of 500 simulation replicates was
229 drawn with spatially-regular sampling along with a set of 500 simulation replicates with spatially-
230 irregular sampling.

231

[Figure 3 about here.]

232

233

234

235

Unless otherwise specified, the number of populations K is drawn from a uniform distribution on $\{1, 2, 3, 4\}$. Because the aim here is to study spatially structured data, there is a simulation of data sets where the number of polygons m in the hidden tessellation giving the population membership is sampled from a uniform distribution on $\{1, \dots, 15\}$.

236

237

238

239

240

241

242

It is important to note that in both the regular and irregular spatial sampling designs the number of individuals per population is random, since the number depends on the spatial location of borders between populations. In addition, the total number of individuals sampled is set to 200 whatever the value of K . These two features make this simulation scheme more realistic than the one previously used in Guillot et al. (2005) where the total number of simulated individuals was proportional to the number of populations.

243

244

245

246

247

248

249

As a reference data-set, simulations are drawn from the model where the groups are panmictic. The allele frequencies are sampled from independent Dirichlet $\mathcal{D}(1, \dots, 1)$ distributions. By processing in this way, population genetic differentiation as measured by pairwise F_{ST} (Weir and Cockerham 1984) has approximately a symmetric empirical distribution with quartiles values equal to 0.04 and 0.16, respectively and a mode of 0.095. Hence, it spans a broad range of F_{ST} values from weak to marked levels of genetic differentiation.

250

251 Files from the model were simulated where the groups display IBD patterns. The scale pa-
 252 rameter β controlling the intensity of IBD was sampled from a uniform distribution in $[0, 5]$.
 253 Another set of files with $K = 2$ for all files was also analyzed to test whether accuracy increased
 254 with the level of genetic differentiation within and among populations. The parameter β governs
 255 the rate at which the correlation between allele frequencies decreases with distance and can be
 256 considered as a distance beyond which the correlations between allele frequencies become weak
 257 (see discussion section on how β could be roughly estimated in practice).

258

259 In order to relate β to more meaningful measures in population genetics, a statistic was in-
 260 troduced that quantifies the level of genetic differentiation between pairs of individuals located
 261 within a certain distance. Such a statistic already exists (i.e. the statistic a_r introduced by
 262 Rousset (1997)). However, being based on allele identity, it has a large variance. Since these
 263 simulations involve the intermediate simulation of allele frequencies at any site of the study do-
 264 main, a new statistic is proposed here based on local allele frequencies, according to the following
 265 formula:

$$D = \frac{1}{n_\varepsilon} \sum_{\substack{s, s' \\ 0 < d(s, s') \leq \varepsilon}} \frac{1}{L} \sum_l \frac{1}{J_l} \sum_j |f_{slj} - f_{s'lj}| \quad (2)$$

266 where f_{slj} denotes the allele frequency of allele j at locus l at site s , $d(s, s')$ denotes the
 267 geographical distance between sites s and s' , ε is a small arbitrary distance (set to 0.1 in com-
 268 putations on the unit square) and n_ε denotes the number of pairs of distinct sampled sites such
 269 that $d(s, s') \leq \varepsilon$. This statistic can be computed either for pairs of individuals located in the

270 same population or for pairs of individuals located in different populations. This leads either to
271 a statistic of local differentiation within populations (D_W) or a statistic of local differentiation
272 between populations across the population borders (D_B). The inverse of D_W relates closely to
273 the scale parameter β , as can be seen from Fig. 4 (middle panel).

274

275 The statistic D_B can be viewed as a local statistic in the sense that it involves pairs of indi-
276 viduals whose mutual distance is small with respect to the size of the domain. This is preferable
277 to a more traditional measure of pairwise population differentiation such as F_{st} (Weir and Cock-
278 erham, 1984) because the latter measure is not spatially explicit as the computation is based on
279 individuals located all over the population domain, regardless of their spatial coordinates. F_{st}
280 therefore tends to be low in the presence of IBD as soon as the sampled area becomes large
281 enough with respect to the scale of variation of allele frequencies. This is illustrated on the left
282 panel of Fig. 4 which shows that F_{st} values decrease considerably when the strength of IBD
283 increases (i.e. when β values decrease). In contrast, the right panel of Fig. 4 shows that there
284 is no dependence between the scale parameter β and the between-population differentiation D_B ,
285 as could be expected since the allele frequencies are sampled independently across populations
286 in this model. This property makes D_B a more suitable statistic than F_{st} in quantifying the
287 intensity (or lack of permeability) of a barrier to gene flow between two IBD populations.

288

[Figure 4 about here.]

289 3.3 Detail on method for inference

290 For each simulated data-set, both the number of populations K and the population memberships
291 for each individual were inferred using the spatial algorithm implemented in Geneland. Although
292 these simulations display a relatively simple spatial organization (populations simulated had spa-
293 tial domains delimited by the same number m of polygons $m = 15$), this information was not
294 used and the maximum rate of the Poisson process was set to $\lambda_{max} = 50$, which amounts to the
295 assumption of a complex spatial organization.

296

297 Inference on all data-sets was also carried out using a non-spatial model. In this option of
298 Geneland, the spatial coordinates are not used and the inference on K and population member-
299 ships is carried out on the basis of individual genotypes only. This makes the fixed K component
300 of this inference algorithm similar to that of the software Structure (Pritchard et al., 2000) with
301 the no-admixture, uncorrelated frequencies and fixed alpha options. Note that in this setting,
302 the determination of the spatial location of borders between populations is not possible.

303

304 MCMC computations include 50000 iterations with a burn-in of 25000 iterations and a thin-
305 ning of 50 iterations. The minimum and maximum numbers of populations considered in the
306 first run were $K_{min}=1$ and $K_{max}=10$, respectively. Using larger values of K_{max} did not affect
307 the result of the inferences (results not shown).

308

309 The accuracy in the inference of K was assessed by computing the proportions of runs for

310 which $\hat{K} \neq K$ and $\hat{K} > K$, respectively. The accuracy in terms of inference of population
 311 membership is assessed through the error rate in co-assignment defined as:

$$\text{ERCA} = \frac{1}{n(n-1)/2} \sum_{i \neq j}^n I_{\{x_{ij} \neq \hat{x}_{ij}\}} \quad (3)$$

312 where x_{ij} (resp. \hat{x}_{ij}) is the true (resp. estimated) clustering matrix (i.e. $x_{ij} = 1$ whenever
 313 individuals i and j belong to the same population, 0 otherwise.) This computes the proportion of
 314 pairs of individuals belonging to the same population that were not correctly co-assigned by the
 315 inference algorithm. It has the advantage over a statistic referring to single individuals in that it
 316 is insensitive to the labeling of populations. ERCA can be computed either for assessing whether
 317 individuals are correctly assigned to their population (i.e. individual population membership)
 318 or for assessing whether borders between populations are correctly located (i.e. pixel population
 319 membership). In the later case, numerical computations are processed using the 50×50 pixels
 320 produced when discretizing the spatial domains. The statistic for individual assignment and
 321 border location are thereafter referred to as ERCA_i and ERCA_p , respectively. Both ERCA
 322 statistics range between 0 and 1. The manner by which a particular value of ERCA should
 323 be interpreted actually depends on the true and inferred numbers of population since the value
 324 particularly tends to be lower for large K than for small K . Hence, ERCA values will not be
 325 interpreted as absolute values but will be used for comparing various simulation scenarios.

326 **4 Results and discussion**

327 **4.1 Sensitivity of clustering algorithms to spatial sampling scheme**

328 In the absence of IBD patterns, the use of data where individuals are spatially sampled in an irreg-
329 ular way does not significantly affect the accuracy of inferences in terms of number of populations
330 and assignment of individuals (see Table 1).

331 [Table 1 about here.]

332 The use of irregular spatial sampling entails a decrease of the accuracy for the exact spatial
333 location of the borders between populations ($ERCA_p$ increases from 4.97 to 20.7 for an irregular
334 spatial sampling). Additional simulations indicate that, at first sight, the use of more irregular
335 samples (e.g. 5 bunches of 40 individuals) significantly decreases the accuracy of inferences on
336 K , which appears to be underestimated (results not shown), simply because with such data sets
337 some genuine populations are simply not present. Even in this case, the algorithm still remains
338 reliable in terms of estimating the number of populations actually sampled as well as in terms of
339 individual assignment in the populations present in the sample, while giving maps with a poor
340 resolution (see supplementary material).

341 **4.2 Sensitivity of clustering algorithms to IBD patterns**

342 IBD had a clear effect, strongly decreasing the accuracy of inferences of the number of popu-
343 lations, the individual population memberships and the spatial localization of borders between
344 populations. For 72.2% of runs, the true K is still properly estimated but there are 27.4% of
345 runs where K is over-estimated while under-estimation occurs very seldomly (0.4% with spatially

346 regularly-sampled data and never with spatially irregularly-sampled data). Hence, the effect of
347 IBD on inference is globally substantial. The sampling design (spatially regular or irregular) does
348 not seem to influence strongly this result.

349

350 As long as the strength of IBD remains weak, the algorithm seeking HWLE populations per-
351 forms well (low error on K and ERCA, see supplementary material), whereas beyond a value of
352 β around 2, the algorithm starts to be perturbed. This occurs always as an over-estimation of
353 K which is consistent with the underlying assumptions of the algorithm: it seeks populations at
354 HWE and LE and hence tends to split some populations into sub-populations if this condition is
355 not fulfilled.

356

357 IBD also tends to bring about spurious populations (and spurious borders between them) even
358 in case of use of spatially regular sampling. This result raises an important question when faced
359 with inferred structures from a real data set where the presence of IBD is suspected: Are the
360 inferred structures genuine or artifacts of the algorithm? Ideally, one would know the value of β
361 or else know the value of population density and dispersal parameters that govern the magnitude
362 of IBD, and check whether these values correspond to a safe situation (IBD of limited magnitude).

363

364 In practice, these parameters are unknown and estimating them is a challenging statistical
365 issue in cases where variations of allele frequencies across space result from both the effect of
366 distance (IBD, hence smooth variations) and the effect of barriers (very limited gene flow, hence

367 more abrupt variations). This is the case considered here and it is believed to be commonly
368 encountered in practice. In this case, identifiability issues (cline being wrongly interpreted as
369 effect of barriers or vice-versa) can be expected.

370

371 In the case where IBD is suspected and the presence of several groups has been inferred, an
372 ad-hoc but informative method has been proposed and implemented by Fontaine et al. (2007). It
373 consists in plotting genetic distances against spatial distances and labeling each pair (by a color)
374 depending on whether individuals in each pair belong to the same inferred group or to different
375 groups. This method can reveal whether the differentiation encountered is due to distance only
376 or to other factors not related to distance. However, this method does not permit the inference
377 of any of the above-mentioned parameters. This stresses the need for new inference tools in this
378 context.

379 **4.3 Directions for future research**

380 Within each barrier-defined group, this model complies with more traditional biologically explicit
381 IBD models and also with empirical observations in the sense that (i) allele frequencies display
382 spatial auto-correlation (Hardy and Wekemans, 1999; Wagner et al., 2005), (ii) the variations
383 in allele frequencies create an excess of homozygous, compared to the expected proportion under
384 HWE, and (iii) the variations in allele frequencies also create identity disequilibrium (as shown
385 e.g. by Epperson (1995)).

386

387 In biologically explicit IBD models such as the classical stepping stone model, the spatial

388 variability of allele frequencies depend on local population density, dispersal, as well as past
389 population history. In this model, this is parametrized through the covariance function of the
390 allele frequencies. This function depends on three parameters: a variance parameter α , a scale
391 parameter β and a smoothness parameter γ . See supplementary material and (Schlather, 1999)
392 or (Lantuéjoul, 2002) for details.

393

394 It is expected that the variance parameter relates to allele diversity, the scale parameter
395 relates to population density and dispersal and the smoothness parameter can be viewed as a
396 nuisance parameter (as implicitly done by Wasser et al. (2004)). It could also be related to local
397 landscape properties such as friction. There is a clear need of further work to relate these three
398 statistical parameters in a precise quantitative manner to biological parameters.

399

400 However, the simulation program described in the present paper should be still useful in
401 helping in the design of spatial sampling or in assessing the effect of IBD on different inference
402 problems in cases where the accuracy of results do not depend on the choice of parameters (as
403 in the case in this study of the effect of the spatial sampling scheme), or in the case, considered
404 by Novembre and Stephens (2008), where only a qualitative assessment is required.

405

406 **Acknowledgment:** *This work benefited from discussions with M. Beaumont, J.M. Cornuet,*
407 *A. Estoup, J.M. Marin, J. Novembre, C.P. Robert, F Rousset and A. le Rouzic. We also thank*
408 *J.F. Cosson, R. Leblois, F. Mortier and G. Roderick for comments on an earlier version of this*

409 *paper and the Associate Editor for various helpful comments. This work was partially supported*
410 *by ANR grant No NT05-4-42230.*

411 **References**

- 412 R.J. Adler. *The geometry of random fields*. Series in Probability and Mathematical Statistics.
413 Wiley, 1984.
- 414 J.P. Chilès and P. Delfiner. *Geostatistics*. Wiley, 1999.
- 415 J.C. Corander, P. Waldmann, and M.J. Sillanpää. Bayesian analysis of genetic differentiation
416 between populations. *Genetics*, 163:367–374, 2003.
- 417 A. Coulon, G. Guillot, J.F. Cosson, J.M.A. Angibault, S. Aulagnier, B. Cargnelutti, M. Galan,
418 and A.J.M Hewison. Genetics structure is influenced by landscape features. Empirical evidence
419 from a roe deer population. *Molecular Ecology*, 15:1669–1679, 2006.
- 420 K.J. Dawson and K. Belkhir. A Bayesian approach to the identification of panmictic populations
421 and the assignment of individuals. *Genetical Research*, 78:59–77, 2001.
- 422 B.K. Epperson. Spatial structure of two-locus genotypes under isolation by distance. *Genetics*,
423 140:365–375, 1995.
- 424 D. Falush, M. Stephens, and J.K. Pritchard. Inference of population structure using multilocus
425 genotype data: Linked loci and correlated allele frequencies. *Genetics*, 164:1567–1587, 2003.
- 426 M. Fontaine, S.J.E. Baird, S. Piry, N. Ray, K. Tolley, S. Duke, A. Birkun, M. Ferreira, T. Jauni-
427 aux, A. Llavona, B. Östürk, A.A. Östürk, V. Ridoux, E. Rogan, M. Sequeira, U. Siebert, G.A.
428 Vikingson, J.M. Bouquegneau, and J.R. Michaux. Rise of oceanographic barriers in continuous

429 populations of a cetacean: the genetic structure of harbour porpoises in old world waters. *BMC*
430 *Biology*, 5(30), 2007.

431 G. Guillot, A. Estoup, F. Mortier, and J.F. Cosson. A spatial statistical model for landscape
432 genetics. *Genetics*, 170(3):1261–1280, 2005.

433 O.J. Hardy and X. Wekemans. Isolation by distance in a continuous population: reconciliation
434 between spatial autocorrelation analysis and population genetics models. *Heredity*, 83:145–154,
435 1999.

436 C. Lantuéjoul. *Geostatistical simulation*. Springer Verlag, 2002.

437 J. Novembre and M. Stephens. Interpreting principal component analyses of spatial population
438 genetic variation. *Nature Genetics*, 40:646 – 649, 2008.

439 M.F. Ostrowski, J. David, S. Santoni, H. Mckhann, Reboud X., V. Le Corre, Camilleri C., Brunel
440 D., Bouchez D., B. Faure, and T. Bataillon. Evidence for a large-scale population structure
441 among accessions of *arabidopsis thaliana*: possible causes and consequences for the distribution
442 of linkage disequilibrium. *Molecular Ecology*, 15(6):1507–1517, 2006.

443 J.K. Pritchard, M. Stephens, and P. Donnelly. Inference of population structure using multilocus
444 genotype data. *Genetics*, 155:945–959, 2000.

445 N.A. Rosenberg, S. Saurabh, S. Ramachandran, C. Zhao, J. Pritchard, and M.W. Feldman.
446 Clines, clusters, and the effect of study design on the influence of human population structure.
447 *Public Library of Science*, 1(6):660–671, 2005.

- 448 F. Rousset. Genetic differentiation between individuals. *Journal of Evolutionary Biology*, 13:
449 58–62, 2000.
- 450 F. Rousset. Genetic differentiation and estimation of gene flow from F-statistics under isolation
451 by distance. *Genetics*, 145:1219–1228, 1997.
- 452 M. Schlather. Introduction to positive definite functions and to unconditional simulations of
453 random fields. Technical Report ST-99-10, Department of Mathematics and Statistics, Faculty
454 of Applied Sciences, Lancaster, UK, 1999.
- 455 D. Serre and S. Pääbo. Evidence for gradients of human genetic diversity within and among
456 continents. *Genome Research*, 14:1679–1685, 2004.
- 457 N.C. Stenseth, A. Shabbar, K.S. Chan, S. Boutin, E.K. Rueness, D. Ehrich, J.W. Hurrell, O.C.
458 Lingjæde, and K.S. Jakobsen. Snow conditions may create an invisible barrier for lynx. *Pro-
459 ceedings of the National Academy of Sciences*, 101:10632–10634, 2004.
- 460 S. Tavaré and O. Zeitouni. *Ancestral inference in population genetics. Proceedings of Saint Flour
461 Summer School in Probability and Statistics*. Lecture Notes in Statistics. Springer Verlag, 2001.
- 462 H. Wackernagel. *Multivariate geostatistics : an introduction with applications*. Springer Verlag,
463 Berlin, 1995.
- 464 H. H. Wagner, R. Holderegger, S. Werth, F. Gugerli, S.E. Hoebee, and C. Scheidegger. Var-
465 iogram analysis of the spatial genetic structure of continuous populations using multilocus
466 microsatellite data. *Genetics*, 169:1739–175, 2005.

- 467 S.K. Wasser, A.M. Shedlock, K. Comstock, E.A. Ostrander, B. Mutayoba, and M. Stephens.
468 Assigning African elephants DNA to geographic region of origin: applications to the ivory
469 trade. *Proceedings of the National Academy of Sciences*, 101(41):14847–14852, 2004.
- 470 B.S. Weir and C.C. Cockerham. Estimating F-statistics for the analysis of population structure.
471 *Evolution*, 38(6):1358–1370, 1984.
- 472 S. Wright. Isolation by distance. *Genetics*, 28:114–138, 1943.

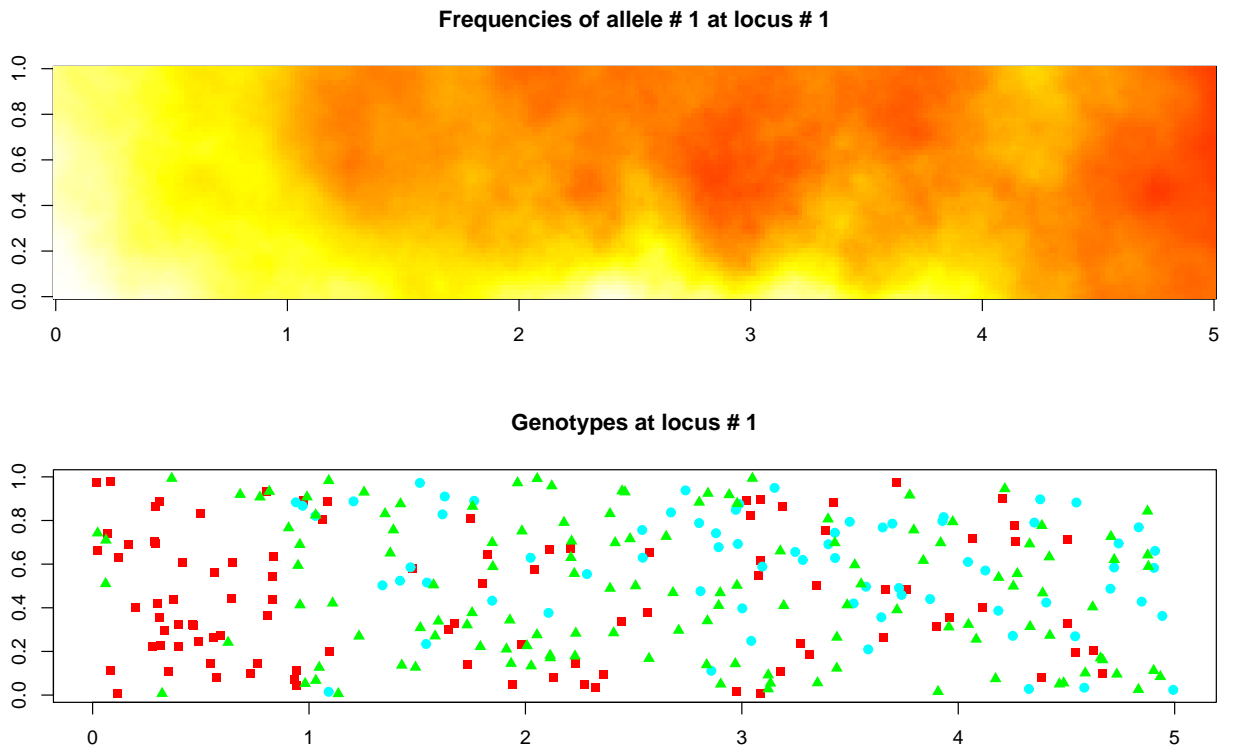
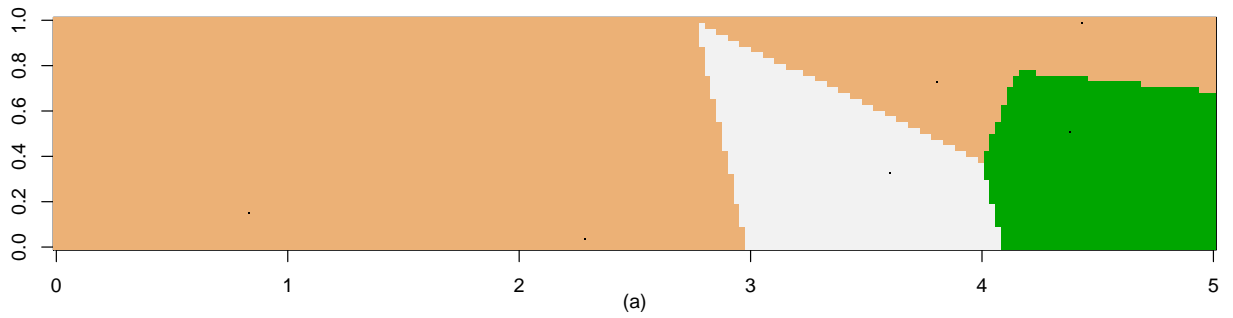
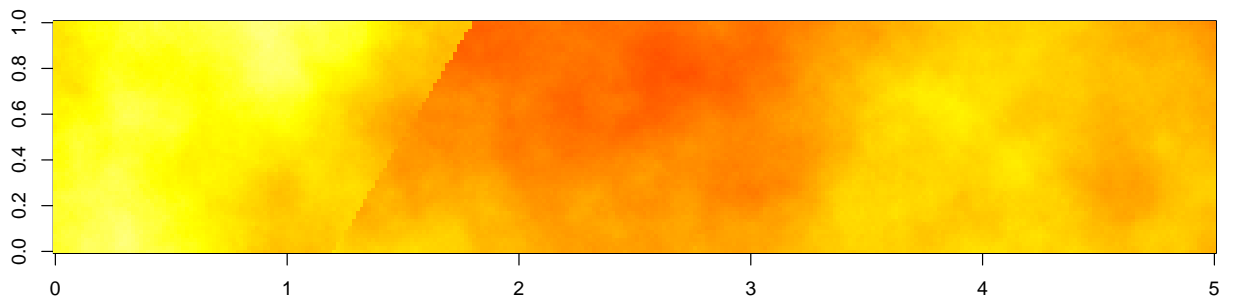


Figure 1: Top: Simulated frequencies and genotypes in a continuous population under moderate IBD. Bottom: Simulated genotypes $\{A,A\}$, $\{a,A\}$ and $\{a,a\}$ are represented by blue circles, green triangles and red squares, respectively. Scale parameter $\beta = 2$, smoothing parameter $\gamma = 1.5$.



Frequencies of allele # 1 at locus # 1



Frequencies of allele # 1 at locus # 1

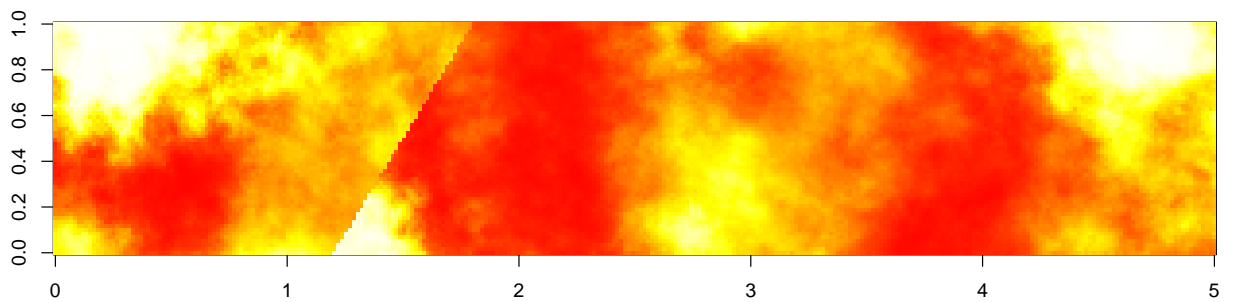


Figure 2: Simulated example of allele frequencies surfaces for two IBD populations. Top panel: spatial domains, middle: weak IBD ($\beta=4$), bottom: strong IBD ($\beta=0.5$).

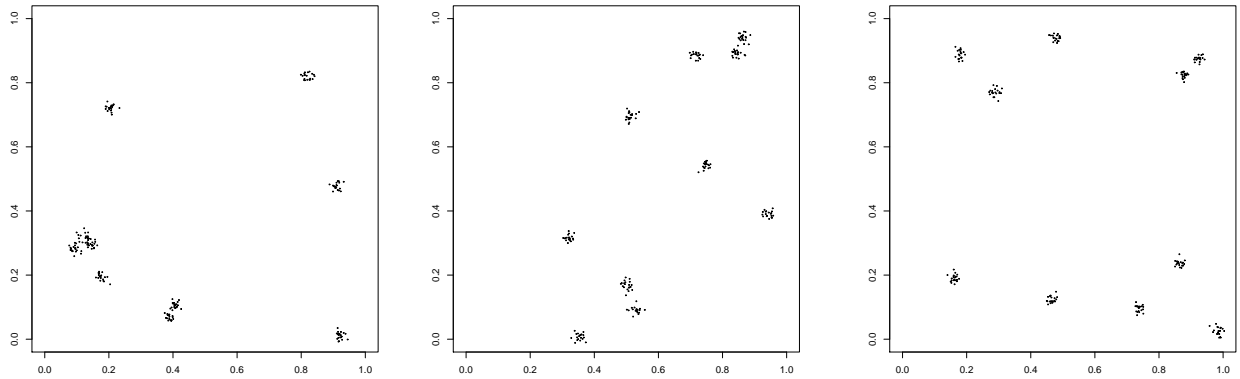


Figure 3: Examples of spatially-irregular sampling pattern with ten bunches of 20 individuals. The center of bunches are sampled uniformly and independently on $[0, 1] \times [0, 1]$ and the individuals within each bunch are sampled from a centered independent bi-variate Gaussian distribution with a standard deviation of 0.01

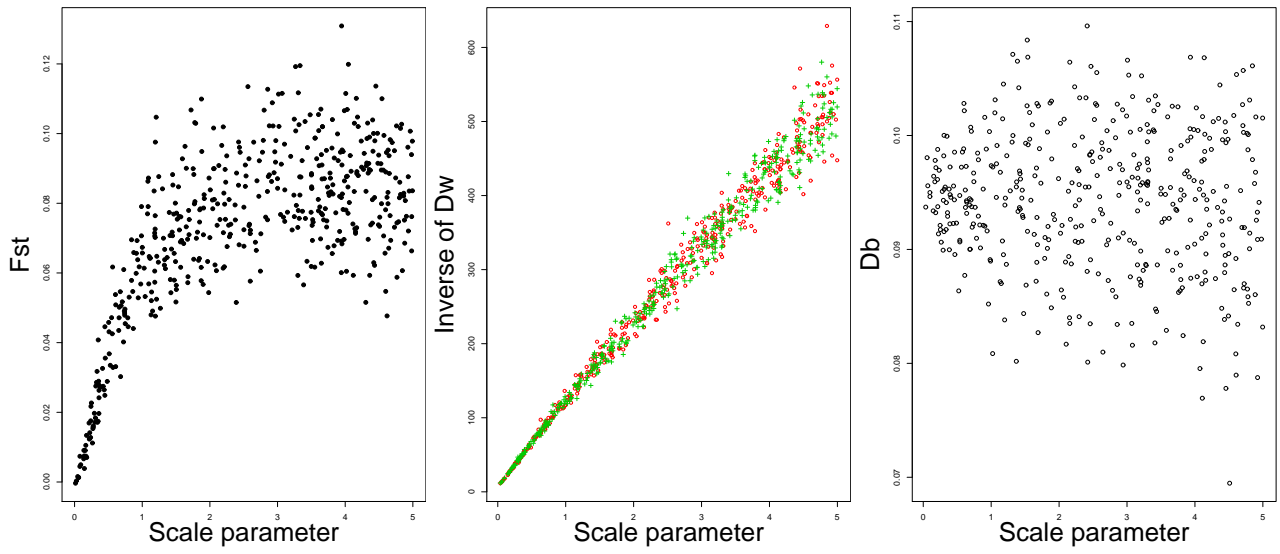


Figure 4: IBD model with two populations: relationships between the scale parameter β and F_{st} or the measure of local differentiation within populations (D_W) and between population (D_B). Each dot corresponds to one of the 500 simulated data-sets.

	$\%_{\hat{K} \neq K}$	$\%_{\hat{K} > K}$	ERCA _i	ERCA _p
Panmictic data - Regular spatial sampling				
spatial inference	0.6	0	1.91	4.97
non spat. inf.	10.4	8.8	3.49	-
Panmictic data - Irregular spatial sampling				
spatial inference	1	0.4	1.49	20.7
IBD data - Regular spatial sampling				
spatial inference	27.8	27.4	12.6	15.4
non spat. inf.	38.2	37	15.2	-
IBD data - Irregular spatial sampling				
spatial inference	28.4	28.4	12.7	27.5

Table 1: Accuracy in terms of inference on number of groups K (percentage of runs where $\hat{K} \neq K$ and where $\hat{K} > K$), and in terms of individual assignment (ERCA_i) and location of borders between populations (ERCA_p). \hat{K} denotes estimated K . ERCA denotes Error Rate in Co-Assignments (proportion of pairs of individuals not correctly co-assigned) which can be computed for sampled individuals (ERCA_i) or for pixels of the domain (ERCA_p). Data simulated for 200 individuals at 10 loci. Each number obtained from 500 independent data-sets. See supplementary material for details.