Prix La Recherche

Modélisation géostatistique en génétique des populations.

Synthèse du travail de
Gilles Guillot[1], *Arnaud Estoup*[2], *Frédéric Mortier*[3]
et Jean-François Cosson.[2]

Juin 2004

(1) UMR Mathématiques et Informatiques Appliquées INRA-INAPG-ENGREF, Paris.
(2) Centre de Biologie et de Gestion des Populations UMR INRA-IRD-CIRAD, Montpellier.
(3) CIRAD Département Forêts, Montpellier.

Flux de gènes, structures spatiales et méthodes de classification en génétique des populations

L'intérêt porté à l'étude des flux géniques date des débuts de la génétique des populations avec en particulier les modèles de Fisher et de Wright portant sur l'analyse des fréquences alléliques. Plus récemment l'amélioration des outils génétiques, combinée aux développements méthodologiques et informatiques ont considérablement renouvelé ce secteur de la génétique des populations, et donné naissance à différentes estimations plus ou moins directes des flux géniques. Ainsi, les procédures d'assignations multilocus permettent sous certaines hypothèses de détecter des migrants de première génération et leur descendants dans des populations, de tester l'appartenance de ces migrants à différentes sources potentielles, ou encore d'associer des probabilités d'appartenance à ces différentes sources de façon indépendante Paetkau et al.[14], Rannala et al. [17], Cornuet et al. [5]. Ces approches statistiques très performantes ont en commun la nécessité de définir *a priori* des sous-populations ou des dèmes, ce qui pose des problèmes pratiques lorsque les individus sont régulièrement distribués dans l'espace, et/ou lorsqu'une structuration insoupçonnée amène à définir des limites erronées autour de populations composites. Les approches « individus-centrées » sont de ce point de vue plus satisfaisantes puisqu'elles font l'économie de cette définition a priori, et limitent ainsi les erreurs potentielles. Des développements analytiques récents ont permis de dériver une méthode pour détecter les populations régies par un patron d'isolement par la distance (IBD), et d'estimer les tailles de voisinages et les distances de dispersion (Rousset [19]). Une hypothèse forte et inhérente à cette dernière approche est la continuité/homogénéité du patron d'IBD dans l'espace et le temps, ce qui peut rendre son application problématique dans le cas où on suspecte des discontinuités spatiales et/ou temporelles. D'un point de vue évolutif, comme d'un point de vue appliqué à la gestion des populations, ce sont cependant ces discontinuités et les facteurs qui en sont responsables qui peuvent se révéler les plus intéressants à mettre en évidence. Dans ce contexte, différentes tentatives pour combiner l'écologie du paysage et la génétique des populations ont donné naissance a une nouvelle discipline, la génétique du paysage (landscape genetics Manel et al. [13]) dont l'un des objectifs clef est la détection des discontinuités génétiques, et leur corrélation spatiale avec certains éléments du paysage qui peuvent jouer le rôle de barrières aux flux de gènes. Dans cette perspective, la structuration spatiale de la variabilité génétique doit être évaluée à partir de données individuelles, sans définition de populations a priori. Un certain de nombre de méthodologies ont été développées pour avancer dans cette direction (voire la syntèse récente de Manel et al. [13]) mais aucune d'entre-elles ne permet d'inférer statistiquement la localisation spatiale des discontinuités génétiques dans l'espace. Parmi les méthodes les plus utilisées, les autocorrélations spatiales permettent au mieux de définir l'échelle spatiale de l'homogénéité génétique (Smouse et al. [20]) mais pas de localiser les discontinuités génétiques. A l'opposé, les méthodes d'agrégation bayesienne ([15], Dawson et al. [6]) permettent de définir des ensembles d'individus, homogènes d'un point de vue génétique, avec un minimum d'hypothèses a priori sur les limites de ces ensembles. Une récente extension de ces modèles par Corander et al. [4] permet de traiter des situations où le nombre de sous-populations est inconnu. Mais ces approches ne sont pas explicitement spatialisées et peu efficaces en définitive pour localiser des discontinuités génétiques sans a priori. Une limite

supplémentaire vient du fait qu'elles requièrent l'hypothèse d'association au hasard des individus au sein des sous-populations, ce qui exclue de fait de nombreux organismes à reproduction partiellement endogame, autofécondante ou asexuée, mais également ceux caractérisés par une dispersion faible (en intensité et en distance) et évoluant selon des modèles en population continue avec isolement par la distance. La méthode d'analyse statistique que nous avons développée est par opposition à toutes celles développées jusqu'à maintenant, spatialement explicite.

Statistique spatiale : de la géostatistique minière aux systèmes hautement structurés

D'un point de vue plus statistique, la question que nous avons traité consiste à regrouper les individus dans un certain nombre de populations sur la base du génotype multilocus et de la localisation spatiale des individus. Ce problème concentre les difficultés techniques :
- On souhaite traiter des situations où les flux géniques sont tels que la similarité génétique (en un sens statistique que l'on précisera plus loin) dépend de la distance géographique qui sépare ces individus. Par conséquent, les données relatives à plusieurs individus doivent être considérées comme statistiquement dépendantes. Or la plupart des méthodes statistiques courament employées sont basées sur une hypothèse d'indépendance statistique des observations (ou sur une paramétrisation très simpliste de la dépendance).
- Les données génétiques sont en général acquises typiquement pour une dizaine de loci. Il faut donc être capable de spécifier un modèle multivariable. Ceci limite le recours aux méthodes statistiques existantes ou nécessite au minimum des adaptations des méthodes monovariables au cas multivariable.
- Les observations pour chaque individu se présentent sous la forme d'un vecteur donnant les formes alléliques observées. Il s'agit donc de variables catégorielles, ce qui une fois encore éloigne notre problème des méthodes statistiques les plus courantes qui traitent en général des données quantitatives. Les modèles de mélanges qui servent à modéliser des populations hétérogènes (en biologie mais aussi en finance, astronomie, analyse d'image) ont fait l'objet d'aucun travaux dans le cas de plusieurs variables catégorielles.
- Enfin, on souhaite regrouper les individus en populations homogènes sans connaitre a priori le nombre de ces populations : il y a des quantités inconnues dans le modèle (la matrice de classification, les fréquences allèliques dans les populations, etc...), et le nombre de quantités inconnues est lui-même inconnu.

Ces difficultés peuvent être mises en regard du développment des méthodes statistiques en relation avec notre problème :
Des méthodes permettant de modéliser la dépendance statistique d'une variable géophysique par rapport aux coordonnées d'espace ont été développées à partir des années 50. La contribution méthodologique essentielle est celle de G. Matheron à l'École des Mines de Paris qui a proposé des modèles (regroupé sous le terme de *géostatistique*) pour l'évaluation des ressources naturelles (ressources minières initialement) à partir d'un petit nombre de sondages. Les développements de la géostatistique sont restés orientés vers la même question centrale, celle de l'évaluation de la valeur d'une variable en un site pour lequel on ne dispose pas de mesures. (Cf Chilès et al. [3] ou Wackernagel [21]).

Parrallèlement, des méthodes pour traiter des données géoréférencées ont commencé à apparaitre à la fin des années 80 dans la communauté "mainstream statistics" Besag [1], à partir notamment de questions qui se posent en analyse d'image. La question centrale étant généralement de reconstituer une image "vraie" à partir d'une image observée qui est une version corrompue de l'image vraie.

Ces modèles de plus en plus nombreux et complexes, très séduisants d'un point de vue théorique sont restés pour la plupart peu utilisés jusqu'à la fin des années 80 car les calculs requis pour estimer les paramètres (optimisation et calcul d'intégrales en très grande dimension) restaient hors de portée avec les algorithmes de l'époque. Un verrou a été levé avec l'article de Gelfand et Smith [8] qui proposait de généraliser l'usage des méthodes de calcul d'intégrale par Monte-Carlo en statistiques.
Entre autres conséquences importantes, l'introduction de méthodes numériques intensives en statistiques (désignées souvent sous le terme de *computationnal statistics*) a permis de faire le lien avec le paradigme Bayesien, dont l'intérêt majeur est de permettre d'injecter une information *a priori* dans un modèle statistique. Le modélisateur spécifie un modèle qui décrit comment ses observations sont distribuées statistiquement et comment elles sont supposées interagir. Sa description repose sur des paramètres qui sont inconnus mais sur lesquels il a toujours un minimum de connaissance. Ces connaissances a priori sont injectées dans le modèle à travers une loi de distribution des paramètres inconnus. P.G. Green et S. Richardson [18] ont mis en oeuvre ces idées pour estimer la densité de probabilités d'un échantillon lorsque les données proviennent de différentes populations (ayant des distributions différentes) et sont d'origines inconnues.

Le livre de Green, Hjort et Richardson [9] intitulé *Highly Structured Stochastic Systems* donne une synthèse récente des modèles développés grâce à la fertilisation croisée de la statistique, de l'algorithmique et de l'analyse Bayesienne. Ce livre montre que l'on peut modéliser et estimer de manière rigoureuse les paramètres d'un système arbitrairement complexe, à condition d'avoir une quantité suffisante d'observations. Toutefois l'exporation de ces idées vers la génétique des populations est restée assez limitée. Le paragraphe suivant résume très rapidement notre contribution dans ce sens.

Le modèle : mosaïque de Voronoï colorée, fréquences allé-liques corrélées, et équilibre de Hardy-Weinberg

On suppose que l'on a observé n individus diploïdes (mais on pourrait considérer n'importe quelle ploïdie sans perte de généralité), on connait leurs génotypes ainsi que leurs coordonnées spatiales à certaine précision près (on peut même considérer le cas d'animaux non sédentaires, à condition d'avoir une information sur leurs déplacements potentiels). On suppose que ces n individus proviennent de K populations différentes, K étant inconnu. On souhaite modéliser la structure dans cet échantillon. Le mot *structure* peut prendre des sens très divers selon la communauté scientifique dans lequel il est employé. En statistique, ce terme désigne en général un écart au désordre total modélisé par des variables aléatoires indépendantes.

Dans notre contexte, nous proposons de distinguer des structurations à trois niveaux différents : au niveau de l'organisation spatiale des populations, au niveau des fréquences allèliques dans les différentes populations et au niveau des génotypes à l'intérieur de chaque

population.

Modélisation de la structure spatiale par une partition aléatoire

L'idée centrale qui guide notre travail est qu'en général, les populations animales sont spatialement structurées, au sens où deux individus géographiquement proches ont une probabilité plus grande d'appartenir à la même population que deux individus géographiquement éloignés. On souhaite pouvoir *injecter cette information* dans le modèle de manière qualitative, et *quantifier* l'intensité de cette dépendance spatiale. Pour cela nous avons recours à un modèle développé initialement en géostatistique pour modéliser des formations géologiques des gisements miniers ou des réservoirs pétroliers [12].
Nous supposons que les aires de répartition de chaque population peuvent se décomposer en un nombre fini de polygones de Voronoi engendrés par un processus de Poisson sur le domaine d'étude (cf figure 1).

Modélisation des fréquences allélique

Pour modéliser les fréquences alléliques dans chaque population, on utilise une idée introduite par Falush et al. [7] qui consiste à faire référence à une population ancestrale. Les fréquences alléliques dans cette population ancestrale sont supposées suivre des lois de Dirichlet indépendantes d'une pop. à l'autre, et d'un locus à l'autre. Les populations présentes sont supposées avoir dérivé de la population ancestrale à une vitesse qui est paramétrisée par un vecteur de dérives $(d_1, ..., d_K)$. Cette représentation permet de paramétriser la dépendance entre les différentes populations en terme de fréquences alléliques.

Génotypes

Nous supposons qu'à l'intérieur de chaque population, les individus sont à l'équilibre d'Hardy-Weinberg sans déséquilibre de liaison. Autrement dit, conditionnellement à la partition de l'espace et aux fréquences allèliques, il y a indépendance de toutes les variables observées. Les individus d'une même sous-population sont "conditionnellement indépendants", en termes moins techniques : une fois les paramètres estimés, il ne reste plus de dépendance résiduelle, les frontières entre populations "expliquent" totalement l'écart à l'indépendance observé. Cette hypothèse peut-être assez restrictive dans un contexte de populations naturelles (nombreuses sources possibles d'écart à l'équilibre d'Hardy-Weinberg, cf. ci-dessus), mais constitue une première approche réaliste pour ce problème très complexe.

Incertitude sur les coordonnées

Enfin nous prenons en compte le fait que les coordonnées spatiales des individus fournies au moment de la capture ne sont pas forcément les coordonnées les plus pertinentes à prendre en compte (déplacement au moment de la capture, animaux non sédentaires, erreurs de positionnement...) C'est pourquoi nous introduisons une différence entre des coordonnées vraies et des coordonnées observées reliées par une équation du type

$$coord.\ \ observées = coord.\ \ vraies + erreur,$$

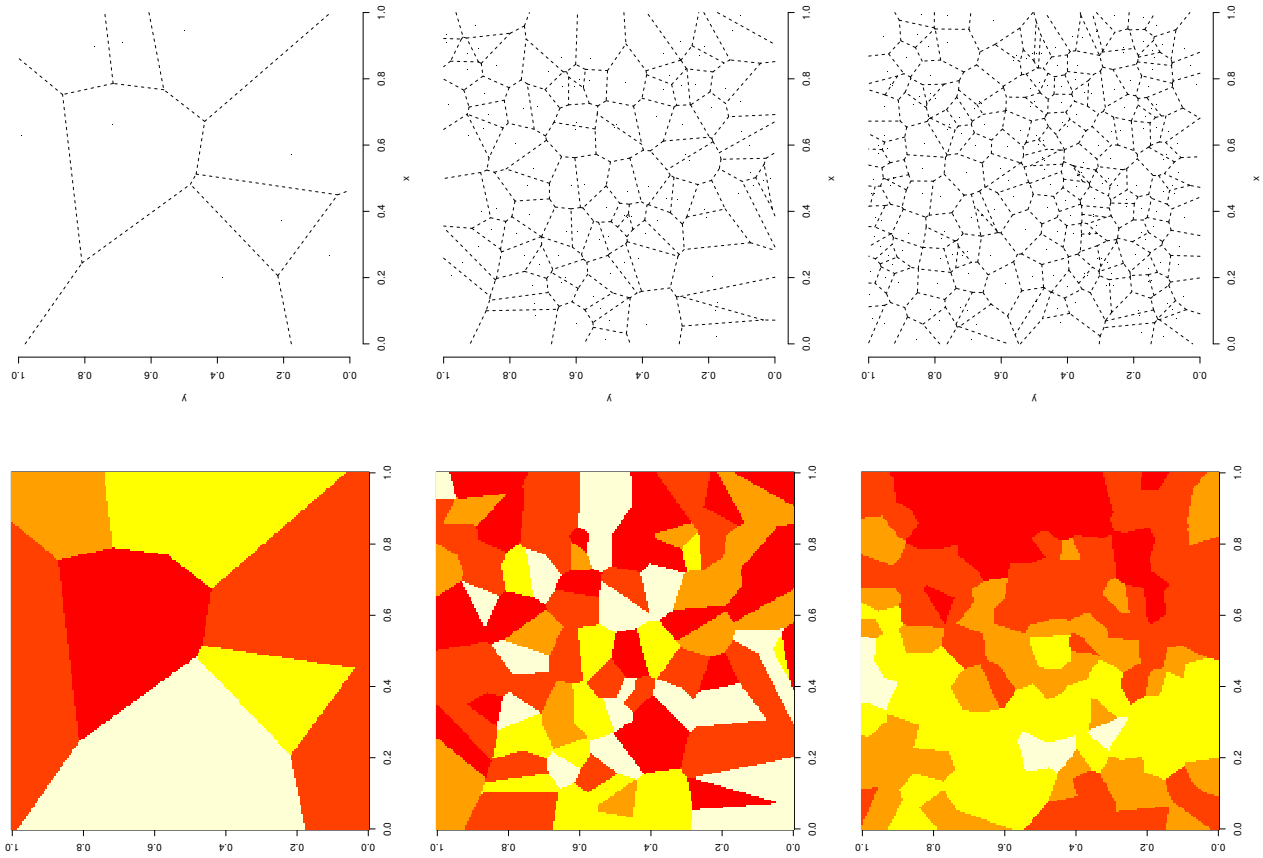

FIG. 1 – Illustration du modèle statistique proposé pour le partitionnement de l'espace géographique. Chaque population est supposée occuper un territoire. Les frontières de ce territoire sont des inconnues du problème. Nous supposons que chaque territoire peut être décomposé en une réunion de polygones convexes engendré par un semis de points poissoniens (points au hasard dans le domaine). Cette hypothèse a surtout pour but de permettre un paramétrage simple des territoires de chaque population et n'admet pas forcément une interprétation écologique. Dans certains contextes toutefois, ces polygones pourront s'interpréter comme des territoires individuels ou familiaux (cf Blackwell [2]). La figure montre trois simulations de tels découpages de l'espace (ici en quatre domaines en niveaux de gris du blanc au noir) selon notre modèle de partitionnement a priori. Ici on suppose que l'espace est occupé par quatre populations ditinctes. La ligne du haut montre les points poissonniens et les polygones de Voronoï qu'ils engendrent, la ligne du bas les mêmes polygones après affectation à une population. La probabilité d'appartenance de deux individus à une même population dépend alors de la distance géographique qui les sépare (à la différence de tous les modèles existants, où cette probabilité ne dépend que des génotypes des individus). La position et la couleur (i.e la population d'appartenance) de ces polygones doit être estimée à partir des données.

la distribution statistique de l'erreur dépendant fortement du contexte.

Les paramètres qui interviennent dans ce modèle sont : $\theta = (K, m, u, c, f_A, d, f, s)$ où K est le nombre de population, m est le nombre de polygones, $(u_1, ..., u_m)$ sont les centres de polygones, $(c_1, ..., c_m)$ sont les variables de classes (codant l'appartenance de chaque polygone à une population), f_{Alj} sont les fréquences alléliques dans la population ancestrale (au locus l, pour l'allèle j), $(d_1, ..., d_K)$ sont les dérives, f_{klj} sont les fréquences alléliques dans la population k (au locus l, pour l'allèle j) et $(s_1, ..., s_n)$ sont les vraies coordonnéees spatiales des individus considérées comme inconnues.

Ils sont estimés en spécifiant des distributions a priori sur chaque bloc de paramètres et en simulant une chaine de Markov $(\theta_t)_{t\in\mathbb{N}}$ dont la loi asymptotique est la distribution a posteriori du vecteur de paramètres. Le vecteur de paramètres est à valeur dans un l'espace $\Theta = \bigcup_{d\in\mathbb{N}} \mathbb{R}^d$ qui n'a pas la structure d'un espace euclidien. On réalise la simulation par un algorithme à sauts réversibles adapté de l'article de Green [10, 18].

Quelques résultats clés

Le modèle permet de
(i) réaliser une estimation précise du nombre de populations présentes dans l'échantillon observé ;
(ii) assigner correctement les individus à leur population d'origine. En particulier, nous améliorons les résultats par rapport à la méthode de référence dans le domaine [15] tandis qu'une autre méthode classique (souvent employée mais qui n'avait pas fait l'objet de comparaison systématique), apparait comme totalement inadaptée ;
(iii) cartographier de manière précise les limites géographiques de chaque population ;
(iv) détecter des migrants de première génération (cf figure 3).

ceci, sans a priori sur la localisation de ces inconnues.

Par ailleurs notre modèle permet de
(v) travailler à partir d'individus dont les coordonnées spatiales sont connues de manières imprécises ;

(vi) estimer correctement les paramètres aussi bien à partir de données individuelles régulièrement espacées (données individuelles) que de données très irrégulièrement espacées (données populationnelles).

Enfin, notre travail permet de
(vii) clarifier la notion de structure en écologie spatiale,
(viii) analyser les potentialités de différents modèles courament utilisés à ce jour
(ix) guider les utilisateurs dans les options de modélisations et le choix des paramètres pour ces modèles.

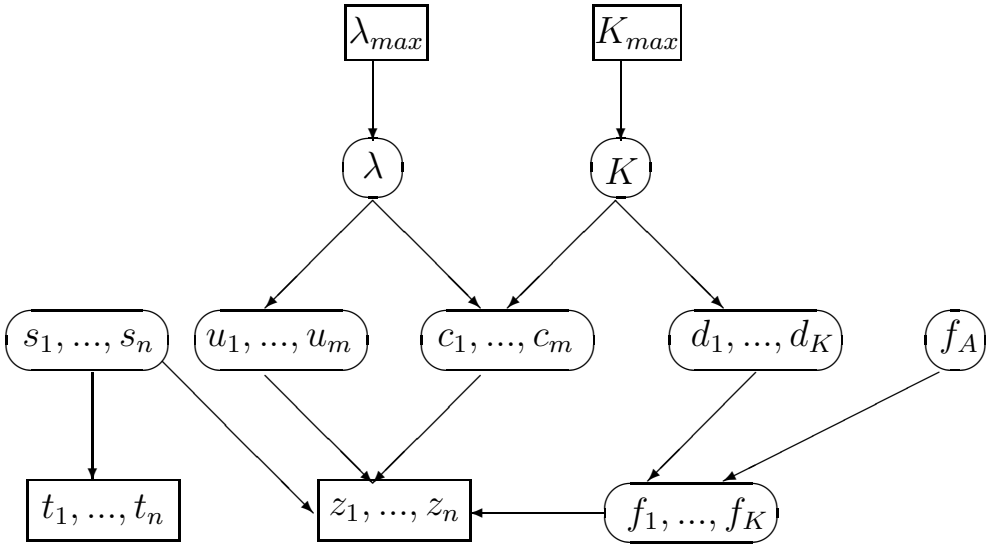

FIG. 2 – Graphe causal du modèle. Selon la convention habituelle dans la théorie des modèles graphiques (Jensen [11]), les quantités inconnues sont dans des boites arrondies, les données et les hyper-paramètres sont dans des boites rectangulaires. Il y a typiquement 1000 paramètres scalaires à estimer, toutefois, l'espace des paramètres Θ est pondéré par des lois a priori et on n'explore qu'une "petite partie" de cet ensemble.

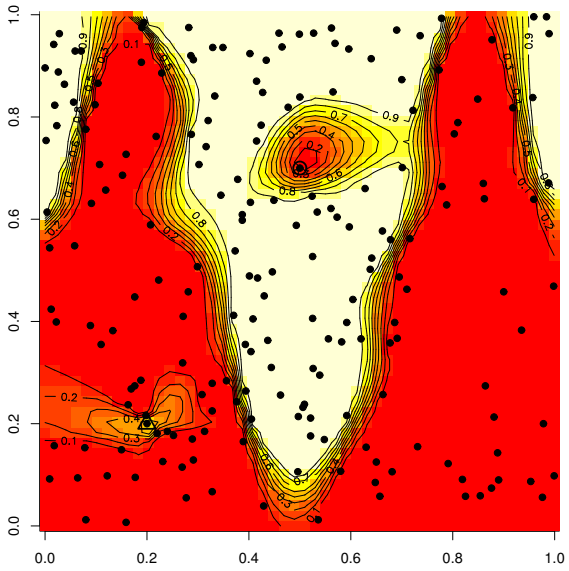

FIG. 3 – Illustration de la capacité de notre modèle à classifier des individus à partir d'un jeu de données simulées. Deux populations panmictiques moyennement différenciées sont simulées (10 loci, 10 allèles par locus, Fst=0.16). Les individus qui composent chacune d'entre elles sont positionnés aléatoirement de part et d'autre d'une frontière sinusoïdale. Un individu de chaque population est placé au hasard de l'autre coté de la frontière par rapport à sa population d'origine. Le graphiqe montre la probabilité a posteriori de chaque pixel du domaine d'appartenir à la première population : la frontière est parfaitement détectée ainsi que la présence des deux migrants (taches de part et d'autre de la frontière).

Perspectives

• Une des limitations forte de notre modèle réside dans le fait que toute la structure spatiale est supposée contenue dans la partition du domaine : à l'intérieur de chaque population, nous supposons que toutes les variables sont indépendantes et de même loi. Ceci correspond à une situation de co-existence de populations panmictiques séparées par des barrières naturelles assez imperméables.

Nous souhaitons considérer des situations plus réalistes dans lesquelles une forme de structure (comme de l'isolement par la distance) serait présente à l'intérieur de chaque population.

D'autre part, on peut imaginer qu'au voisinage des zones de contact entre populations se trouvent des *hybrides* dont le génotype ne sera correctement modélisé par aucun des modèles intra-population. La prise en compte explicite de tels individus aux frontières des domaines devrait permettre d'obtenir des algorithmes plus stables et des estimation plus précises. Ceci fait l'objet d'un travail un cours.

• Par ailleurs nous travaillons à une comparaison des sorties du modèle à celles obtenus par des méthodes non spatiales sur différentes espèces pour lesquelles des structurations spatiales fortes sont attendues : campagnols, chevreuil.

• Enfin tous les outils informatiques développés en Fortran et interfacé avec le logiciel statistique R [16] feront l'objet d'un package disponibe sur le réseau CRAN.

Références

[1] J. Besag. On the statistical analysis of dirty pictures. *Journal of the Royal Statistical Society, serie B*, 48(3) :259–302, 1986.

[2] P.G. Blackwell. Bayesian inference for a random tesselation process. *Biometrics*, 57(2) :502–507, 2001.

[3] J.P. Chilès and P. Delfiner. *Geostatistics*. Wiley, 1999.

[4] J.C. Corander, P. Waldmann, and M.J. Sillanpää. Bayesian analysis of genetic differentiation between populations. *Genetics*, 163 :367–374, 2003.

[5] J.M. Cornuet, S. Piry, G. Luikart, A. Estoup, and M. Solignac. New methods employing multilocus genotypes to select or exclude populations as origins of individuals. *Genetics*, 153 :1989–2000, 1999.

[6] K.J. Dawson and K. Belkhir. A bayesian approach to the identification of panmictic populations and the assignment of individuals. *Genet. Res.*, 78 :59–77, 2001.

[7] D. Falush, M. Stephens, and J.K. Pritchard. Inference of population structure using multilocus genotype data : Linked loci and correlated allele frequencies. *Genetics*, 164 :1567–1587, 2003.

[8] A.E. Gelfand and A.F.M. Smith. Sampling based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85 :389–409, 1980.

[9] P. J. Green, N.L. Hjort, and S. Richardson, editors. *Highly Structured Stochastic System*. Oxford University Press, 2003.

[10] P.J. Green. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82(4) :711–732, 1995.

[11] F.V. Jensen. *Bayesian networks and decision graphs*. Springer Verlag, 2201.

[12] C. Lantuéjoul. *Geostatistical simulation*. Springer, 2002.

[13] S. Manel, M.K. Schwartz, G. Luikart, and P. Taberlet. Landscape genetics : combining landscape ecology and population genetics. *Trends in Ecology and Evolution*, 18(4) :189–197, 2003.

[14] D. Paetkau, W. Calvert, I. Stirling, and C. Strobeck. Microsatellite analysis of population structure in canadian polar bears. *Molecular Ecology*, 4 :347–354, 1995.

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

[16] http ://cran.r-project.org/.

[17] B. Rannala and J.L. Moutain. Detecting immigration by using multilocus genotypes. *Proceedings of the National Academy of Sciences USA*, 94 :9197–9201, 1997.

[18] S. Richardson and P.J. Green. On Bayesian analysis of mixtures with an unknown number of components. *Journal of the Royal Statistical Society, serie B*, 59(4) :731–792, 1997.

[19] F. Rousset. *Handbook of Statistical Genetics*, chapter Inferences from spatial population genetics, pages 239–269. John Wiley & Sons, 2001.

[20] P.E. Smouse and R. Peakall. Spatial autocorrelation analysis of multi-allele and multi-locus genetic microstructure. *Heredity*, 82 :561–573, 1999.

[21] H. Wackernagel. *Multivariate geostatistics : an introduction with applications*. Springer Verlag, third edition, 2003.