Exploring the bayesian hierarchical approach for the statistical modeling of spatial structures: application in population ecology

Sophie Ancelet

N°/__/__/__/__/__/__/__/__/__/__/

THÈSE

pour obtenir le grade de

Docteur

de

L'Institut des Sciences et Industries du Vivant et de l'Environnement (Agro Paris Tech)

Spécialité : Biostatistique

présentée et soutenue publiquement
par

Sophie ANCELET

le 1 Juillet 2008

EXPLOITER L'APPROCHE HIERARCHIQUE BAYESIENNE POUR LA MODELISATION STATISTIQUE DE STRUCTURES SPATIALES

APPLICATION EN ECOLOGIE DES POPULATIONS

Directeur de thèse : ***Eric PARENT***

Codirecteur(s) de thèse : ***Marie-Pierre ETIENNE, Gilles GUILLOT***

Travail réalisé : AGRO PARIS TECH, UMR518 Mathématiques et Informatique Appliquées,

F-75231 Paris

Devant le jury :

M. Gilles CELEUX, Directeur de recherches, **INRIA** ……**Président**
Mme Verena TRENKEL, Chargée de recherches, **IFREMER** ……**Rapporteur**
M. Joël CHADOEUF, Directeur de recherches, **INRA** ……**Rapporteur**
M. Hugues BENOIT, Chargé de recherches, **FISHERIES AND OCEANS CANADA** ……**Examinateur**
M. Avner BAR-HEN, Professeur, **UNIVERSITE PARIS 5** ……**Examinateur**
M. Eric PARENT, Ingénieur en chef, **AGRO PARIS TECH** ……**Directeur de thèse**
Melle. Marie-Pierre ETIENNE, Maître de conférences, **AGRO PARIS TECH** ……**Codirecteur de thèse**
M. Gilles GUILLOT, Chargé de recherches, **UNIVERSITE D'OSLO** ……**Codirecteur de thèse**

Remerciements

Tout d'abord, je tiens à dédier cette thèse à ma grand-mère, Yvette, qui, j'en suis sûre, aurait été très fière du travail accompli et aurait, comme à son habitude, précieusement conservé ce document.

Un grand merci à Eric Parent pour avoir accepté de diriger cette thèse, pour l'aide compétente qu'il m'a apportée, ses grandes qualités humaines et surtout, son incroyable patience. Merci pour m'avoir permis d'entrer et de travailler dans un environnement incroyablement riche d'idées.

Je remercie chaleureusement Marie-Pierre Etienne qui, bien que cela ne soit pas officiel, a assumé à merveille toutes les responsabilités d'un encadrant de thèse. Merci pour son enthousiasme permanent et son aide précieuse sur le plan scientifique et administratif. Merci pour avoir accepté de m'accompagner aux Reflexives, un des séminaires auquel j'ai assisté pendant mon parcours de thèse et dont je conserve d'excellents souvenirs.

Merci à vous deux, Eric et Marie-Pierre, pour m'avoir encouragé à vivre des expériences formidables. Je pense en particulier à mes séjours au Canada et à Londres.

Merci également à Gilles Guillot pour m'avoir accordé sa confiance à l'origine de ce projet de thèse. Merci pour m'avoir initié avec beaucoup de patience à la programmation en Fortran et aux algorithmes d'inférence MCMC. Merci pour avoir accepté de participer à mon jury de thèse.

Je tiens à exprimer ma profonde reconnaissance à Mme Verena Trenkel et Mr Joel Chadoeuf qui ont accepté de juger ce mémoire avec beaucoup d'attention et de participer au jury de cette thèse. Ils ont notamment contribué par leurs nombreuses remarques et suggestions à améliorer la qualité de ce mémoire et je les en remercie. Merci également à Avner Barhen et Gilles Celeux pour avoir accepté le rôle d'examinateur de cette thèse.

Je souhaite exprimer toute ma gratitude et toute mon amitié à Hugues Benoît pour sa gentillesse et son accueil chaleureux lors de mon séjour à Moncton. Merci pour m'avoir fait découvrir la richesse écologique du Golfe du Saint-Laurent. Merci pour n'avoir jamais renoncé à participer à mon jury de thèse malgré les difficultés techniques posées par la mise en place de la vidéo-conférence. Merci pour avoir contribué à l'amélioration de ce mémoire par tes nombreuses remarques et suggestions.

Je remercie Françoise Launay, Corinne Fiers et Claude Pingeon pour leur gentillesse et l'aide administrative précieuse qu'elles m'ont apportée lors de l'organisation de mon séjour au Canada en août 2006.

Je tiens également à remercier mes amis qui, eux aussi, ont vécu cette thèse du début à la fin et m'ont aidé dans l'aboutissement de ce travail : Carole pour son soutien sans faille, ses encouragements et ses conseils avisés, Benjamin pour son humour incomparable et ses conseils précieux, Rozenn pour sa confiance et son inconditionnelle présence dans les moments importants, Julia pour son optimisme à tout épreuve. Un remerciement particulier à toute l'équipe du département d'épidémiologie et de santé publique de l'Imperial

College (Londres) qui m'a beaucoup soutenue pendant la rédaction des derniers chapitres de ce mémoire. Je pense en particulier à Sylvia Richardson, Léa, Juanjo, Jassy, Lawrence, Sonia et Arturo.

J'adresse un grand merci à toute ma famille qui a toujours été présente lorsque j'en ai eu besoin. Merci à toi maman pour avoir courageusement supporté les nombreux états d'âme à travers lesquels je suis passée au long de ces trois années de cette thèse. Merci à toi papa pour ton soutien téléphonique quotidien dans les derniers moments difficiles de la rédaction du mémoire : il s'est avéré déterminant pour mener ce travail à terme. Merci à vous, Guillaume, Angélique et Bruno pour vos innombrables encouragements et pour avoir contribué à faire de ma soutenance de thèse une journée... inoubliable !

Enfin, je remercie Thierry pour son amour inconditionnel et irremplaçable, sans lequel cette thèse n'aurait peut-être pas été ce qu'elle est aujourd'hui. C'est dur de partager sa vie avec une thésarde qui travaille tard tous les soirs, reste accroché à son ordinateur des week-end entiers et a pour sujet de conversion favori... sa thèse. Merci à toi Thierry pour avoir toujours fait preuve de patience à mon égard et avoir conservé cet humour qui m'a toujours remonté le moral dans les moments difficiles.

Résumé

Dans la plupart des questions écologiques, les phénomènes aléatoires d'intérêt sont spatialement structurés et issus de l'effet combiné de multiples variables aléatoires, observées ou non, et inter-agissant à diverses échelles. En pratique, dès lors que les données de terrain ne peuvent être directement traitées avec des structures spatiales standards, les observations sont généralement considérées indépendantes. Par ailleurs, les modèles utilisés sont souvent basés sur des hypothèses simplificatrices trop fortes par rapport à la complexité des phénomènes étudiés. Dans ce travail, la démarche de modélisation hiérarchique est combinée à certains outils de la statistique spatiale afin de construire des structures aléatoires fonctionnelles "sur-mesure" permettant de représenter des phénomènes spatiaux complexes en écologie des populations. L'inférence de ces différents modèles est menée dans le cadre bayésien avec des algorithmes MCMC. Dans un premier temps, un modèle hiérarchique spatial (Geneclust) est développé pour identifier des populations génétiquement homogènes quand la diversité génétique varie continûment dans l'espace. Un champ de Markov caché, qui modélise la structure spatiale de la diversité génétique, est couplé à un modèle bivarié d'occurrence de génotypes permettant de tenir compte de l'existence d'unions consanguines chez certaines populations naturelles. Dans un deuxième temps, un processus de Poisson composé particulier,appelé loi des fuites, est présenté sous l'angle de vue hiérarchique pour décrire le processus d'échantillonnage d'organismes vivants. Il permet de traiter le délicat problème de données continues présentant une forte proportion de zéros et issues d'échantillonnages à efforts variables. Ce modèle est également couplé à différents modèles sur grille (spatiaux, régionalisés) afin d'introduire des dépendances spatiales entre unités géographiques voisines puis, à un champ géostatistique bivarié construit par convolution sur grille discrète afin de modéliser la répartition spatiale conjointe de deux espèces. Les capacités d'ajustement et de prédiction des différents modèles hiérarchiques proposés sont comparées aux modèles traditionnellement utilisés à partir de simulations et de jeux de données réelles (ours bruns de Suède, invertébrés épibenthiques du Golfe-du-Saint-Laurent (Canada)).

Mots-clés : algorithmes MCMC, champs de Markov cachés, champs gaussiens bivariés, données génotypiques multilocus, données *zero-inflated*, loi des fuites, modélisation hiérarchique, modèles *delta*, processus de Poisson composé, variables latentes

Abstract

For most ecological questions, the random processes studied are spatially structured and come from the combined effect of several observed or unobserved random variables interacting at various scales. In practice, when data can't be directly treated with traditional spatial structures, observations are often considered as independent. Moreover, the usual models are often based on hypotheses that are too simple with regards to the complexity of the studied phenomena. In the present work, the hierarchical modelling framework is combined with some spatial statistics tools to build specific functional random structures for complex and spatially structured phenomena in population ecology. Model inference is done under the bayesian framework using MCMC algorithms. In the first part, a spatial hierarchical model (called Geneclust) is developed to identify genetically homogeneous populations when genetic diversity varies continuously in space. A hidden Markov random field, used to model the spatial structure of genetic diversity, is combined with a bivariate model for the occurrence of genotypes to take into account the possible occurrence of inbreeding in some natural populations. In the second part of the thesis, a particular compound Poisson process, called law of leaks, is presented from the hierarchical point of view. The goal was to describe the process of sampling living organisms. This approach explicitly confronts the technical issue of modelling continuous zero-inflated data from sampling characterized many zero values and variable sampling effort. This model is combined with different area-based models to add spatial dependencies between geographical units then with a bivariate gaussian random field built by process convolutions to model the joint spatial distribution of two species. The fitting and predictive capacities of the different hierarchical models are compared to the traditional models from simulated and real data (Scandinavian brown bears, epibenthic invertebrates in Saint-Lawrence Gulf (Canada)).

Keywords : MCMC algorithms, hidden Markov random fields, bivariate gaussian random fields, multilocus genotype data, zero-inflated data, law of leaks, hierarchical modelling, delta models, compound Poisson process, latent variables.

Table des matières

Publications scientifiques issues du travail de thèse

Bayesian Clustering using Hidden Markov Random Fields in Spatial Population Genetics
Olivier François, Sophie Ancelet, Gilles Guillot.
Paru dans *Genetics*

Hidden Markov random fields and the genetic structure of the scandinavian brown bear population
Sophie Ancelet, Gilles Guillot, Olivier François
Paru dans *le Journal de la Société Française de Statistique*

Modelling spatial zero-inflated continuous data with an exponentially compound Poisson process
Sophie Ancelet, Marie-Pierre Etienne, Hugues Benoît, Eric Parent
Soumis à *Environmental and Ecological Statistics*

Notations et principales densités de probabilité

Notations

Y,X,Z :	Majuscules désignant des variables aléatoires
y,x,z :	Minuscules latines désignant les réalisations respectives des variables Y,X,Z
θ :	Lettre grecque désignant par défaut les paramètres inconnus d'un modèle
$[Y]$:	Distribution de probabilité de la variable aléatoire Y
$[Y\|Z]$:	Distribution conditionnelle de la variable Y, conditionnellement à la variable Z
$[Y = y]$:	Probabilité de l'événement Y=y lorsque Y est une variable aléatoire discrète
$[y]$:	Densité de probabilité en y d'une variable continue Y
$[y\|z]$:	Densité de probabilité en y d'une variable continue Y conditionnellement à une valeur z de la variable Z
$X \sim \text{Loi}(\theta)$:	X suit une loi de probabilité de paramètres θ
$X_i \overset{i.i.d}{\sim} \text{Loi}(\theta)$:	Les variables X_i (i=1,2,...,n) sont indépendantes et identiquement distribuées selon une loi Loi(θ)

L'utilisation de la notation [], introduite en 1990 par A. Gelfand et A.F.M. Smith, est justifiée dans (Parent & Bernier, 2007). Dans ce document, nous nous plaçons sous le paradigme Bayésien. Aussi, nous noterons par $[y|\theta]$ la vraisemblance d'un modèle paramétré par θ car les paramètres sont considérés comme des variables aléatoires conditionnantes.

Principales densités de probabilités utilisées

Densités discrètes

Poisson(λ) : Loi de Poisson de paramètre d'intensité $\lambda > 0$
$[X = k|\lambda] = e^{-\lambda}\frac{\lambda^k}{k!}$

pour tout $k \in \mathbb{N}$

Binomiale(N,p) : Loi Binomiale issue de N tirages de Bernoulli avec une probabilité p $\in [0, 1]$
$[X = k|N, p] = \frac{N!}{p!(N-p)!}p^k(1-p)^{N-k}$

pour tout $k = 0, 1, ..., N$

Multinomiale(N,p) : Loi multinomiale de probabilités de tirage $p = (p_1, ..., p_k)$
telles que $p_i > 0$ pour tout i=1,...,k et $\sum_{i=1}^{k} p_i = 1$:

$[X_1 = x_1, ..., X_K = x_K|N, p_1, p_2, \ldots, p_K] = \frac{N!}{\prod_{i=1}^{k} x_i!}\prod_{i=1}^{K} p_i^{x_i}$

pour tout $x_1, .., x_K \geq 0$ satisfaisant $\sum_{i=1}^{K} x_i = N$

Dirichlet(α_1,..,α_K) : Loi de Dirichlet d'ordre $K \geq 2$ et de paramètres α_1,..,$\alpha_K > 0$:

$[x_1, ..., x_K|\alpha_1, \alpha_2, \ldots, \alpha_K] = \frac{\Gamma(\alpha_1+...+\alpha_K)}{\Gamma(\alpha_1)...\Gamma(\alpha_K)}x_1^{\alpha_1-1}...x_K^{\alpha_K-1}$

pour tout $x_1, ..., x_K > 0$; $\sum_{i=1}^{K} x_i = 1$

Densités continues

Exponentielle(ρ) : Loi exponentielle de paramètre ρ avec $\rho > 0$:
$[x|\rho] = \begin{cases} \rho e^{-\rho x} & \text{si} \quad x \geq 0 \\ 0 & \text{sinon} \end{cases}$

Gamma(a,b) : Loi gamma de paramètres (a, b) avec $a > 0$ et $b > 0$:
$[x|a, b] = \begin{cases} \frac{b^a}{\Gamma(a)}x^{a-1}e^{-bx} & \text{si} \quad x > 0 \\ 0 & \text{sinon} \end{cases}$

Remarque : Si $X \sim Gamma(a, b)$ alors $\widetilde{X} = 1/X$ est dite Gamma-inverse(a,b).

Normal(m, σ^2) : Loi normale de paramètres (m, σ^2) :

$$[x|m,\sigma^2] = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(x-m)^2} \qquad x \in \mathbb{R}$$

Beta(α,β) : Loi béta de paramètres (α, β) avec $\alpha > 0$ et $\beta > 0$:

$$[x|\alpha,\beta] = \frac{1}{B(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1} \qquad \text{avec } x \in [0,1] \text{ et } B(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

Première partie

Objectifs et concepts

Chapitre 1

Coupler modélisation hiérarchique et statistique spatiale : une idée à exploiter en écologie des populations

1.1 Deux problèmes pratiques en écologie des populations

Tout au long de ce travail de thèse, ma réflexion a été guidée par deux problèmes pratiques liés à l'écologie des populations. Le premier, basé sur des données génotypiques multi-locus, a porté sur la détection et la localisation de lignes de discontinuités génétiques au sein d'un ensemble d'individus géoréférencés. Le second, basé sur des données de relevés d'abondance, a consisté à modéliser la distribution d'invertébrés marins du Golfe-du-Saint-Laurent. Dans ce chapitre introductif, je commence par décrire ces deux problèmes. Dans un deuxième temps, j'en extrais des difficultés techniques communes, auquel le statisticien est souvent confronté lorsqu'il cherche à modéliser des phénomènes écologiques : les données sont généralement spatialement structurées mais ne peuvent être directement modélisées avec les structures aléatoires spatiales standards et les phénomènes aléatoires d'intérêt sont souvent la manifestation de l'effet combiné de multiples variables aléatoires. Puis, je montre que les modèles statistiques, usuels en écologie des populations, présentent certaines limites car, pour contourner ces difficultés, ils sont souvent basés sur des hypothèses simplificatrices trop fortes par rapport au phénomène étudié : absence de structure spatiale, ignorance des aspects multivariés, données supposées gaussiennes... Enfin, je définis la question méthodologique ainsi que l'objectif principal de ce travail de thèse : proposer des modèles statistiques, appliqués à des problèmes d'écologie des populations, qui permettent de palier simultanément aux difficultés techniques précédemment décrites en combinant les performances respectives de la statistique spatiale et de la modélisation hiérarchique.

1.1.1 Problème 1 : Détecter et localiser des lignes de discontinuités génétiques à partir de données génotypiques multi-locus géoréférencées

Tous les concepts génétiques énoncés dans cette partie sont rappelés dans l'Annexe A.

Une ligne de discontinuités génétiques est une zone géographique où intervient un changement brutal des fréquences alléliques pour un ou plusieurs marqueurs génétiques caractérisant un ensemble d'individus. Elle délimite une barrière aux flux de gènes entre les individus. Il peut s'agir d'une barrière géographique (ex : océan, montagne, forêt...) ou d'une barrière dûe à des préférences d'habitat différentes à l'intérieur d'un même domaine géographique.

En biologie évolutive ou en biologie de la conservation, la détection et la localisation de lignes de discontinuités génétiques sont une étape clé à l'analyse de patterns génétiques. En effet, ces zones particulières délimitent des populations plus ou moins isolées aux patterns génétiques distincts. Selon leurs specificités génétiques, ces populations peuvent définir des unités de gestion et/ou de conservation d'espèces rares. D'autre part, la détection et la localisation de lignes de discontinuités génétiques sont une étape préalable à l'étude de l'influence du paysage et autres caractéristiques environnementales (ex : montagne, rivière, déforestation, gradient d'humidité...) sur la structure génétique d'une population d'individus (Manel & al., 2003). Un enjeu possible est de prédire l'évolution de la diversité génétique des espèces sous différents scénarii (ex : fragmentation et/ou modification des paysages).

Deux cas d'applications ont été traités au cours de cette thèse. Tout d'abord, l'ours brun de Scandinavie *sp. Ursus Arctos* dont la survie préoccupe de nombreux écologues (Swenson & al., 1994), (Waits & al., 2000). Cette population sauvage a connu une histoire évolutive particulière. En 1930, elle a connu une quasi-extinction suite à une période d'extermination massive. Grâce à l'instauration de mesures de protection et à la survie de quatre foyers de femelles, elle a ensuite pu se reconstituer partiellement. Cependant, les brassages génétiques sont très limités car ces quatre foyers sont géographiquement séparés (cf figure 1.1) en raison de mouvements de dispersion très limités des ours femelles. Celles-ci ont tendance à retourner toujours en un même site pour se nourrir et se reproduire (Blum & al., 2004). Une telle situation augmente les risques de pertes de diversité génétique et de consanguinité d'autant que l'effectif des reproducteurs a diminué. Les populations humaines ont constitué le deuxième cas d'étude. Un enjeu d'intérêt est de vérifier si les populations délimitées par des critères arbitraires (ex : linguistiques, morphologiques, culturels, politiques, géographiques etc...) sont consistentes par rapport à des frontières définies par l'information génétique (Pritchard & al., 2000).

Du point de vue statistique, les problèmes de détection et de localisation de lignes de discontinuités génétiques se ramènent à un problème de classification d'individus en populations génétiquement homogènes i.e., dont les génotypes sont issus d'un même jeu de fréquences alléliques. Cette classification s'effectue à partir de données génotypiques multi-locus géo-référencées. Pour chaque individu observé, nous disposons d'un vecteur donnant les 2 formes alléliques observées pour L marqueurs microsatellites d'intérêt. Ces données génotypiques sont dites géo-référencées car elles sont chacune associées à un site géographique dont la localisation spatiale est connue.

Concernant les ours bruns de Scandinavie, le jeu de données disponible indique le

Figure 1.1 – La répartition spatiale de l'ours brun en Scandinavie. Les quatres sous-populations (North-North (NN), North-South (NS), Middle (M), and South (S)) ont été préalablement identifiées comme des foyers de concentration de femelles ayant survécu au programme d'extermination massive de 1930 (Image extraite du site web : www.quantum-conservation.org)

génotype de 366 ours en 19 marqueurs microsatellites. En chaque locus, les allèles observés sont indiqués par des variables catégorielles prenant leurs valeurs entre 1 et le nombre de versions alléliques possibles pour le marqueur associé. Les ours étant des individus diploïdes, le génotype d'un ours en un locus donné est constitué de deux versions alléliques, identiques ou non, d'un même marqueur. Les coordonnées spatiales des ours ont été relevées au moment de l'échantillonnage. Le tableau 1.1 est un extrait du tableau de données génétiques disponibles.

Le jeu de données humaines HGDP-CEPH (Cann & al., 2002) a été utilisé pour le deuxième cas d'étude : il contient le génotype de 1056 humains associé à 377 marqueurs microsatellites autosomals (i.e., situés sur des chromosomes non-sexuels). Ces individus ont été échantillonnés dans 52 populations humaines réparties sur l'ensemble des continents. Afin de considérer uniquement des individus vivant sur un même continent, l'étude a été limitée aux populations asiatiques (Eurasie et Asie de l'Est) ce qui forme un échantillon de 451 individus répartis dans 27 populations : 8 du Pakistan, 16 de Chine, 1 de Sibérie, 1 du Japon et 1 du Cambodge (cf figure 1.2). Contrairement au cas des ours bruns, les coordonnées spatiales des individus génotypés ne sont pas connues explicitement. Seuls des intervalles d'échantillonnage sont disponibles et les coordonnées individuelles ont été générées aléatoirement à l'intérieur de chaque intervalle.

1.1.2 Problème 2 : Modéliser la distribution des invertébrés marins du Golfe-du-Saint-Laurent à partir de données de relevés au chalut de fond

Le ministère fédéral canadien des Pêches et Océans (MPO) veille à la conservation et à la gestion durable des ressources halieutiques canadiennes par l'élaboration et la mise en

	L1	L2	L3	L4	L5	L6	L7	L8	L9	L10	L11	L12	L13	L14	L15	L16	L17	L18	L19
Ours 1	1,2	3,3	1,4	1,6	4,5	6,6	1,5	5,6	3,5	2,6	6,7	4,4	2,4	1,6	7,8	3,5	1,1	6,8	3,6
Ours 2	2,2	1,1	1,1	6,6	4,5	5,6	3,3	2,5	5,5	2,6	7,7	2,5	1,1	5,8	2,2	5,8	1,1	7,7	5,9
Ours 3	2,2	1,2	2,4	5,6	2,4	3,6	4,4	5,6	4,7	3,6	2,7	3,5	1,5	1,5	3,5	5,5	2,9	5,8	3,9
Ours 4	2,2	2,2	2,4	4,5	2,2	6,6	2,4	1,1	5,7	3,6	6,6	3,3	1,4	5,7	3,3	3,3	1,9	5,8	7,7
Ours 5	2,2	1,1	1,4	1,5	2,4	6,6	3,3	6,6	3,3	3,7	7,6	5,5	1,4	3,7	6,6	5,6	1,2	2,5	8,9
Nombre de versions alléliques	3	3	5	6	6	6	7	7	7	7	7	7	8	8	8	8	10	10	10

Tableau 1.1 – Extrait du jeu de données des ours bruns de Scandinavie. Chaque ligne correspond aux génotypes d'un ours en 19 marqueurs microsatellites. La dernière ligne indique le nombre de versions alléliques possibles pour chaque marqueur microsatellite. Chaque colonne indique les 2 allèles (séparées par une virgule) constituant un génotype en un locus donné.

oeuvre de politiques et programmes répondant aux intérêts économiques et environnementaux du Canada. La région du Golfe du Saint-Laurent, une des 6 régions administratives de MPO, est gérée par le Centre des Pêches du Golfe.

Depuis 1971, le Centre des Pêches du Golfe organise annuellement des relevés au chalut de fond, dans la partie sud du Golfe-du-Saint-Laurent (Hurlbut & Clay, 1990) (cf figure 1.3). La région d'étude a été divisée en 27 strates d'échantillonnage selon la partition indiquée dans la figure 1.5. Ces strates ont été définies en fonction de la profondeur des fonds marins puis rafinées afin d'obtenir des zones d'échantillonnage plus petites. Les sites d'observation sont choisis aléatoirement au début de chaque campagne et leur nombre varie proportionnellement à la surface de chaque strate. Autour de chaque site pré-choisi, un large filet, appelé chalut, balaye les fonds marins pendant une durée de trait ciblée de 30 minutes à une vitesse de 3.5 noeuds (i.e., distance équivalente parcourue de 1.75 miles). Une fois le chalut remonté, les espèces ramassées sont identifiées, triées puis comptées ou pesées. (cf figure 1.4). Au cours de ces campagnes, l'abondance de toutes les espèces de poissons (commerciales ou non) est quantifiée et, depuis 1989, l'abondance en invertébrés *épibenthiques** (e.g., oursins, anémones de mer,...).

L'objectif principal de ces campagnes est de suivre la dynamique temporelle des principales espèces marines vivant dans cette région d'étude. L'enjeu est ensuite d'étudier l'influence de divers facteurs environnementaux (e.g., prédation, relations inter-espèces, anomalies climatiques) ou humains (e.g., pêche commerciale,..) sur cette dynamique. En pratique, les écologues synthétisent les données de biomasse recueillies en divers indicateurs de biodiversité (indice de richesse, indice de rareté, indice de conservation des espèces) et/ou interpolent ces données pour cartographier la distribution des espèces dans la région d'étude. Cette approche a pour principale faiblesse de ne pas tenir compte des diverses sources de variabilité susceptibles d'expliquer la répartition observée de la bio-

∗. invertébrés vivants à la surface de substrats meubles ou rocheux des fonds marins

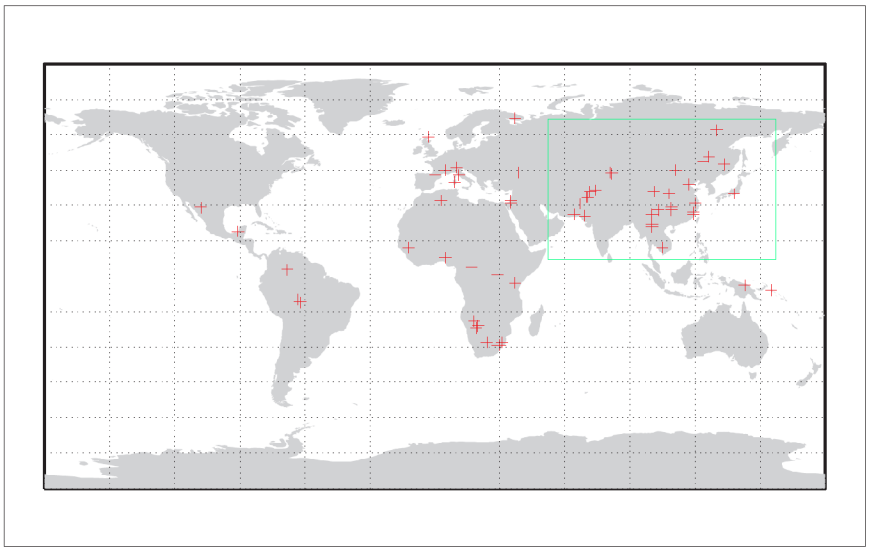


Figure 1.2 – Localisation spatiale (désignée par une croix rouge) des 52 populations humaines échantillonnées dans le jeu de données HGDP-CEPH. Notre étude s'est restreinte aux 27 populations situées dans le carré vert.

masse.

Au cours de cette thèse, je me suis intéressée au problème de la construction de modèles statistiques pertinents pour représenter la répartition spatiale de la biomasse dans le Golfe-du-Saint-Laurent. En effet, en écologie appliquée, la spécification de modèles statistiques est souvent rendue incontournable par les multiples avantages offerts. Ils permettent de mieux comprendre les mécanismes de fonctionnement des ecosystèmes par la synthèse de données issues du suivi de populations dans le milieu naturel et la prise en compte de leurs incertitudes. Par ailleurs, ils permettent de réaliser des diagnostics sur l'état des écosystèmes, de faire des prédictions quantifiées de leur évolution sous divers scenarii environnementaux ou encore de comparer l'effet de plusieurs actions de gestions humaines.

Dans le cadre de cette thèse, l'étude a été limitée à la distribution des invertébrés épibenthiques. Ces espèces ont pour caractéristique commune de se déplacer seulement sur de très courtes distances (quelques centaines de mètres au maximum par an) voire de rester immobiles. Ces espèces jouent un rôle important pour l'écosystème marin canadien car elles aident à former l'habitat physique utilisé par les poissons, en plus de servir de source de nourriture. Nous disposons des quantités de biomasse, en kilogrammes, mesurées en chaque trait de chalut réalisé depuis 1989 pour 14 espèces d'invertébrés dont les oursins, les anémones de mer, les étoiles de mer, les concombres de mer, les bigorneaux etc... Le tableau 1.2 est un extrait du tableau de données disponibles pour 3 espèces épibenthiques particulières : les oursins, les anémones et les concombres de mer. Dans ce tableau, nous pouvons remarquer que, comme cela est souvent le cas pour des mesures d'abondance, la plupart des quantités de biomasse recueillies sont nulles. Chaque zéro peut être vu comme le résultat d'un trait de chalut réalisé dans une zone abritant un habitat non propice à l'existence de ces espèces ou comme le fruit de la variabilité naturelle liée à l'échantillonnage : erreurs d'observation, faibles taux de capturabilité, etc.

Classer des individus en populations génétiquement homogènes à partir de données génotypiques multilocus géo-référencées et modéliser la distribution d'invertébrés marins

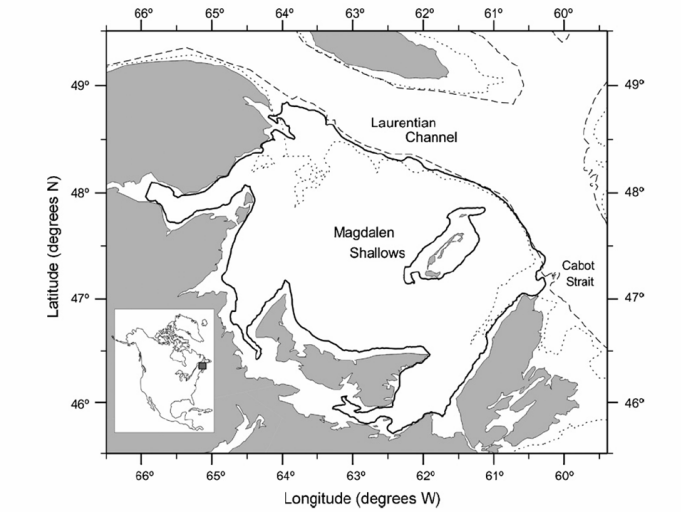


Figure 1.3 – Le sud du Golfe du Saint-Laurent, domaine de suivi ($73182km^2$) couvert par les relevés au chalut de fond du Centre des Pêches du Golfe.

à partir de relevés au chalut de fond sont deux problèmes *a priori* très différents de par leurs finalités respectives et le type de données disponibles. Pourtant, ils partagent des difficultés techniques communes, fréquentes dans la plupart des questions de modélisation en écologie des populations.

1.2 Des phénomènes aléatoires spatialisés

1.2.1 La répartition des organismes vivants est spatialement structurée

Résoudre les deux problèmes de modélisation précédemment décrits suppose de faire des hypothèses sur la répartition des organismes vivants dans leur environnement naturel. Pour l'un, il s'agit d'imaginer comment varie, dans le domaine d'étude, l'information génétique portée par ces organismes. Pour l'autre, il s'agit d'imaginer comment varie leur biomasse respective.

Par expérience, les écologues savent que les organismes vivants ne se répartissent pas aléatoirement et uniformément dans leur environnement naturel. Des processus physiques (e.g., géologie, vents, courants marins,...) et biologiques (e.g., dispersion, reproduction, compétition inter-espèces,...) ont un pattern spatial caractérisé par une forte régularité locale qui induit de la structure spatiale dans la répartition des organismes (Legendre & Legendre, 1998). Les quantités de biomasse observées en deux sites géographiquement proches ont tendance à se *ressembler* ou se *différencier* davantage par rapport à deux sites aléatoirement choisis dans le domaine d'étude. De même, la variabilité génétique peut être spatialement structurée c'est-à-dire dépendre de la distance géographique qui sépare les individus (Manel & al., 2003) (Rosenberg & al., 2005). En particulier, chez les espèces aux mouvements limités, comme les ours bruns de Scandinavie, les brassages génétiques sont réduits, ce qui favorise les similarités génétiques entre individus géographiquement proches.

Dans la plupart des questions de modélisation en écologie, il est donc pertinent de considérer comme statistiquement dépendantes les données relatives à des points de me-

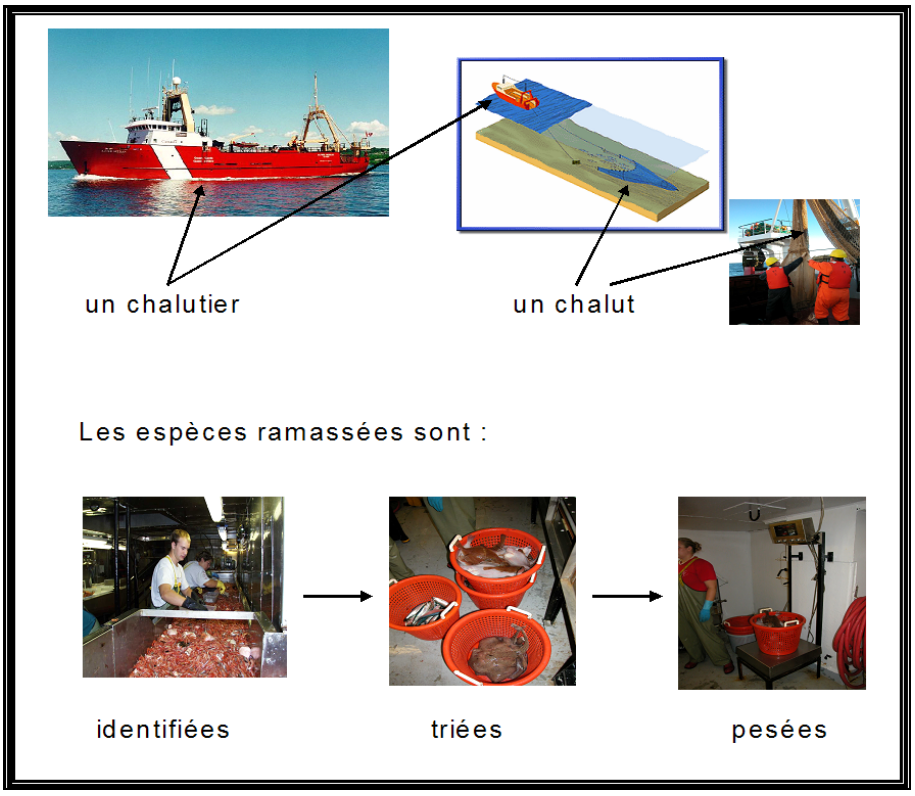


Figure 1.4 – Les relevés d'abondance du Centre des Pêches du Golfe (Images fournies par le Centre des Pêches du Golfe)

sure géographiquement proches. Dans cette optique, quelques modèles statistiques de référence, très bien formalisés, existent pour représenter des variables aléatoires spatialement structurées.

1.2.2 Quelques modèles statistiques spatiaux usuels

Dans cette sous-section, je propose un rapide tour d'horizon des modèles spatiaux univariés les plus utilisés. Pour plus de détails, je renvoie le lecteur vers l'ouvrage de Cressie (1993).

Les modèles spatiaux peuvent être classés en trois grandes catégories :

- Les modèles sur grille permettent de représenter la répartition d'une variable aléatoire dans un espace discrétisé. On suppose disposer d'une partition, régulière ou non, du domaine d'étude en petites zones (ou régions). L'intérêt se porte sur la modélisation des ressemblances (autocorrélation positive) ou dissemblances (autocorrélation négative) entre les valeurs de la variable en des zones proches (voisines).
- Les modèles géostatistiques décrivent un processus spatial continu sur un domaine d'étude. L'intérêt se porte sur la modélisation des covariations spatiales entre paires de lieux et de leurs transformations en fonction de la distance séparant ces lieux. Les questions de prédiction de la mesure en un point non observé et d'interpolation spatiale sont au coeur des méthodes usuelles de *krigeage*.
- Les processus ponctuels modélisent des données dont les localisations géographiques forment les quantités aléatoires d'intérêt. L'analyse porte sur l'absence éventuelle d'une répartition aléatoire de ces localisations et sur le caractère uniforme ou non

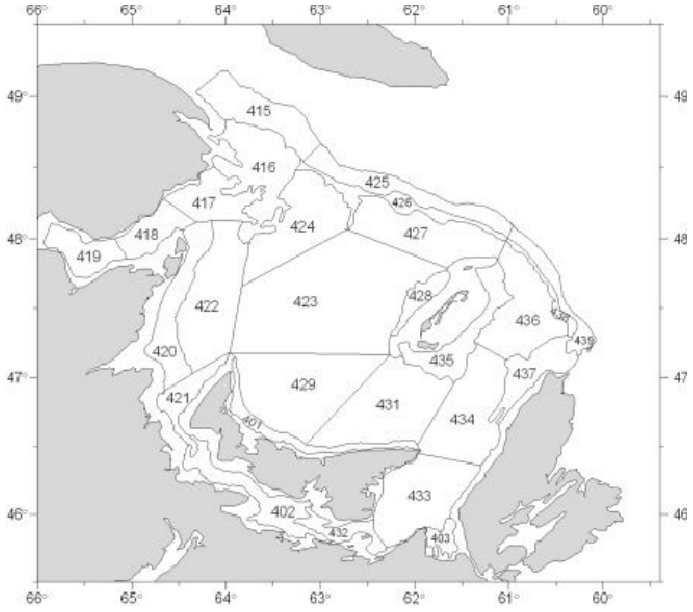


Figure 1.5 – Le domaine de suivi a été divisé en strates

d'une répartition aléatoire de ces points.

Ce travail de thèse a été principalement axé sur l'utilisation de modèles sur grille et de modèles géostatistiques.

Les champs aléatoires markoviens sont traditionnellement utilisés pour modéliser des dépendances spatiales entre sites (ou régions) géographiques lié(e)s par une structure de voisinage (Besag, 1974). Cette dernière est formalisée à l'aide d'une matrice de voisinage W définie par $\omega_{ij} = 1$ si les sites (ou régions) i et j sont géographiquement voisins et $\omega_{ij} = 0$ sinon. Un champ aléatoire $Y = \{Y_i, i = 1, 2, \ldots, I\}$ est dit markovien lorsque ses distributions conditionnelles complètes vérifient la propriété markovienne suivante :

$$[Y_i|Y_j, j \neq i] = [Y_i|Y_j, j \in \delta_i] \tag{1.2.1}$$

où $\delta_i = \{j; \omega_{ij} = 1\}$ désigne l'ensemble des voisins géographiques du site (ou région) i. La structure de voisinage permet de récupérer de nombreuses propriétés d'indépendance conditionnelle entre sites (régions) si bien que la conditionnelle complète associée à chaque Y_i dépend uniquement des voisins géographiques du site (ou région) i.

En pratique, dans le cas continu, les distributions conditionnelles complètes de l'équation 1.2.1 sont supposées gaussiennes :

$$[Y_i|Y_j, j \in \delta_i] \sim \mathcal{N}\left(\frac{\rho}{\omega_{i+}} \sum_{j \in \delta_i} Y_j, \frac{\sigma^2}{\omega_{i+}}\right) \tag{1.2.2}$$

où ω_{i+} désigne le nombre de voisins de i, σ^2 est le paramètre de variance locale du processus et ρ est un paramètre contrôlant la force de l'intéraction spatiale. Un tel choix s'explique principalement par la simplicité des modélisations qu'il engendre : les conditionnelles complètes ont l'avantage d'être totalement déterminées par leur moment d'ordre 1 et 2. Ce modèle gaussien appartient à la classe des modèles Autorégressifs Conditionnels ou CAR, acronyme de l'anglais *Conditional AutoRegressive model*. La loi conditionnelle de Y_i est une loi normale centrée en une certaine proportion de la moyenne des effets

strate échantillonnée	longitude	latitude	distance chalutée (en miles)	biomasse oursins (en kg)	biomasse anemones (en kg)	biomasse concombres (en kg)
429	-63.325	46.806	1.76	2.227	**0**	0.094
432	-63.153	45.938	1.68	**0**	**0**	**0**
414	-62.423	47.299	1.77	15.819	**0**	**0**
430	-60.302	47.217	1.86	**0**	**0**	**0**
426	-61.689	48.114	1.75	4	**0**	**0**
408	-63.082	46.738	1.75	8.03	0.08	**0**
412	-62.346	47.783	1.77	15.819	**0**	**0**
423	-63.733	47.6	1.76	0.572	**0**	**0**
430	-60.123	47.298	1.78	**0**	0.034	**0**
407	-62.153	48.037	1.74	17.098	**0**	0.156
425	-61.21	48.145	1.74	**0**	8.046	**0**
405	-64.3	47.528	1.8	4.919	**0**	0.258
408	-62.441	47.125	1.77	41.229	0.049	5.784
433	-61.413	46.368	1.8	8.079	**0**	4.103
435	-61.364	47.343	1.79	**0**	**0**	**0**
401	-62.49	46.522	1.72	3.561	**0**	**0**
411	-61.131	46.959	1.78	26.643	**0**	3.736
421	-64.36	46.907	1.71	**0**	0.056	0.563
416	-63.678	48.64	1.8	5.513	0.282	**0**
438	-60.952	47.943	1.78	0.983	**0**	**0**

Tableau 1.2 – Extrait du jeu de données du Centre des Pêches du Golfe pour la campagne de septembre 1999. Chaque ligne correspond à un trait de chalut. Parmi les données de biomasses mesurées, les valeurs nulles sont nombreuses.

de ses voisins géographiques et de variance inversement proportionnelle au nombre de ses voisins. La contrainte $\sum_i Y_i = 0$ assure l'identifiabilité du modèle. Par ailleurs, le théorème d'Hammersley-Clifford assure que les distributions conditionnelles complètes 1.2.2 définissent une loi jointe valide, analytiquement calculable grâce au lemme de Brook :

$$[y_1, y_2, \ldots, y_I|\sigma^2] \propto (\frac{1}{\sigma^2})^{-I/2} exp\left(-\frac{\sigma^2}{2} y^T \Sigma_\rho^{-1} y\right) \tag{1.2.3}$$

où :

$$\Sigma_\rho(i,j)^{-1} = \begin{cases} \omega_{i+} & \text{si i=j} \\ -\rho & \text{si i est voisin de j} \\ 0 & \text{sinon} \end{cases}$$

Soit H_ω la matrice diagonale définie par $(H_\omega)_{ii} = \omega_{i+}$ pour tout i=1,2,..,I. La contrainte $\rho \in [1/\lambda_1, 1/\lambda_n]$ où $\lambda_1 < \lambda_2 < \ldots < \lambda_n$ sont les valeurs propres ordonnées de $H_\omega^{-1/2} W H_\omega^{-1/2}$ est nécessaire pour que Σ_ρ^{-1} soit non singulière et, ainsi, que l'expression 1.2.3 définisse une loi propre (Banerjee & al., 2004).

Parmi les champs de Markov pour données catégorielles les plus utilisés se trouvent le modèle d'Ising dans le cas binaire (Cipra, 1987) et son extension pour les variables aléatoires à plus de deux catégories appelée modèle de Potts (Wu, 1982). A l'origine, ces modèles ont été développés en physique statistique pour modéliser différents phénomènes

dans lequels des effets collectifs sont produits par des intéractions locales entre particules à deux ou plusieurs états, comme le ferromagnétisme. Depuis, ces modèles sont fréquemment utilisés en analyse d'images (Geman & Geman, 1984) pour modéliser la corrélation entre couleurs de pixels voisins. Nous reviendrons plus en détails sur ces classes de modèles dans le chapitre 3.

D'autres modèles sur grille comme les modèles Autorégressifs simultanés (SAR) ou les modèles Moyennes Mobiles (MA) sont également particulièrement employés dans la littérature de statistique spatiale.

Les processus gaussiens sont les modèles les plus utilisés pour traiter des données géostatistiques définies sur un domaine d'étude D. Si $Y = \{Y_s, s \in D \subset \mathbb{R}^2\}$ désigne un processus gaussien continu alors, pour tout $n \geq 1$ et tout ensemble de sites $\{s_1, \ldots, s_n\}$:

$$(Y_{s_1}, Y_{s_2}, \ldots, Y_{s_n})^T \sim \mathcal{N}(\mu, \Sigma)$$

où $\mu = (\mu_{s_1}, \mu_{s_2}, \ldots, \mu_{s_n})^T$ et $\Sigma_{s_i,s_j} = c_{ij}$. La matrice de variance-covariance Σ doit être définie positive i.e., $x^T \Sigma x > 0$ pour tout vecteur x non nul. La spécification d'un modèle géostatistique repose essentiellement sur le choix d'une fonction de covariance c(.) valide. Celle-ci est généralement une fonction décroissante de la distance inter-sites : la corrélation entre sites géographiquement proches est supposée plus forte qu'entre des sites géographiquement éloignés. Les deux hypothèses de modélisation classiques sont la stationnarité et l'isotropie. Un processus est dit stationnaire d'ordre 1 si la moyenne du processus est constante sur D i.e., $\mu_s = \mu$ pour tout site s de D. Un processus stationnaire d'ordre 1 est dit stationnaire d'ordre 2 si la covariance du processus entre deux sites s_i et s_j dépend uniquement du vecteur liant ces deux sites :

$$cov(Y_{s_i}, Y_{s_j}) = c(s_i - s_j)$$

Un processus est dit isotropique si la covariance du processus entre deux sites dépend uniquement de la distance séparant ces deux sites i.e.,

$$cov(Y_{s_i}, Y_{s_j}) = c(|s_i - s_j|)$$

Sous l'hypothèse d'isotropie, la fonction c(.) est traditionnellement choisie parmi un ensemble de fonctions de covariance paramétriques de référence, appelées covariogrammes, qui assurent la définie-positivité de Σ. Quelques covariogrammes de référence sont donnés dans le tableau 1.3. Trois paramètres principaux caractérisent ces covariogrammes. τ^2 désigne l'effet pépite : il peut-être vu comme une variance résiduelle non-spatialisée intervenant au niveau de chaque observation. σ^2 désigne le palier, vu comme la variance de l'effet spatial. Enfin, la portée ϕ désigne la distance à partir de laquelle il n'y a presque plus de dépendance entre deux sites.

1.2.3 Une application directe de ces modèles n'est pas toujours possible en écologie

En écologie, lorsqu'on s'intéresse à de *vrais* problèmes pratiques, on s'éloigne très souvent du cas simple où les données peuvent être directement modélisées avec l'une des structures spatiales décrites dans la section précédente 1.2.2.

Dans le cas de données catégorielles, prenons l'exemple de la classification d'individus en populations génétiquement homogènes. Elle s'effectue à partir de données génotypiques

Modèle	Fonction de covariance c(h)
Exponentiel	$c(h) = \begin{cases} \sigma^2 exp(-\phi h) & \text{si } h > 0 \\ \tau^2 + \sigma^2 & \text{si h=0} \end{cases}$
Gaussien	$c(h) = \begin{cases} \sigma^2 exp(-\phi^2 h^2) & \text{si } h > 0 \\ \tau^2 + \sigma^2 & \text{si h=0} \end{cases}$
Exponentiel puissance $(0 < p \leq 2)$	$c(h) = \begin{cases} \sigma^2 exp(-\lvert\phi h\rvert^p) & \text{si } h > 0 \\ \tau^2 + \sigma^2 & \text{si h=0} \end{cases}$
Sphérique	$c(h) = \begin{cases} \sigma^2[1 - \frac{3}{2}\phi h + \frac{1}{2}(\phi h)^3] & \text{si } 0 < h \leq \frac{1}{\phi} \\ \tau^2 + \sigma^2 & \text{si h=0} \end{cases}$
Matérn	$c(h) = \begin{cases} \frac{\sigma^2}{2^{\nu-1}\Gamma(\nu)}(2\sqrt{\nu}h\phi)^\nu K_\nu(2\sqrt{\nu}h\phi) & \text{si } h > 0 \\ \tau^2 + \sigma^2 & \text{si h=0} \end{cases}$

Tableau 1.3 – Les fonctions de covariance (covariogrammes) de référence pour les modèles paramètriques isotropiques.

indiquant les 2 formes alléliques possédées par chaque individu échantillonné en L loci d'intérêt. Ces formes alléliques sont des variables aléatoires catégorielles prenant leurs valeurs entre 1 et le nombre de versions alléliques possibles pour chaque locus. Les données traitées sont multivariées à deux niveaux : l'information génétique de chaque individu est caractérisée par plusieurs loci et 2 allèles sont mesurés par locus. Aussi, l'existence de plusieurs niveaux de variabililité rend notamment impossible l'utilisation directe d'un unique modèle de Potts pour introduire des dépendances entre les génotypes des individus géographiquement proches.

Dans le cas de données continues, prenons l'exemple de la modélisation de la distribution des invertébrés marins du Golfe-du-Saint-Laurent. La figure 1.6 contient les histogrammes empiriques des quantités de biomasses collectées entre 1999 et 2001 par le Centre des Pêches du Golfe pour 4 espèces d'invertébrés particulières. Les distributions sont surdispersées et cette surdispersion se manifeste par une très forte occurrence de valeurs nulles. Ces zéros ne forment pas une part naturelle d'une même loi continue comme celle produite par les valeurs positives. Le recours à la loi normale ou lognormale est donc exclu pour modéliser de telles données dites *zero-inflated* (Heilbron, 1994). En effet, ces lois diffuses supposent que la probabilité d'occurrence d'un zéro est nulle. Ainsi, en particulier, les champs markoviens ou géostatistiques gaussiens classiques ne sont pas directement utilisables pour introduire des dépendances entre points de mesure géographiquement proches. Des structures aléatoires spécifiques doivent donc être définies pour générer des données continues *zero-inflated* et spatialisées.

1.3 Des phénomènes aléatoires complexes

La plupart des phénomènes aléatoires étudiés en écologie des populations sont complexes : ils sont la manifestation de l'effet combiné de multiples variables aléatoires observées ou non et inter-agissant à diverses échelles.

Revenons, pour exemple, aux deux problèmes de modélisation précédemment décrits. De multiples sources de variabilité influent sur le pattern génétique des organismes vivants. Ainsi, les données génétiques acquises en vue de la classification en populations génétiquement homogènes sont multivariées à deux niveaux. Tout d'abord, elles concernent

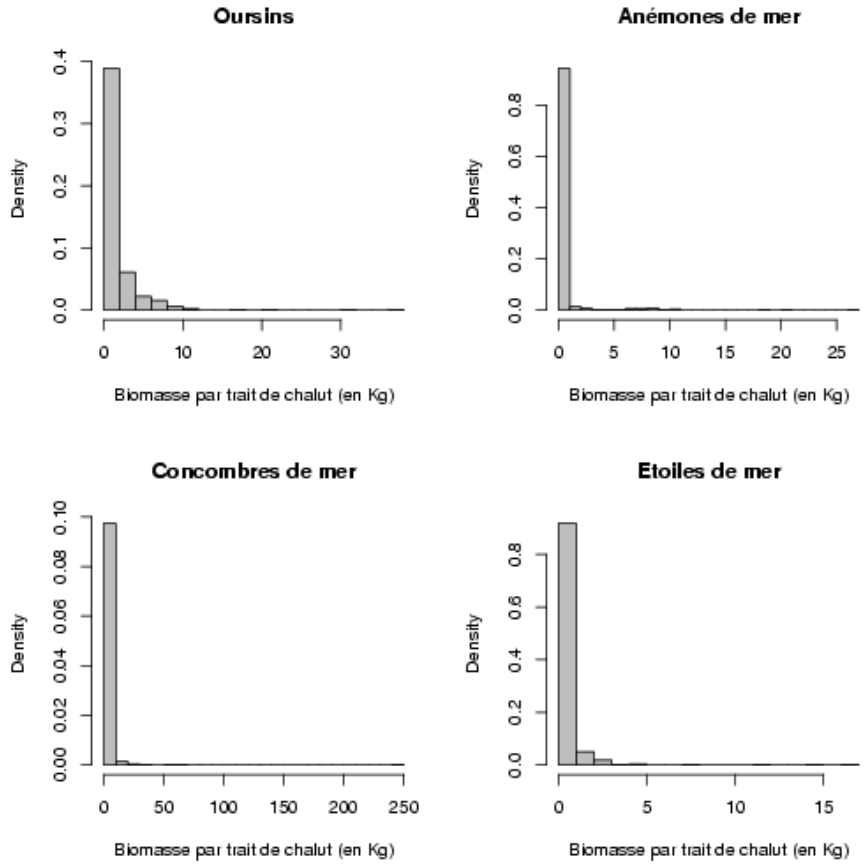


Figure 1.6 – Biomasses en oursins, anémones de mer, concombres de mer et étoiles de mer recueillies dans le Golfe-du-Saint-Laurent par le Centre des Pêches du Golfe en 1999,2000 et 2001

généralement une dizaine de loci. Les ours de Scandinavie, par exemple, ont été génotypés en 19 loci. De plus, comme les individus d'intérêt sont diploïdes, une grandeur bivariée est associée à chaque locus. Il faut donc être capable de spécifier un modèle bivarié pour chaque locus et un modèle multivariable pour traiter les différents loci. Ceci limite le recours aux méthodes statistiques existantes ou nécessite au minimum une adaptation des méthodes monovariables au cas multivariable. Par ailleurs, la localisation géographique des individus, les fréquences alléliques, les mécanismes génétiques de mutation, de sélection de gènes et de dérive génétique sont autant de phénomènes aléatoires observables ou non, susceptibles d'expliquer la variabilité génétique des individus dans un environnement donné.

Ecosystems are increasingly seen as the product of huge numbers of interacting forces
(Clark & Gelfand, 2006)

De multiples facteurs environnementaux (i.e., température, géologie, lumière, pH,...), biologiques (relations trophiques,...), physiques (courants marins, vents,...) et humains (modification ou destruction des habitats,...) affectent la répartition des organismes vivants dans leur environnement naturel. Ces facteurs se manifestent généralement à différentes échelles spatiales : l'échelle de l'observation, l'échelle d'une population d'individus, l'échelle d'un écosystème formé par une association ou communauté d'êtres vivants et son environnement naturel etc. Par ailleurs, plusieurs de ces facteurs covarient (e.g., température et profondeur,..).

Les données recueillies sont souvent multivariées : plusieurs observations sont réalisées en chaque site d'échantillonnage. Cela est notamment le cas, au cours de relevés d'abondance : les quantités de biomasse sont généralement mesurées pour plusieurs espèces d'intérêt en chaque site échantillonné. Par exemple, au cours des campagnes scientifiques du Centre des Pêches du Golfe, les quantités de biomasse en plus de 40 espèces d'invertébrés marins sont mesurées en chaque site chaluté. En théorie, on s'attend à ce qu'il existe des dépendances spatiales à la fois entre les mesures réalisées en un site particulier en raison des relations inter-espèces et entre les mesures réalisées en plusieurs sites en raison de l'existence de structure spatiale dans la répartition de chaque espèce.

La plupart des facteurs influant sur la variabilité génétique ou la répartition des organismes vivants ne sont pas directement mesurables (e.g., dérive génétique, mutation, appartenance à une population, erreurs de mesure, ..). Pourtant, en théorie, les écologues savent qu'ils jouent souvent un rôle non-négligeable sur le fonctionnement interne et caché du phénomène aléatoire étudié.

Much of nature is unmeasurable,unobservable or both (Clark, 2007)

Ainsi, la plupart des questions de modélisation liées à l'écologie des populations conduit à des problèmes de modélisation spatiale et/ou multivariée nécessitant de recourir à des structures aléatoires spécifiques.

1.4 Faiblesses des modèles statistiques usuels en écologie

1.4.1 Ils ne sont généralement pas spatialement explicites

Malgré l'existence de modèles spatiaux très bien formalisés (cf section 1.2.2), les modèles statistiques couramment utilisés en écologie des populations sont souvent basés sur une hypothèse d'indépendance statistique ou sur une paramétrisation très simpliste de la dépendance entre observations (McCarthy, 2007).

Ainsi, les principaux modèles statistiques utilisés jusqu'à présent pour classer des individus en populations génétiquement homogènes ne sont pas spatialement explicites (Pritchard & al., 2000), (Dawson & Belkhir, 2001), (Falush & al., 2003). Les coordonnées géographiques des individus échantillonnés ne sont utilisées qu'une fois la classification réalisée pour visualiser les frontières entre sous-populations. En ne tenant pas compte explicitement de la nature spatiale du problème de détection de discontinuités génétiques, ces modèles peuvent perdre en efficacité. En particulier, dans le cas où le nombre de loci ou d'individus échantillonnés est faible, on s'attend à ce que la robustesse des classifications diminue. Tirer profit de la localisation géographique des individus pour estimer leur pattern génétique peut permettre de renforcer les performances de classification dans de telles situations.

Les modèles Delta sont classiquement utilisés pour représenter des données de biomasse continues *zero-inflated* (Aitchison, 1955), (Stefansson, 1996). Ils sont définis par une probabilité discrète d'occurrence de zéros et une densité continue pour les valeurs strictement positives. Une de leurs principales faiblesses est de ne pas introduire de dépendances spatiales entre les quantités de biomasses mesurées en des sites géographiquement proches : ils supposent que les organismes vivants se répartissent aléatoirement et uniformément dans leur environnement naturel. Ainsi, ces modèles négligent l'hypothèse théorique selon laquelle la distribution des organismes vivants est spatialement structurée (cf section

1.2.1). Le chapitre 4 de ce document sera entièrement consacré à cette classe de modèles statistiques.

Comme vu dans la section 1.2.3, le type de données disponibles en écologie des populations rend souvent inappropriée l'utilisation directe de structures spatiales standards (cf section 1.2.2). Cela peut expliquer, en partie, pourquoi de nombreux modèles ne tiennent pas compte explicitement de la structure spatiale du phénomène aléatoire traité.

De manière générale, l'existence d'autocorrélation spatiale entre points de mesure crée de la redondance dans l'information disponible. Dans ce cas, le recours à des modèles basés sur une hypothèse d'indépendance des observations engendre une sous-estimation de la variabilité du processus et de possibles erreurs décisionnelles ou de prédiction (Cressie, 1993).

1.4.2 Ils sont souvent basés sur des hypothèses simplificatrices trop fortes

Dans sa boîte à outils, le statisticien dispose d'un certain nombre de lois de probabilités usuelles (e.g., gaussienne, gamma, beta, Poisson, géométrique). Elles correspondent à des phénomènes aléatoires simples : lancés de dés, jeu de pile ou face, comptage d'événements rares, erreurs de mesure. Malheureusement, un recours à de tels modèles s'avère souvent inadapté, soit parce qu'ils ne sont pas assez sophistiqués par rapport aux données disponibles (e.g., données de biomasse *zero-inflated*), soit parce qu'ils incitent à faire des hypothèses simplificatrices trop fortes par rapport à la complexité des phénomènes étudiés : indépendance, normalité, homoscédasticité, stationnarité etc... Le modèle n'est alors plus construit en fonction de sa pertinence écologique ou fonctionnelle mais en fonction de son adéquation avec l'outil statistique disponible (Clark & Gelfand, 2006).

Pour modéliser des phénomènes aléatoires complexes, le modélisateur adopte généralement un raisonnement global. Il cherche à doter directement les observations d'une loi jointe multivariée. Malheureusement, cette tâche s'avère souvent fastidieuse. En effet, la panoplie des modèles multivariés standard est pauvre si bien que le modélisateur n'y trouve pas toujours sont bonheur. Parmi les structures les plus utilisées, citons les modèles gaussiens multivariés, Wishart, multinomial et Dirichlet. Dès que les données disponibles obligent à sortir du cadre gaussien, la spécification de fonctions de covariances spatiales (ou spatio-temporelles) devient rapidement impossible. Par ailleurs, ces modèles multivariés ne sont pas suffisamment flexibles pour permettre de tenir compte de sources de variabilité multiples qui covarient et agissent à diverses échelles sur le phénomène aléatoire étudié. Ainsi, le modélisateur se heurte à des problèmes de modélisation souvent insolubles : Comment modéliser des relations de dépendances multiples et/ou non-linéaires ? Comment modéliser des phénomènes non stationnaires ? Comment tenir compte de multiples sources d'information ? La solution la plus simple est alors de contourner ces difficultés en ignorant les aspects multivariés ou en se ramenant à des hypothèses simples (e.g., gaussianité, linéarité, stationnarité). Mais la non-pertinence du modèle spécifié sera sans conteste un frein à son utilisation.

1.5 Objectifs et stratégie adoptée dans la thèse

Les deux problèmes de modélisation décrits dans la section 1.1 laissent apparaître deux enjeux actuels de la modélisation statistique en écologie des populations :

– Tenir compte de la nature spatiale des phénomènes écologiques étudiés.
– Proposer des modèles statistiques plus "réalistes" par rapport à la complexité des phénomènes écologiques étudiés.

Il ressort que les outils classiques de modélisation statistique montrent des limites pour faire face à ces enjeux (cf section 1.4). Dans ce contexte, la question de recherche méthodologique abordée dans cette thèse est la suivante :

Comment construire un cadre de modélisation statistique, flexible et transposable, permettant d'atteindre ces deux objectifs dans un même cadre cohérent ?

Il y a quelques années, une nouvelle démarche de modélisation statistique est née : la modélisation hiérarchique. Elle consiste à décomposer la loi jointe d'une collection de variables aléatoires en une série de modèles usuels simples puis à les combiner par conditionnements probabilistes successifs pour élaborer des structures fonctionnelles plus complexes. Malheureusement, l'essor de cette technique de modélisation a longtemps été freiné par l'absence de méthodes d'estimation adéquates permettant d'intégrer des centaines voire des milliers de quantités aléatoires inconnues. Récemment, le développement des algorithmes Monte Carlo par Chaines de Markov (MCMC), utiles aux Bayésiens, et de l'algorithme EM, utile aux classiques, a permis de lever en partie les difficultés d'estimation de ces modèles. Ainsi, la modélisation hiérarchique a commencé à faire son nid dans des disciplines variées comme l'épidémiologie (Green & Richardson, 2002), l'économie (Gelfand & al., 1998) ou encore les sciences environnementales (Wickle, 2003).

L'objectif principal de cette thèse est d'enrichir la panoplie des modèles statistiques actuellement employés en écologie des populations pour modéliser la structuration spatiale de phénomènes aléatoires complexes. En particulier, l'idée est de fournir des solutions "sur-mesure" à deux problèmes pratiques liés à l'écologie des populations : le problème de classification d'individus en populations génétiquement homogènes et le problème de modélisation de la répartition spatiale de la biomasse en invertébrés épibenthiques du Golfe-du-Saint-Laurent (cf section 1.1).

Les modèles hiérarchiques sont adaptés au traitement de phénomènes aléatoires complexes. La statistique spatiale propose un certain nombre de structures aléatoires possibles pour induire des dépendances spatiales entre points de mesure (cf section 1.2.2). Aussi, la stratégie adoptée dans ce travail est de combiner les performances respectives offertes par ces deux cadres de modélisation. Plus précisément, je me suis attachée à mettre en oeuvre différentes façons d'introduire de la structure spatiale à l'intérieur d'un schéma de construction hiérarchique.

Les objectifs annexes de cette thèse sont de proposer un modèle statistique alternatif aux modèles Delta pour le traitement de données continues *zero-inflated*, de réaliser des inférences statistiques en articulant des modèles d'observations complexes avec des modèles spatiaux et enfin, de mettre en évidence les avantages offerts par l'approche hiérarchique avec composante spatiale.

La modélisation hiérarchique est une approche prometteuse mais qui soulève de nombreux défis pour les statisticiens. Les différents problèmes de modélisation traités pendant cette thèse m'ont également permis d'aborder les problèmes d'inférence et de sélection de modèles liés à l'utilisation de structures hiérarchiques. Le problème de la sélection de modèles sera abordé dans le chapitre 6.

Tous les modèles hiérarchiques spatiaux construits au cours de cette thèse ont été inférés dans le cadre statistique bayésien via des algorithmes MCMC. Ce document est volontairement centré sur la présentation de structures hiérarchiques originales dont la

couche phénoménologique latente est spatialement structurée. Pour cette raison, j'ai volontairement placé en annexes tous les aspects méthodologiques utilisés ou développés au cours de cette thèse : point sur l'inférence bayésienne, détails sur l'inférence des modèles hiérarchiques proposés, facteur de Bayes, etc...

J'ai souhaité écrire un document qui s'adresse tout d'abord à un public de statisticiens. Celui-ci pourra être intéressé par les problèmes de modélisation traités et par les structures hiérarchiques développées. J'adresse également ce document à un public d'écologues qui pourra trouver des solutions opérationnelles aux deux problèmes écologiques traités au cours de cette thèse.

Avant d'aborder de front les deux problèmes de modélisation de la section 1.1, je propose de décrire, dans le chapitre 2, le cadre de modélisation statistique hiérarchique dans lequel tous les modèles construits au cours de cette thèse ont été développés. Le plan général de ce document, qui donne l'intérêt spécifique de chaque chapitre en fonction des structures hiérarchiques spatialisées qui y sont proposées, est donné à la fin du chapitre 2.

Chapitre 2

La modélisation statistique hiérarchique

Les deux questions de modélisation, décrites dans le chapitre précédent, sont deux situations typiques pour lesquelles les modèles statistiques usuels ne sont pas adaptés. Des structures probabilistes plus riches sont nécessaires pour être pertinentes, notamment aux yeux des écologues, face à la complexité des phénomènes étudiés. Comme vu dans le chapitre précédent, la modélisation statistique hiérarchique apparaît comme une voie d'approche prometteuse pour enrichir la panoplie des structures stochastiques et ainsi s'accommoder de la complexité de certains phénomènes aléatoires. Dans ce chapitre, je commence par décrire, à partir d'exemples simples, le principe général de la modélisation statistique hiérarchique qui constitue le noyau thématique central de ce document. Puis, je présente la spécificité des structures hiérarchiques construites au cours de ce travail de thèse.

2.1 La vision conditionnelle ou comment pousser l'interprétation au-delà de la vraisemblance

En modélisation statistique, la manipulation de structures construites par conditionnements probabilistes est fréquente. Toutefois, à partir du moment où la vraisemblance associée est calculable analytiquement (i.e., la probabilité des observations sachant les paramètres inconnus), il n'est plus jugé utile de faire apparaître leur structure hiérarchique. La difficulté de spécification d'une telle structure est aussi parfois un handicap. Voici deux exemples simples à travers lesquels le lecteur pourra mieux percevoir ce qu'est un modèle hiérarchique.

2.1.1 Exemple 1 : les modèles de mélange ZIP et ZINB

Comme énoncé dans le chapitre précédent, les données écologiques issues de relevés d'abondance (i.e., comptages, biomasses) sont souvent caractérisées par une très forte occurrence de zéros. Les données de comptage d'oiseaux étudiées par Martin et al. (2005) en sont un exemple typique. La figure 2.1 contient la répartition empirique des comptages, réalisés en 96 sites d'échantillonnage, pour une espèce d'oiseau particulière (sp. *Malurus cyaneus*) vivant dans les régions boisées d'Australie sub-tropicale. Comme pour les

données de biomasse en invertébrés épibenthiques du Golfe-du-Saint-Laurent décrites dans le chapitre 1, l'histogramme obtenu est surdispersé et très piqué en zéros. Nous sommes en présence de données discrètes *zero-inflated*.

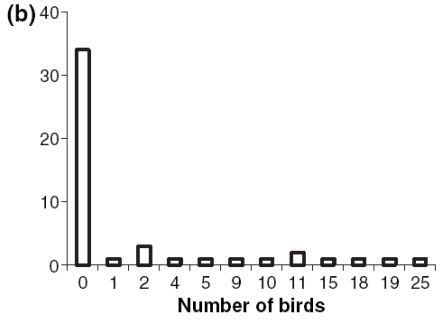


Figure 2.1 – Un exemple de distribution empirique *zero-inflated* discrète (extrait de l'article de Martin et al. (2005)). Données de comptage d'oiseaux (*Malurus cyaneus*) échantillonnés en Australie sub-tropicale. On observe une forte occurrence de zéros.

Pour traiter de telles données, la piste la plus intuitive est de recourir aux lois de probabilités classiques pour le comptage d'événements rares : lois de Poisson ou binomiale négative. Malheureusement, ces modèles ne sont généralement pas adaptés. Prenons l'exemple des données de comptage d'oiseaux de la figure 2.1. Une loi de Poisson, ajustée par maximum de vraisemblance ($\widehat{p} = 0.001$), sous-estime nettement la probabilité d'occurrence d'un zéro ($p_{obs} = 0.375$). Il en est de même pour un ajustement basé sur une loi binomiale négative.

Comme approche alternative, le statisticien a souvent recours aux modèles de mélange basés sur une combinaison linéaire pondérée de deux lois standards : une Dirac en zéro, notée δ_0, et une loi discrète, notée $\mathcal{L}(\lambda)$ (Poisson ou binomiale négative). Le nombre d'oiseaux observés au site i (i=1,2,...,n) est représenté par une variable aléatoire Y_i dont la distribution de probabilité s'écrit sous la forme :

$$[Y_i = y_i|\lambda, \pi] = \pi\delta_0(y_i) + (1-\pi)\mathcal{L}(\lambda)(y_i) \tag{2.1.1}$$

π et $1-\pi$ sont appelés les poids du mélange et sont à valeurs dans $[0,1]$. λ désigne le(s) paramètre(s) caractérisant la loi paramétrique $\mathcal{L}$. Ce modèle de mélange est usuellement appelé modèle ZIP (*Zero-Inflated Poisson*) ou modèle ZINB (*Zero-Inflated Negative-Binomial*) lorsque la loi discrète $\mathcal{L}$ est respectivement une loi de Poisson ou une loi Binomiale Negative (Johnson & Kotz, 1969). Ces modèles de mélange sont des structures mathématiques adaptées pour *gonfler* la probabilité d'occurrence de la valeur nulle. Ainsi, la probabilité d'obtenir zéro sous le modèle ZIP est supérieure à celle donnée par une loi de Poisson de paramètre $\lambda > 0$:

$$[Y_i = 0|\lambda, \pi] = \pi + (1-\pi)e^{-\lambda} \geq e^{-\lambda} \qquad \forall\lambda > 0 \qquad \forall\pi \in [0,1] \tag{2.1.2}$$

Les modèles de mélange ZIP et ZINB peuvent aussi être interprétés avec un raisonnement conditionnel. D'un point de vue écologique, les zéros obtenus lors du processus

d'échantillonnage d'organismes vivants peuvent être classés en deux catégories (Martin & al., 2005) :

- les *vrais* zéros issus d'un échantillonnage dans une zone où l'espèce cible est absente ou très rare (i.e., présente selon une très faible densité). Une des causes est que le type d'habitat présent en ce site est non ou très peu propice à la survie de l'espèce.
- les *faux* zéros considérés comme des erreurs d'observation dûes au protocole expérimental : faible niveau de capturabilité et/ou effet d'échappement de l'espèce cible et/ou espèce absente au moment de l'échantillonnage.

Les zéros observés proviennent de deux sources de zéros qu'il convient de modéliser. La sur-abondance de valeurs nulles observée dans la figure 2.1 peut être attribuée à une occurrence importante de *vrais* zéros car l'espèce d'oiseau étudiée est rare. Aussi, les *vrais* zéros s'obtiennent avec probabilité 1 et sont issus d'une Dirac en zéro, notée δ_0. Quant aux *faux* zéros, ils sont issus d'un tirage selon une loi de Poisson (ou Binomiale négative). Celle-ci permet également de quantifier l'incertitude sur les comptages strictement positifs.

Soit π la probabilité d'absence de l'espèce en un site donné, uniquement liée à l'occurrence de *vrais* zéros. Soit c_i la variable de classe, à valeurs dans $\{0, 1\}$, indiquant si l'espèce est présente ou absente au site i. Conditionnellement au paramètre π, c_i suit une loi de Bernoulli de paramètre π :

$$[c_i = 0|\pi] = \pi$$
$$[c_i = 1|\pi] = 1 - \pi$$

Chaque variable de classe c_i contrôle l'origine distributionnelle de l'observation y_i (i=1,2,..,n). Si c_i est nulle, l'espèce est absente au site i : Y_i suit une loi de Dirac en zéro. Si c_i vaut un, Y_i suit la loi $\mathcal{L}$. L'espèce est présente en ce site. L'occurrence de zéros est possible mais il s'agit uniquement de *faux* zéros. Les relations de conditionnement entre les variables latentes c_i et les observables Y_i s'écrivent :

$$[Y_i = 0|c_i = 0, \lambda] = 1$$
$$[Y_i = y|c_i = 1, \lambda] = \mathcal{L}(\lambda)(y) \qquad \forall y \geq 0$$

Dans une vision conditionnelle, les variables de classe c_i ont le statut de variables aléatoires latentes (encore appelé *états cachés*). Elles sont conditionnées par le paramètre π et conditionnent l'occurrence des comptages y_i. Ces variables ne sont pas directement observables mais elles peuvent s'avérer utiles pour décrire et formaliser le phénomène à modéliser. Leur explicitation permet notamment de donner une interprétation écologique à la forte occurrence des *zéros* observée dans le cas de relevés d'abondance et de distinguer deux catégories de valeurs nulles.

Le modélisateur peut considérer les modèles de mélange ZIP ou ZINB sous deux angles de vue différents. Les deux structures probabilistes associées conduisent à une même expression de la vraisemblance. En effet, une intégration sur les valeurs des variables c_i permet de retrouver la formule intégrée 2.1.1 d'un modèle de mélange ZIP ou ZINB :

$$\begin{aligned}[Y_i = y|\lambda, \pi] &= [Y_i = y|c_i = 0, \lambda][c_i = 0|\pi] + [Y_i = y|c_i = 1, \lambda][c_i = 1|\pi] \\ &= \pi\delta_0(y) + (1 - \pi)\mathcal{L}(\lambda)(y)\end{aligned}$$

Dans une vision intégrée, le modèle lie directement la variable aléatoire observable Y_i à un vecteur de paramètres inconnus : $\theta = (\lambda, \pi)$. Dans une vision conditionnelle, les variables de classe c_i apparaîssent dans l'explicitation du modèle et jouent un rôle pivot pour la description et la formalisation du phénomène aléatoire étudié.

2.1.2 Exemple 2 : les modèles à espaces d'états

Jonsen et al. (2003) ont étudié la trajectoire suivie par une tortue luthe (sp. *Dermochelys coriacea*). Il s'agit de la plus grande espèce de tortue vivant dans les océans tropicaux et sub-tropicaux (cf figure 2.2). Comment modéliser une telle trajectoire ?

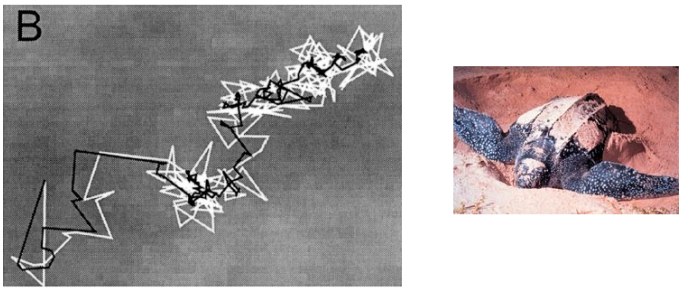


Figure 2.2 – (A gauche) Simulation d'une trajectoire suivie par une tortue luthe (extrait de l'article de Jonsen et al. (2003)). En blanc, trajectoire observée et en noir, trajectoire *réelle* (i.e., débruitée) (A droite) Une tortue luthe. (Image extraite de la page web :www.cybelle-planete.org/newsletter-7.html)

Soit Y_t le processus temporel des positions géographiques observées de l'animal ($t \in T$). Y_t peut être un vecteur bi-dimensionnel composé des coordonnées spatiales (e.g., longitude, latitude) du site occupé par l'animal à l'instant t.

Supposons que les observations soient erronnées i.e., que la trajectoire observée ne soit qu'une réalisation bruitée de la trajectoire réellement suivie par l'animal (cf figure 2.2). Cette hypothèse est l'étape de base à un raisonnement conditionnel. Elle suppose l'existence d'un processus temporel non observé α_t qui décrit la *vraie* trajectoire suivie par l'animal. La première étape est de modéliser cette trajectoire. Supposons l'indice de temps t discret. Une marche aléatoire gaussienne est une structure aléatoire simple mais plausible pour modéliser les α_t :

$$\alpha_{t+1} = \alpha_t + \eta_t \qquad \text{avec} \qquad \eta_t \overset{i.i.d}{\sim} \mathcal{N}(0, \sigma_\eta^2) \tag{2.1.3}$$

Ce modèle dynamique suppose que la position de l'animal à l'instant t+1 dépend uniquement de sa position à l'instant t. Les innovations η_t permettent de tenir compte des incertitudes liées aux déplacements *réels* de la tortue. Les variables η_t et α_t sont supposées indépendantes.

Une structure de dépendance conditionnelle doit ensuite être spécifiée pour relier la *vraie* position α_t et la position observée y_t. Une solution simple est de supposer que la *vraie* trajectoire est bruitée par un bruit blanc gaussien de moyenne 0 et de variance σ_ϵ^2 :

$$Y_t = \alpha_t + \epsilon_t \qquad \text{avec} \qquad \epsilon_t \overset{i.i.d}{\sim} \mathcal{N}(0, \sigma_\epsilon^2) \tag{2.1.4}$$

Les variables ϵ_t et α_t sont supposées indépendantes.

La structure probabiliste décrite par les équations 2.1.3 et 2.1.4 est appelée modèle à espace d'états (Jonsen & al., 2003). Elle est basée sur des relations de dépendances conditionnelles probabilistes. Les observations Y_t dépendent de quantités aléatoires non observées α_t qui elles-même dépendent d'un paramètre inconnu σ_η. Les variables α_t ont le statut de variables aléatoires latentes.

Le modèle à espace d'états décrit permet de modéliser la trajectoire de la tortue luthe en tenant compte des incertitudes liées aux déplacements *réels* de la tortue et celles liées aux erreurs d'observation (et autres sources de variabilité naturelle). Chaque modèle élémentaire apporte sa pierre à l'édifice pour construire une structure complexe : la marche aléatoire gaussienne (cf eq. 2.1.3) modélise la première source d'incertitudes (i.e., bruit d'état) puis le modèle linéaire gaussien (cf eq. 2.1.4) modélise la seconde (i.e., bruit d'observations).

2.2 Tous les modèles hiérarchiques sont identiquement structurés

2.2.1 Les trois ingrédients d'un modèle hiérarchique

Comme les modèles statistiques élémentaires, une structure hiérarchique est composée de grandeurs observables et de paramètres. Par la suite, nous désignerons les variables observables par la lettre majuscule Y. Elles apportent une information partielle par rapport au phénomène étudié. Dans l'exemple 1, $Y = (Y_1, Y_2, \ldots, Y_n)$ est un vecteur de variables aléatoires discrètes dont les réalisations observées sont des comptages d'oiseaux. Dans l'exemple 2, $Y = (Y_t)_{t\in T}$ est un processus temporel discret à valeurs réelles dont une réalisation est l'ensemble des positions géographiques successivement occupées par la tortue luthe.

Par défaut, nous désignerons les paramètres inconnus par la lettre grecque θ. Ces derniers décrivent un état récapitulatif des propriétés du phénomène étudié. Dans l'exemple 1, $\theta = (\pi, \lambda)$, où π est la probabilité d'absence en oiseaux en un site et λ le nombre moyen d'oiseaux observés quand l'espèce est présente. Dans l'exemple 2, $\theta = (\sigma_\epsilon, \sigma_\eta)$ où σ_ϵ indique la variabilité de l'erreur d'observation autour de la *vraie* trajectoire suivie par la tortue et σ_η la variabilité naturelle liée au déplacement de la tortue dans l'espace.

Les modèles hiérarchiques ont pour spécificité d'être basés sur des variables aléatoires non observables dites latentes (ou cachées). Par la suite, elles seront désignées par la lettre majuscule Z. Ces variables définissent généralement un mécanisme aléatoire non observable. Dans l'exemple 1, $Z = (c_1, c_2, \ldots, c_n)$ est un vecteur de labels de classes qui permettent de distinguer les *vrais* des *faux* zéros. Dans l'exemple 2, $Z = (\alpha_t)_{t\in T}$ est un processus temporel discret à valeurs réelles qui formalise la trajectoire *réelle* suivie par la tortue luthe.

2.2.2 Une même démarche de construction brique par brique

La démarche de modélisation hiérarchique consiste à décomposer un phénomène aléatoire complexe en processus aléatoires élémentaires qui peuvent être modélisés avec des structures probabilistes standards. Cette idée n'est pas neuve. Elle reprend les idées de la méthode cartésienne :

Diviser chacune des difficultés que j'examinerais, en autant de parcelles qu'il se pourrait et qu'il serait requis pour les mieux résoudre. Descartes (1637) Discours de la méthode. Pour bien conduire sa raison et chercher la vérité dans les sciences.

En pratique, cette décomposition s'appuie sur *une* unique formule probabiliste : la formule des probabilités composées. Celle-ci permet d'écrire la distribution jointe d'un vecteur aléatoire $(U_1, U_2, \ldots, U_n)$ comme un produit de modèles conditionnels :

$$[U_1, U_2, \ldots, U_n] = [U_1][U_2|U_1][U_3|U_1, U_2]...[U_n|U_1, U_2, \ldots, U_{n-1}] \tag{2.2.5}$$

Ainsi, la formule des probabilités composées permet d'écrire la distribution jointe de grandeurs observables et de variables aléatoires latentes, notées $[Y, Z|\theta]$, sous la forme :

$$[Y, Z|\theta] = [Y|Z, \theta][Z|\theta] \tag{2.2.6}$$

où

- Le modèle $[Y|Z, \theta]$, dit *modèle des observations*, décrit l'occurrence des données conditionnellement aux variables latentes et à certains paramètres.
- Le modèle $[Z|\theta]$, dit *modèle du processus interne*, décrit l'occurrence des variables latentes conditionnellement aux paramètres.

Comme un jeu de "LEGO", la démarche de construction hiérarchique consiste ensuite à emboîter judicieusement ces différents modèles élémentaires par conditionnements probabilistes successifs pour élaborer des structures fonctionnelles plus complexes.

Chaque source d'incertitude est décrite dans une couche différente du modèle hiérarchique (cf figure 2.3). Au premier niveau de la hiérarchie se situe le modèle des observations $[Y|Z, \theta]$. Il permet de modéliser la variabilité naturelle de l'observation et de tenir compte des erreurs de mesure. Au second niveau est décrit, par l'intermédiaire de variables latentes, comment fonctionne le processus interne, conditionnellement à d'autres paramètres, $[Z|\theta]$. La variabilité du processus non observé ainsi que les erreurs de modélisation sont prises en compte à ce niveau. On l'appelle la couche phénoménologique.

2.2.3 Un même graphe orienté acyclique pour tous les modèles hiérarchiques

Dans une structure hiérarchique, les variables observables, les variables latentes et les paramètres entretiennent des relations de dépendances conditionnelles toujours orientées dans le même sens. Les paramètres sont toujours des grandeurs conditionnantes. Certains conditionnent l'occurrence des variables latentes, d'autres celle des observations. Dans l'exemple 1, le paramètre π influent sur les variables de classe c_i et le paramètre λ conditionne directement les observations. Dans l'exemple 2, σ_η contrôle le degré de variabilité des *vraies* coordonnées spatiales α_t pendant que σ_ϵ contrôle celui des observations. Les observations dépendent conditionnellement des variables latentes et de certains paramètres. Ainsi, les variables latentes jouent un double rôle : celui de variables conditionnantes et de variables conditionnées.

La modélisation graphique permet de visualiser ces relations de dépendances conditionnelles. En particulier, les graphes acycliques orientés (ou DAG pour Direct Acyclic Graph selon la terminologie de Spiegelhalter et al. (1996)) sont traditionnellement utilisés dans le cas de modèles hiérarchiques (Parent & Bernier, 2007). Ils sont constitués de noeuds

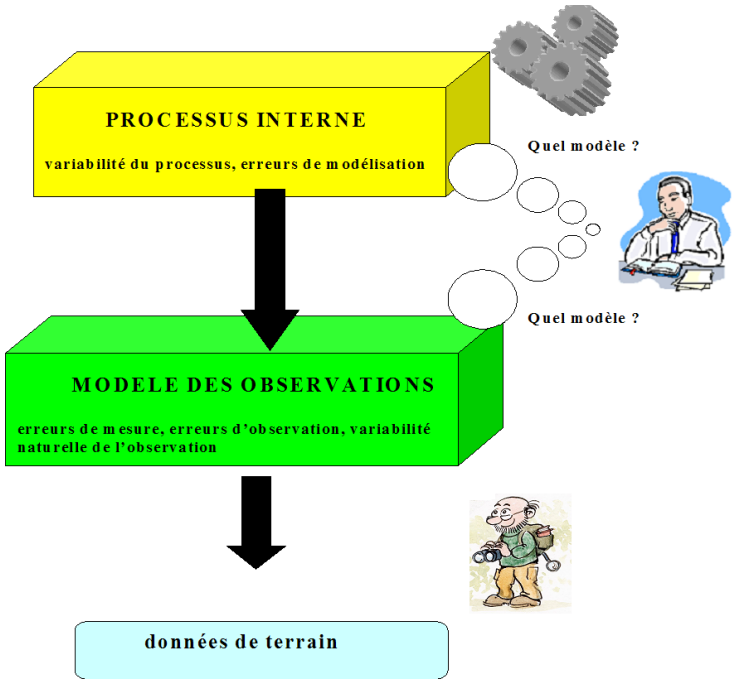


Figure 2.3 – La modélisation statistique hiérarchique : une démarche de construction brique par brique

ν désignant soit une variable observable, soit une variable latente soit un paramètre. Les relations de dépendance sont représentées par des flèches orientées qui partent des grandeurs conditionnantes (i.e., les parents $pa(\nu)$) et pointent vers les grandeurs conditionnées (i.e., les enfants). Les noeuds initiaux qui n'ont pas de parents sont les paramètres : ils ne dépendent d'aucune autre variable aléatoire. Les noeuds terminaux sans enfants, encore appelés feuilles, sont les observables. Enfin, tous les noeuds qui ne sont ni des paramètres ni des observables désignent des variables aléatoires latentes. Les graphiques sont représentés avec les conventions suivantes. Les noeuds aléatoires sont représentés sous la forme d'ellipses tandis que les variables observables apparaissent sous la forme de rectangles. Les flèches en trait plein indiquent les relations stochastiques conditionnelles alors que les doubles flèches indiquent les relations logiques déterministes.

La figure 2.4 contient les DAG des 2 modèles hiérarchiques décrits dans les exemples du paragraphe précédent. Bien qu'ils aient leurs caractéristiques propres, il est clair que ces 2 modèles possèdent une structure générale commune. Ceci est généralisable. Tout modèle hiérarchique est construit à partir de relations de dépendances conditionnelles qu'il est possible de décrire simplement sous la forme d'un même DAG, celui de la figure 2.5.

Les DAG donnent tout son sens à la notion de hiérarchie. Un modèle hiérarchique se présente comme un empilement par niveaux successifs de différents *sous-modèles* d'où l'appellation fréquente de modèle mixte à niveaux ou modèle à couches.

Par ailleurs, les DAG permettent de visualiser l'étroite solidarité entretenue par la démarche de modélisation hiérarchique et la démarche d'inférence (cf figure 2.6). La démarche de modélisation consiste à parcourir le DAG dans le sens des flèches qui partent des paramètres θ, traversent l'espace des variables latentes Z pour aboutir au niveau des données Y. L'inférence statistique parcourt le DAG dans le sens inverse. Il s'agit de partir des données et de faire remonter l'information vers les paramètres par l'intermédiaire des variables latentes.

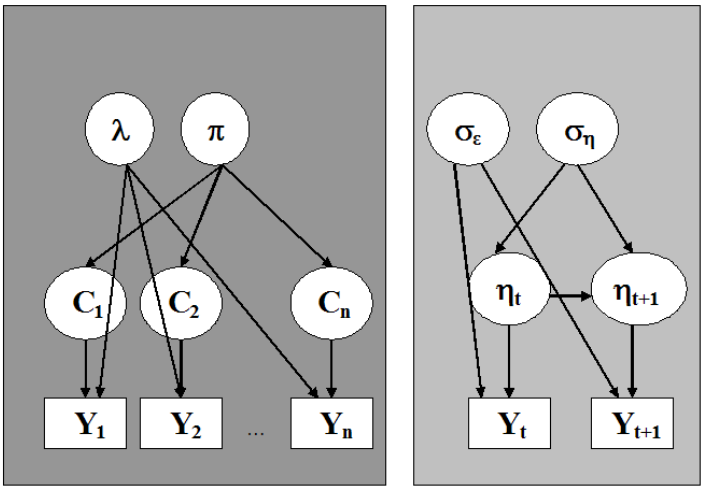


Figure 2.4 – Graphes acycliques orientés des deux modèles définis en exemple. De gauche à droite : (1) Modèle de mélange pour données discrètes avec surabondance de zéros (2) Modèle à espace d'états. Les variables observables sont représentées par un carré. Les variables latentes et les paramètres sont représentés par des cercles. Les relations de dépendances conditionnelles entre noeuds sont représentées avec des arcs orientés.

Enfin, les DAG montrent que chaque noeud unidimensionnel d'un modèle hiérarchique est entièrement défini grâce à la connaissance de ses parents et enfants directs. En d'autres termes, conditionnellement à ses parents et enfants directs, chaque noeud est indépendant de tous les autres noeuds. C'est la propriété d'*indépendance conditionnelle*.

2.2.4 Les variables latentes jouent un rôle pivot

Dans une structure hiérarchique, les variables latentes sont l'outil de soutien au raisonnement conditionnel : elles font le pont entre les paramètres et les variables observables. Elles jouent un double rôle : tantôt elles jouent le rôle de paramètres quand elles conditionnent la naissance d'observations, tantôt elles jouent le rôle d'un résultat potentiellement observable quand elle sont générées dans les couches internes de la structure du modèle.

Quand le statisticien veut donner à son modèle une interprétation conceptuelle, il doit avant tout faire apparaître des variables latentes pour décrire et formaliser le phénomène aléatoire étudié. En effet, bien qu'elles n'aient de sens que par rapport au modèle et à sa finalité, les variables latentes aident à renforcer le sens conceptuel d'un modèle. Ainsi, les variables de classe c_i de l'exemple 1 permettent de rendre compte de l'existence conceptuelle de deux sources de valeurs nulles dans le cas de données de comptage d'espèces : les *vrais* et les *faux* zéros. Une telle démarche améliore souvent la compréhension du modélisateur et le dialogue avec le non-statisticien.

Enfin, nous verrons dans les chapitres suivants que les variables latentes permettent de

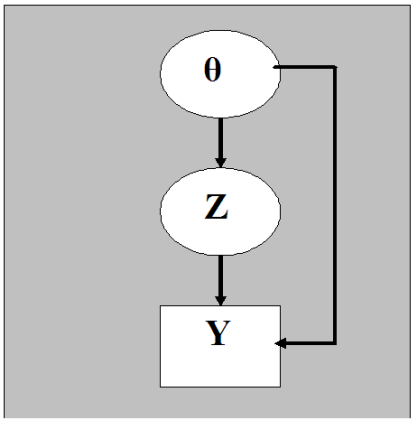


Figure 2.5 – Un même DAG pour une infinités de modèles hiérarchiques. Les variables latentes Z conditionnent l'occurrence des observables Y et sont conditionnées par les paramètres θ.

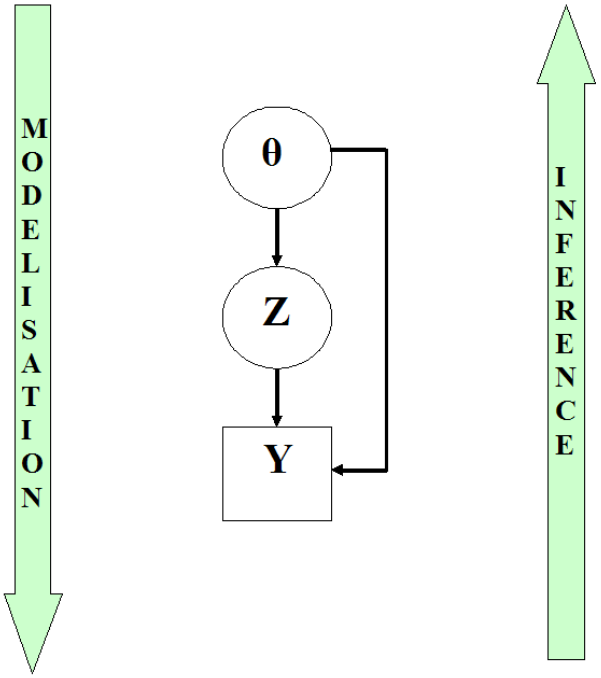


Figure 2.6 – Le conditionnement probabiliste : direct pour la modélisation, inverse pour l'inférence

construire des modèles beaucoup plus riches (au prix d'une vraisemblance non explicite) et évitent ainsi les hypothèses de commodité (normalité, stationnarité..) qui président à l'adoption d'un modèle ad hoc.

2.2.5 La loi *a priori* : une nouvelle brique de la construction hiérarchique

Sous le paradigme bayésien, les paramètres sont considérés comme des variables aléatoires au même titre que les grandeurs observables. Au-delà de la spécification du modèle des observations et du modèle du processus interne, la construction d'un modèle hiérarchique nécessite donc d'assigner une loi *a priori* sur les paramètres inconnus θ. L'appellation *a priori* exprime le fait que cette distribution de probabilité a été établie préalablement à l'observation des données y. Elle donne un niveau de crédibilité aux différentes valeurs possibles de θ. Elle peut-être établie sur la base d'autres données similaires ou refléter l'avis d'experts avec des techniques dite d'élicitation de lois *a priori* (Parent & Bernier, 2007).

L'assignation d'une loi *a priori* sur les paramètres θ s'intègre naturellement dans la démarche de construction hiérarchique. Elle ajoute une "brique" supplémentaire à l'échafaudage hiérarchique déjà constitué du modèle des observations et du modèle du processus interne (cf figure 2.7). L'association du modèle d'occurrence, du modèle du processus interne et de la loi *a priori*, encore appelé modèle d'expertise, permet d'intégrer dans une même structure nos connaissances sur toutes les grandeurs du phénomène en présence (observables ou non). La formule des probabilités composées permet d'écrire la distribution jointe des grandeurs observables, des variables latentes et des paramètres sous la forme :

$$[Y, Z, \theta] = [Y|Z, \theta][Z|\theta][\theta] \tag{2.2.7}$$

L'approche bayésienne abolit la différence de nature mathématique entre les paramètres et les variables latentes. Leur distinction se fait uniquement en fonction de leur place dans la structure hiérarchique. Dans un graphe acyclique orienté, les paramètres sont les noeuds sans parents (i.e., pas conditionnés par une variable aléatoire) alors que les variables latentes sont les noeuds qui possèdent des parents et des enfants (cf figure 2.7).

2.3 D'une structure commune à la spécificité des structures hiérarchiques

Les modèles simples présentés jusqu'à présent sont loin d'illustrer toutes les possibilités offertes par la modélisation hiérarchique. En particulier, aucun d'entre eux n'illustre la possibilité de lier des variables latentes avec des relations de dépendances spatiales.

Pendant mon travail de thèse, je me suis attachée à mettre en oeuvre différentes façons d'introduire de la structure spatiale à l'intérieur du schéma de construction hiérarchique de la figure 2.5. Les structures de dépendances spatiales ne sont pas spécifiées directement sur les observations. Elles sont introduites dans la couche phénoménologique par l'intermédiaire de variables aléatoires latentes. Celles-ci sont généralement supposées gaussiennes ce qui permet de recourir aux modèles spatiaux traditionnels. Dans les modèles à espace d'états par exemple, la trajectoire *réelle* est formalisée dans la couche latente

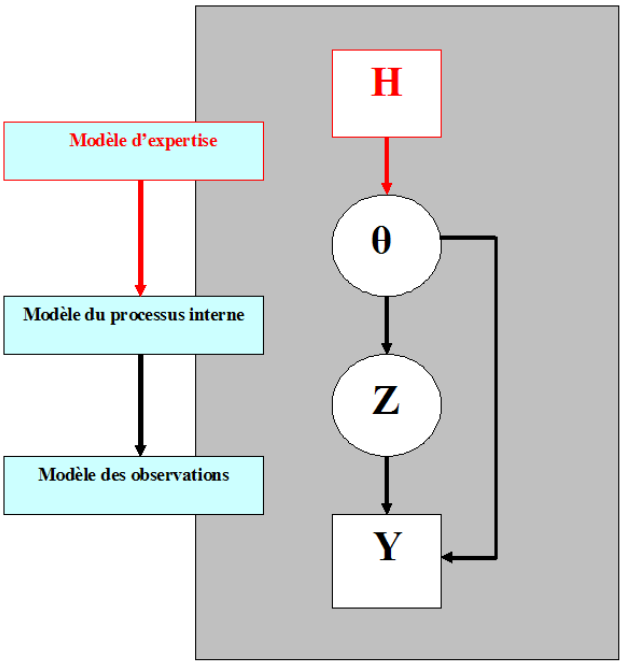


Figure 2.7 – La loi *a priori* : une nouvelle brique de la structure hiérarchique

avec un modèle dynamique gaussien. Conditionnellement aux variables latentes Z, les observations sont supposées indépendantes. Cela permet d'utiliser un modèle d'observations simple mais de récupérer des liens de dépendance entre observations par propagation des incertitudes d'une couche à l'autre du modèle.

Même si la construction d'un modèle hiérarchique s'effectue toujours selon une même démarche, les classes de sous-modèles conditionnels choisies à chaque étape de la construction sont spécifiques. Nous allons voir dans les chapitres à venir que ces choix sont guidés par le problème de modélisation rencontré et par le type des observations disponibles et des dépendances spatiales que le modélisateur cherche à formaliser.

Un des objectifs de cette thèse est de fournir au lecteur un catalogue de structures hiérarchiques spécifiques (θ, Z, Y), répertoriées en fonction du type des données disponibles, des variables latentes et des dépendances spatiales à prendre en compte (cf figure 2.8). Le problème de la classification d'individus en populations génétiquement homogènes ainsi que le problème de modélisation de la biomasse en invertébrés épibenthiques du Golfe-du-Saint-Laurent ont servi de base à la construction de ce tableau.

A l'exception du chapitre 4, chaque chapitre de ce document s'articule autour de la description d'un ou de plusieurs modèles hiérarchiques possibles permettant de répondre aux deux problèmes de modélisation décrits dans le chapitre précédent. Tous les modèles construits ont un point commun : le sous-modèle du processus interne $[Z|\theta]$ est spatial c'est-à-dire que la force de liaison entre variables latentes dépend d'un éloignement géographique. Tous les modèles construits selon le schéma 2.5 ont aussi leurs propres spécificités vis-à-vis du type de variables rencontrées.

Le chapitre 3 présente un modèle hiérarchique bayésien pour la classification d'individus en populations génétiquement homogènes, à partir de données génotypiques multilocus géoréférencées. L'homogénéité génétique des individus est prise en compte dans le modèle des observations et les dépendances spatiales dans la couche cachée du modèle. Les va-

	Y	Z	θ
Chapitre 3	catégorielles bivariées coordonnées spatiales	C= variables de classe à valeurs dans {1,K} avec K≥1 indicées sur une grille irrégulière	f= fréquences d'occurrence d'une observation dans chaque groupe ψ= paramètre de dispersion qui module le degré d'organisation spatiale φ= paramètre de corrélation des observations
	cas d'étude: données génotypiques multilocus	sens conceptuel: populations d'appartenance de chaque individu	Processus latent [c|ψ] = Champs de Markov caché
Chapitre 5	continues positives surabondance de 0	1/ N= variables discrètes positives, nombre de points d'un semi de points 2/ X= variables continues positives, marques d'un processus ponctuel	μ= intensité d'un processus de Poisson (μ >0) ρ= intensité de la marque (ρ>0)
	cas d'étude: Biomasses issues de relevés au chalut de fond dans le Sud du Golfe-du-Saint-Laurent	sens conceptuel: 1/ nombre de *gisements* de biomasse ramassés 2/ quantités de biomasse dans un *gisement*	Processus latent [N,X|μ,ρ]=[N|μ][X|ρ] = Processus de Poisson homogène
Chapitre 6	continues positives surabondance de 0 coordonnées spatiales Domaine d'échantillonnage découpé en m régions	1/ N et X comme dans Chapitre 5 2/ $\mu=(\mu_1,\mu_2,..,\mu_m)$: intensités de m processus de Poisson homogènes indicées sur une grille irrégulière 3/ $\rho=(\rho_1,\rho_2,...,\rho_m)$: intensités de la marque de m processus de Poisson homogènes indicées sur une grille irrégulière	θ = paramètres qui contrôlent le degré de cohérence entre les régions homogènes
	cas d'étude: Biomasses issues de relevés au chalut de fond réalisés dans 38 strates dans le Sud du Golfe-du-Saint-Laurent	sens conceptuel: 1/ idem que Chapitre 5 2/ Nombre moyen de gisements de biomasse ramassés dans la région i 3/ Quantité moyenne de biomasse dans un *gisement* de la région i	2 structures latentes sur grille: régionale et CAR gaussien [N,X,μ,ρ |θ]=[N|μ][X|ρ][μ,ρ|θ]
Chapitre 7	comptages bivariées coordonnées spatiales	$\{Z_1(s),Z_2(s), s\in D\}$=un champs spatial bivarié variables continues indiçage continu dans IR^2	$X_1(j)$= variables continues indicées au m nœuds d'une grille régulière $X_2(j)$= variables continues indicées au m nœuds d'une grille régulière $\varsigma\in[-1,1]$ caractérise la force de liaison entre les champs Z_1 et Z_2 θ_1,θ_2 = paramètres de cohérence spatiale
	cas d'étude: Comptages d'espèces issus de relevés au chalut de fond dans le Sud du Golfe-du-Saint-Laurent	1/ $Z_1(s)$=nombres moyens d'organismes de l'espèce 1 présents au site s (à une transformation près) 2/ $Z_2(s)$=nombres moyens d'organismes de l'espèce 2 présents au site s (à une transformation près)	Processus latent $[Z_1,Z_2\|X_1,X_2,\varsigma,\theta_1,\theta_2]$ = Champs gaussien bivarié construit par convolution de fonctions noyaux sur grille régulière

Figure 2.8 – Spécificités des modèles hiérarchiques construits au cours de la thèse

riables latentes Z sont des variables de classe qui affectent chaque individu à un groupe donné. Dans l'exemple 2 du modèle de mélange, les variables latentes étaient également des variables de classe mais nous étions dans un cas simple : celles-ci, à valeurs dans $\{1, 2\}$, suivaient un modèle de Bernoulli. Dans ce chapitre, nous traiterons le cas plus complexe où les variables de classe sont à valeurs dans $\{1, K\}$ avec $K \geq 1$ et liées par des relations de dépendances spatiales. L'originalité du modèle repose principalement sur l'introduction d'un champ de Markov pour données catégorielles appelé modèle de Potts dans la couche latente. Une alternative à la loi multinômiale sera également proposée pour modéliser des corrélations entre données catégorielles bivariées. Dans ce chapitre, le statisticien bénéficiera d'un aperçu des avantages et des difficultés posés par l'introduction d'une structure markovienne de type Potts dans un schéma hiérarchique. L'écologue, quant à lui, trouvera une approche possible pour modéliser des phénomènes de consanguinité ou encore pour créer des configurations de groupes plus ou moins organisées spatialement. Nous appliquerons ce modèle de classification pour décrire la structure génétique de l'ours brun de Scandinavie. Nous l'illustrerons également avec le jeu de données humain HGDP-CEPH .

Dans le chapitre 4, je décris et mets en évidence les principales faiblesses des modèles Δ traditionnellement utilisés pour modéliser des données continues *zero-inflated*.

Dans le chapitre 5, je propose d'adopter une vision conditionnelle probabiliste pour décrire le processus d'échantillonnage de données de biomasses caractérisées par une surabondance de zéros. Nous supposons qu'une espèce se répartit dans son environnement selon un processus ponctuel marqué. Le semi de points associé est assimilé à des *gisements* de biomasse et les marques du processus à la quantité de biomasse dans chaque *gisement*. L'originalité du modèle hiérarchique construit repose sur cette hypothèse purement conceptuelle. La couche latente est construite autour d'un processus de Poisson homogène marqué. Contrairement aux structures hiérarchiques présentées jusqu'à présent, les variables latentes sont ici de deux natures différentes et indépendantes. Il y a des variables de comptage N qui indiquent le nombre de points (ou *gisements*) du processus ponctuel et des variables latentes continues et positives X qui représentent les marques associées. Deux paramètres composent ce modèle : l'intensité du processus de Poisson μ et l'intensité de la marque ρ. Nous sommes dans une situation où la couche latente se décompose sous la forme :

$$[Z|\theta] = [N, X|\mu, \rho] = [N|\mu][X|\rho]$$

Dans ce chapitre, une structure probabiliste peu connue pour modéliser des données caractérisées par une surabondance de zéros sera présentée au statisticien. Les avantages de la vision conditionnelle pour la modélisation de processus d'échantillonnage seront également mis en avant. De son coté, l'écologue trouvera une approche nouvelle pour modéliser des données de relevés d'abondance qui, comme nous le verrons, a l'avantage de rester cohérente par rapport à l'effort d'échantillonnage. Ce modèle sera appliqué aux données de biomasses en invertébrés épibentiques du Golfe-du-Saint-Laurent issues de relevés au chalut de fond.

Dans le chapitre 6, je reprends le modèle du chapitre précédent auquel j'ajoute une couche latente supplémentaire qui introduit des relations de dépendances spatiales entre les composantes d'un vecteur latent. Ainsi, le modèle présenté dans ce chapitre a la particularité d'avoir une couche phénoménologique cachée elle-même décomposée en deux

sous-modèles sous la forme :

$$[Z|\theta] = [N, X, \mu, \rho|\theta_1, \theta_2] = [N|\mu][X|\rho][\rho|\theta_1][\mu|\theta_2]$$

avec $Z = (N, X, \mu, \rho)$ et $\theta = (\theta_1, \theta_2)$. Nous retrouvons la couche latente du modèle défini dans le chapitre 5 : $[N|\mu][X|\rho]$ auquel est ajoutée une deuxième couche latente $[\rho|\theta_1][\mu|\theta_2]$. L'objectif de la modélisation est le même que dans le chapitre précédent : décrire le processus d'échantillonnage de données de biomasses caractérisées par une surabondance de zéros. Cependant, je considère désormais le cas où le domaine d'échantillonnage est découpé en m régions homogènes. Dans chaque région i, la biomasse se répartit selon un processus de Poisson homogène latent d'intensité μ_i et dont la marque est contrôlée par la grandeur aléatoire ρ_i (i=1,2,...,m). L'idée est de modéliser l'existence d'une possible cohérence comportementale entre ces régions afin de tirer profit de l'information disponible dans toutes les régions. Les paramètres μ et ρ définis dans le chapitre précédent sont désormais des vecteurs latents : $\mu = (\mu_1, ..., \mu_m)$ et $\rho = (\rho_1, \rho_2, ..., \rho_m)$. A noter que les grandeurs aléatoires N et X du chapitre 5 restent des variables latentes. Contrairement au chapitre 4 dans lequel les variables latentes sont de type catégorielles, les μ_i et les ρ_i sont des variables latentes continues à valeurs strictement positives. Ainsi, je proposerai deux structures sur grille possibles pour données continues à valeurs strictement positives. La première repose sur une distribution dite régionale qui introduit une cohérence entre plusieurs variables latentes sans notion de distances géographiques. La seconde repose sur un champ de Markov gaussien de type CAR (Conditional AutoRegressive) afin de modéliser des dépendances spatiales entre variables latentes. L'une des originalités de ce modèle est d'intégrer un modèle CAR gaussien dans la couche latente plutôt que dans celle des observations, comme cela est souvent le cas. Ce modèle sera de nouveau appliqué aux données de biomasse en invertébrés marins du Golfe du Saint-Laurent.

Dans le chapitre 7, je montre que la modélisation hiérarchique est une approche prometteuse pour la représentation de structures spatiales bivariées. En particulier, je me suis intéressée à la construction d'un modèle hiérarchique qui permette de décrire la répartition spatiale conjointe de deux espèces. Ce travail a consisté à définir le pattern spatial propre de chaque espèce et à intégrer des liens de dépendance entre ces deux répartitions. Les grandeurs observables considérées sont des comptages bivariés indicés en n sites géoréférencés. Contrairement aux chapitres précédents dans lesquels chaque modèle intègre une unique structure spatiale latente sur grille, le modèle hiérarchique défini dans ce chapitre intègre deux structures spatiales latentes variant continûment dans $\mathbb{R}^2$ et liées par des relations de dépendance. Inspiré des idées récentes de Barry et Ver Hoef (1996), ce modèle a la particularité d'être basé sur un champ gaussien bivarié latent $Z = \{(Z_1(s), Z_2(s)), s \in D \subset \mathbb{R}^2\}$ construit par approximation de la convolution de fonctions noyaux sur une grille discrète latente. La structure de chaque champ Z_1 et Z_2 est contrôlée par un ensemble de m variables non observées $X_1(j)$ et $X_2(j)$ (j=1,2,...m), indicées aux m noeuds d'une même grille régulière. La convolution confère une cohérence spatiale à chaque champ Z_1 et Z_2 tandis qu'un même jeu de variables latentes (e.g., $X_1(j)$) contrôle simultanément les deux champs générant ainsi des liens de dépendances. Contrairement aux méthodes usuelles dans lesquelles le modélisateur cherche à spécifier directement une structure spatiale bivariée au niveau des observations, l'originalité est d'intégrer cette structure dans la couche phénoménologique et d'expliciter par conditionnement les liens de dépendance entre champs. La couche latente se décompose ainsi :

$$[Z|\theta] = [Z_1, Z_2|X_1, X_2, \theta_1, \theta_2, \zeta] = [Z_1|X_1, \theta_1][Z_2|X_1, X_2, \theta_2, \zeta]$$

avec θ_1 et θ_2 les paramètres respectivement associés aux champs Z_1 et Z_2 et ζ un paramètre gérant la force de liaison entre ces deux champs. Ce travail n'a pas encore été conduit à terme : l'implémentation du modèle est en cours. De ce fait, aucune simulation ni analyse sur données réelles n'a pu être réalisée. Cependant, le statisticien trouvera dans ce chapitre la description d'une approche constructive pour la spécification de fonctions de covariance croisée ainsi qu'une réponse aux problèmes d'inférence liés à la grande dimension des modèles bivariés usuels. Quant à l'écologue, il s'enrichira d'une structure originale et flexible pour la modélisation souvent difficile de la répartition spatiale conjointe de deux espèces.

Deuxième partie

Détecter et localiser des lignes de discontinuités génétiques

Chapitre 3

La modélisation hiérarchique spatiale pour la classification d'individus en populations génétiquement homogènes

Actuellement, le modèle hiérarchique le plus utilisé pour classer des individus en populations génétiquement homogènes est celui de Pritchard et al. (2000), implémenté sous le logiciel Structure. Ce modèle a pour principale faiblesse de négliger l'hypothèse selon laquelle des similarités génétiques peuvent exister entre individus géographiquement proches (cf section 1.2.1). Or, la prise en compte de ces dépendances spatiales peut permettre d'améliorer les performances de classification en tirant profit des dépendances génotypiques entre individus géographiquement proches. Un tel transfert d'informations peut être d'autant bénéfique que les données génotypiques disponibles sont peu abondantes et/ou peu informatives. Aussi, Guillot et al. (2005) ont récemment proposé une extension spatialement explicite du modèle Structure. Le modèle hiérarchique developpé, nommé Geneland, tient compte de la structure spatiale de la diversité génétique dans sa couche phénoménologique latente. Il suppose que les différentes sous-populations occupent des territoires géographiquement disjoints, séparés par des frontières rectilignes simples qui délimitent une forte et brutale discontinuité des patterns génétiques.

Chez certaines espèces, la diversité génétique peut posséder un pattern spatial plus complexe que celui défini par le modèle Geneland : les dissimilarités génétiques entre individus croient continûment en fonction de leur éloignement géographique. En génétique des populations, une telle structure spatiale est appelée : une cline. Dans ce cas, la détection de populations aux patterns génétiques distincts est plus difficile car seules de petites discontinuités existent entre les patterns génétiques observés. Nous proposons un modèle hiérarchique spatial adapté à des situations pour lesquelles la diversité génétique est continûment structurée spatialement. L'originalité de ce modèle, baptisé Geneclust, repose sur l'introduction, dans la couche latente, d'un champ de Markov pour données catégorielles. Une alternative à la loi de Hardy-Weinberg est également proposée pour modéliser l'occurrence des génotypes dans les populations où se produisent des unions consanguines.

Dans ce chapitre je décris puis compare les trois structures hiérarchiques Structure, Geneland et Geneclust. Plutôt que de présenter un à un ces différents modèles, j'ai choisi

d'adopter une démarche de dé-construction hiérarchique. L'objectif est double. Il s'agit d'alléger la description des modèles qui possèdent des "briques" élémentaires communes. Par ailleurs, il s'agit d'illustrer, à partir d'un problème plus complexe que ceux décrits dans le chapitre 2, la démarche de modélisation hiérarchique ainsi que les avantages d'une telle construction.

Ce chapitre s'articule en 4 parties. La première partie donne les éléments de formalisation statistique communs aux trois modèles. Dans les deux parties suivantes sont présentées, pour chaque niveau respectif de la modélisation (processus interne et modèle des observations), les différentes structures probabilistes proposées dans les modèles Structure, Geneland et Geneclust en indiquant les hypothèses biologiques liées à chacune d'entres elles. Enfin, dans la dernière partie, j'invite le lecteur à lire les sections de l'article paru dans Genetics, dans lesquelles sont décrites des simulations comparatives entre les trois modèles ainsi que les analyses menées à partir du modèle Geneclust sur le jeu de données des ours bruns de Scandinavie et le jeu de données humaines HGDP-CEPH (cf section 1.1.1) .

3.1 Vers une démarche de modélisation hiérarchique...

3.1.1 Notations

Soit $y = (y_1, y_2, ..., y_n)$ le vecteur des génotypes de n individus diploïdes observés en L loci (cf section 1.1.1). Chaque composante y_i est un vecteur composé d'une collection de L paires d'allèles $y_i^l = (\alpha_i^l, \beta_i^l)$ pour l=1,2,..L. y_i^l désigne le génotype de l'individu i au locus l, composé de deux versions alléliques. J_l dénote le nombre d'allèles possibles au locus l. Enfin, $s = (s_1, s_2, ..., s_n)$ désigne le vecteur des coordonnées spatiales bi-dimensionnelles des n individus échantillonnés.

3.1.2 Explicitation des variables aléatoires latentes

Comme vu dans le chapitre 2, adopter une démarche de modélisation hiérarchique suppose d'expliciter des variables latentes qui vont conditionner l'occurrence des variables observables et dépendre de paramètres inconnus.

Les variables observables sont les vecteurs bi-dimensionnels des génotypes de chaque individu en L loci. Notons Y_i le vecteur des génotypes de l'individu i et $Y_{i,l}$ son génotype au locus l. Ce vecteur catégoriel prend ses valeurs dans $\{1, J_l\} \times \{1, J_l\}$.

Nous souhaitons regrouper les individus en populations homogènes sans connaître a priori le nombre $K \geq 1$ de ces populations. Nous sommes donc dans une situation de classification non supervisée : K est un paramètre inconnu.

Une population est un ensemble d'individus génétiquement similaires. Cette similarité génétique est contrôlée par un ensemble de paramètres inconnus : les fréquences alléliques.

Pour chaque population k, nous noterons $f_{k,l,.}$ le vecteur de dimension J_l des fréquences d'occurrence de chacun des J_l allèles possibles au locus l :

$$f_{k,l,.} = \{f_{k,l,1}, f_{k,l,2}, ..., f_{k,l,J_l}\} \qquad \text{avec} \qquad \sum_{j=1}^{J_l} f_{k,l,j} = 1$$

Par ailleurs, nous noterons f l'ensemble des fréquences alléliques inconnues $\{f_{k,l,j}; k = 1, 2, .., K, l = 1, 2, ..., L, j = 1, 2, ..., J_l\}$.

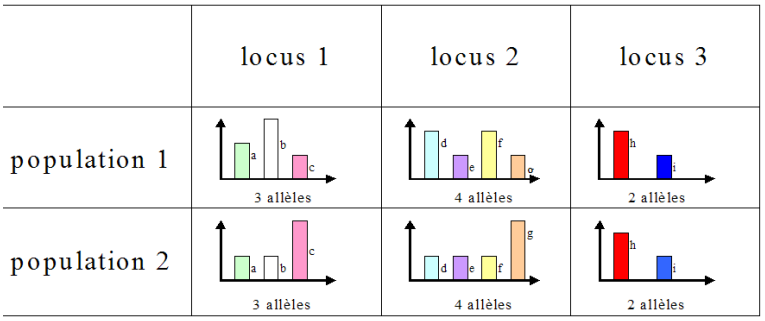


Figure 3.1 – Les génotypes de chaque population sont régis par un jeu de fréquences alléliques. Exemple pour deux populations et 3 loci d'intérêt. Pour chaque population k et chaque locus l, chaque allèle j est présente avec une fréquence $f_{k,l,j}$

La figure 3.1 donne une interprétation visuelle de ce qu'est une fréquence allélique. Elle représente l'histogramme des fréquences alléliques, supposées connues ici, pour 3 loci dans 2 populations distinctes. Au locus 1, 3 allèles (a,b,c) sont possibles. L'allèle b est plus fréquent dans la population 1 que dans la population 2 et inversement pour l'allèle c. Le locus 3 pour lequel deux allèles (h,i) sont possibles illustre le cas où les patterns génétiques sont les mêmes dans chaque population.

Nous supposons que chaque individu i provient d'une population unique, notée c_i. Cette grandeur inconnue présente un intérêt pour le biologiste qui cherche à identifier des populations génétiquement homogènes en vue de la localisation de lignes de discontinuités génétiques. Cette variable catégorielle, à valeurs dans $\{1, 2, ..K\}$, aura, dans les modèles présentés, le statut de variable aléatoire latente. L'ensemble des labels de population $c = \{c_i, i = 1, 2, .., n\}$ est appelé configuration de classes.

3.2 Comment modéliser la structure spatiale de la diversité génétique ?

Intéressons-nous tout d'abord à la couche cachée $Z|\theta$ (cf chapitre 2 section 2.2.2) du modèle hiérarchique que nous souhaitons construire pour classer des individus en populations homogènes. Cette couche phénoménologique doit permettre de modéliser la structure spatiale de la diversité génétique c'est-à-dire, plus simplement, l'organisation spatiale des populations. Le rôle des variables latentes Z est tenu par les variables de classe c_i (i=1,2,..,n) indiquant la population d'origine de chaque individu i.

Dans la littérature biologique sont cités cinq patterns spatiaux possibles de la diversité génétique : les clines, l'isolation par la distance, les barrières génétiques aux flux de gènes, les métapopulations et les patterns aléatoires (*Manel &al.*,2003).

Dans cette partie, je décris comment trois de ces situations biologiques (i.e., patterns aléatoires, barrières génétiques aux flux de gènes et clines) peuvent être formalisées en terme de classes latentes $c = \{c_i, i = 1, 2, .., n\}$ des n individus génotypés.

3.2.1 Le cas simple du pattern génétique spatial aléatoire

Pritchard et al. (2000) ont considéré la situation biologique du pattern spatial aléatoire : les individus se répartissent indépendamment dans l'espace et la localisation géographique d'un individu apporte aucune information sur sa population d'appartenance.

Formellement parlant, cela revient à supposer que, conditionnellement au nombre de populations K, les variables de classe latentes c_i sont indépendantes et uniformément distribuées selon une loi uniforme sur $\{1, K\}$:

$$[c_i = k|K] = \frac{1}{K} \qquad \forall i \in \{1, 2, ..n\}$$

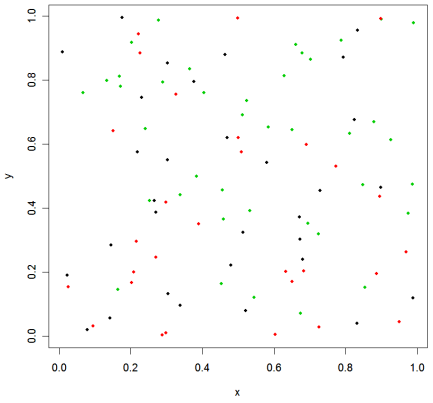


Figure 3.2 – Un exemple de répartition aléatoire de 100 individus issus de 3 populations. Chaque couleur (rouge, noir, vert) désigne une population.

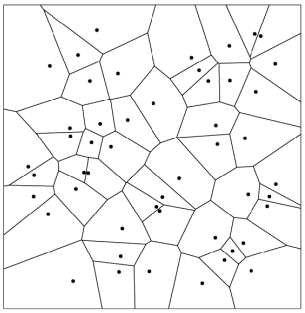
Cette hypothèse est simple et peu coûteuse en paramètres. Par contre, son utilisation nécessite que les populations aient été échantillonnées équitablement. Par ailleurs, elle est inadaptée dans les situations, fréquentes en pratique, où la diversité génétique est structurée spatialement.

3.2.2 Les tessellations de Voronoï permettent de représenter des populations isolées par barrières aux flux de gènes

En 2005, Guillot et al. ont proposé un modèle statistique permettant de traiter les situations biologiques dans lesquelles les populations sont isolées par des barrières écologiques et/ou humaines limitant les mouvements des individus (rivière, montagne, forêt..). Ces barrières limitent les flux de gènes et créent ainsi des frontières imperméables qui empêchent les mélanges entre individus de populations distinctes.

Pour caricaturer le plus simplement possible cette situation, ils supposent que les populations animales (ou végétales) occupent des territoires issus de l'union de polygonaux disjoints. Les frontières entre populations ou lignes de discontinuités génétiques ont alors une forme simple de type ligne brisée. Mathématiquement parlant, ils supposent que les K populations inconnues se répartissent dans des sous-domaines disjoints Δ_1,Δ_2,...,Δ_K formant une partition du domaine d'échantillonnage Δ :

$$\Delta = \Delta_1 \cup \Delta_2 \cup ... \cup \Delta_K \qquad \text{et} \qquad \Delta_k \cap \Delta_l = \emptyset \qquad \text{pour} \qquad k \neq l$$



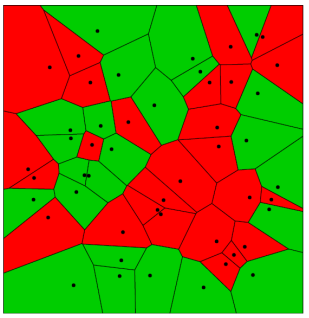


Figure 3.3 – Gauche : Réalisation d'un processus de Poisson et tessellation de Voronoï induite Droite : Tessellation de Voronoï colorée : partition obtenue après coloration des cellules de Voronoï en fonction de la marque de leur noyau. Ici K=2.

Chaque Δ_k peut avoir une forme plus ou moins complexe. Une approche pertinente pour couvrir un maximum de situations est de considérer chaque territoire comme une union de polygones convexes.

Construire la tessellation de Voronoï (ou tessellation de Dirichlet) associée à un ensemble de points du plan est une méthode usuelle pour obtenir une partition d'un domaine géographique D en sous-domaines polygonaux convexes. Etant donné un ensemble de points $S = (u_1, u_2, ..., u_m)$ du plan, la tessellation de Voronoï associée est la partition du plan formée par les polygones V_j (j=1,2,...,m) dont les points sont plus proches de u_j que des autres points de S :

$$V_j = \{s \in D; dist(s, u_j) \leq dist(s, u_g) \qquad \forall g = 1, 2, ..., m\}$$

Chaque polygone V_j s'appelle une cellule de Voronoï et les centres respectifs de ces cellules, notés u_j, sont appelés noyaux. La partition de D s'obtient en traçant les médiatrices de toutes les paires de points de S (cf figure 3.3).

Dans notre problème, le nombre et la localisation spatiale des noyaux $(u_1, u_2, ..., u_m)$ des polygones sont des grandeurs inconnues. Guillot et al.. supposent que ces noyaux forment un semi de points issus d'un processus de Poisson homogène latent. Ainsi, le nombre m de cellules de Voronoï induites par ces noyaux est une quantité aléatoire issue d'une loi de Poisson de paramètre λ.

$$[m|\lambda] = \frac{e^{-\lambda}\lambda^m}{m!}$$

Quant aux noyaux $(u_1, u_2, ..., u_m)$, ils sont aléatoirement répartis dans le domaine d'échantillonnage Δ selon une loi uniforme :

$$(u_1, u_2, .., u_m) \sim^{i.i.d} \mathcal{U}niforme(\Delta)$$

Chaque cellule de Voronoï contient des individus issus d'une même population inconnue. A chaque noyau u_j (j=1,2,..,m) est ainsi associée une marque catégorielle latente $c(u_j)$ à valeurs dans $\{1, 2, .., K\}$. L'ensemble $(u_j, c(u_j))$ forme un processus ponctuel marqué. Guillot et al. supposent que les marques $c(u_j)$ sont indépendantes et identiquement distribuées selon une loi uniforme sur $\{1, 2, ..., K\}$:

$$[c(u_j) = k|K] = \frac{1}{K} \qquad \forall j \in \{1, 2, ..m\}$$

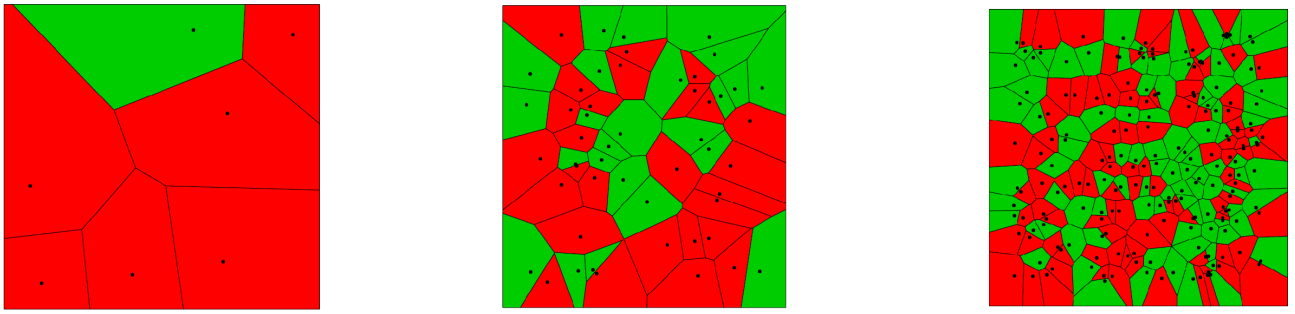


Figure 3.4 – Le degré d'organisation spatiale des populations dépend du paramètre λ. Tessellations de Voronoi colorées pour K=2 avec de gauche à droite λ=10, 50, 200

Le territoire Δ_k occupé par la population k est supposé issu de l'union de cellules de Voronoï induites par un processus ponctuel latent et affectées à cette même population k :

$$\Delta_k = \bigcup_{c(u_j)=k} V_j$$

Dans ce modèle, les cellules de Voronoï n'ont pas d'interprétation biologique. Elles sont juste un outil mathématiques commode pour créer des configurations de classes spatialement structurées avec des populations délimitées par des frontières rectilignes simples. Ainsi, deux individus i et i' qui appartiennent à une même cellule de Voronoï ont une probabilité plus élevée d'appartenir à la population k que deux individus appartenant à des cellules différentes :

$$[c_i = k \cap c_{i'} = k|K] = [c(u_{\delta_i}) = k \cap c(u_{\delta_{i'}}) = k] = \begin{cases} \frac{1}{K} & \text{si} \quad u_{\delta_i} = u_{\delta_{i'}} \\ \frac{1}{K^2} & \text{si} \quad u_{\delta_i} \neq u_{\delta_{i'}} \end{cases}$$

où u_{δ_i} désigne le noyau de plus proche de l'individu i.

Chaque individu i est affectée à la population d'appartenance de son noyau le plus proche δ_i. Ainsi, le lien entre le label de classe de l'individu i, c_i, et le label de classe de son noyau le plus proche $c(u_{\delta_i})$ est déterministe. Par conséquent, le vecteur des labels de classe latents se réduit au vecteur $(c(u_1), c(u_2), ..., c(u_m))$.

Pour visualiser graphiquement la population d'appartenance d'un individu ou encore l'organisation spatiale des K populations présentes, il suffit d'attribuer une couleur à chaque cellule de Voronoï en fonction de la marque de son noyau. Le graphe obtenu est appelé : tessellation de Voronoï colorée. Un exemple est donné dans la figure 3.3 pour K=2 et m=50.

Le paramètre d'intensité λ contrôle le nombre moyen de polygones dans Δ et ainsi le degré d'organisation spatiale des populations. Plus λ est petit, plus les populations sont fortement structurées spatialement et occupent des territoires aux formes simples (e.g., frontières rectilignes). Au contraire, plus λ est grand, plus la dépendance spatiale entre individus diminue et tend vers une configuration aléatoire de type 3.2.1. Le territoire occupé par chaque population prend alors des formes plus complexes caractérisées par des frontières irrégulières et/ou par l'union de régions non connexes (cf figure 3.4).

3.2.3 Les modèles de Markov cachés permettent de représenter des organisations spatiales complexes

Les tessellations de Voronoï sont bien adaptées pour formaliser des situations assez courantes où les populations sont isolées spatialement c'est-à-dire réparties sur des territoires disjoints de formes relativement simples. Toutefois, il existe des situations biologiques plus complexes dans lesquelles les fréquences alléliques ont une variation géographique continue perturbée par de petites discontinuités à la frontière entre populations. Cette continuité implique que les populations se chevauchent localement. Cela est notamment le cas de populations récemment entrées en contact par un processus de colonisation ou marquées par l'émergence d'une sous-population au sein même d'une population d'origine suite à l'apparition de barrières comportementales intraspécifiques (ex : unions consanguines). Modéliser un tel pattern génétique spatial semble adapté pour décrire la structure génétique des populations humaines (Serre & Paabo, 2004),(Rosenberg & al., 2005) connues pour être faiblement différentiées génétiquement (Bamshad & al., 2003). En effet, seule 10 à 15% de la variabilité génétique globale concerne des individus issus de populations différentes. On s'attend ainsi à devoir délimiter de petites discontinuités génétiques entre populations.

Une variation continue des fréquences alléliques implique l'existence de similarités génétiques locales entre individus. En effet, une hypothèse de continuité signifie que les fréquences alléliques en un site géographique donné sont plus proches des fréquences alléliques aux sites voisins que de celles des sites éloignés. Une approche possible est de modéliser cette dépendance spatiale locale par l'intermédiaire des variables de classe latentes c_i (i=1,2,..,n). Pour cela, nous nous sommes orientés vers les champs aléatoires markoviens pour données catégorielles, aussi appelés *modèles de Potts* (Wu, 1982). Depuis 1970, ces modèles sont souvent utilisés en analyse d'images (Geman & Geman, 1984) pour modéliser la corrélation entre couleurs de pixels voisins. Les modèles de Potts permettent de tenir compte du fait que deux pixels adjacents sont plus susceptibles d'avoir la même couleur que deux pixels non-adjacents. Par analogie, ces modèles peuvent donc traduire l'hypothèse selon laquelle deux individus voisins sont plus susceptibles d'appartenir à la même population que deux individus géographiquement éloignés. Ces modèles semblent donc pertinents pour étudier des populations avec une variation continue des fréquences alléliques.

L'utilisation d'un modèle de Potts nécessite la spécification d'une structure de voisinage entre individus. Comme illustré par le jeu de données des ours bruns de Scandinavie, les individus géo-référencés sont souvent irrégulièrement espacés dans le domaine d'échantillonnage. Aussi, nous avons opté pour le graphe de voisinage naturel induit par une tessellation de Voronoï appelé graphe de Delaunay. Chaque site s_i (i=1,2,...,n) définit le noyau d'une cellule de Voronoï composée des points les plus proches de ce site par rapport aux autres sites d'échantillonnage $\{s_j\}_{j\neq i}$. A noter que cette tessellation de Voronoï est différente de celle considérée dans le modèle Geneland pour laquelle les noyaux de cellules sont supposés aléatoires.

Definition 1. *Deux individus i et j sont dits voisins, noté $i \sim j$, si leurs cellules de Voronoï respectives partagent une arête commune.*

Dans certains cas, l'utilisation d'un graphe de Delaunay n'est pas naturelle pour définir une structure de voisinage. Je pense par exemple aux situations dans lesquelles le domaine

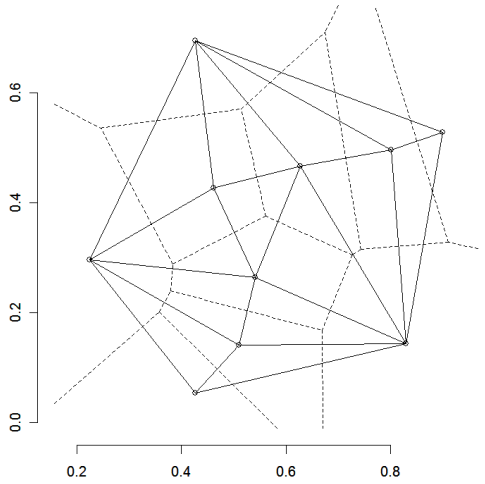


Figure 3.5 – Exemple de graphe de Delaunay (traits pleins) pour 10 individus (les points) échantillonnés aléatoirement dans le carré $[0,1]\times[0,1]$. Les traits en pointillés représentent les cellules de Voronoï centrées au niveau des individus échantillonnés

d'échantillonnage n'est pas convexe. Cela est notamment le cas du jeu de données humaines HGDP-CEPH (cf section 1.1.1) pour lequel l'Amérique et l'Europe sont séparées par l'océan Atlantique. Je pense également aux domaines dans lesquels apparaissent des barrières écologiques/humaines (montagne, pont, rivière) infranchissables par les individus. La liaison entre certains individus n'est alors pas réaliste. Dans ce cas, il suffit soit de restreindre le domaine d'échantillonnage à un sous-domaine convexe (ce que nous avons fait pour traiter les données HGDP-CEPH) soit de définir une structure de voisinage plus adaptée. En effet, l'utilisation d'un graphe de Delaunay n'est pas une condition sinéquanone à l'utilisation d'un modèle de Potts. Le modèle de Potts construit à partir d'une tessellation de Voronoï est appelé modèle de *Potts-Dirichlet*.

Chaque configuration de classes $c = (c_1, c_2, .., c_n)$ des n individus échantillonnés est supposée issue d'un modèle de Potts, défini par :

$$[c|\psi] = \frac{1}{Z(\psi, K)} exp(\psi U(c)) \qquad c \in \{1, ..., K\}^n \tag{3.2.1}$$

avec

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

$$Z(\psi, K) = \sum_{c \in \{1,2,..,K\}^n} exp(\psi U(c))$$

Le symbole de Kronecker δ_{c_i,c_j} vaut 1 ssi $c_i = c_j$ et 0 sinon. $Z(\psi, K)$ est une constante de normalisation appelée fonction de partition. Nous reviendrons plus en détail sur cette constante dans la section suivante car celle-ci pose des difficultés pour l'inférence du modèle.

$U(c)$, appelée fonction potentiel, désigne le nombre de paires de points voisins appartenant à une même population. Plus U(c) est petit, plus l'organisation spatiale des

populations est faible. Au contraire, plus U(c) est grand, plus les populations sont clusterisées. Le modèle de Potts affecte à chacune des K^n configurations spatiales possibles une probabilité qui croit avec $U(c)$.

ψ est un paramètre positif appelé paramètre d'intéraction du champ aléatoire. Il permet de moduler le degré d'organisation spatiale des populations. En effet, plus ψ est grand, plus le modèle favorise les configurations spatialement organisées i.e., pour lesquelles le nombre de paires de points voisins appartenant à une même population est grand. Si $\psi = 0$ alors toutes les configurations c ont la même probabilité : $[c] = \frac{1}{K^n}$. Cela revient à une situation de répartition aléatoire des individus.

Tel que défini par la relation 3.2.1, le modèle de Potts est une distribution de Gibbs (cf chapitre 4 (Banerjee & al., 2004) pour plus de détails). D'après Geman et Geman (1984) il possède donc la propriété markovienne suivante :

$$[c_i|c_{-i}] = [c_i|c_{\partial i}] \tag{3.2.2}$$

où c_{∂_i} désigne l'ensemble des populations d'appartenance des voisins de i i.e., $(c_j)_{i\sim j}$. Cette relation signifie que la probabilité pour que l'individu i appartienne à une population donnée ne dépend que de l'influence des labels de classe respectifs de ses voisins. La structure de dépendance entre individus est donc locale.

La propriété 3.2.2 s'obtient facilement en isolant les quantités dépendant de c_i dans l'expression de la loi jointe 3.2.1. Il vient alors :

$$[c_i|c_{-i}] = \frac{exp(\psi \sum_{j;j\sim i} \delta_{c_i,c_j})}{T_i}$$

$$\text{avec} \qquad T_i = \sum_{k=1}^{K} exp(\psi \sum_{j;j\sim i} \delta_{k,c_j})$$

La loi conditionnelle $[c_i|c_{-i}]$ dépend uniquement des sites voisins de i à travers la somme $\sum_{j;j\sim i} \delta_{c_i,c_j}$. Celle-ci correspond au nombre de voisins de l'individu i qui appartienne à la même population que i. Le mode de cette distribution conditionnelle correspondra à la population la plus fréquemment occupée par les voisins de i.

Cette propriété markovienne facilite la simulation de configurations de classes sous le modèle de Potts (cf figure 3.6). En effet, les distributions conditionnelles complètes sont des lois discrètes calculables analytiquement ce qui permet de recourir à un algorithme MCMC de type Gibbs (cf annexe B). La difficulté est de définir le nombre d'itérations nécessaires à l'algorithme pour atteindre la convergence. En effet, ce nombre dépend fortement de la valeur du paramètre d'intéraction spatial ψ mais aussi du nombre de points considérés. Après avoir suivi l'évolution de la proportion des différents états possibles au cours de la chaine de Gibbs sur différentes simulations, j'ai trouvé que 50000 itérations semblaient suffir à atteindre la convergence dans la plupart des cas.

3.3 Comment modéliser des données génotypiques multilocus ?

Dans cette section, je présente deux structures probabilistes qui permettent de décrire l'occurrence des génotypes de chaque individu à l'intérieur d'une population. Par référence

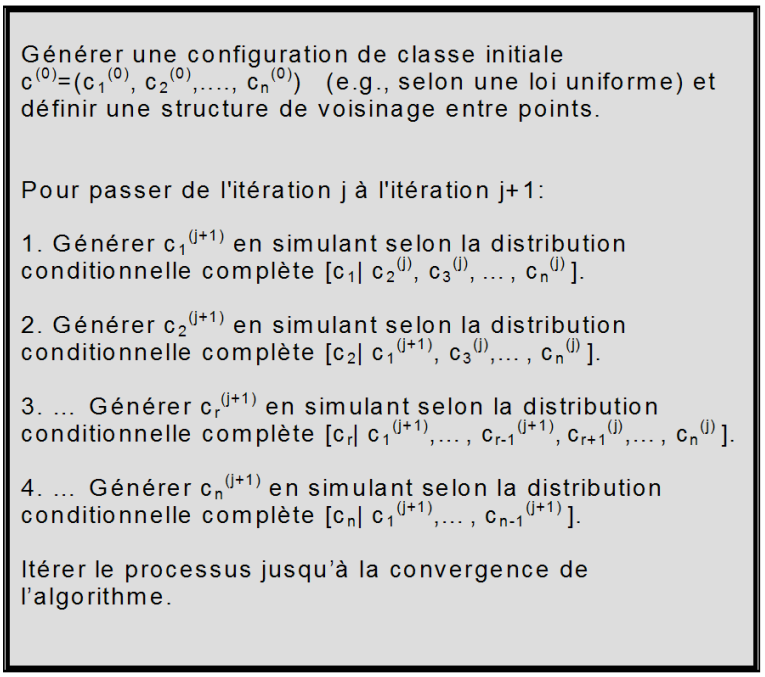


Figure 3.6 – Description de l'algorithme de Gibbs permettant de simuler une configuration de classes selon le modèle de Potts

à une démarche de modélisation hiérarchique, cela correspond à la formalisation du modèle des observations $Y|Z, \theta$ dans le cas où $Z = c$ et $\theta = (f, K)$.

La variabilité génétique est un phénomène complexe car elle est liée à l'interaction de nombreux facteurs biologiques (mutations de gènes, migrations d'individus....) et aléatoires (dérive génétique...). Ainsi, la modélisation des propriétés génétiques d'une population pousse généralement le biologiste à recourir à des hypothèses simples.

Une première hypothèse, commune aux deux structures probabilistes ci-après, est de considérer que, dans une population, il n'y a pas de déséquilibres de liaison entre loci. Un déséquilibre de liaison se produit lorsque deux allèles correspondant à deux locus distincts d'un même chromosome sont plus fréquemment associés sur une même branche que ne le voudrait le hasard. D'un point de vue statistique, supposer qu'il n'y a pas de déséquilibre de liaison entre loci revient simplement à supposer l'indépendance des génotypes en chaque locus de chaque individu :

$$[y_i|c_i = k, f_{k,.,.}] = \prod_{l=1}^{L} [y_{i,l}|c_i = k, f_{k,l,.}] \qquad \forall i \in \{1, ..., n\} \tag{3.3.3}$$

Cette hypothèse est raisonnable dès lors que les loci considérés sont physiquement éloignés sur les chromosomes. Les ours bruns de Scandinavie et les individus du jeu de données humains HGDP-CEPH ont été génotypés au niveau de microsatellites (cf section 1.1.1). Ces marqueurs génétiques ont la propriété d'être dispersés sur tout le génome. Dans ce cas, il est donc raisonnable de supposer l'absence de déséquilibres de liaison entre loci.

pattern spatial	Hypothèse biologique	Hypothèse statistique	Modèle $[Z\|\theta]$	θ	Z
aléatoire	répartition aléatoire et indépendante des individus	uniformité indépendance	loi uniforme sur [[1,K]]	K	c
barrières génétiques aux flux de gènes	populations isolées par des barrières écologiques/humaines	chaque population occupe un territoire issu de l'union de polygones (frontières rectilignes)	tessellation de Voronoï colorée	K λ u	m c
clines	variation géographique continue des fréquences alléliques	dépendances spatiales locales entre individus	champ aléatoire markovien : Potts-Dirichlet	K, ψ	c

Tableau 3.1 – A chaque hypothèse biologique son modèle statistique

3.3.1 Le modèle classique : la loi de Hardy-Weinberg

Pour simplifier la modélisation de la variabilité génétique à l'intérieur d'une population, les biologistes font généralement les hypothèses de référence suivantes :

1. la population est panmictique c'est-à-dire que les couples se forment aléatoirement (panmixie) et indépendamment des génotypes et des liens de parenté.
2. La population est "infinie" afin de minimiser les variations d'échantillonnage
3. il n'y a ni migration, ni sélection, ni mutation (pas de perte/gain d'allèle)
4. Les générations successives sont discrètes (pas de croisement entre générations différentes)

Sous de telles hypothèses, une population est dite à l'équilibre de Hardy-Weinberg (encore appelé équilibre panmictique). La diversité génétique de la population se maintient et doit tendre vers un équilibre stable de la distribution génotypique (Stern, 1943).

L'équilibre de Hardy-Weinberg suppose que les deux allèles observés en un locus soient statistiquement indépendants. Etant donné un individu de la population k et les fréquences alléliques au locus l $f_{k,l,.}$, la probabilité d'observer le génotype $y_i^l = (\alpha_i^l, \beta_i^l)$ est donnée par la loi multinômiale (ou loi de Hardy-Weinberg en génétique des populations) de paramètres $f_{k,l,.}$:

$$[y_i^l | c_i = k, f_{k,l,.}] = \begin{cases} f_{k,l,\alpha_i^l}^2 & \text{si} \quad \alpha_i^l = \beta_i^l \\ 2 f_{k,l,\alpha_i^l} f_{k,l,\beta_i^l} & \text{sinon} \end{cases} \tag{3.3.4}$$

3.3.2 Un modèle bivarié pour tenir compte des dépendances alléliques en situation de consanguinité

Dans les populations naturelles, divers écarts à l'équilibre de Hardy-Weinberg peuvent subvenir. Des écarts à la panmixie (hypothèse 1) apparaissent dans les populations dans lesquelles les unions entre apparentés sont pratiquées ou pour lesquelles le mode naturel de reproduction impose ce type d'union (ex : autofécondation). Ces écarts apparaissent aussi dans les populations dont les effectifs de reproducteurs sont limités (i.e., populations subdivisées). C'est pourquoi nous proposons un modèle alternatif d'occurrence des génotypes qui tienne compte de l'existence possible d'unions consanguines dans chaque population.

Une union est dite consanguine lorsqu'elle concerne deux individus partageant des liens de parenté. A titre d'exemple, l'arbre généalogique de la figure 3.7 montre que Zoé est le fruit de l'union consanguine d'Irène et Jules qui partagent pour ancêtre commun Amédée.

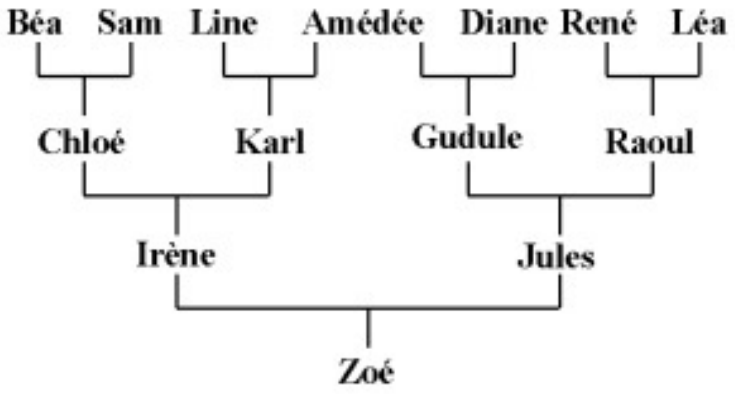


Figure 3.7 – Un exemple d'union consanguine : Irène et Jules ont pour ancêtre commun Amédée

La consanguinité a pour conséquence d'augmenter la fréquence des génotypes homozygotes et de diminuer celle des hétérozygotes. Un génotype est dit homozygote lorsqu'il est composé de deux allèles identiques. Un génotype est dit hétézygote lorsqu'il est composé de deux allèles différentes. Ainsi, modéliser un tel phénomène suppose de spécifier une structure de covariation positive entre allèles.

A chaque population k (k=1,2,..,K) est associé un coefficient de consanguinité ϕ_k qui représente la probabilité pour que deux gènes soient identiques par descendance c'est-à-dire copiés d'un même gène ancestral. ϕ_k est donc à valeurs dans $[0, 1]$. Etant donné un individu de la population k, les fréquences alléliques $f_{k,l,.}$ au locus l et le coefficient de consanguinité ϕ_k, la probabilité d'observer le génotype $y_i^l = (\alpha_i^l, \beta_i^l)$ est :

$$[y_i^l | c_i = k, \phi_k, f_{k,l,.}] = \begin{cases} f_{k,l,\alpha_i^l}^2 + \phi_k f_{k,l,\alpha_i^l}(1 - f_{k,l,\alpha_i^l}) & \text{si} \quad \alpha_i^l = \beta_i^l \\ 2 f_{k,l,\alpha_i^l} f_{k,l,\beta_i^l}(1 - \phi_k) & \text{sinon} \end{cases} \tag{3.3.5}$$

Avec ce modèle d'occurrence, la fréquences des homozygotes est supérieure et la fréquences des hétérozygotes est inférieure à ce que suppose la loi de Hardy-Weinberg (cf equation 3.3.4).

Si $\phi_k = 0$, on retrouve le modèle d'occurrence des génotypes donné par la loi de Hardy-Weinberg.

Si $\phi_k = 1$ alors il n'y a que des génotypes homozygotes :

$$[y_i^l | c_i = k, \phi_k, f_{k,l,.}] = 0 \qquad \text{si} \qquad \alpha_i^l \neq \beta_i^l$$

3.4 Re-construction hiérarchique des modèles Structure et Geneland

Dans la section précédente, j'ai présenté quelques modèles de base possibles pour décrire la structure génétique spatiale d'individus dans un espace géographique donné.

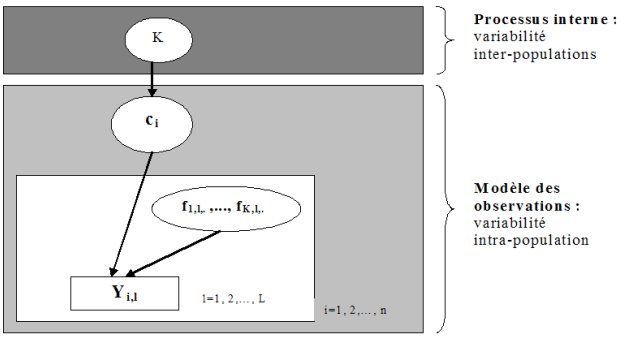


Figure 3.8 – Graphe acyclic orienté du modèle Structure

Afin de poursuivre dans une démarche de modélisation hiérarchique, je vais assembler certains de ces modèles par couches successives afin de *re-construire* deux structures hiérarchiques connues en génétique des populations : les modèles Structure et Geneland.

3.4.1 Le modèle Structure

Actuellement, le modèle statistique le plus utilisé pour classer des individus en populations génétiquement homogènes à partir de données génotypiques multilocus est le modèle de Pritchard et al. (2000), implémenté sous un logiciel nommé Structure. Dans sa version la plus simple, ce modèle hiérarchique est construit par couplage d'une loi uniforme pour le processus interne (cf section 3.2.1) et de la loi de Hardy-Weinberg ajoutée à une hypothèse d'équilibre de liaisons entre loci pour le modèle de observations (cf section 3.3.1). Dans ce modèle, les variables latentes sont les labels de classe c et les paramètres inconnus sont les fréquences alléliques f et le nombre de populations K. Le graphe acyclic orienté de la figure 3.8 permet de visualiser les relations de dépendances conditionnelles induites par cet assemblage.

Par ailleurs, ce modèle suppose que les individus d'une même sous-population sont *conditionnellement indépendants* dans le sens où, une fois les fréquences alléliques et labels de classe estimés, il ne reste plus de dépendance résiduelle entre individus : les frontières entre populations expliquent totalement l'écart à l'indépendance observé. Ajoutée à l'hypothèse d'équilibre de liaisons entre les loci qui suppose leur indépendance mutuelle (cf section 3.3.1), la probabilité d'observer l'ensemble des génotypes y conditionnellement aux fréquences alléliques et aux labels de classe est donnée par le double produit suivant :

$$[y|f,c] = \prod_{i=1}^{n}\prod_{l=1}^{L}[y_{i,l}|c_i = k, f_{k,l,.}]$$

avec $[y_{i,l}|c_i = k, f_{k,l,.}]$ défini par la relation 3.3.4.

Le point faible principal de ce modèle est qu'il ne tient pas compte explicitement d'une possible structure spatiale de la diversité génétique. Ainsi, il n'utilise pas l'information disponible dans les coordonnées spatiales des individus échantillonnés et suppose que ces derniers se répartissent indépendamment et uniformément dans les différentes populations.

L'inférence MCMC du modèle Structure est détaillée dans l'article (Pritchard & al., 2000).

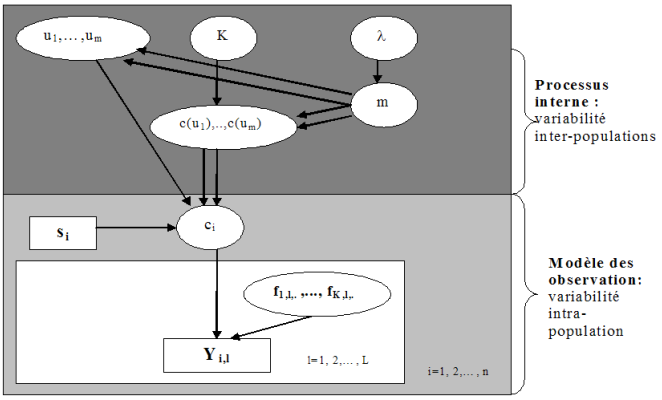


Figure 3.9 – Graphe acyclic orienté du modèle Geneland. Les doubles flèches représentent des relatiosn déterministes.

3.4.2 Le modèle Geneland

Récemment, un modèle hiérarchique spatialement explicite a été proposé et implémenté par Guillot et al. (2005) sous le package R appelé Geneland. Le modèle conditionnel choisi pour représenter l'occurrence des génotypes est le même que celui du modèle Structure : la loi de Hardy-Weinberg ajoutée à une hypothèse d'équilibre de liaisons entre loci et à une hypothèse d'indépendance conditionnelle des individus à l'intérieur d'une même sous-population. Il vient :

$$[y|f,c] = \prod_{i=1}^{n}\prod_{l=1}^{L}[y_{i,l}|c_i = k, f_{k,l,.}]$$

avec $[y_{i,l}|c_i = k, f_{k,l,.}]$ défini par la relation 3.3.4.

La spécificité du modèle Geneland provient de la modification apportée à la couche phénoménologique latente qui, désormais, est décrite à l'aide d'une structure plus complexe induisant des relations de dépendances spatiales entre individus : une tessellation de Voronoï (cf section 3.2.2).

Dans ce modèle, les paramètres inconnus sont $\theta = (K, \lambda, f, u)$ et les variables latentes sont le nombre de cellules de Voronoï, m, et leur label de classe respectif $(c(u_1), c(u_2), ..., c(u_m))$. Le graphe acyclic orienté de ce modèle est représenté dans la figure 3.9.

La loi jointe du modèle Geneland s'écrit :

$$[y, z, \theta] = [y, m, c, u, f, K, \lambda] = [K][\lambda][f][u][m|\lambda][c|K, u][y|c, f]$$

avec :

$$\begin{aligned} m|\lambda &\sim \mathcal{P}oisson(\lambda) \\ c|K, u &\sim \mathcal{U}niforme(\{1, 2, .., K\}) \end{aligned}$$

Les lois *a priori* suivantes ont été choisies pour les paramètres inconnus K, f, λ, u :

$$\begin{aligned} K &\sim \mathcal{U}niform(\{K_{min}, K_{max}\}) \\ \lambda &\sim \mathcal{U}niform([0, \lambda_{max}]) \\ f &\sim \mathcal{D}irichlet(1, 1, ..., 1) \\ u &\sim \mathcal{U}niform(\Delta) \end{aligned}$$

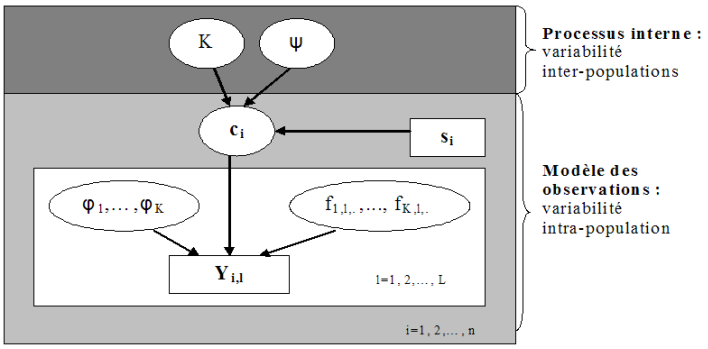


Figure 3.10 – Graphe acyclic orienté du modèle Geneclust

Concernant les choix de K_{min}, K_{max} et λ_{max}, Guillot et al. préconisent de choisir $K_{min} = 1$, et K_{max} suffisamment grand par rapport au nombre de populations attendues. Quant à λ_{max}, ils suggèrent de le prendre égal au nombre d'individus échantillonnés afin de couvrir une large gamme de configurations spatiales possibles.

L'inférence MCMC du modèle Geneland est détaillée dans l'article (Guillot & al., 2005).

3.5 Le modèle Geneclust : une construction hiérarchique alternative basée sur un champ de Markov caché

3.5.1 Description du modèle

Le modèle que nous proposons, baptisé Geneclust, s'écrit comme l'assemblage d'un modèle de Potts-Dirichlet au niveau de la couche phénoménologique latente (cf section 3.2.3) et d'un modèle d'observation tenant compte de l'existence possible d'unions consanguines à l'intérieur d'une population (cf section 3.3.2). L'originalité du modèle Geneclust intervient aux deux niveaux de la hiérarchie avec l'utilisation d'un modèle alternatif à la loi de Hardy-Weinberg pour l'occurrence des génotypes et l'utilisation d'un outil statistique jusqu'alors non utilisé en génétique des populations pour modéliser une continuité spatiale de la diversité génétique : les champs de Markov.

Comme pour les modèles Structure et Geneland, nous supposons qu'il n'y a pas de déséquilibre de liaisons entre les loci et que les individus sont conditionnellement indépendants à l'intérieur d'une même sous-population. Il vient :

$$[y|f, c, \phi] = \prod_{i=1}^{n} \prod_{l=1}^{L} [y_{i,l}|c_i = k, \phi_k, f_{k,l,.}]$$

avec $[y_{i,l}|c_i = k, \phi_k, f_{k,l,.}]$ définie par la relation 3.3.5.

Le graphe acyclic orienté de ce modèle est représenté dans la figure 3.10.

3.5.2 Le choix des lois *a priori*

Dans la liste des grandeurs à inférer, il y a les variables de classes latentes c et les paramètres inconnus $\theta = (f, \phi, \psi)$ avec $f = (f_{k,l,j})$ (k=1,2,..,K, l=1,2,..,L, j=1,2,..,J_l) les fréquences alléliques, ψ le paramètre d'intéraction spatial, $\phi = (\phi_1, \phi_2, ..., \phi_K)$ le vecteur des coefficients de consanguinité. Mener l'inférence bayésienne du modèle nécessite au préalable de définir des lois *a priori* sur les paramètres inconnus θ.

En génétique des populations, une hypothèse biologiquement pertinente ((Pritchard & al., 2000), (Guillot & al., 2005)) est de supposer que les fréquences alléliques sont indépendantes et identiquement distribuées selon une loi de Dirichlet sous la forme :

$$f_{k,l,\cdot} \sim \mathcal{D}irichlet(\alpha, \alpha, .., \alpha) \qquad \text{k=1,2,..,K} \qquad \text{l=1,2,..,L}$$

où $f_{k,l,\cdot}$ désigne le vecteur des fréquences alléliques au locus l dans la population k. Ne disposant d'aucune connaissance quantitative *a priori* sur les fréquences alléliques, nous avons choisi $\alpha = 1$ afin de définir un prior non-informatif.

Les coefficients de consanguinité ϕ_k (k=1,2,..,K) sont des paramètres inconnus à valeurs dans $[0, 1]$. Nous supposons que ces coefficients sont indépendants et identiquement distribués selon une loi Béta :

$$\phi_k \sim \mathcal{B}eta(4, 40)$$

Un tel prior semble en effet pertinent d'un point de vue biologique. (cf figure 3.11)

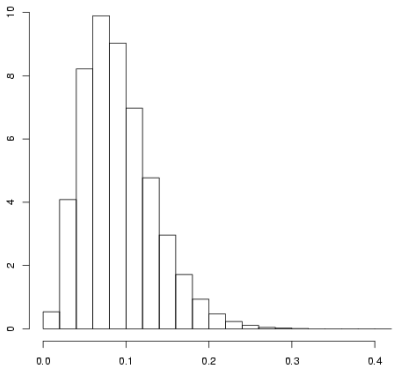


Figure 3.11 – Loi *a priori* sur les coefficients de consanguinité : loi Béta(4,40)

Pour des raisons de commodités calculatoires, nous considérons une loi *a priori* discrète uniforme pour le paramètre d'intéraction spatiale ψ :

$$\psi \sim \mathcal{U}\{0, 0.1, 0.2, ..., \psi_{max}\}$$

Nous posons $\psi_{max} = 1$ car cette valeur correspond à un fort niveau d'intéraction spatiale dans le cas où le nombre de populations inconnu K vaut 3,4,5 ou 6. En effet, dans ce cas, la probabilité pour que deux sites adjacents soient affectés à la même population est proche de 1. Au contraire, $\psi = 0.5$ correspond à un faible niveau d'intération pour $K = 3, 4, 5, 6$. Aussi, des valeurs de ψ autour de 0.6-0.7 semblent raisonnables pour détecter K=3,4,5,6 populations spatialement cohérentes. Un prior qui accorde de forts degrés de crédibilité aux valeurs 0.6 et 0.7 sera donc également envisagé.

3.5.3 Le nombre de populations est évalué par régularisation

Nous sommes dans un problème de classification non supervisée. Le nombre de populations K est inconnu et doit donc, en pratique, être estimé. Deux grands types d'approche existent :

1. Calculer, pour différentes valeurs de K, un critère de sélection de modèles pénalisé en fonction de la complexité du modèle (ex : BIC, AIC, DIC...). Puis, estimer $\widehat{K}$ comme la valeur de K associée au meilleur modèle c'est-à-dire à celui pour lequel le critère est le plus faible.
2. Choisir une loi *a priori* sur K et générer un échantillon MCMC issu de sa loi *a posteriori*. Malheureusement, une telle approche nécessite de recourir à des techniques d'inférence MCMC, type *reversible jump*, qui permettent de passer d'une dimension à l'autre. Ces algorithmes sont souvent longs et difficiles à implémenter et à utiliser.

Nous avons choisi une approche alternative connue sous le nom de *regularisation* en statistiques. Pour plus de détails sur cette terminologie, je renvoie le lecteur vers l'ouvrage de Ripley (1996, Chapter 4.3, p.136). Cette méthode consiste à partir d'un nombre de populations K_{max} suffisamment grand par rapport au *vrai* nombre de populations cherché. Au début de l'algorithme MCMC, les labels de classe c_i sont initialisés avec des valeurs arbitraires comprises entre 1 et K_{max}. Puis, au fil des itérations MCMC, l'algorithme va éliminer naturellement les populations superflues. En pratique, des valeurs croissantes de K_{max} sont testées jusqu'à ce que la classification obtenue se stabilise. Chaque individu est classé dans la population qui lui est la plus probable *a posteriori* :

$$\widehat{c_i} = argmax_{\{1,2,..,K\}}[c_i = k|y]$$

où $\widehat{c_i}$ désigne la classe d'affectation *a posteriori* de l'individu i. $\widehat{K}$ est alors défini comme le nombre de labels de classe différents contenus dans $\widehat{c}$.

3.5.4 Inférence bayésienne

La loi jointe du modèle Geneclust s'obtient facilement grâce aux multiples relations d'indépendances conditionnelles entre grandeurs inconnues illustrées sur le DAG de la figure 3.10 :

$$[Y, Z, \theta] = [Y, c, \phi, \psi, f] = [Y|f, c, \phi][c|\psi][f|c][\phi][\psi]$$

J'ai implémenté la mise à jour successive des variables latentes et paramètres inconnus du modèle Geneclust selon un algorithme de type Metropolis-Hastings codé sous un package R interfacé avec du langage Fortran. Les détails de l'algorithme MCMC implémenté sont indiqués dans l'Annexe E. La difficulté majeure a concerné l'estimation du paramètre d'intéraction spatiale ψ du modèle de Potts-Dirichlet. En effet, le calcul du ratio de Metropolis-Hastings, impliqué dans la mise à jour de ce paramètre, nécessite de pouvoir calculer la constante de normalisation d'un champ de Markov. Or, celle-ci est souvent non-calculable analytiquement si bien que, par simplicité, le paramètre d'intéraction est généralement fixé. Pour résoudre cette difficulté, j'ai utilisé la méthode d'intégration thermodynamique (Green & Richardson, 2002) qui consiste à approximer numériquement la constante de normalisation d'un champ de Markov pour chaque valeur discrétisée du paramètre ψ. Les détails de la méthode sont donnés dans l'Annexe E.

3.6 Simulations et analyses de données réelles

L'analyse des performances du modèle Geneclust pour la classification d'individus en populations génétiquement homogènes a été menée par simulations puis avec les deux jeux de données réelles décrits dans le chapitre 1.

Dans un premier temps, j'ai réalisé une étude par simulations dont les objectifs étaient :

- de tester la capacité du modèle Geneclust à retrouver le nombre inconnu de populations K et la configuration de classes latente c à partir de données simulées sous ce même modèle.
- de comparer les capacités de classification respectives des modèles Structure, Geneclust et Geneland sous l'hypothèse du modèle Geneclust selon laquelle la variabilité génétique est spatialement continue et seulement caractérisée par de petites discontinuités génétiques.

Dans un deuxième temps, le modèle Geneclust a été appliqué aux données d'ours bruns de Scandinavie et aux données humaines HGDP-CEPH afin de comparer les configurations de classe *a posteriori* estimées par ce modèle à celles obtenues avec les modèles Structure et Geneland. Plusieurs runs MCMC ont été réalisés afin de tester la sensibilité du modèle au nombre initial de populations K_{max} et au nombre de marqueurs génétiques considérés.

Le modèle Geneclust ainsi que les objectifs visés par la construction de ce modèle ont fait l'objet de deux articles. Le premier, paru dans *Genetics*, est intitulé *Bayesian Clustering using Hidden Markov Random Fields in Spatial Population Genetics*. La plupart des idées de cet article ont été détaillées dans ce chapitre. Le second, paru dans *le Journal de la Société Française de Statistique*, s'intitule *Hidden Markov random fields and the genetic structure of the scandinavian brown bear population*. Il est davantage dédié à un public de statisticiens et montre très brièvement comment la modélisation hiérarchique bayésienne et les outils de statistique spatiale peuvent être couplés pour fournir des moyens de mesure utiles à l'évaluation de la biodiversité.

Aucune analyse supplémentaire n'a été réalisée depuis la publication de ces deux articles. Aussi, je les introduis à ce stade de la rédaction afin d'inviter le lecteur à lire tout particulièrement les sections *Simulation study* et *Real data analysis* de l'article 1 paru dans *Genetics*. Ce dernier est en effet plus détaillé que le second article. A noter que l'abréviation HRMF, acronyme de "Hidden Random Markov Field", désigne le modèle Geneclust dans les articles.

3.6.1 Article 1 paru dans Genetics

DOI: 10.1534/genetics.106.059923

Bayesian Clustering Using Hidden Markov Random Fields in Spatial Population Genetics

Olivier François,*[,1] Sophie Ancelet[†] and Gilles Guillot[†]

**TIMC, TIMB (Department of Mathematical Biology), F38706 La Tronche, France and [†]Unité de Mathématiques et Informatique Appliquées, ENGREF, 75732 Paris Cedex 15, France*

Manuscript received April 25, 2006
Accepted for publication July 26, 2006

ABSTRACT

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

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

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

[1]*Corresponding author:* Faculty of Medicine, Grenoble, F38706 La Tronche, France. E-mail: olivier.francois@imag.fr

In the latter and other applications of clustering algorithms, the spatial data are actually treated off line and are not part of the modeling. Bayesian models such as those developed by Pritchard *et al.* (2000), Dawson and Belkhir (2001), or Corander *et al.* (2003) nevertheless offer a natural and appropriate framework for including spatial prior information when assigning an individual to a fixed number of clusters. For example, a recent study by Guillot *et al.* (2005) used spatial explicit priors in a full-Bayes perspective and successfully identified genetic barriers in a wolverine population. An assignment method was also used by Wasser *et al.* (2004) to infer the spatial origin of African elephants. Here we argue that modified Bayesian algorithms can provide additional evidence to solve cline/cluster dilemmas such as those discussed in Rosenberg *et al.* (2005). A natural way to proceed is to include priors on continuous variation of genetic diversity in the Bayesian model used by STRUCTURE and check whether or not the previously discussed clusters are robust.

In this study, we present a new hierarchical Bayes algorithm that incorporates models for geographical continuity of allele frequencies. This is achieved by using hidden Markov random fields (HMRFs) as prior distributions on cluster membership. An informal definition of HMRFs states that allele frequencies at a specific geographical site are more likely to be close to the allele frequencies at neighboring sites than at distant sites. The problem of local differentiation may also be studied in terms of change in correlation with distance as considered by Malécot (1948), where "individuals living nearby tend to be more alike than those

living far apart" (Kimura and Weiss 1964, p. 561). The HMRF is basically another formulation of the same idea with statistical correlation hidden at the cluster membership level.

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

THE POTTS–DIRICHLET MODEL

In this study we borrow from the toolbox of statistical physics the concept of Markov random field (MRF), also called the Potts model (Potts 1952; Preston 1974; Wu 1982. The model has been coined to handle stochastic networks where particles in identical states evolve in patches larger than expected under an absence of interactions. Guttorp (1995) gives a recent review of the Potts model at a fairly introductory level. Since the 1970s, MRFs have a long tradition in image analysis, where the color of pixels is correlated to the color of neighboring pixels (see, *e.g.*, Geman and Geman 1984; Besag 1986; Ripley 1988). In this context MRFs account for the property that adjacent pixels are more likely to be of the same color than nonadjacent pixels. HMRFs are relatively recent, but they have been successfully applied in several domains (Zhang *et al.* 2001; Green and Richardson 2002; Destrempes *et al.* 2005). Ideas from Bayesian spatial genetics were also used in association studies (Thomas *et al.* 2003). In analogy with image analysis, MRF can model the fact that individuals from spatially continuous populations are more likely to share cluster membership with their close neighbors than with distant representatives. They seem therefore relevant to study populations for which continuous variation of allele frequencies may be used as a postulate.

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

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

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

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

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

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

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

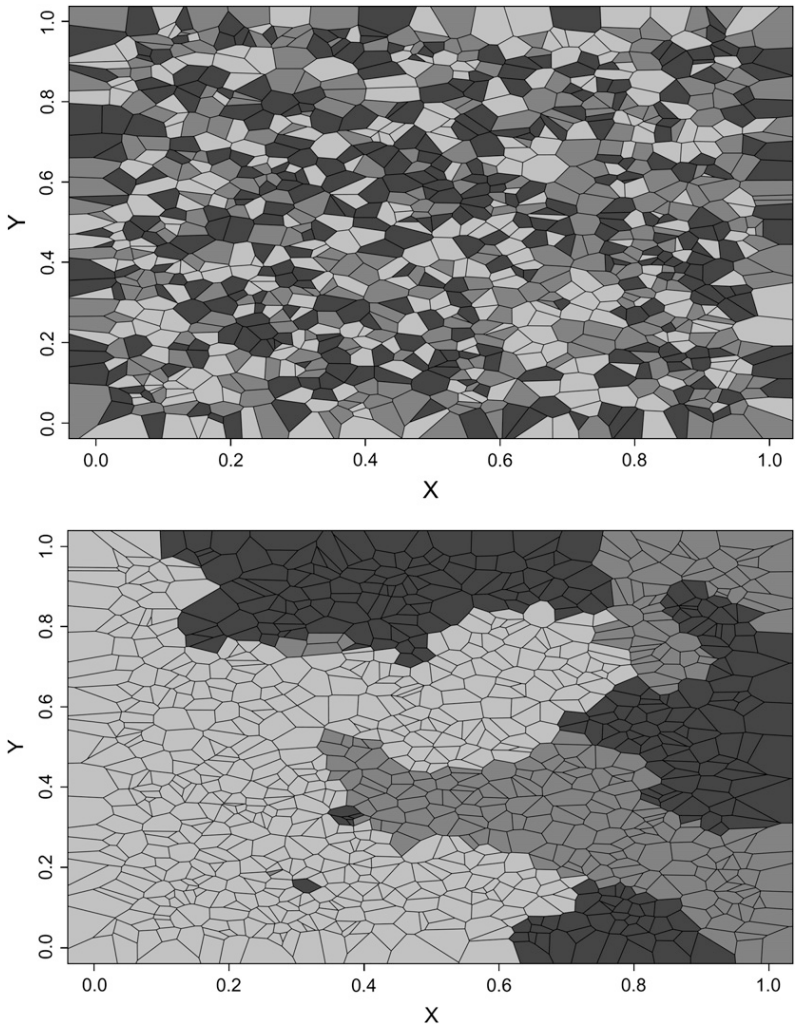


FIGURE 1.—Two cluster configurations from the three-states Potts–Dirichlet model. For $\psi = 0.1$, no spatial structure can be observed (the situation is close to the noninformative prior used by STRUCTURE). For $\psi = 0.9$, a number of nonnecessarily connected random clusters can be observed.

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

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

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

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

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

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

$$\pi(c_i \mid c_{-i}) = \pi(c_i \mid c_{\partial i}). \tag{5}$$

More specifically, we have

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

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

HIERARCHICAL BAYES

Model: In this section, we present the hierarchical Bayes model based on an HMRF. With $\psi = 0$, the HMRF model assumes a noninformative spatial prior and then

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

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

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

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

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

Assuming linkage equilibrium between loci, the likelihood is defined as

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

where $\mathcal{L}_k$ is defined in Equation 7.

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

Estimating the number of clusters: As other Bayesian clustering methods do, the HMRF model refers implicitly to an unknown number of clusters K. In practice this number K has to be estimated. Previous approaches typically fall into two categories: (1) maximizing the likelihood modified with a penalty that decreases with model complexity (*e.g.*, Bayes information criteria and deviance information criteria) and (2) choosing a prior distribution on K and maximizing the posterior distribution using transdimensional MCMC computations (which are usually time-consuming to develop and to run). Although these methods have proved effective in many cases, we use an alternative approach known as regularization in statistics. For this terminology, we refer to the book by Ripley (1996, Chap. 4.3, p. 136). The rationale for regularization and the relationship with the algorithm implemented in STRUCTURE can be explained as follows. Let $L_s(z, f, c)$ denote the log-probability for the complete data (observed plus unobserved) in the original approach of Pritchard *et al.* (2000). When we refer to this approach, we mean the no-admixture model with uncorrelated allele frequencies. Assuming absence of inbreeding, the log-probability of the HMRF model can be expressed as

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

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

In the regularization approach, $K_{\max}$ is a value presumed larger than the true number of clusters K. When the algorithm is started, the cluster configuration c spans arbitrary values between 1 and $K_{\max}$. As the chain runs, the program attempts to reduce the number of nonempty clusters that is finally considered as an estimate of K. In practice, one starts with runs with small

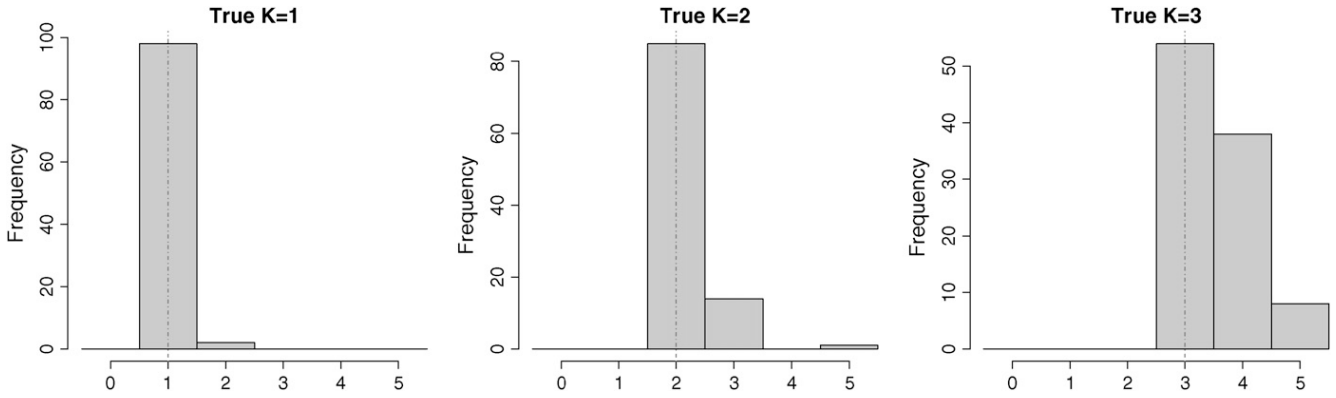


FIGURE 2.—Distributions of the number of clusters estimated by the HRMF model. Data sets were simulated from the prior distributions of the HMRF model. The vertical lines indicate the true number of populations.

values of $K_{\max}$ and increases $K_{\max}$ unless the estimated K is strictly lower than $K_{\max}$. Then, one checks that the result remains identical when higher values of $K_{\max}$ are used. Practice also shows that repeating shorter runs and performing estimation from the runs with the highest likelihood is a reasonable strategy.

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

SIMULATION STUDY

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

Estimating the number of clusters: To check the validity of the HMRF model, we performed inferences for 300 simulated data sets obtained as replicates from the model prior distributions. Individual geographical coordinates were generated from a two-dimensional uniform distribution on a square domain. Genotypes with 10 loci and 10 alleles per locus were simulated using multinomial sampling from the Dirichlet $\mathcal{D}(1, \ldots, 1)$ distribution. The interaction parameter ψ was simulated according to a uniform distribution on $\{0, 0.1, \ldots, 1\}$. The inbreeding coefficients were simulated according to a beta $\mathcal{B}(4, 40)$ distribution. The hidden cluster configurations c were generated from the Potts–Dirichlet model with $K = K_{\max} = 1, 2, 3$ classes. Replicates with $K = 1, 2, 3$ classes were simulated for $n = 50, 100, 150$ individuals, respectively.

In the full-Bayes inference method (inference of ψ), the computation of the partition function $Z(\psi, K_{\max})$ involved preliminary off-line runs. They were carried out with 20,000 cycles of a Gibbs sampler with a thinning period of 10 cycles. The maximal number of clusters was fixed to $K_{\max} = 5$, and 30,000 cycles, a burn-in period of 20,000, and a thinning period of 10 cycles were used. The parameter ψ was kept equal to 0 during the first 5000 cycles (see *Updating the interaction parameter* ψ in the APPENDIX for more details).

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

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

TABLE 1

Proportions of individuals assigned to extra clusters given the number of estimated clusters $\hat{K}$ and their true number K

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

— indicates cases that never occurred during the simulation study.

accuracy of cluster assignment through the error rate in coassignment (ERCA) defined as

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

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

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

We simulated 200 data sets from the HMRF model prior distributions with $K_{\max} = 2$, using simulations from the MCMC program without data (1000 cycles). Running the program for a fixed number of cycles did not warrant the convergence of the MCMC sampler. As the aim of the simulation study was the retrieval of previously stored allele frequencies and cluster memberships, this shortcoming did not affect the performance study. In the sampled data, individuals were occasionally grouped in a single cluster (for values of $\psi > 0.8$). The clusters had no predefined size and might consist of very few (<10) individuals. The ERCA rates are reported in Table 2. In this table, the rates were averaged either over all data sets or over subsets of data that corresponded to different levels of pairwise F_{ST}, interaction parameter ψ, and inbreeding coefficients (ϕ_1, ϕ_2).

The results provided evidence that the HMRF model increased the number of correct assignments compared to the nonspatial model. A more detailed look at subsets of simulated data revealed that the HMRF model always performed better than the other models whatever the levels of spatial interaction or inbreeding. The highest

TABLE 2

Error rate in coassignments (ERCA) for 200 simulated data sets ($n = 100$, $L = 10$, $J_\ell = 10$) with $K_{\max} = 2$

Genetic structure: F_{ST}	Spatial structure: ψ	Inbreeding (ϕ_1, ϕ_2)	Nonspatial model	HMRF model	GENELAND
All	All	All	16.1	0.7	3.2
$F_{ST} \le 0.08$	All	All	26.3	1.6	6.6
$0.08 < F_{ST} \le 0.09$	All	All	7.6	0.6	1.4
$0.09 < F_{ST} \le 0.1$	All	All	8	0.6	1.4
$F_{ST} > 0.1$	All	All	8.3	0.2	1.1
All	$\psi \le 0.2$	All	1.1	1	1.1
All	$0.2 < \psi \le 0.4$	All	1	0.8	1.6
All	$0.4 < \psi \le 0.6$	All	2.7	0.7	0.9
All	$0.6 < \psi \le 0.8$	All	28.2	0.4	4.7
All	$\psi > 0.8$	All	42.4	0.5	6.9
All	All	(<0.06, <0.06)	17.2	0.3	0.7
All	All	(<0.06, >0.1) or (>0.1, <0.06)	10	0.5	1.9
All	All	(>0.1, >0.1)	12.3	1	1.5
$F_{ST} \le 0.08$	$\psi \le 0.4$	All	2.7	2.1	2.8
$F_{ST} \le 0.08$	$0.6 < \psi \le 1$	All	41.8	0.9	9.4
$F_{ST} > 0.1$	$\psi \le 0.4$	All	0.2	0.1	0.4
$F_{ST} > 0.1$	$0.6 < \psi \le 1$	All	23.7	0.3	2.4

The three models were initialized at $K_{\max} = 2$.

improvements were obtained at low levels of differentiation ($F_{ST} \leq 0.08$) and high levels of spatial structure ($\psi > 0.6$). The HMRF model achieved the smallest improvements over the other models for high levels of inbreeding, although it still gave very accurate results. In these cases, the inbreeding coefficients were correctly estimated (results not shown).

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

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

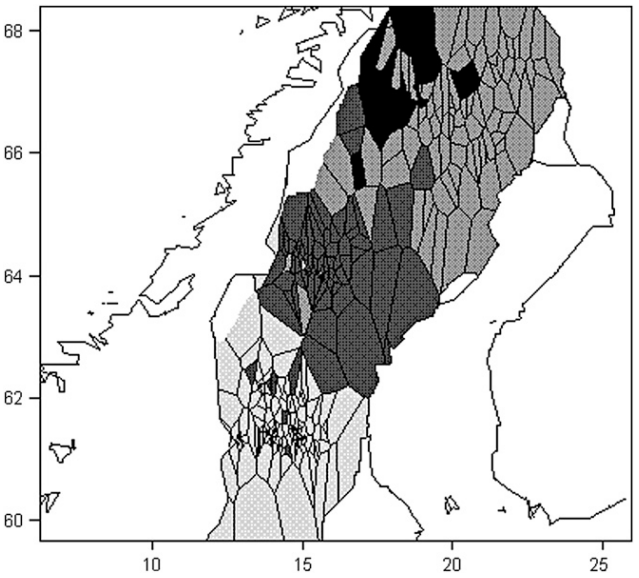


FIGURE 3.—Estimated cluster configuration for the Scandinavian brown bear data set in North Sweden using the HMRF model (four clusters).

REAL DATA ANALYSIS

Scandinavian brown bears: The Scandinavian brown bear (*Ursus arctos*) is an example of a wild population with strong female phylopatry and male-mediated gene flows. We analyzed the same data set as in two previous studies (WAITS *et al.* 2000; MANEL *et al.* 2004) from 366 geo-referenced individuals genotyped at 19 microsatellite loci. We first used the full-Bayes HMRF model implemented with the same prior distributions as in the simulation study and ran the algorithm with $K_{max} = 4$–7. After 30,000 cycles, the HMRF model with $K_{max} = 4$ converged to the same clusters as described in the previous study. We referred to these clusters as the south (S), middle (M), north–west–north (NWN), and north–north (NN) areas. With $K_{max} = 5$–7, the HMRF model yielded five clusters, three of which coincided with the $K_{max} = 4$ run while the fourth (S) was split into two subsets with random shapes. The spatial interaction parameter ψ had posterior mode within the range (0.6, 0.8) (95% credible interval). However, the random shapes of the two S subclusters were an indicator that the MCMC runs might have not converged, perhaps due to the large amount of computational resource spent in the estimation of ψ. Therefore we performed 10 additional runs of the algorithm for two values of the interaction parameter $\psi = 0.7$–0.8. The runs that reached the highest likelihood resulted in the same four clusters as previously observed (see Figure 3). Inferences carried out under a fixed large value of ψ usually favor cluster configurations made of few large clusters. The fact that the HRMF model obtained the same clusters as STRUCTURE gave evidence that these original clusters were robust to the inclusion of a spatial prior. A by-product of the HMRF model is its ability to infer inbreeding coefficients. The inbreeding coefficients posterior estimates were computed as $\phi_{NN} = 0.022$, $\phi_{NWN} = 0.006$, $\phi_M = 0.013$, and $\phi_S = 0.007$. These small values were consistent with the observation that STRUCTURE worked well for this data set. The HMRF model with fixed parameter setting converged faster than the full-Bayes version. (We used 1000 cycles for $K_{max} = 4$ and 20,000 cycles for $K_{max} = 7$.) GENELAND runs at fixed $K = 4$–6 produced the same assignment results as the HMRF model (5000 cycles). Using reversible jumps, the posterior distribution of K exhibited a mode at $K = 5$ and a 95% credible interval $K \in (4, 8)$ (50,000 cycles).

Human data: We used the Human Genome Diversity Panel–Centre d'Etude du Polymorphisme Humain (HGDP–CEPH) (CANN *et al.* 2002) to further assess the influence and the benefit of including spatial continuity prior hypotheses in the analysis of multilocus genotypes. The HGDP–CEPH diversity panel data set contains 1056 individuals genotyped at 377 autosomal microsatellite loci. It was first studied with the software STRUCTURE by ROSENBERG *et al.* (2002). Without using predefined populations, six main genetic clusters were identified, five of which corresponded to major geographic regions. Here we restricted the study to the Eurasian and East Asian populations, including samples with distinct origins, 8 from Pakistan, 16 from China, and 1 from Siberia, Japan, and Cambodia (451 individuals). Two reasons could be given for limiting the study to Eurasian and East Asian populations. First, these populations contained two of the five main clusters as well as the sixth cluster found by ROSENBERG *et al.* (2002). Second, the 27 populations live on a same mainland, which justified using the Dirichlet tiling without modifying the neighborhood structure (although our computer program makes this

TABLE 3

Latitudes and longitudes for the eight Pakistan samples (from CANN *et al.* 2002)

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

possible). Coordinates of individuals in each sample were not known explicitly. Instead they were available as sample intervals from CANN *et al.* (2002). For instance, the Kalash from Pakistan have longitudes in the range 35°–37° E and latitudes in the range 71°–72° N. Individual coordinates were generated randomly within the specified intervals. We checked that the results presented here were rather independent of the individual coordinates within each sample (not reported).

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

To further highlight the potential of the HMRF model, we focused on the Pakistan data set and the retrieval of the Kalash cluster. The Kalash sample contains 25 of the 200 individuals from the eight Pakistan samples. Ranges for sample spatial coordinates are reported in Table 3 (CANN *et al.* 2002), and a representation of the resampled individual locations is displayed in Figure 4. In this study, data sets with 40, 30, and 20 randomly chosen loci were extracted from the Pakistan data set. The idea here is to use the results from a large number of loci as the "correct" answer and then see which methods are able to get this correct answer with fewer loci. Because all the extracted data sets did not contain the same amount of information about genetic structure, we distinguished three distinct levels of potential difficulty (strong clustering, SC; weak clustering, WC; and no cluster, NC) according to the following classification. For each subset, we preliminarily

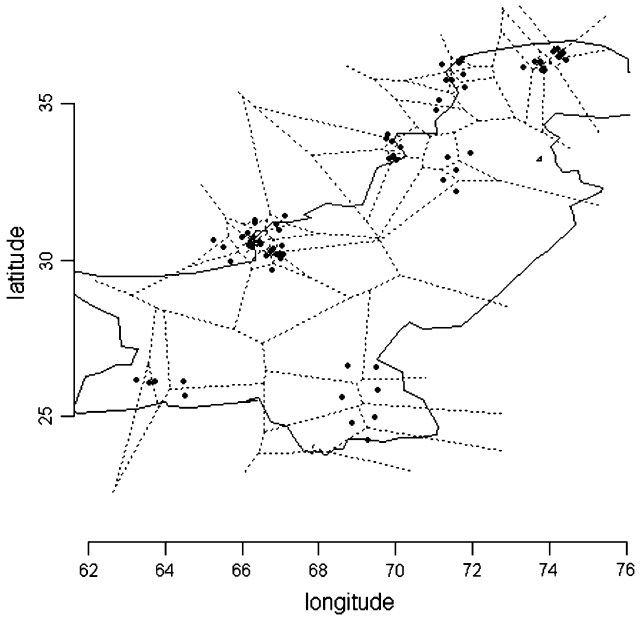


FIGURE 4.—Sampled geographical coordinates of 70 individuals from the Pakistan data set and the associated Dirichlet tiling. (The full sample was not shown but a similar spatial distribution was assumed for the 200 individuals.)

computed a neighbor-joining (NJ) tree using the shared allele distance (see NEI and KUMAR 2000), which separated the Pakistan samples in two sister clades. Data sets for which one clade contained >20 Kalash grouped against the remaining Pakistan representatives were classified as SC. Such data sets were expected to be easy for Bayesian clustering algorithms, because a more basic analysis gives a correct answer. As well there were data sets for which no obvious clusters could be directly inferred from the NJ tree. These data sets were classified as NC, and they were expected to be difficult for Bayesian clustering algorithms. We added an intermediate class, WC, for which the Kalash sample generally formed a cluster in the NJ tree, but this was done in association with other samples such as Pathan or Balochi/Brahui. With 40 randomly chosen loci, ~38% of all data sets were in the SC category, 24% were classified as WC, and the remaining 38% were NC. One NJ tree clustered the Balochi/Brahui against the rest of Pakistan. With 30 loci, these numbers changed to SC + WC = 42% and NC = 58%, and NC increased to 76% in the 20-loci data sets. These ratios were obtained from 300 distinct data sets.

We performed 10 runs of the HMRF model for 42 subsets (21 subsets with 40 loci and 21 subsets with 30 loci). The HMRF model was first run for 200 cycles at $\psi = 0.4$, and these cycles were followed by a further 500–1000 cycles at $\psi = 0.6$. For the CEPH diversity panel data set, this strategy appeared more efficient than the full-Bayes approach, which was statistically unable to identify the Kalash (we attributed this failure to the algorithmic complications and the approximations made

in estimating ψ). The run with the highest likelihood was saved as the final result. The same strategy was also used at $\psi = 0$ with a larger total number of cycles (up to 2000). Small burn-in (10 cycles) and thinning (1 cycle) periods were implemented. We first used $K_{max} = 2$ in both the HMRF and nonspatial versions. To compare with published results, we also assumed absence of inbreeding.

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

DISCUSSION

Detecting population subdivision is a subject of great interest to population geneticists, and a large body of approaches have been developed for this. In this study, we have presented a Bayesian clustering algorithm that incorporates hidden Markov random fields as prior distributions on cluster configurations. Markov random fields are mathematical models that account for the "continuity" of discrete random variables on a graph or a network (for a rigorous definition of continuity in this context, refer to Preston 1974). The term hidden means that the cluster configuration is unobserved and is instead reconstructed from an MCMC algorithm. In spatial population genetics the term continuous population usually refers to S. Wright's famous concept of isolation by distance (Wright 1943), which can in turn be understood in terms of the stepping-stone model (Kimura and Weiss 1964; Rousset 2004). Because it considers interacting demes on a lattice, the stepping-stone model exhibits the same type of spatial Markov property as does the Potts model. Inserting the stepping-stone model into a Bayesian framework generates conceptual difficulties because its stationary distribution has no known formulation. However, the HMRF model may capture its essential properties.

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

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

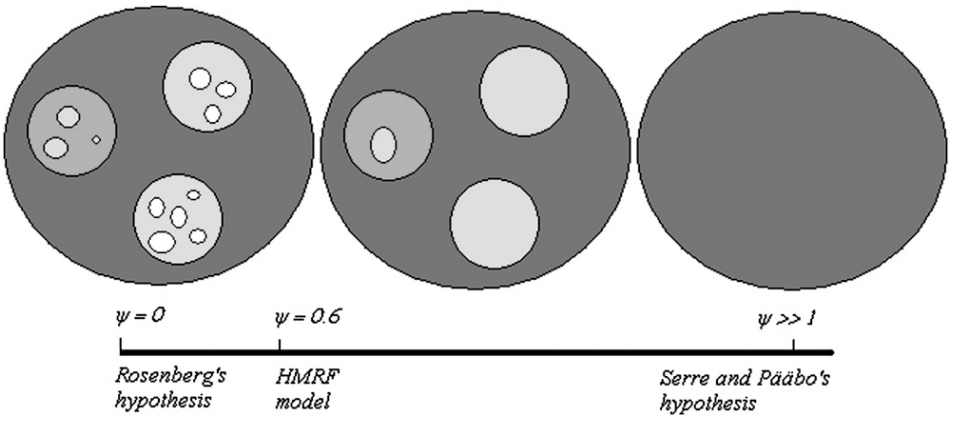


FIGURE 5.—The reconciliation illustrated. At the left of the ψ-axis, a clustering analysis does not account for the spatial continuity of allele frequencies and may detect more clusters than actually exist. At the right, the pure continuity hypothesis assumes no cluster. Here the vision is intermediate, with the main discontinuities confirmed, but some small clusters may be considered nonsignificant.

The use of the HMRF model has been illustrated in two previously published data sets. The Scandinavian brown bear is an example of a population with a strong female phylopatry. Scandinavian bears were almost exterminated at the beginning of the 20th century. After efforts to protect the species in Sweden, the bear population has recovered from four female concentration areas. Until recently these areas were believed to represent the surviving relict subpopulations after the 1930s bottleneck (see, *e.g.*, WAITS *et al.* 2000). Using two independent methods (neighbor-joining trees and the Bayesian clustering algorithm STRUCTURE), MANEL *et al.* (2004) found four genetic clusters that matched with geographical clusters, but two of them were distinct from the original female concentration areas. Using a coalescent approach, BLUM *et al.* (2004) computed the female dispersal rate and found an estimate of 9 km per generation. Because of the low dispersal rate in this population, local genetic similarities can be considered as a reasonable assumption to be included in a Bayesian model for brown bear genetic diversity. The HMRF model has been used for detecting geographical discontinuities in allele frequencies. The results confirmed previously published results and provided reasonable estimates for the number of clusters.

Using the human CEPH diversity panel data set, we checked whether the clusters obtained without spatial priors were robust to the hypothesis of continuous geographical variation in allele frequencies. The results presented here reconciled the two apparently divergent perspectives of ROSENBERG *et al.* (2002, 2005) and SERRE and PÄÄBO (2004), which brought into conflict clines and clusters regarding variation of human diversity. Restricting to Eurasian and Asian populations and working with a prior on continuous variation ($\psi \approx 0.6$), we recovered the three main clusters found by the algorithm STRUCTURE. Some important facts must be mentioned at this stage:

1. The two main clusters (Pakistan/non-Pakistan) were identified with <20 randomly chosen loci. The Kalash cluster was identified using <50 loci.
2. More importantly, the algorithm was unable to confirm the presence of other clusters in the Pakistan and East Asia areas, perhaps due to the simultaneous effects of reducing the number of loci (<120 loci) and imposing the continuity prior. The combination of these effects may have led to the neglect of some very small discontinuities that were previously detected when STRUCTURE was used with large values of K and a larger number of loci. We performed 10 additional runs of the HMRF model using the full set of loci. Regarding the Pakistan data, we were also not able to retrieve other clusters. Regarding the East Asia data set, we identified one additional cluster in the northeastern area that matched with the Yakut–Japanese samples. This cluster was also apparent in the NJ tree.
3. The weight given to the prior distribution was a moderate value that also corresponded to the posterior mean estimated from the full-Bayes algorithm when it converged [$\psi \approx 0.6$, 95% credible interval (0.5, 0.9)].
4. A stronger level of prior interaction (*e.g.*, $\psi \approx 1$) led to a unique cluster and gave strong support to Serre and Pääbo's hypothesis of clinal variation within a unique cluster.
5. Weaker levels of prior interaction (*e.g.*, $\psi \approx 0.2$) led to the same results as STRUCTURE and supported Rosenberg's small discontinuities hypothesis.
6. Here we supported the intermediate view of clinal variation of allele frequencies with a number of discontinuities smaller than those estimated by ROSENBERG *et al.* (2002). See Figure 5 for a picture of the reconciliation.

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

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

We are grateful to Noah Rosenberg for his suggestions on an early version of this manuscript. We thank Stephanie Manel, Oscar Gaggiotti, and Chibiao Chen for fruitful discussions and Mathieu Emily for his help with simulations of the Potts model on a Dirichlet tiling. We are also grateful to two anonymous reviewers for their constructive comments. O.F. was supported by grants from the Algorithmes et populations biologiques-Institut Informatique et Mathématiques Appliquées de Grenoble project and the French ministry of research Action Concertée Incitative-Interface Mathématique physique Biologie project.

LITERATURE CITED

Bamshad, M., S. Wooding, W. Watkins, C. Ostler and M. e. a. Batzer, 2003 Human population genetic structure and inference of group membership. Am. J. Hum. Genet. **72:** 578–589.

Besag, J., 1986 On the statistical analysis of dirty pictures. J. R. Stat. Soc. Ser. B **48**(3): 259–302.

Bishop, C., 1995 *Neural Networks for Pattern Recognition*. Clarendon Press, Oxford.

Blum, M. G. B., C. Damerval, S. Manel and O. François, 2004 Brownian models and coalescent structures. Theor. Popul. Biol. **65:** 249–261.

Cann, H., C. Toma, L. Cazes, M. Legrand, V. Morel *et al.*, 2002 A human genome diversity cell line panel. Science **296:** 261–262.

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

Coulon, A., G. Guillot, J. Cosson, J. Angibault, S. Aulagnier *et al.*, 2006 Genetics structure is influenced by lansdcape features. Empirical evidence from a roe deer population. Mol. Ecol. **15:** 1669–1679.

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

Destrempes, F., M. Mignotte and J.-F. Angers, 2005 A stochastic method for Bayesian estimation of hidden Markov random field models with application to a color model. IEEE Trans. Image Proc. **14:** 1097–1108.

Gelman, A., and X. Meng, 1998 Simulating normalizing constants: from importance sampling to bridge sampling to path sampling. Stat. Sci. **13:** 163–185.

Geman, S., and D. Geman, 1984 Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. IEEE Trans. Patt. Anal. Machine Intell. **6:** 721–741.

Green, P., and S. Richardson, 2002 Hidden Markov models and disease mapping. J. Am. Stat. Assoc. **97**(460): 1055–1070.

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

Guillot, G., F. Mortier and A. Estoup, 2005 Geneland: a computer package for landscape genetics. Mol. Ecol. Notes **5**(3): 708–711.

Guttorp, P., 1995 *Stochastic Modelling of Scientific Data*. Chapman & Hall, London/New York.

Hartl, D., and G. Clark, 1997 *Principles of Population Genetics*. Sinauer Associates, Sunderland MA.

Hoerl, A., and R. Kennard, 1970 Ridge regression: biased estimation for nonorthogonal problems. Technometrics **12:** 55–67.

Hurn, M., O. Husby and H. Rue, 2003 A tutorial in image analysis, pp. 87–141 in *Spatial Statistics and Computational Methods* (Lecture Notes in Statistics), edited by J. Møller. Springer, Berlin/Heidelberg, Germany/New York.

Kimura, N., and G. Weiss, 1964 The stepping stone model of population structure and the decrease of genetic correlation with distance. Genetics **49:** 561–575.

Malécot, G., 1948 *Les Mathématiques de l'Hérédité*. Masson, Paris.

Manel, S., E. Bellemain, J. Swenson and O. François, 2004 Assumed and inferred spatial structure of populations: the Scandinavian brown bears revisited. Mol. Ecol. **13:** 1327–1331.

Metropolis, N., A. Rosenbluth, M. Rosenbluth, A. Teller and E. Teller, 1953 Equation of state calculations by fast computing machines. J. Chem. Phys. **21:** 1087–1092.

Nei, M., and S. Kumar, 2000 *Molecular Evolution and Phylogenetics*. Oxford University Press, London/New York/Oxford.

Potts, R., 1952 Some generalized order-disorder transformations. Proc. Camb. Philos. Soc. **48:** 106–118.

Preston, C., 1974 *Gibbs States on Countable State Space*. Cambridge University Press, Cambridge, UK.

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

Ramachandran, S., N. Rosenberg, L. Zhivotovsky and M. Feldman, 2004 Robustness of the inference of human population structure: a comparison of X-chromosomal and autosomal microsatellites. Hum. Genomics **1:** 87–97.

Ripley, B., 1988 *Statistical Inference for Spatial Processes*. Cambridge University Press, Cambridge, UK.

Ripley, B., 1996 *Pattern Recognition and Neural Networks*. Oxford University Press, Oxford.

Rosenberg, N., J. Pritchard, J. Weber, H. Cann, K. Kidd *et al.*, 2002 Genetic structure of human populations. Science **298:** 2981–2985.

Rosenberg, N., S. Saurabh, S. Ramachandran, C. Zhao, J. Pritchard *et al.*, 2005 Clines, clusters, and the effect of study design on the influence of human population structure. PLoS Genet. **1**(6): 660–671.

Rousset, F., 2004 *Genetic Structure and Selection in Subdivided Populations*. Princeton University Press, Princeton, NJ.

Serre, D., and S. Pääbo, 2004 Evidence for gradients of human genetic diversity within and among continents. Genome Res. **14:** 1679–1685.

Thomas, D., D. Stram, D. Conti, J. Molitor and P. Marjoram, 2003 Bayesian spatial modeling of haplotype associations. Hum. Hered. **56:** 32–40.

Tibshirani, R., 1996 Regression shrinkage and selection via the lasso. J. R. Stat. Soc. Ser. B **58:** 267–288.

Waits, L., P. Taberlet, J. Swenson, F. Sandegren and R. Franzen, 2000 Nuclear DNA microsatellite analysis of genetic diversity and gene flow in the Scandinavian brown bear *Ursus arctos*. Mol. Ecol. **9:** 610–621.

Wasser, S., A. Shedlock, K. Comstock, E. Ostrander, B. Mutayoba *et al.*, 2004 Assigning African elephants DNA to geographic region of origin: applications to the ivory trade. Proc. Natl. Acad. Sci. USA **101**(41): 14847–14852.

Wright, S., 1943 Isolation by distance. Genetics **28:** 114–138.

Wu, F., 1982 The Potts model. Rev. Mod. Phys. **54:** 235–268.

Zhang, Y., M. Brady and S. Smith, 2001 Segmentation of brain MR Images through a hidden Markov random field model and the Expectation-Maximization algorithm. IEEE Trans. Med. Imag. **20:** 45–57.

Communicating editor: M. Nordborg

APPENDIX: DETAILS OF MARKOV CHAIN MONTE CARLO COMPUTATIONS

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

Updating allele frequencies $f_{k\ell j}$: We used a componentwise Metropolis–Hastings Markov chain simulation algorithm. For the cluster labeled k and locus labeled ℓ, an update of $(f_{k\ell 1}, \ldots, j_{k\ell J_\ell})$ selected two alleles at random with indexes j and j' and proposed to change their frequencies $f_{k\ell j}$ and $f_{k\ell j'}$ as follows. Denoting $a = 1 - \sum_{\substack{m\neq j \\ m\neq j'}} f_{k\ell m}$, new frequencies $f^*_{k\ell j}$ and $f^*_{k\ell j'}$ are

proposed as $f^*_{k\ell j} = aB_f$ and $f^*_{k\ell j'} = a - f^*_{k\ell j}$, where B_f is sampled from a beta $\mathcal{B}(\alpha, \alpha)$ distribution (often $\alpha = 1$). This move was accepted with probability

$$1 \wedge \frac{\pi(z \mid \theta^*)}{\pi(z \mid \theta)} \frac{f_{k\ell j}(1 - f_{k\ell j})}{f^*_{k\ell j}(1 - f^*_{k\ell j})}. \tag{A1}$$

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

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

$$1 \wedge \frac{\pi(z \mid \theta^*)\phi^{*3}_k(1 - \phi^*_k)^{39}}{\pi(z \mid \theta)\phi^3_k(1 - \phi_k)^{39}} \tag{A2}$$

as we assumed a uniform prior on ϕ_k and made a symmetric proposal.

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

$$1 \wedge \frac{\pi(z \mid \theta^*)}{\pi(z \mid \theta)} \frac{\pi(c^*)}{\pi(c)} \tag{A3}$$

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

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

where

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

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

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

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

3.6.2 Article 2 paru dans le Journal de la Société Française de Statistique

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

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

ABSTRACT

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

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

RÉSUMÉ

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

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

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

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

* Corresponding author: Olivier.Francois@imag.fr

1. Introduction

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

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

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

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

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

2. Hierarchical Bayes model

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

3. Scandinavian brown bears

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

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

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

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

Because of the low dispersal rate, continuity can be considered as a reasonable assumption to be included in a Bayesian model for Scandinavian brown bear genetic diversity. We analysed the same data set as did the two previous studies. We first used a full-Bayes approach where the prior on ψ was uniform over $(0, 1)$. Values of ψ in this range allowed the prior coexistence of several clusters (simulations not reported), and we ran the algorithm with $K_{\max} = 4 - 7$. After 30,000 cycles, the runs with $K_{\max} = 4$ led to the same clusters as described by Manel *et al.* (2004). We referred to these clusters as the S (South), M (Middle), NWN (North West North) and NN (North North) areas. With $K_{\max} = 5 - 7$, the HMRF model yielded 5 clusters, three of which coincided with the $K_{\max} = 4$ run and the fourth (S) was splitted into two subsets with random shapes. The spatial interaction parameter ψ had posterior mode in the range $(0.6, 0.8)$ (95% credible interval). However, the irregular shapes of the two S subclusters indicated that the MCMC might have not been run long enough for warranting convergence, perhaps due to the large amount of computational resource spent into the estimation of ψ. Therefore we performed 10 additional runs of the algorithm for two values of interaction parameter $\psi = 0.7 - 0.8$. The runs that reached the highest likelihood resulted in the same four clusters as previously observed (see Fig. 1).

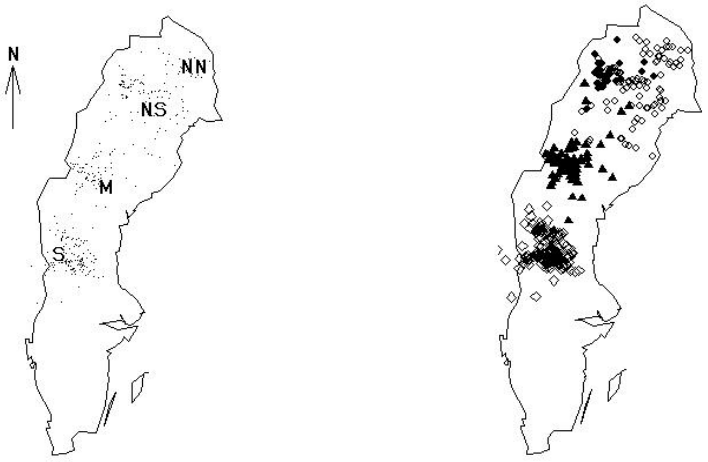


FIG 1. — Left: The spatial distribution of brown bears in Scandinavia. The four subpopulations (NN, NS, M, S) were defined as areas of female concentration. Right: Estimated cluster configuration using the HMRF model. $\psi = 0.7$ and $K_{\max} = 6$. Two clusters (diamonds, triangles) coincided with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

4. Discussion

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

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

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

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

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

References

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

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

Dawson K. and K. Belkhir (2001). A Bayesian approach to the identification of panmictic populations and the assignment of individuals. Genetical Research **78:** 59-77.

Destrempes F., M. Mignotte, and J.-F. Angers (2005). A Stochastic Method for Bayesian Estimation of Hidden Markov Random Field Models With Application to a Color Model. IEEE Transactions on Image Processing **14:** 1097-1108.

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

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

KIMURA N. and G. WEISS (1964). The stepping stone model of population structure and the decrease of genetic correlation with distance. Genetics **49:** 561-575.

MALÉCOT G. (1948). *Les Mathématiques de l'Hérédité*. Paris: Masson.

MANEL S., E. BELLEMAIN, J. SWENSON, and O. FRANÇOIS (2004) Assumed and inferred spatial structure of populations: the Scandinavian brown bears revisited. Molecular Ecology **13:** 1327-1331.

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

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

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

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

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

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

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

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

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

WAITS L., P. TABERLET, J. SWENSON, F. SANDEGREN, and R. FRANZEN (2000). Nuclear DNA microsatellite analysis of genetic diversity and gene flow in the Scandinavian brown bear *Ursus arctos*. Molecular Ecology **9:** 610-621.

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

3.7 Conclusions

La détection et la localisation de lignes de discontinuités génétiques est une étape clé pour comprendre et prédire les processus d'évolution de la diversité génétique. Elle nécessite le développement de modèles pertinents pour la classification d'individus en populations génétiquement homogènes à partir de données génotypiques multilocus géo-référencées ce qui pose pas mal de difficultés techniques. Dans ce contexte, l'approche hiérarchique semble apporter des solutions car elle traite ces difficultés une à une. Elle facilite la description de la structure génétique d'un ensemble d'individus qu'elle décompose en deux étapes : la première permet de décrire l'organisation spatiale de la diversité génétique et la seconde permet de décrire les propriétés génétiques des individus à l'intérieur d'une même population. Autrement dit, elle permet de distinguer deux niveaux de variabilité génétique : la variabilité inter-populations qui est modélisée dans la couche phénoménologique latente et la variabilité intra-populations qui est modélisée au niveau des observations.

Ce chapitre donne au lecteur un aperçu de la grande flexibilité offerte par l'approche hiérarchique en modélisation statistique. Structure, Geneclust et Geneland sont trois modèles dédiés à la classification d'individus en population génétiquement homogènes en vue de la détection de lignes de discontinuités génétiques dans un espace géographique donné. Ce qui distingue ces trois modèles ce sont les *briques* de base qui les composent et auxquelles sont associées des hypothèses biologiques spécifiques. Dans ce chapitre, j'ai décrit cinq briques élémentaires et trois combinaisons possibles de certaines de ces briques. Toutes les autres combinaisons sont également possibles (e.g., consanguinité + tessellation de Voronoï). Par ailleurs, les modèles de base peuvent être complexifiés. Pour exemple, diverses extensions au modèle Structure ont été proposées (Falush & al., 2003) dont notamment une complexification du modèle d'occurrence des génotypes. Ainsi, au lieu de supposer que chaque individu provient d'une des K populations (*no-admixture model*), une hypothèse alternative est que les individus peuvent avoir des ancêtres dans plus d'une population (*admixture model*). Par ailleurs, plutôt que de supposer les fréquences alléliques indépendantes d'une population à l'autre, une hypothèse alternative est de supposer qu'elles se ressemblent. Le modèle de Falush et al. (2003) supposent que les fréquences alléliques en un locus donné dans une population k dérivent toutes d'un même jeu de fréquences alléliques lié à une hypothétique population ancestrale :

$$f_{k,l,.} \sim \mathcal{D}irichlet\left(f_{Al1}\frac{1-d_k}{d_k}, ..., f_{AlJ_l}\frac{1-d_k}{d_k}\right) \qquad \forall k \in \{1,2,..,K\} \qquad \forall l \in \{1,2,..,L\} \tag{3.7.6}$$

où f_{Alj} désigne la fréquence de l'allèle j au locus l dans la population ancestrale et d_k un facteur de dérive propre à la population k. La prise en compte du degré de ressemblance entre les patterns génétiques de chaque population s'obtient facilement en donnant aux fréquences alléliques le statut de variables aléatoires latentes conditionnées par des fréquences alléliques ancestrales f_{Alj} (l=1,2,..,L, j=1,2,..,J_l) et des facteurs de dérives $(d_1, d_2, .., d_K)$. Plus de détails sur ce modèle sont donnés dans l'article (Guillot & al., 2005) dans lequel est également testée une version de Geneland où les fréquences alléliques suivent le modèle de Falush.

Le modèle de Potts inclus dans la couche latente du modèle Geneclust permet d'introduire des dépendances entre les génotypes des individus échantillonnés, à partir d'un graphe de Delaunay (cf section 3.2.3). Un des inconvénients majeur à l'utilisation de ce type de graphe de voisinage est qu'il est très sensible à la procédure d'échantillonnage.

Aussi, une perspective intéressante serait de réaliser une étude par simulations permettant d'apprécier la sensibilité des classifications d'individus obtenues avec le modèle Geneclust en fonction de l'échantillonnage.

Au cours de ce travail, j'ai implémenté un package R interfacé fortran commenté qui permet de réaliser la classication de données génotypiques multilocus géoréférencées avec le modèle Geneclust. Une de mes perspectives est de soumettre ce package sur le site CRAN de R afin de promouvoir l'utilisation de ce modèle auprès de la communauté des statisticiens et des généticiens.

Au delà de la génétique des populations, les modèles Structure, Geneland et Geneclust peuvent s'adapter à la classification de n'importe quel jeu de données catégorielles multivariées et spatialement structurées. Par exemple, ils peuvent servir à la détection de communautés d'espèces. Considérons un domaine D dans lequel la présence/absence respective de L espèces a été observée en n sites d'échantillonnage. L'objectif est de définir les espèces qui ont tendance à coexister aux mêmes sites. Autrement dit, cela revient à regrouper les sites qui partage un pattern de présence/absence similaire. Par analogie avec la génétique des populations, on peut assimiler les populations à délimiter aux communautés d'espèces, les individus aux sites, les génotypes aux espèces et les allèles aux états absence/présence. En chaque site , une variable catégorielle à 2 états possibles, présence (1) ou absence (0), est associée à chaque espèce. Supposons que chaque communauté d'espèces k soit régie par un jeu de probabilités de présence particulier des L espèces $(p_{k,1,1}, p_{k,2,1}, ..., p_{k,L,1})$ et que les patterns de présence entre sites géographiques proches se ressemblent davantage que ceux entre sites éloignés et le tour est joué. En effet, on se retrouve dans des conditions quasi-analogues à un problème de classifications d'individus en population génétiquement homogènes permettant l'utilisation des modèles Geneland ou Geneclust.

Troisième partie

Modéliser des données de biomasse *zero-inflated* et géo-référencées

Les données issues de relevés d'abondance (e.g. comptages, biomasses) ont souvent la propriété d'être *zero-inflated* i.e., caractérisées par une très forte occurrence de valeurs nulles (cf section 2.1.1). Le traitement et la prédiction de telles données requièrent l'utilisation de modèles statistiques spécifiques. En effet, dans le cas de données de comptage, les lois de probabilités traditionnellement utilisées pour modéliser des événements rares (e.g., Poisson, binomiale négative) s'ajustent mal car leurs hypothèses distributionnelles ne sont pas vérifiées. En particulier, elles ont tendance à sous-estimer la probabilité d'occurrence de zéros (cf section 2.1.1). De même, la relation moyenne-variance de ces modèles n'est généralement pas vérifiée. L'unique paramètre de la loi de Poisson n'est pas suffisant pour décrire les données. Par ailleurs, dans le cas de données continues, nous avons vu dans le chapitre 1 que le recours à des lois diffuses (normal, gamma, lognormal) est impossible car ces modèles supposent que la probabilité d'occurrence d'un zéro est nulle.

Ridout et al. (1998) proposent une revue bibliographique détaillée des différents modèles utilisés pour la modélisation de données *zero-inflated* discrètes. Les modèles ZIP et ZINB décrits dans la section 2.1.1 font partie des structures les plus courantes. Diverses extensions à ces modèles auxquels ont été ajoutées des variables explicatives sous forme de modèles linéaires généralisés et/ou des effets aléatoires (Hall, 2000) ont été proposées. Dans cette partie, l'approche hiérarchique est exploitée pour la modélisation de données de biomasses *zero-inflated* continues.

Cette étude est le fruit d'une collaboration avec Hugues Benoît, écologue au sein du Centre des Pêches du Golfe (Moncton, Canada). Tous les modèles définis sont basés sur l'application pratique dont les besoins sont nés. L'objectif principal est d'analyser la distribution de certains invertébrés marins du Golfe du Saint-Laurent à partir des relevés au chalut de fond réalisés par le Centre des Pêches du Golfe (cf section 1.1.2). Deux sous-objectifs principaux sont visés :

- Valider ou invalider l'hypothèse selon laquelle chacune de ces espèces a une répartition spatialement structurée dans le sud du Golfe-du-Saint-Laurent.
- Déterminer des facteurs d'impact susceptibles d'expliquer l'existence et la distribution de chacune de ces espèces

Ma démarche de travail s'est composée de quatre étapes. La première a consisté à proposer un modèle hiérarchique alternatif aux modèles *delta*, classiquement utilisés pour modéliser des données continues *zero-inflated*, qui soit cohérent par rapport à tout changement d'effort d'échantillonnage. Puis, pour valider l'hypothèse d'existence de structuration spatiale dans la répartition des espèces, je me suis attachée à modéliser différents liens de "ressemblance" entre les régions d'échantillonnage. Ces liens ont été introduits dans la couche phénoménologique latente du modèle *delta* et du nouveau modèle proposé. J'ai comparé les différentes structures hiérarchiques proposées. En particulier, j'ai calculé des critères de sélection de modèles bayésiens et me suis intéressée au problème de la sensibilité du facteur de Bayes par rapport aux choix de lois *a priori*. Enfin, j'ai travaillé en collaboration avec Vincent Garreta, doctorant au CEREGE d'Aix-en-Provence, sur la spécification d'un modèle hiérarchique permettant de représenter la répartition spatiale conjointe de deux espèces à partir de données de biomasse *zero-inflated* multivariées.

Quatre chapitres composent cette partie. Dans le premier chapitre, je décris les modèles *delta* et mets en évidence leurs principales limites dans le cas de relevés d'abondance à efforts d'échantillonnage variables. Dans le deuxième chapitre, je propose l'utilisation d'un modèle hiérarchique alternatif, baptisé loi des fuites, qui comble les lacunes des modèles *delta*. Dans le troisième chapitre, j'ajoute à la loi des fuites et aux modèles *delta* une couche

latente supplémentaire pour la prise en compte d'effets aléatoires liés à l'hétérogénéité spatiale du phénomène modélisé. Enfin, le dernier chapitre est consacré au problème de la modélisation hiérarchique spatiale multivariée.

L'inférence bayésienne des modèles hiérarchiques décrits dans cette partie a été effectuée avec des algorithmes MCMC implémentés dans OpenBUGS (Spiegelhalter & al., 2002) (Version 2.1) couplé au package BRugs du logiciel R Project for Statistical Computing. La convergence des algorithmes a été vérifiée par un examen visuel des chaines MCMC et par le calcul de la statistique de Brooks et Gelman (1998) décrite dans l'Annexe B.

Chapitre 4

Les modèles Delta : description et limites dans le cas de relevés à efforts d'échantillonnage variables

4.1 Hypothèses et notations

Soit D un domaine d'échantillonnage et $Y = \{Y_k; k = 1, .., r\}$ les variables aléatoires désignant les biomasses respectives en r sites de D. Elles prennent leurs valeurs dans $[0, +\infty]$. Nous notons $y = \{y_k, k = 1, 2, ..., r\}$ une réalisation de Y.

S_k désigne l'effort d'échantillonnage associé à la mesure y_k (k=1,..,r). Il peut s'agir d'une aire d'échantillonnage, d'une durée d'observation, d'un volume d'eau ou d'air filtré. Dans le cas particulier des relevés du Centre des Pêches du Golfe (cf section 1.1.2), il s'agit de la distance (en miles) parcourue par le chalut lors de chaque trait.

Plaçons-nous dans le cas simple où le domaine d'échantillonnage D est suffisamment petit pour abriter un milieu de vie homogène et adapté à l'existence de l'espèce étudiée. Cela suppose que les conditions de vie (e.g., topographie, température, nourriture...) soient comparables en chaque site. Dans notre cas d'étude par exemple, chaque strate est supposée définir une région aux conditions de vie homogènes. Compte-tenu de cette hypothèse, nous supposons que la probabilité d'occurrence d'un zéro et la distribution des biomasses strictement positives sont indépendantes du site d'échantillonnage et identiquement distribuées dans D.

Dans le cas des relevés du Centre des Pêches du Golfe, l'échantillonnage est aléatoire et les distances entre chaque trait de chalut sont grandes. Ainsi, nous supposons que les variables observables Y_k (k=1,..,r) sont indépendantes.

Enfin, par simplicité, le niveau de capturabilité du chalut est considéré constant dans l'espace et pour chaque espèce d'intérêt. Il est confondu avec les autres sources de variabilité dûes au processus d'échantillonnage.

4.2 Le modèle

Pour modéliser des données continues *zero-inflated*, une approche usuelle simple, à l'origine proposée par Aitchison (1955), consiste à spécifier une structure mathématique composée d'une probabilité discrète d'occurrence de zéros et d'une densité continue pour

les valeurs strictement positives. Les modèles de ce type, baptisés modèles *delta*, s'écrivent comme une combinaison linéaire pondérée d'une masse de Dirac en zéro et d'une loi paramétrique continue définie sur $]0, +\infty[$.

Compte-tenu des hypothèses précédemment énoncées (cf section 4.1), considérons une version simple des modèles *delta* pour laquelle la fonction de répartition de Y_k (k=1,2,..,r) s'écrit :

$$[Y_k = 0] = \delta \tag{4.2.1}$$

$$[Y_k \leq y_k] = \delta + (1 - \delta) G_\theta \left(\frac{y_k S}{S_k} \right) \qquad \forall y_k > 0 \tag{4.2.2}$$

où δ désigne la probabilité d'occurrence d'un zéro ($0 \leq \delta \leq 1$) pour un effort d'échantillonnage de référence S et G_θ une fonction de répartition continue qui décrit l'occurrence des valeurs strictement positives pour un même effort S. G_θ est définie sur $]0, +\infty[$ et régie par les paramètres θ.

Les modèles *delta* peuvent aussi être décrits sous un angle de vue conditionnel. Ils proposent une description probabiliste de l'occurrence ou non d'une valeur nulle en chaque site et, conditionnellement à l'occurrence d'une quantité de biomasse strictement positive, ils spécifient la distribution des niveaux d'abondance. Soit X_k la variable observable binaire qui prend la valeur 0 si aucune espèce est observée au site k et 1 sinon. Les modèles *delta* supposent que :

$$X_k \overset{i.i.d}{\sim} \mathcal{B}ernoulli(1 - \delta) \qquad k = 1, ..., r$$

Conditionnellement aux variables X_k, les Y_k sont ensuite supposées indépendantes et distribuées comme suit :

$$Y_k | X_k = 0 \sim \delta_0$$
$$Y_k | X_k = 1 \sim g_\theta$$

où δ_0 désigne la loi de Dirac en 0 et g_θ la densité de probabilité conditionnelle associée à la fonction de répartition $G(\theta)$. La formule des probabilités totales appliquée, à partir des deux hypothèses ci-dessus, permet de retrouver facilement les relations 4.2.1 et 4.2.2.

Comme mis en évidence par la description conditionnelle ci-dessus, les modèles *delta* font partis de la catégorie des *two-parts models* car ils traitent séparément les zéros issus d'une loi de Bernoulli de paramètre δ et les données strictement positives issues de la loi G_θ. Les zéros n'ont qu'une seule origine distributionnelle possible.

Dans la littérature, G_θ désigne le plus souvent une loi lognormale ou une loi gamma. Aitchison (1955) propose d'utiliser une structure delta-lognormale pour modéliser des dépenses particulières de ménages caractérisées par une forte proportion de zéros. En écologie marine, Pennington (1983), Lo et al. (1992) et plus récemment Fletcher (2005) utilisent une distribution lognormale pour modéliser des abondances strictement positives en poissons ou planctons marins. Stefansson (1996) utilise plutôt une loi gamma pour modéliser les abondances strictement positives en haddocks islandais*. Enfin, en météorologie, Coe et Stern (1982) utilise un modèle delta pour modéliser des données de pluie et choisissent une loi gamma pour traiter les quantités de pluie strictement positives.

Myers et Pepin (1990) recommandent l'utilisation d'une loi gamma lorsque la probabilité d'occurrence de faibles valeurs non-nulles est forte. Ceci est notamment le cas pour les données d'invertébrés du Saint-Laurent (cf figure 1.6). Aussi, dans ce qui suit, G_θ désigne

∗. Le haddock ou aiglefin ou églefin (Melanogrammus aeglefinus) est un poisson de fond

la fonction de répartition d'une loi gamma. En notant $a > 0$ son paramètre d'échelle et $b > 0$ l'inverse de son paramètre de forme, la fonction de répartition de Y_k devient :

$$[Y_k = 0] = \delta$$
$$[Y_k \leq y_k] = \delta + (1-\delta)\frac{b^a}{\Gamma(a)}\left(\frac{y_k S}{S_k}\right)^{a-1} exp\left(-b\left(\frac{y_k S}{S_k}\right)\right)$$

Le modèle *delta* associé est appelé modèle *delta-gamma* et noté ($\Delta\Gamma$) dans la suite du document.

L'espérance et la variance du modèle $\Delta\Gamma$ sont facilement calculables en chaque site. Elles sont données par :

$$\mathbb{E}(Y_k) = (1-\delta)\frac{a}{b}$$
$$\mathbb{V}ar(Y_k) = (1-\delta)\frac{a}{b^2} + \frac{a^2}{b^2}(1-\delta)\delta$$

Le modèle $\Delta\Gamma$ compte trois paramètres inconnus (δ, a, b). Les rapports $\frac{a}{b}$ et $\frac{a}{b^2}$ désignent respectivement l'espérance et la variance des quantités de biomasse non nulles mesurées lorsque l'espèce est présente. δ désigne la probabilité d'occurrence d'un zéro lors d'une procédure d'échantillonnage à effort fixé S.

L'inférence bayésienne du modèle $\Delta\Gamma$ a été menée avec des lois *a priori* non-informatives. Nous avons choisi une loi *a priori* uniforme sur $[0, 1]$ pour le paramètre δ, la probabilité d'occurrence d'une valeur nulle. Pour les paramètres a et b de la loi gamma, nous avons opté pour des lois normales tronquées en zéro de moyenne nulle et de variance 10^6. D'autres choix auraient été possibles. L'un des avantages du modèle $\Delta\Gamma$ est que les paramètres inconnus δ, $\frac{a}{b}$ et $\frac{a}{b^2}$ ont un sens physique simple : probabilité de ne pas ramasser l'espèce à effort d'échantillonnage fixé, espérance et variance des quantités de biomasse non nulles mesurées lorsque l'espèce est présente. Aussi, il aurait été possible de recourir à des experts ou d'utiliser des données de biomasse issues de relevés d'abondance réalisés dans le même domaine de suivi mais autres que celui du Centre des Pêches du Golfe afin de définir des priors informatifs pour ces paramètres.

4.3 Les inconvénients principaux des modèles Delta

4.3.1 Traitement séparé des zéros et non-zéros

La modélisation séparée des zéros et non-zéros facilite l'inférence du modèle $\Delta\Gamma$ qui peut s'effectuer en deux temps. Un recodage en 0/1 (zéro/non-zéro) des données y_k permet d'estimer le paramètre δ. Quant aux données y_k non nulles, elles permettent d'inférer le couple de paramètres (a, b)

En contrepartie, le modèle $\Delta\Gamma$ ne tient pas compte d'une possible cohérence entre la probabilité d'occurrence d'un zéro et la distribution des quantités de biomasses strictement positives pour un effort S. Or, intuitivement, on s'attend à ce que la probabilité d'occurrence d'un zéro soit forte lorsque les quantités de biomasse recueillies sont principalement faibles. Aussi, une modélisation séparée des zéros et non-zéros empêche de tirer parti de l'information contenue dans les données d'abondance non nulles de l'espèce étudiée pour inférer la probabilité d'occurrence d'un zéro.

4.3.2 Le modèle $\Delta\Gamma$ n'est pas stable par addition

Soient Y_i et $Y_{i'}$ les quantités de biomasses recueillies dans un domaine homogène lors de deux traits de chalut indépendants (non nécessairement contigus) dont les efforts d'échantillonnage respectifs sont les aires A_i et $A_{i'}$ balayées par le chalut (supposées disjointes). Supposons que Y_i et $Y_{i'}$ suivent une même loi de probabilité $\mathcal{L}$ de paramètres respectifs $\theta(A_i)$ et $\theta(A_{i'})$. Une propriété naturellement souhaitée est que $Y_i + Y_{i'}$ suive la même loi $\mathcal{L}$, régie par les paramètres $\theta(A_i \cup A_{i'})$. Plus généralement, cette propriété doit être vérifiée pour tout type de relevé d'abondance et d'effort d'échantillonnage. Le modèle $\Delta\Gamma$ a pour inconvénient majeur de ne pas vérifier cette propriété d'additivité. La démonstration s'écrit facilement avec le calcul de la fonction caractéristique du modèle, notée $\phi_{\Delta\Gamma}$.

Definition 2. *La fonction caractéristique d'une variable aléatoire absolument continue X est donnée par la transformée de Fourier de sa densité de probabilité f :*

$$\phi_X(\omega) = \mathbb{E}(e^{i\omega X}) = \int_{-\infty}^{+\infty} e^{i\omega x} f(x)dx$$

La fonction caractéristique d'une loi gamma de paramètres a et b , notée ϕ_g, est donnée par :

$$\phi_g(\omega) = \int_0^{+\infty} e^{i\omega x} \frac{b^a}{\Gamma(a)} x^{a-1} e^{-bx} dx = \frac{1}{(1 - \frac{i\omega}{b})^a} \tag{4.3.3}$$

La fonction caractéristique d'une loi $\Delta\Gamma$ se déduit de la relation 4.3.3 et s'écrit :

$$\phi_{\Delta\Gamma}(\omega) = \delta + (1 - \delta)\phi_g(\omega) \tag{4.3.4}$$

Considérons à présent deux variables aléatoires indépendantes, Y_1 et Y_2, obéissant à une même loi $\Delta\Gamma$ de paramètres (δ, a, b). L'indépendance des variables implique que la fonction caractéristique de la somme $Y_1 + Y_2$ s'écrit comme le produit de leur fonction caractéristique respective, donnée par la relation 4.3.4. Il vient :

$$\phi_{Y_1+Y_2}(\omega) = \delta^2 + 2\delta(1 - \delta)\phi_g(\omega) + (1 - \delta)^2(\phi_g(\omega))^2 \tag{4.3.5}$$

L'écriture 4.3.5 n'est pas la fonction caractéristique d'une loi $\Delta\Gamma$ donnée par 4.3.4 : la somme $Y_1 + Y_2$ ne suit pas un modèle $\Delta\Gamma$.

Par conséquent, le modèle $\Delta\Gamma$ n'assure pas de cohérence distributionnelle entre des quantités de biomasse mesurées à effort d'échantillonnage variable. Par exemple, la quantité de biomasse recueillie sur un trait de chalut de 2 miles n'obéit pas au même modèle statistique que la quantité de biomasse recueillie sur un trait de chalut de 1 mile. Le modèle $\Delta\Gamma$ ne permet donc pas de comparer des relevés d'abondance basés sur des efforts d'échantillonnage distincts. Mais comment justifier que le modèle $\Delta\Gamma$ soit plus légitime pour un effort d'échantillonnage donné que pour un autre ? En effet, quel que soit l'effort d'échantillonnage considéré, la quantité de biomasse recueillie est le résultat d'un même mécanisme aléatoire : elle doit donc intuitivement conserver les mêmes propriétés distributionnelles.

4.3.3 Les données sont préalablement standardisées pour se ramener à un effort de référence

Nous avons vu que le modèle $\Delta\Gamma$ n'a pas la capacité à s'adapter structurellement en cas de variabilité de l'effort d'échantillonnage. Pourtant, en pratique, les efforts d'échantillonnage sont souvent variables. Prenons le cas des relevés d'abondance du Golfe-du-Saint-Laurent. L'effort standard ciblé est de 1.75 miles (cf chapitre 1) mais, en pratique, il varie comme l'illustre l'histogramme de la figure 4.1 des distances chalutées entre 1999 et 2001. En effet, la vitesse du chalutier fluctue en fonction des vents et des courants marins ce qui influe sur la distance parcourue en 30 minutes. Par ailleurs, les traits de chalut réalisés dans des sites aux fonds marins accidentés sont souvent écourtés pour éviter de déchirer le filet du chalut.

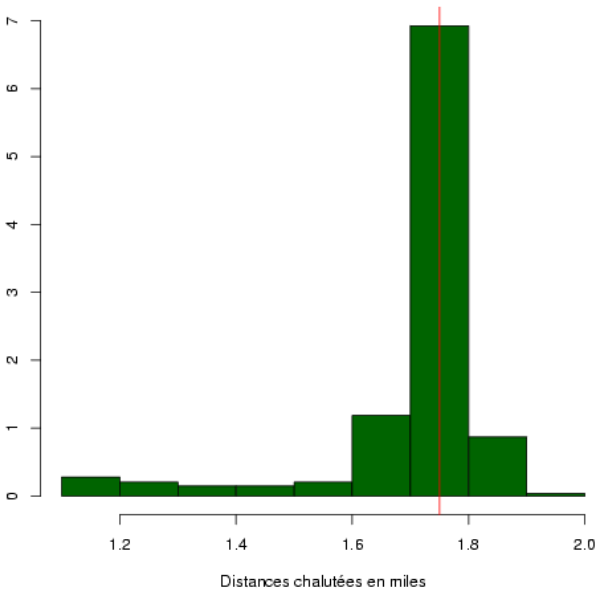


Figure 4.1 – Distances chalutées lors des 540 traits de chalut réalisés dans le Golfe-du-Saint-Laurent entre 1999 et 2001. Le trait vertical rouge indique la distance d'échantillonnage ciblée : 1.75 miles.

Une approche courante consiste à standardiser au préalable les données d'abondance afin de se ramener à un même effort S. Dans le cas de données d'abondance non nulles, cette standardisation permet d'augmenter ou de diminuer linéairement l'abondance mesurée en fonction de l'effort d'échantillonnage. Par contre, elle ne permet pas de modifier l'occurrence des zéros observée en fonction de l'effort d'échantillonnage réalisé.

Plutôt que de standardiser les données, Stefansson (1996) a proposé une extension au modèle $\Delta\Gamma$ qui permette de prendre en compte l'effort d'échantillonnage. Il suggère de recourir à un modèle linéaire généralisé et d'inclure l'effort de chalutage comme variable explicative de l'espérance des quantités de biomasse strictement positives. Le modèle

s'écrit :

$$[Y_k = 0] = \delta_k$$
$$[Y_k \leq y_k] = \delta_k + (1 - \delta_k)\frac{b_k^{a_k}}{\Gamma(a_k)}\left(\frac{y_k S}{S_k}\right)^{a_k - 1} exp\left(-b_k\left(\frac{y_k S}{S_k}\right)\right)$$

La probabilité d'occurrence d'un zéro δ_k et l'espérance des quantités de biomasse strictement positives $\eta_k = \frac{a_k}{b_k}$ dépendent désormais du site d'échantillonnage k. Une fonction de lien log-linéaire permet de relier η_k à l'effort d'échantillonnage :

$$log(\eta_k) = \alpha_1 + \beta_1 l_k$$

où l_k désigne l'effort d'échantillonnage au site k (e.g., longueur d'un trait de chalut), α_1 la quantité moyenne de biomasse strictement positive (à échelle log) susceptible d'être ramassée sur l'ensemble du domaine d'étude et β_1 l'effet de la covariable indiquant l'effort d'échantillonnage. Malheureusement, comme le modèle $\Delta\Gamma$ induit un traitement séparé des zéros et non-zéros, la probabilité d'occurrence de zéros reste indépendante de l'effort d'échantillonnage. Considérons deux traits de chalut indépendants Y_1 et Y_2 d'aires d'échantillonnage respectives A_1 et A_2. Supposons que $A_1 < A_2$. Le modèle suppose que :

$$[Y_1 = 0] = [Y_2 = 0]$$

alors qu'en pratique, dans le cas où le milieu est propice à l'existence de l'espèce, on s'attend plutôt à ce que la probabilité de ne pas ramasser de biomasse diminue quand l'aire chalutée augmente.

Dans le chapitre suivant, je décris une structure aléatoire alternative aux structures $\Delta\Gamma$ pour le traitement de données *zero-inflated*. Contrairement aux modèles $\Delta\Gamma$, ce modèle est hiérarchique : sa couche phénoménologique latente est basée sur un processus de Poisson homogène dont l'avantage est de produire une structure stable par addition avec laquelle il est possible de traiter directement des données de biomasse brutes.

Chapitre 5

Un modèle hiérarchique alternatif : la loi des fuites

Le modèle $\Delta\Gamma$, décrit dans le chapitre précédent, est la structure aléatoire actuellement utilisée pour représenter des données continues *zero-inflated*. Ce modèle a l'avantage d'être simple à comprendre et à manipuler (cf section 4.2). Toutefois, il existe une limite théorique principale à son utilisation : il n'assure pas de cohérence distributionnelle entre des observations réalisées dans un même domaine homogène mais liées à des efforts d'échantillonnage distincts. Cela implique une standardisation préalable des données qui rend impossible toute comparaison d'analyses basées sur des efforts d'échantillonnage distincts. De plus, on s'attend à ce qu'une telle standardisation biaise les résultats fournis par le modèle dans le cas de relevés à efforts d'échantillonnage variables.

L'objectif de ce chapitre est de présenter une structure aléatoire alternative au modèle $\Delta\Gamma$ pour la représentation de données de biomasses continues *zero-inflated*. Décrit avec un raisonnement conditionnel, ce modèle permet de décrire le processus d'échantillonnage d'organismes vivants, dans une région homogène. Ce modèle est hiérarchique à deux niveaux et fait partie de la classe des processus de Poisson composés. Sa couche latente est basée sur un processus de Poisson homogène marqué ce qui lui permet d'être stable par addition (Feller, 1968).

Ce chapitre s'articule en 5 parties. La première partie décrit le modèle proposé en parcourant de manière descendante le schéma de construction hiérarchique (θ, Z, Y) de la figure 2.5 du chapitre 2. La deuxième partie liste les différentes propriétés du modèle LOL afin de justifier en quoi ce modèle constitue une alternative intéressante au modèle $\Delta\Gamma$. Dans les trois dernières parties sont décrites des études par simulation réalisées pour étudier les performances d'estimation du modèle LOL et les comparer à celles du modèle $\Delta\Gamma$. J'ai considéré diverses situations dans lesquelles varient les paramètres du modèle ou le nombre d'observations ou l'effort d'échantillonnage.

5.1 Description du modèle

Dans ce chapitre, je reprends les mêmes notations et me place sous les mêmes hypothèses que celles décrites dans la section 4.1 du chapitre précédent.

5.1.1 Une idée née d'une analogie

Bernier et Fandeux (1970) ont baptisé *loi des fuites* le modèle des marques exponentielles d'un processus de Poisson associé à l'observation d'une somme de marques sur une période donnée T. Jusqu'à présent, ce modèle est utilisé dans le contexte temporel pour modéliser des processus stochastiques en temps continu (cf figure 5.1). Le nom *loi des fuites* (LOL pour *Law of Leaks*) provient d'une étude des distributions des débits de fuites sur une conduite de gaz (Morlat, n.d.). Un processus de Poisson localise les trous le long d'un tronçon de conduite. Si chaque fuite locale donne une quantité aléatoire exponentielle de gaz alors le débit total sur le tronçon suit le modèle LOL. Si aucune fuite ne se produit alors le débit total de gaz est nul. Le modèle LOL permet donc de décrire des données positives continues caractérisées par une probabilité non nulle d'occurrence de zéros.

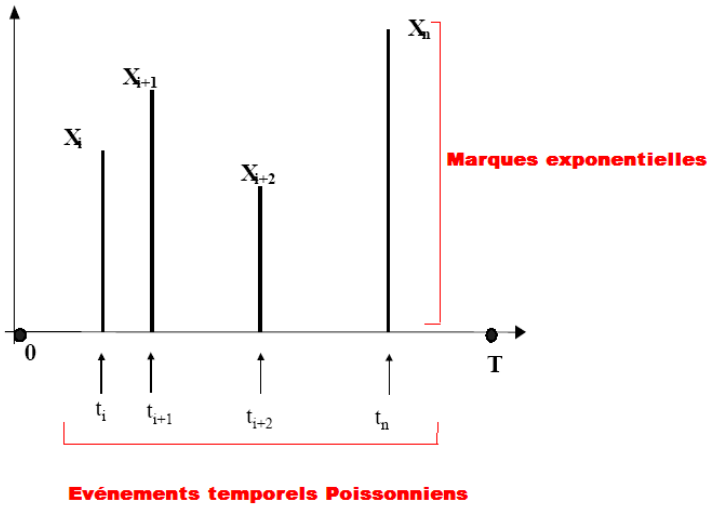


Figure 5.1 – La loi des fuites est à l'origine utilisée pour modéliser des processus temporels en temps continu

Nous proposons de transposer le modèle LOL vers l'écologie des populations en lui donnant une interprétation écologique liée au processus d'échantillonnage d'organismes vivants. A la différence des applications temporelles réalisées jusqu'à présent, nous proposons une extension du modèle LOL permettant de représenter un processus spatial. Plutôt que d'observer des marques exponentielles sur une période de temps T, nous considérons des marques exponentielles associées à un ensemble de points irrégulièrement distribués dans un sous-domaine spatial de $\mathbb{R}^2$. Nous passons des processus ponctuels marqués temporels aux processus ponctuels marqués spatiaux. Pour chaque couche du modèle sont définies les hypothèses conceptuelles et mathématiques sous-jacentes à chaque modèle élémentaire.

5.1.2 Un processus ponctuel marqué latent décrit l'organisation spatiale de la biomasse

La couche latente du modèle LOL, étendue dans $\mathbb{R}^2$, est basée sur un processus ponctuel spatial marqué. Celui-ci permet de modéliser l'organisation spatiale de la biomasse

en une espèce donnée.

Considérons le domaine d'échantillonnage borné D de $\mathbb{R}^2$ et une espèce d'intérêt (cf section 4.1). Imaginons que cette espèce se répartisse aléatoirement dans D sous forme de *gisements* contenant chacun une quantité aléatoire inconnue de biomasse. Ces *gisements* forment le semi de points du processus ponctuel spatial dont les marques sont les quantités de biomasse, supposées aléatoires, contenues dans chaque *gisement*. Une telle hypothèse répond à l'hypothèse écologique selon laquelle les processus biologiques comme la dispersion, la reproduction et la compétition entre organismes vivants créent des structures spatiales en *patch* des organismes vivants.

Nous supposons le domaine D suffisamment petit pour abriter des conditions de vie homogènes (cf section 4.1) et adaptées à l'existence de l'espèce étudiée. Sous cette hypothèse, la répartition spatiale des *gisements* dans D est supposée suivre un processus de Poisson homogène :

Definition 3. *Dans D, un processus de Poisson homogène est défini comme suit :*
- *Pour $\mu > 0$ et A un sous-domaine de D, le nombre de points du semi tombant dans A, noté N(A), suit une loi de Poisson d'intensité $\mu|A|$ où $|A|$ est l'aire de A*
- *Pour deux régions disjointes A et A', les variables aléatoires N(A) et N(A') sont indépendantes*

Tout processus d'échantillonnage réalisé dans D revient à échantillonner selon ce processus de Poisson homogène. Considérons par exemple un trait de chalut qui, rappelons-le, consiste à balayer les fonds marins avec un filet afin de ramasser les espèces qui y vivent. Au cours d'un trait de chalut réalisé autour du site k, un nombre aléatoire N_k de *gisements*, uniformément répartis sur la trajectoire d'échantillonnage, est collecté et vérifie :

$$N_k \sim Poisson(S_k\mu) \qquad \forall k \in \{1, ..., r\} \tag{5.1.1}$$

où μ désigne le nombre moyen de *gisements* collectés dans D par unité d'effort d'échantillonnage et S_k l'effort d'échantillonnage associé à la mesure d'abondance y_k. Plus le paramètre μ prend une valeur importante, plus la région D abrite un nombre important de *gisements* de l'espèce d'intérêt.

D'après la définition 3, les variables N_k (k=1, 2, ..., r) sont supposées indépendantes. Cela signifie que les divers échantillonnages réalisés dans D sont considérés comme des expériences aléatoires échangeables. Quel que soit l'éloignement géographique des sites d'échantillonnage, il n'existe aucune forme de dépendance entre les nombres de *gisements* collectés. Par ailleurs, les variables N_k ne sont pas identiquement distribuées car elles dépendent linéairement de l'effort d'échantillonnage fourni. Ainsi, pour revenir à l'exemple d'un trait de chalut, cela signifie que plus la surface chalutée est importante, plus la probabilité de ramasser un grand nombre de *gisements* est élevée.

Chaque *gisement* ramassé $g_k = 1, .., N_k$ contient une quantité aléatoire inconnue de biomasse M_{g_k}. Les variables latentes M_{g_k} sont les marques du processus ponctuel spatial. Nous reprenons la version la plus simple du modèle LOL et supposons que les M_{g_k} sont indépendantes et identiquement distribuées selon une loi exponentielle de paramètre ρ :

$$M_{g_k} \sim^{i.i.d} Exp(\rho) \qquad g_k = 1, ..., N_k \tag{5.1.2}$$

La loi exponentielle a l'avantage d'être parcimonieuse puisqu'elle ne dépend que d'un unique paramètre ρ. Ce choix permet également de récupérer une propriété de conjugaison

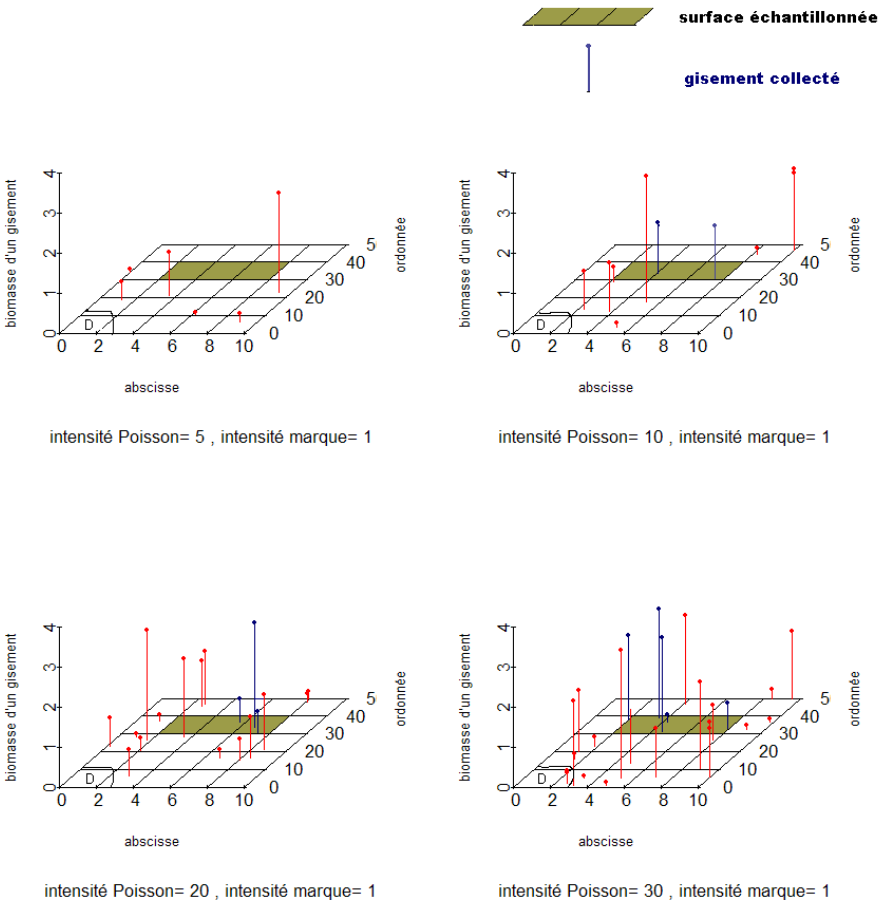


Figure 5.2 – 4 réalisations de processus de Poisson homogène marqués exponentiellement, dans un domaine D de $\mathbb{R}^2$. L'intensité de la marque est constante égale à 1. Seule l'intensité du nombre moyen de *gisements* varie.

qui facilite l'inférence bayésienne du modèle (cf Annexe B). D'un point de vue écologique, la relation 5.1.2 signifie que la probabilité qu'un *gisement* contienne une grande quantité de biomasse décroît exponentiellement et que la quantité moyenne de biomasse contenue dans un *gisement* est $\frac{1}{\rho}$.

La figure 5.2 donne une illustration de la couche cachée du modèle LOL. Sous ce modèle, les données de biomasses issues de relevés d'abondance proviennent d'un échantillonnage selon un processus de Poisson homogène à marques exponentielles.

5.1.3 Le modèle des observations

La quantité de biomasse Y_k observée au site k (k=1,2,..,r) s'écrit comme la somme des quantités de biomasse contenues dans chaque *gisement* collecté :

$$Y_k = \begin{cases} \sum_{g_k=1}^{N_k} M_{g_k} & \text{si} \quad N_k \geq 1 \\ 0 & \text{si} \quad N_k = 0 \end{cases}$$

Par définition, une absence de *gisement* (i.e., $N_k = 0$) induit une valeur nulle et l'occurrence d'au moins un *gisement* (i.e., $N_k \geq 1$) induit une valeur strictement positive. Celle-ci est distribuée comme la somme aléatoire de variables aléatoires indépendantes de

loi exponentielle de paramètre ρ : il s'agit donc d'une loi gamma de paramètre de forme N_k et de paramètre d'échelle $\gamma = \frac{1}{\rho}$. Ainsi défini, le modèle des observations est finalement un processus de Poisson composé d'intensité μ et de fonction de saut exponentielle.

Conditionnellement au nombre inconnu de *gisements* collectés N_k et au paramètre ρ, la quantité de biomasse Y_k observée au site k suit :

$$Y_k|N_k,\rho \sim \begin{cases} \Gamma(N_k,\rho) & \text{si} \quad N_k \geq 1 \\ \delta_0 & \text{si} \quad N_k = 0 \end{cases} \tag{5.1.3}$$

où ρ désigne l'inverse de la quantité moyenne de biomasse contenue dans un *gisement*.

La relation ci-dessus indique que seules les quantités inconnues N_k et ρ sont nécessaires pour définir la loi de Y_k. Les variables latentes M_{g_k} facilitent la description conditionnelle du modèle LOL mais leur quantification n'est pas nécessaire pour définir la loi des Y_k.

Conditionnellement aux paramètres μ et ρ, une intégration sur les différentes valeurs possibles de la variable latente N_k permet d'obtenir, pour chaque observation y_k (k=1,2, ..,r), l'expression de la densité de probabilité du modèle LOL :

$$\begin{aligned}[y_k|\mu,\rho] &= \sum_{n=0}^{+\infty}[y_k|n,\rho][n|\mu,\rho] \\ &= \begin{cases} \sum_{n=1}^{\infty}\left(e^{-\mu S_k}\frac{(\mu S_k)^n}{n!}\right)\frac{\rho^n}{\Gamma(n)}y_k^{n-1}e^{-\rho y_k} & \text{si} \quad y_k > 0 \\ e^{-\mu S_k} & \text{si} \quad y_k = 0 \end{cases}\end{aligned}$$

Les fonctions de Bessel permettent d'obtenir une écriture compacte de cette densité de probabilité sous la forme :

$$[y_k|\mu,\rho] = 1_{y_k=0}e^{-\mu S_k} + 1_{y_k>0}\frac{\mu S_k\rho}{\sqrt{\mu S_k\rho y_k}}e^{-S_k\mu-\rho y_k}I_1(2\sqrt{\mu S_k\rho y_k}) \tag{5.1.4}$$

où I_1 désigne la fonction de Bessel modifiée d'ordre 1 *.

Les valeurs nulles n'ont qu'une seule origine distributionnelle possible : la loi de Poisson du processus ponctuel latent. Aussi, l'occurrence de zéros est uniquement contrôlée par le paramètre d'intensité μ.

Tel que défini, le modèle LOL possède une structure hiérarchique (Y, Z, θ) à deux niveaux. Le graphe acyclique orienté du modèle LOL (cf figure 5.3) permet de distinguer clairement les variables observables $\{Y_k, k = 1, 2, .., r\}$, les variables explicatives $\{S_k, k = 1, 2, .., r\}$, les paramètres $\theta = \{\mu, \rho\}$ et les variables latentes $Z = \{N_k, k = 1, 2, .., r\}$, prises en sandwich entre les variables observables et les paramètres inconnus.

5.1.4 Le choix des lois *a priori*

L'inférence bayésienne du modèle LOL nécessite de définir au préalable une loi *a priori* sur le vecteur de paramètres $\theta = (\mu, \rho)$. Comme pour le modèle $\Delta\Gamma$ (cf section 4.2), nous avons attribué des lois *a priori* non-informatives aux paramètres μ et ρ : des lois normales tronquées en zéro et de variance 10^6. Des lois *a priori* gamma ont également été testées pour ces deux paramètres afin de bénéficier de propriétés de conjugaison. Toutefois, par

∗. Les fonctions de Bessel modifiées d'ordre α sont définies par : $I_\alpha(x) = \sum_{m=0}^{\infty}\frac{x^{2m+\alpha}}{m!(m+\alpha)!}\frac{1}{2^{2m+\alpha}}$ où α désigne un réel ou un complexe arbitraire.

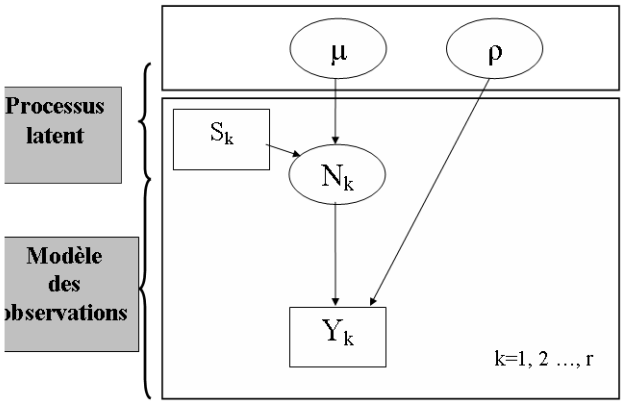


Figure 5.3 – Graphe acyclique orienté du modèle hiérarchique LOL

rapport aux lois normales tronquées, la convergence des algorithmes MCMC s'est avérée moins bonne.

Contrairement au modèle $\Delta\Gamma$ (cf section 4.2), il est difficile de spécifier une loi *a priori* informative pour les paramètres du modèle LOL. Ceci est une limite du modèle LOL par rapport au modèle $\Delta\Gamma$. En effet, le nombre attendu de *gisements* collectés pendant un processus d'échantillonnage (i.e., μ) et la quantité moyenne de biomasse contenue dans un *gisement* (i.e., $\frac{1}{\rho}$) sont des grandeurs inconnues conceptuelles. Comme elles n'ont pas de sens physique, elles sont difficiles à quantifier *a priori* d'autant plus que la définition de *gisement* dépend de l'échelle d'observation.

5.2 Pourquoi choisir le modèle LOL ?

5.2.1 Il possède une interprétation conceptuelle et physique

La notion de *gisements* utilisée pour décrire le modèle est une simple vue de l'esprit qui permet de donner une interprétation conceptuelle au modèle statistique. Elle n'a pas nécessairement un sens écologique mais elle permet de faciliter la discussion avec le biologiste. Elle permet notamment de donner un sens conceptuel aux paramètres et variables latentes du modèle LOL :

- N_k désigne le nombre aléatoire de *gisements* collectés au site k pour un effort d'échantillonnage S_k
- μ désigne le nombre moyen de *gisements* qu'on peut espérer ramasser dans D pour un effort d'échantillonnage standard
- ρ désigne l'inverse de la quantité de biomasse contenue dans un *gisement*.

Décrit avec un raisonnement conditionnel, le modèle LOL permet de décrire les différentes étapes du processus d'échantillonnage d'organismes vivants. Revenons à l'exemple d'un trait de chalut. La première étape consiste à collecter dans le filet un nombre aléatoire de

gisements le long de la trajectoire du chalutier. Ceci est modélisé dans la couche cachée du modèle à l'aide des variables aléatoires N_k et M_{g_k} (k=1,2,..,r). Puis la seconde étape du processus d'échantillonnage consiste à peser la quantité de biomasse totale contenue dans le filet à sa remontée. La variabilité des observations est ainsi modélisée à l'aide d'un processus de Poisson composé.

5.2.2 Les quantités écologiques d'intérêt sont facilement calculables

Le modèle LOL est plus parcimonieux que le modèle $\Delta\Gamma$: il possède 2 paramètres inconnus au lieu de 3 pour $\Delta\Gamma$. Par ailleurs, il permet de calculer facilement les quantités aléatoires d'intérêt pour l'écologue.

La probabilité d'occurrence d'un zéro au site k, pour un effort d'échantillonnage S_k, est donnée par :

$$[Y_k = 0|\mu, \rho] = [N_k = 0|\mu] = e^{-S_k\mu} \tag{5.2.5}$$

Contrairement au modèle $\Delta\Gamma$, l'effort d'échantillonnage joue un rôle direct sur la probabilité d'occurrence d'un zéro. Le modèle suppose une décroissance exponentielle de la probabilité d'occurrence d'un zéro en fonction de l'effort d'échantillonnage et cette décroissance est d'autant plus forte que le paramètre μ est grand. A μ fixé, plus l'effort d'échantillonnage est important (en durée, en distance ou en surface parcourue...), plus la probabilité de ramasser au moins un *gisement* augmente et ainsi, a fortiori, la probabilité de ne rien ramasser diminue. Il faut remarquer que cette hypothèse est écologiquement plausible en théorie sauf dans la situation particulière où le milieu de vie n'est pas du tout adapté à l'espèce.

L'espérance et la variance des quantités de biomasse observées sont calculables en chaque k (k=1,2,..,r) et dépendent linéairement de l'effort d'échantillonnage S_k :

$$\mathbb{E}(Y_k|\rho, \mu) = \frac{\mu S_k}{\rho}$$
$$\mathbb{V}ar(Y_k|\rho, \mu) = \frac{2\mu S_k}{\rho^2}$$

Plus l'effort d'échantillonnage est important, plus la quantité de biomasse collectée a tendance à être grande et variable à μ et ρ fixés. Le coefficient de variation $\sqrt{\left(\frac{2}{\mu S_k}\right)}$ croît vers l'infini quand le nombre moyen de *gisements* tend vers zéro.

Le paramètre μ a une double influence. Il contrôle l'occurrence de zéros et les quantités de biomasse strictement positives collectées par l'intermédiaire des variables latentes N_k :

$$\mathbb{E}(Y_k|N_k > 0, \rho) = \frac{N_k}{\rho}$$
$$\mathbb{V}ar(Y_k|N_k > 0, \rho) = \frac{N_k}{\rho^2}$$

Ainsi, contrairement au modèle $\Delta\Gamma$, le modèle LOL ne propose pas un traitement séparé des zéros et non-zéros.

5.2.3 Il possède une cohérence distributionnelle additive

La fonction caractéristique du modèle LOL de paramètres μ et ρ est donnée par :

$$\phi_{LOL}(\omega) = e^{\mu\left(\frac{i\omega S_k}{\rho - i\omega}\right)} \tag{5.2.6}$$

Soient Y_1 et Y_2 les quantités de biomasses recueillies lors de deux processus d'échantillonnage, réalisés dans D, et dont les efforts d'échantillonnage respectifs sont S_1 et S_2. Dans D, Y_1 et Y_2 suivent le modèle LOL de paramètres (μ, ρ). Leur propriété d'indépendance implique que la fonction caractéristique de la somme $Y_1 + Y_2$ s'écrit comme le produit de leur fonction caractéristique respective, donnée par la relation 5.2.6. Il vient :

$$\phi_{Y_1+Y_2}(\omega) = e^{\mu\left(\frac{i\omega(S_1+S_2)}{\rho - i\omega}\right)} \tag{5.2.7}$$

L'écriture 5.2.7 est la fonction caractéristique du modèle LOL donnée par l'équation 5.2.6. Par conséquent, la quantité de biomasse totale $Y_1 + Y_2$ dont l'effort d'échantillonnage associé est $S_1 + S_2$ suit également le modèle LOL de paramètres (μ, ρ). La figure 5.4 illustre cette propriété d'additivité.

Contrairement au modèle $\Delta\Gamma$, le modèle LOL assure une cohérence distributionnelle entre des observations liées à des efforts d'échantillonnage distincts. En effet, le modèle prévoit que, dans un domaine homogène D, la loi de probabilité suivie par les observations est le modèle LOL quelque soit l'effort d'échantillonnage fourni. Le processus de Poisson homogène latent qui compose le modèle LOL lui permet de s'adapter linéairement en fonction de l'effort d'échantillonnage fourni (cf equation 5.2.7). Ainsi, contrairement au modèle $\Delta\Gamma$, le modèle LOL permet de travailler directement avec les données de biomasse brutes.

Afin d'analyser plus en détails les avantages et les limites du modèle LOL puis de comparer ses performances d'estimation à celles du modèle $\Delta\Gamma$ en fonction du nombre d'observations ou de la variabilité de l'effort d'échantillonnage, j'ai réalisé trois études par simulations que je décris dans les sections suivantes.

5.3 Simulation 1 : performances d'estimation du modèle LOL en fonction des paramètres μ et ρ

Selon les valeurs respectives des paramètres μ et ρ, les hypothèses du modèle LOL changent et, a fortiori, la forme de la distribution (cf figure 5.5) :

1. Paramètre μ "petit" et paramètre ρ "petit" - peu de *gisements* mais beaucoup de biomasse dans chaque *gisement* - Densité LOL dissymétrique caractérisée par un grand pic en zéro et une variance forte
2. Paramètre μ "petit" et paramètre ρ "fort" - peu de *gisements* et peu de biomasse dans chaque *gisement* - Densité LOL caractérisée par un grand pic en zéro, une faible valeur moyenne et une faible variance
3. Paramètre μ "fort" et paramètre ρ "petit" - beaucoup de *gisements* et beaucoup de biomasse dans chaque *gisement* - Densité LOL peu piquée en zéro avec une forte valeur moyenne et une forte variance
4. paramètre μ "fort" et paramètre ρ "fort" - beaucoup de *gisements* mais peu de biomasse dans chaque *gisement* - Densité LOL peu piquée en zéro et à faible variance

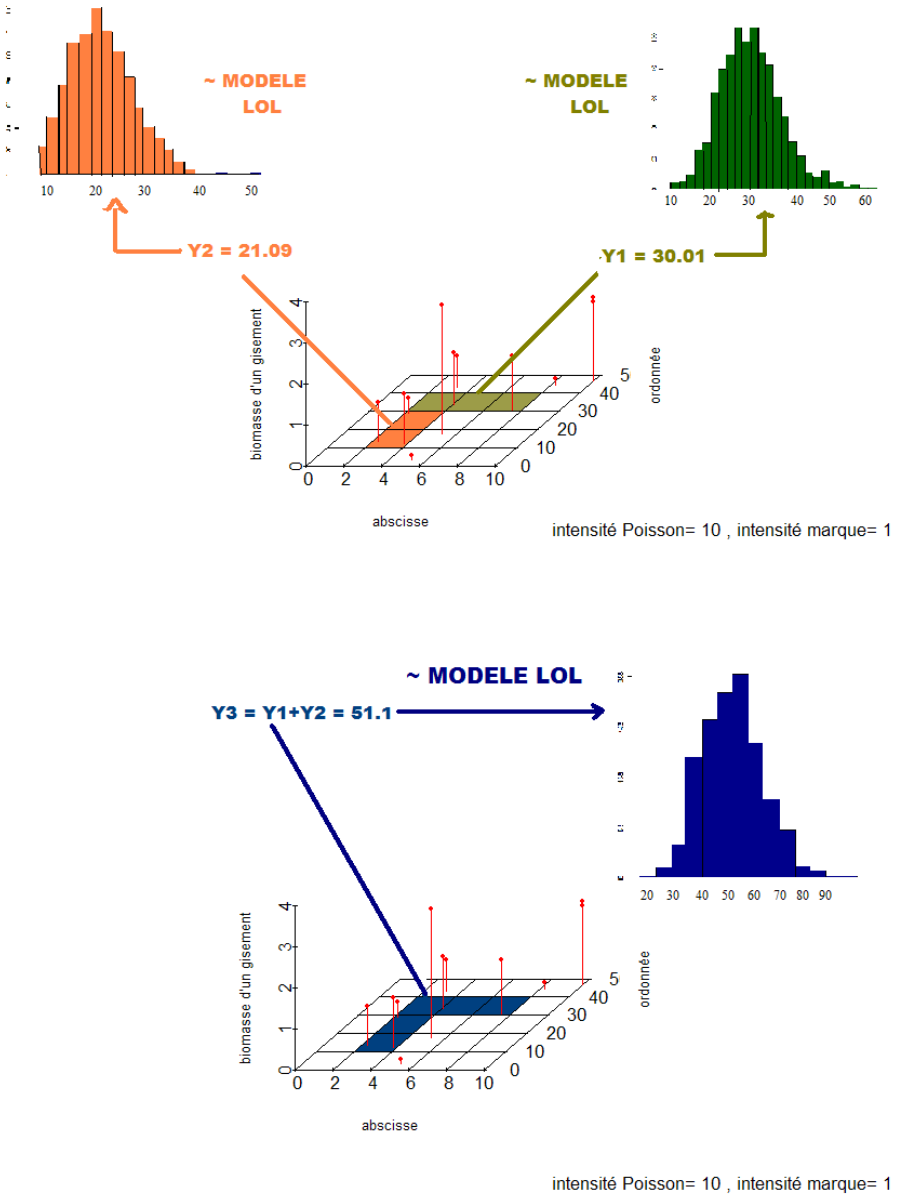


Figure 5.4 – Le modèle LOL assure une cohérence distributionnelle par rapport à tout changement d'effort d'échantillonnage

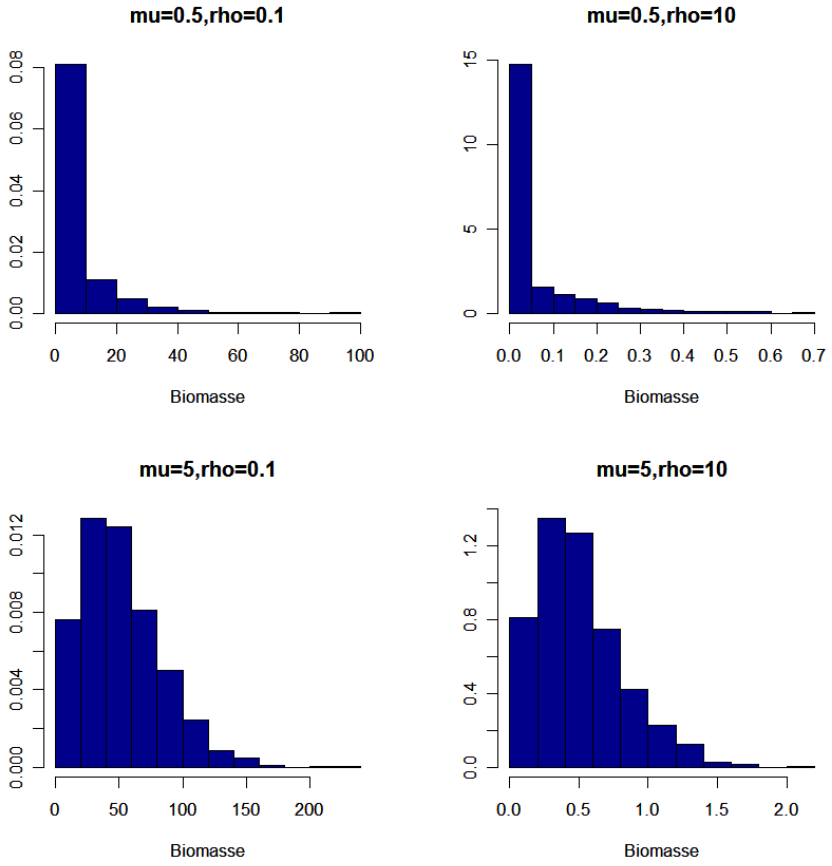


Figure 5.5 – Histogrammes empiriques de 1000 valeurs simulées selon le modèle LOL avec différentes valeurs des paramètres (μ,ρ)

Cette première simulation a pour objectif de comparer les performances d'estimation du modèle LOL sous différentes hypothèses définies en fonction des paramètres μ et ρ. J'ai montré que :

- la performance des estimations sur les paramètres μ et ρ dépend de l'occurrence de zéros.
- L'estimation du paramètre ρ est globalement plus difficile que celle du paramètre μ.
- L'estimation de la biomasse moyenne s'effectue correctement et est précise quelque soit la valeur des paramètres μ et ρ.
- il existe souvent une corrélation linéaire positive entre les chaines MCMC relatives aux paramètres μ et ρ. Cette corrélation diminue quand l'occurrence de zéros augmente. Elle joue sur la vitesse de convergence des algorithmes MCMC mais n'influe pas sur les qualités d'estimation de la biomasse moyenne.

5.3.1 Description de la simulation

J'ai considéré 8 couples de paramètres (μ,ρ). Les valeurs de ces paramètres sont indiquées dans le tableau 5.1. Elles ont été choisies afin de couvrir les 4 situations précédemment décrites et de générer des quantités de biomasse en accord avec les données de biomasses en invertébrés épibenthiques récoltées par le Centre des Pêches du Golfe (cf section 1.1.2). La valeur maximale du paramètre μ a été fixée à 5 car au-delà de cette valeur, les données

simulées ne contiennent systématiquement plus de valeurs nulles ce qui n'est plus cohérent avec l'étude. De même, la valeur minimale de μ a été fixée à 0.1 pour garantir au maximum la présence simultanée de 0 et de valeurs strictement positives dans l'échantillon de biomasses simulées.

	Config N°	μ	ρ
μ "petit", ρ "petit"	1	0.1	0.1
	2	0.5	0.1
μ "petit", ρ "fort"	3	0.5	10
	4	0.1	10
μ "fort", ρ "petit"	5	3	0.1
	6	5	0.1
μ "fort", ρ "fort"	7	5	10
	8	5	100

Tableau 5.1 – Valeurs simulées pour le couple de paramètres (μ, ρ) du modèle LOL

J'ai simulé des quantités de biomasse selon le modèle LOL, défini par les équations 5.1.1, 5.1.2 et 5.1.3, en parcourant, de manière descendante, le graphe acyclique orienté de la figure 5.3.

Pour chacune des 8 configurations testées, j'ai fixé à 20 le nombre d'observations de biomasse et 100 simulations indépendantes ont été réalisées. Par ailleurs, j'ai exclu toute variabilité dans l'effort d'échantillonnage (i.e., fixé S_k à 1 pour tout k).

Les échantillons *a posteriori* ont été calculés sur la base de 40000 itérations MCMC après une période de chauffe de 20000 itérations. Afin de réduire les autocorrélations au sein des échantillons générés, ces derniers ne sont composés que des valeurs simulées toutes les 10 itérations.

Sous le modèle LOL de paramètre (μ, ρ) et en supposant que l'effort d'échantillonnage est constant, la biomasse moyenne est définie comme le rapport des paramètres μ et ρ (cf section 5.2.2) :

$$Q = \frac{\mu}{\rho}$$

Il est ainsi facile de définir un échantillon *a posteriori* de valeurs pour Q à partir des échantillons *a posteriori* obtenus pour μ et ρ.

5.3.2 Critères de comparaison

Plusieurs critères ont été calculés pour examiner les performances d'estimation du modèle LOL dans les 8 cadres de simulation considérés.

Pour chaque paramètre θ (monodimensionnel), j'ai calculé la médiane *a posteriori* moyenne sur 100 simulations, notée $\widehat{\theta}_{100}$. J'ai choisi ces estimateurs bayésiens ponctuels car ils sont théoriquement plus adaptés que les moyennes *a posteriori* lorsque les distributions *a posteriori* sont dissymétriques, ce qui est notamment le cas pour les paramètres μ, ρ et Q. Sous l'approche bayésienne, la médiane *a posteriori* minimise le coût *a posteriori* associé à la fonction de coût absolu $c(d, \theta) = |d - \theta|$ où d désigne un estimateur ponctuel de θ (cf Annexe C). Afin de quantifier le degré de "fiabilité" (ou niveau de "risque") associé à une estimation de θ par la médiane *a posteriori*, j'ai calculé l'approximation suivante du

coût *a posteriori* moyen à partir des échantillons *a posteriori* obtenus sur 100 simulations indépendantes :

$$C_{100}(\theta) = \frac{1}{100}\sum_{l=1}^{100} C_l(\widehat{\theta}^{(l)}) \qquad \text{avec} \qquad C_l(\widehat{\theta}^{(l)}) = \frac{1}{G}\sum_{g=1}^{G} |\widehat{\theta}^{(l)} - \theta_g^{(l)}|$$

où G désigne la taille de l'échantillon *a posteriori* (G=2000) , $\widehat{\theta}^{(l)}$ la médiane *a posteriori* et $\theta_g^{(l)}$ la g-ième valeur générée pour l'échantillon *a posteriori* l (l=1,2,..,100). Considérons un paramètre monodimensionnel θ. Un coût *a posteriori* élevé pour l'estimation de θ signifie que l'échantillon *a posteriori* associé est fortement dispersé autour de la médiane *a posteriori*. Ainsi, la forte incertitude *a posteriori* sur l'estimation de ce paramètre augmente le risque d'erreurs d'estimation et/ou le risque décisionnel si on se réfère à la médiane *a posteriori* comme estimateur ponctuel. Par la suite, je dirai que l'estimation de θ est peu "fiable" lorsque le coût *a posteriori* associé est élevé. A noter que les coûts *a posteriori* associés à la fonction de coût absolu dépendent de la valeur de θ. Comparer des coûts *a posteriori* n'a donc de sens que pour des valeurs similaires de θ.

Pour chaque paramètre μ, ρ et Q, j'ai également calculé le pourcentage de recouvrement, noté R, défini comme le pourcentage d'intervalles de crédibilité à 95% contenant la valeur simulée du paramètre. Il constitue un indicateur de la qualité d'ajustement du modèle.

J'ai calculé le coefficient de variation *a posteriori* moyen sur 100 simulations, noté $CV_{100}(\theta)$, des échantillons *a posteriori* obtenus pour les paramètres μ, ρ et Q. Les coefficients de variation *a posteriori* mesurent la dispersion relative des échantillons *a posteriori* par rapport à leur moyenne : ils sont définis comme le rapport de l'écart-type par la moyenne *a posteriori*. Par rapport à des écarts-types *a posteriori*, ils ont l'avantage de permettre de comparer des précisions d'estimation basées sur des valeurs simulées différentes. Ils constituent des indicateurs de la précision des estimations obtenues : plus CV_{100} est petit, plus les estimations sont précises.

Enfin, pour chaque paramètre inconnu θ (monodimensionnel), j'ai examiné la position des distributions *a posteriori* par rapport à la valeur simulée $\widetilde{\theta}$. Un indicateur possible de ce positionnement est donné par la probabilité *a posteriori* suivante :

$$[\theta \geq \widetilde{\theta}|Y] \tag{5.3.8}$$

Elle permet de savoir si les valeurs les plus probables *a posteriori* ont tendance à se répartir autour de $\widetilde{\theta}$ ou à se répartir principalement "au-dessus" ou "en-dessous" de $\widetilde{\theta}$. Dans cette optique, j'ai calculé l'approximation suivante de 5.3.8 à partir des échantillons *a posteriori* obtenues sur 100 simulations indépendantes :

$$\mathbb{P}_{100}(\theta) = \frac{1}{100}\sum_{l=1}^{100} p_l(\theta) \qquad \text{avec} \qquad p_l(\theta) = \frac{|g \in \{1,2,\ldots,G\}; \theta_g^{(l)} \geq \widetilde{\theta}|}{G}$$

où || désigne le cardinal d'un ensemble, G la taille de l'échantillon *a posteriori* et $\theta_g^{(l)}$ la g-ième valeur générée pour l'échantillon *a posteriori* l (l=1,2,...,100). $\mathbb{P}_{100}(\theta)$ désigne la proportion moyenne de valeurs *a posteriori* supérieures ou égales à la valeur simulée $\widetilde{\theta}$. En particulier, elle indique si la médiane *a posteriori* a plutôt tendance à être proche de $\widetilde{\theta}$ ou au contraire, à le sur ou sous estimer. Trois cas de figure sont envisageables (cf figure 5.6) :

– Si $\mathbb{P}_{100}(\theta) \simeq 0.5$ alors les valeurs les plus probables *a posteriori* se répartissent en moyenne autour de $\widetilde{\theta}$. En particulier, la médiane *a posteriori* est très proche de la valeur simulée.
– Si $\mathbb{P}_{100}(\theta) < 0.5$ alors les valeurs les plus probables *a posteriori* sont en moyenne strictement inférieures à $\widetilde{\theta}$. En particulier, la médiane *a posteriori* aura tendance à sous-estimer $\widetilde{\theta}$
– Si $\mathbb{P}_{100}(\theta) > 0.5$ alors les valeurs les plus probables *a posteriori* sont en moyenne strictement supérieures à $\widetilde{\theta}$. En particulier, la médiane *a posteriori* aura tendance à sur-estimer $\widetilde{\theta}$

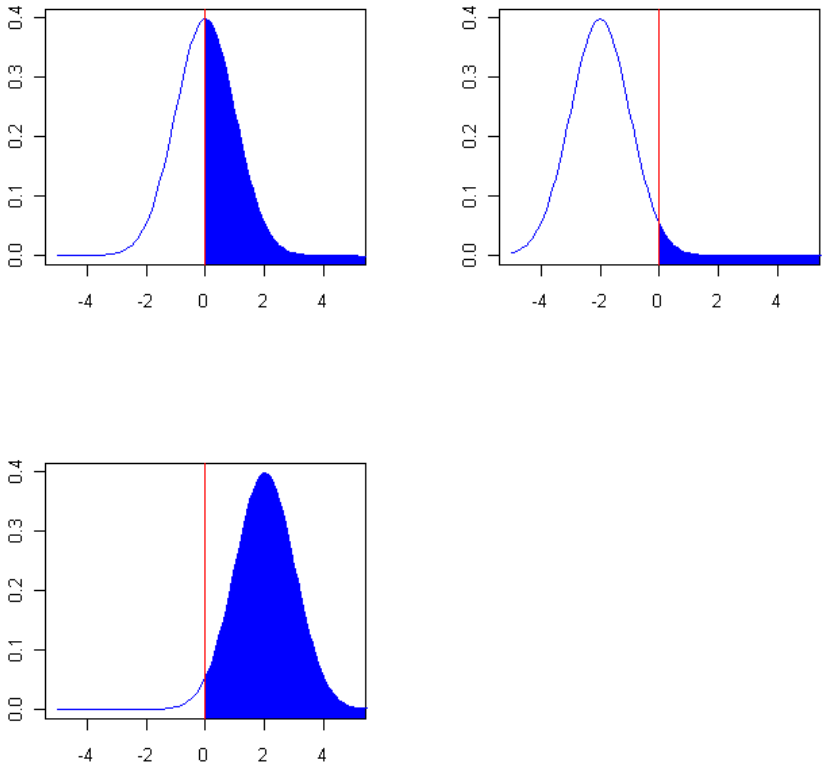


Figure 5.6 – Trois répartitions possibles d'une distribution *a posteriori* normale par rapport à une valeur simulée $\widetilde{\theta}$ (ligne verticale rouge). L'aire de la partie bleue représente la probabilité $[\theta \geq \widetilde{\theta}|Y]$.

5.3.3 Résultats

La figure 5.7 contient les boxplots à 95% (calculés sur 100 simulations) des médianes *a posteriori* (en haut) et des coefficients de variation *a posteriori* (en bas) obtenus pour l'estimation des paramètres μ,ρ et Q pour chacune des 8 configurations testées. De manière générale, elle montre que ces paramètres sont correctement estimés via la médiane *a posteriori* : les boxplots contiennent systématiquement la valeur simulée des paramètres. Par ailleurs, elle montre que la précision des estimations augmente quand l'occurrence de zéros

augmente. En effet, les coefficients de variation sont plus faibles sur les dernières configurations pour lesquelles μ prend une plus forte valeur. Les fortes incertitudes d'estimation obtenues pour la configuration 1 i.e., ($\mu = 0.1, \rho = 0.1$) s'expliquent par un manque de données de biomasse strictement positives dans certains jeux de données simulées, conséquence directe d'une très faible valeur de μ. Ce manque d'information conduit à une dégradation globale de la qualité des estimations et à une forte variabilité des performances d'estimation selon le jeu de données simulées.

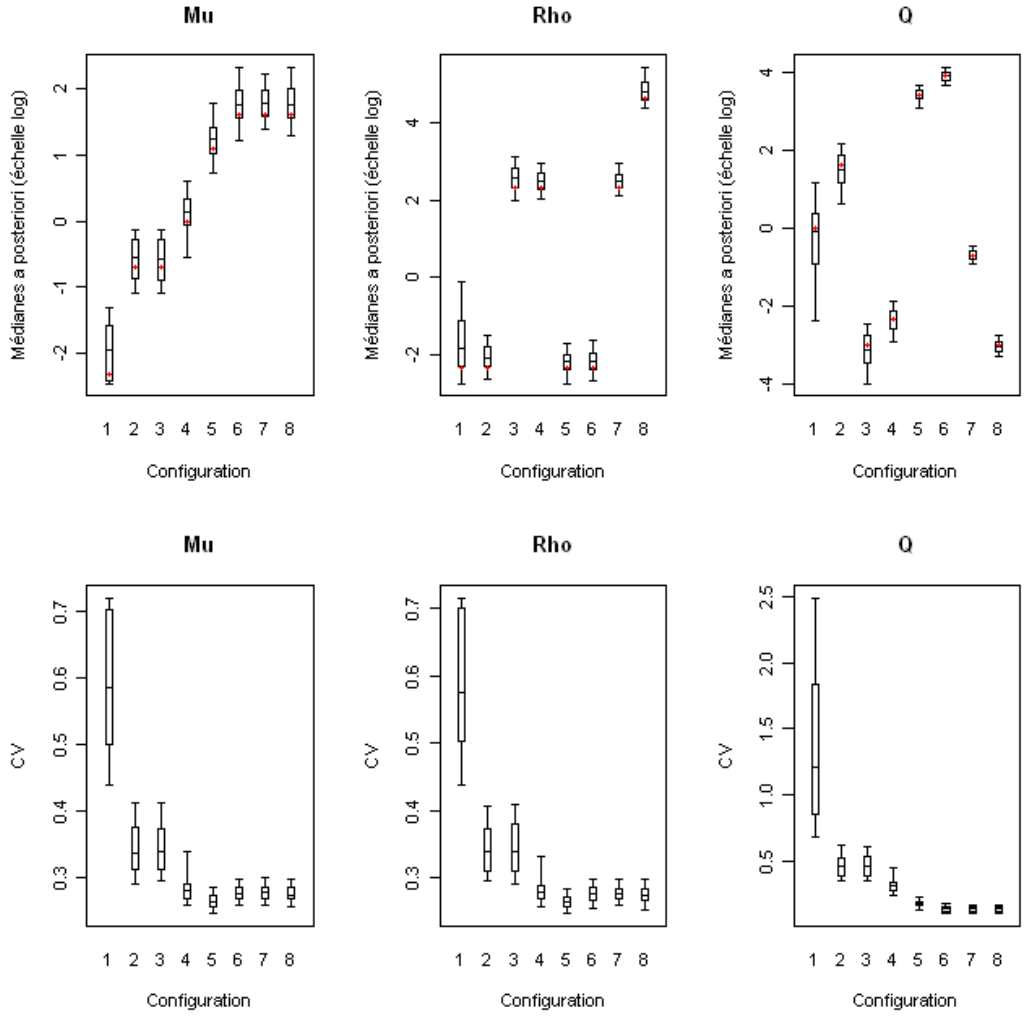


Figure 5.7 – Boxplots à 95% des médianes *a posteriori* (en haut) et des coefficients de variation *a posteriori* (en bas) obtenus pour l'estimation des paramètres μ, ρ et Q. L'axe des abscisses désigne les 8 configurations de paramètres (μ, ρ) simulées, détaillées dans le tableau 5.1. Les points rouges indiquent les valeurs simulées.

Le tableau 5.2 indique, pour chacune des 8 configurations (μ, ρ), les valeurs simulées ainsi que les différents critères précédemment décrits pour évaluer la qualité des ajustements sur les paramètres μ, ρ et Q.

Ce tableau montre que les médianes *a posteriori* ont tendance à sur-estimer les paramètres μ et ρ quelque soit la configuration (μ, ρ). En moyenne, les valeurs les plus probables *a posteriori* de μ et ρ sont supérieures aux valeurs simulées puisque $\mathbb{P}_{100}$ est clairement supérieur à 0.5. Cela signifie que les distributions *a posteriori* ont tendance à

N° config	θ	Valeur simulée	$\widehat{\theta}_{100}$	$R(\theta)$ (en %)	$C_{100}(\theta)$	$CV_{100}(\theta)$	$\mathbb{P}_{100}(\theta)$
1	μ	0.1	0.15	96	0.07	0.63	0.64
	ρ	0.1	0.39	80	0.21	0.59	0.71
	Q	0.01	1.11	89	0.92	1.44	0.45
2	μ	0.5	0.59	97	0.16	0.34	0.61
	ρ	0.1	0.14	91	0.04	0.34	0.69
	Q	5	4.88	92	1.69	0.46	0.44
3	μ	0.5	0.59	97	0.16	0.34	0.61
	ρ	10	13.94	92	3.94	0.34	0.69
	Q	0.05	0.05	93	0.02	0.46	0.44
4	μ	1	1.18	90	0.27	0.28	0.62
	ρ	10	12.41	90	2.84	0.28	0.65
	Q	0.1	0.1	93	0.02	0.32	0.47
5	μ	3	3.69	89	0.79	0.26	0.62
	ρ	0.1	0.12	91	0.03	0.26	0.61
	Q	30.0	30.78	95	4.27	0.18	0.53
6	μ	5	6.20	93	1.40	0.28	0.64
	ρ	0.1	0.12	89	0.03	0.28	0.65
	Q	50	49.79	95	5.30	0.13	0.48
7	μ	5	6.24	92	1.41	0.28	0.66
	ρ	10	12.38	93	2.79	0.28	0.66
	Q	0.50	0.51	93	0.05	0.13	0.52
8	μ	5	6.46	88	1.45	0.27	0.65
	ρ	100	133.17	86	29.92	0.27	0.68
	Q	0.05	0.05	94	0.005	0.13	0.44

Tableau 5.2 – Statistiques *a posteriori* relatives à l'estimation des paramètres (μ, ρ, Q) pour les 8 configurations simulées.

se répartir majoritairement au-delà des valeurs simulées. Par ailleurs, le tableau montre que l'estimation du paramètre ρ semble globalement plus difficile que celle du paramètre μ. En effet, les pourcentages de recouvrement $R(\theta)$ sont généralement plus faibles pour ce paramètre. Enfin, nous pouvons constater que l'estimation de la biomasse moyenne Q est globalement correcte pour toutes les configurations comme l'indique la figure 5.7 (sauf la configuration 1 pour les raisons évoquées précédemment). Le tableau 5.2 indique que plus de 92% des intervalles de crédibilité à 95% contiennent la valeur simulée. Enfin, aucune tendance à la sur ou sous estimation de la biomasse moyenne par la médiane *a posteriori* ne se dégage de cette simulation puisque $\mathbb{P}_{100}(Q)$ est proche de 0.5.

Une analyse plus détaillée des résultats montre que les performances d'estimation du modèle LOL dépendent fortement du nombre de valeurs nulles présentes dans les données simulées. Rappelons que ce nombre dépend directement de la valeur choisie pour μ.

Quand μ est petit (e.g., $\mu = 0.1$, $\mu = 0.5$), le pourcentage de recouvrement de la "vraie" valeur de μ est généralement plus élevé. Ainsi, plus de 96% des intervalles de crédibilité à 95% contiennent la valeur simulée du paramètre contre au plus 93% des intervalles quand μ prend une forte valeur. Par ailleurs, les valeurs les plus probables *a*

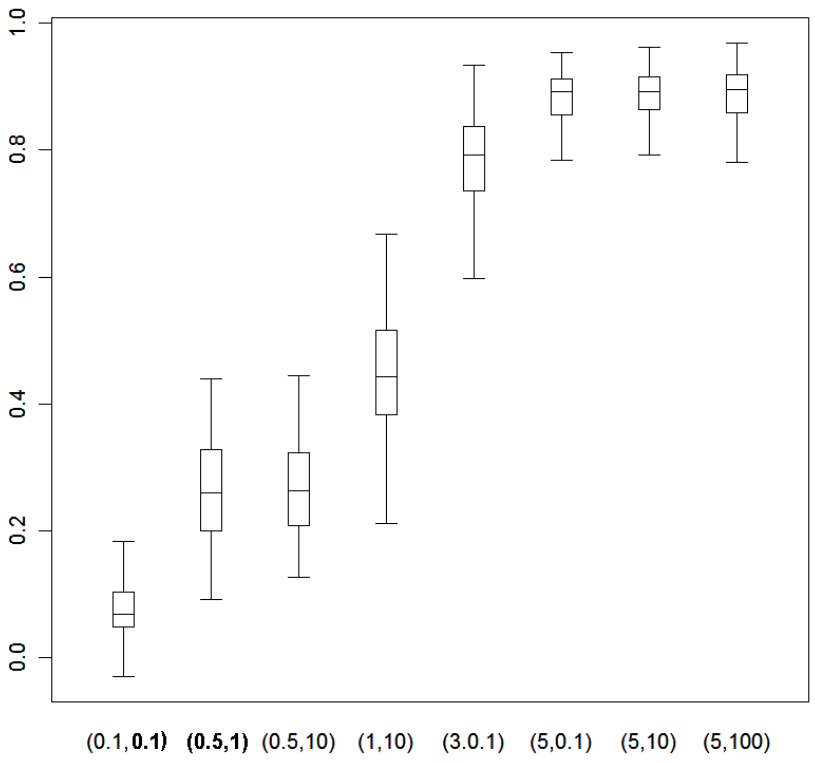


Figure 5.8 – Distribution empirique (sur 100 simulations) des corrélations linéaires entre les échantillons *a posteriori* des paramètres μ et ρ. Lecture en abscisses = (μ, ρ).

posteriori de μ (et en particulier les médianes) sont en moyenne plus proches de la valeur simulée puisque les indicateurs $\mathbb{P}_{100}$ sont plus faibles. Quand μ est petit, la forte occurrence de zéros constitue une information essentielle qui oriente l'algorithme d'inférence vers la "vraie" valeur de μ. Considérons les configurations 2 et 3 pour lesquelles μ est petit, fixé à 0.5. Nous pouvons remarquer que la valeur de ρ n'a aucune influence sur la qualité d'estimation de μ : les coefficients de variation, pourcentages de recouvrement et coûts *a posteriori* relatifs à μ sont identiques quand le paramètre ρ est multiplié par 10. Cela indique que seule l'occurrence de valeurs nulles joue sur la qualité d'estimation de μ. Quand μ est petit, l'estimation du paramètre ρ est généralement plus difficile. Prenons comme exemple les configurations 3, 4 et 7 pour lesquelles ρ est constant fixé à 10. Le coût *a posteriori* moyen associé à ρ est plus élevé quand $\mu = 0.5$ avec $C_{100}(\rho) = 3.94$ que lorsque $\mu = 5$ avec $C_{100}(\rho) = 2.79$. De même, les indicateurs $\mathbb{P}_{100}(\rho)$ sont plus élevés indiquant que les médianes *a posteriori* ont tendance à sur-estimer la valeur simulée. Il en va de même pour les configurations 1, 2, 5 et 6 pour lesquelles ρ est constant fixé à 0.1. Cela signifie qu'une estimation de ρ par la médiane *a posteriori* est moins "fiable" quand μ est petit. La faible quantité de valeurs non-nulles nécessaires à l'estimation de ρ, défini comme l'inverse de la quantité de biomasse strictement positive contenue dans un *gisement*, explique la pauvre qualité d'ajustement observée pour ce paramètre lorsque μ est petit.

Quand μ prend une forte valeur (e.g., $\mu = 3$, $\mu = 5$), l'estimation de μ devient plus

sensible aux valeurs de ρ. Considérons les configurations 6, 7 et 8 pour lesquelles μ est fixé à 5. Moins il y a de biomasse par *gisement* (i.e., ρ grand), plus la qualité d'estimation sur μ se dégrade avec notamment, une diminution des pourcentages de recouvrement et une augmentation des coûts *a posteriori* sur μ. Ainsi, bien que les échantillons *a posteriori* soient moins dispersés (i.e., CV_{100} plus petits), l'algorithme a beaucoup plus de difficultés à s'orienter vers la valeur simulée de μ et les estimations sont donc *a fortiori* moins "fiables". En revanche, l'estimation du paramètre ρ devient plus "fiable" avec des coûts *a posteriori* plus faibles à ρ constant et plus précise avec des coefficients de variation plus faibles.

La figure 5.8 indique, pour chaque configuration testée, la distribution des corrélations linéaires entre les échantillons *a posteriori* respectifs des paramètres μ et ρ. Cette distribution empirique est obtenue à partir des corrélations associées aux 100 simulations générées pour chaque configuration. La figure 5.8 montre qu'il existe souvent une corrélation linéaire positive entre les échantillons *a posteriori* des paramètres μ et ρ. Cette corrélation est d'autant plus forte que l'occurrence de valeurs nulles est faible (i.e. μ grand). Le paramètre ρ n'influe pas sur le niveau moyen de cette corrélation. Lorsque l'occurrence de zéros est forte (μ petit), l'information est suffisamment précise pour guider l'algorithme MCMC vers la "vraie" valeur de μ qui s'estime relativement indépendamment (d'où des corrélations plus faibles). En revanche, lorsque l'occurrence de zéros est faible (i.e, μ fort), l'information portée par les données est moins déterminante pour l'estimation de μ d'où des corrélations plus fortes. L'existence d'une corrélation linéaire positive entre μ et ρ est assez intuitive et induite explicitement par l'écriture du modèle (cf équation 5.3.1). Une même quantité moyenne de biomasse peut être obtenue soit avec beaucoup de *gisements* contenant chacun peu de biomasse soit avec peu de *gisements* contenant chacun beaucoup de biomasse. La figure 5.9 montre que des phénomènes de compensation se produisent entre les chaînes MCMC associées à ces deux paramètres. Les chaînes suivent généralement les mêmes variations : lorsque μ prend une forte valeur, ρ prend également une forte valeur et inversement.

Pour mettre en évidence l'impact possible de la corrélation linéaire entre μ et ρ sur la vitesse de convergence des algorithmes MCMC, j'ai utilisé le diagnostic de convergence proposé par Brooks et Gelman (1998). Son principe général est détaillé dans l'annexe B. Ce diagnostic est en fait un critère de mélange entre plusieurs chaînes MCMC : il consiste à monitorer la convergence d'un algorithme MCMC en comparant des longueurs d'intervalles de crédibilité intra et inter chaînes, calculés à partir d'échantillons de valeurs de plus en plus grands. A partir de la statistique de Brooks et Gelman, j'ai comparé la vitesse de convergence de la configuration 2 (i.e., $(\mu, \rho) = (0.5, 0.1)$) pour laquelle le niveau de corrélation moyen est faible (i.e., 0.26) à celle de la configuration 6 (i.e., $(\mu, \rho) = (5, 0.1)$) pour laquelle le niveau de corrélation moyen est fort (i.e., 0.88). La figure 5.10 indique l'évolution de la statistique de Brooks et Gelman pour les paramètres μ et ρ pour un jeu de données simulé selon la configuration 2 (en haut) et un jeu de données simulé selon la configuration 6 (en bas). Il apparaît clairement que l'existence d'une corrélation entre μ et ρ influe sur la vitesse de convergence. Ainsi, pour la configuration 2, la convergence est extrêment rapide : elle est atteinte au bout de 100 itérations pour μ avec un ratio de Brooks et Gelman stabilisé de 1.017 et au bout de 200 itérations pour ρ avec un ratio de 1.013. En revanche, pour la configuration 6, la convergence est plus lente : 900 itérations pour μ et 550 itérations pour ρ sont nécessaires pour obtenir une stabilisation du ratio de Brooks et Gelman alors égal à 1.013 pour μ et 1.03 pour ρ. La figure 5.11 fournit une explication

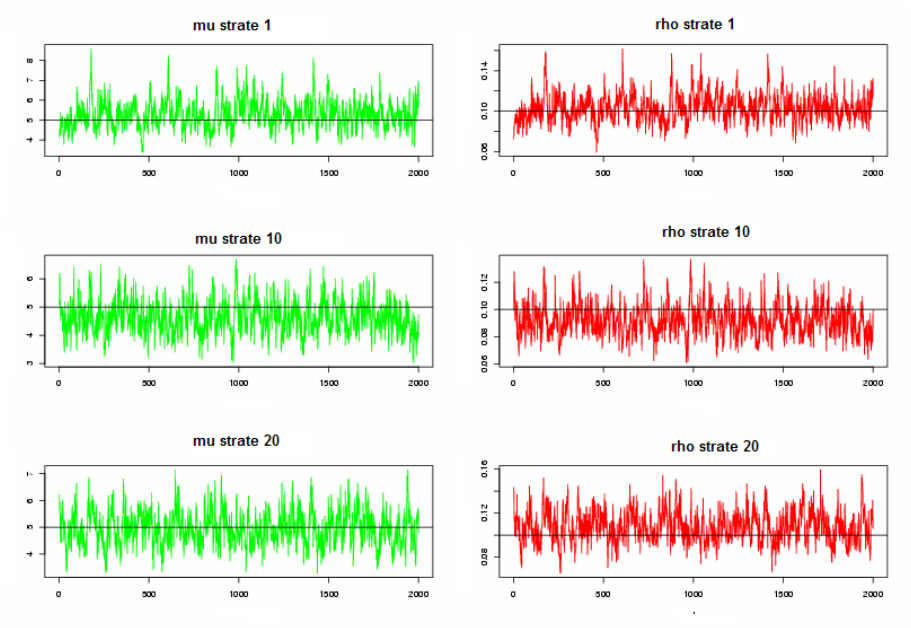


Figure 5.9 – Exemple de chaînes MCMC dans le cas où $\mu = 5$ et $\rho = 0.1$. Les chaînes possèdent des variations similaires témoignant d'une corrélation positive dans l'estimation des paramètres μ et ρ.

possible à cette lenteur de convergence. En effet, nous pouvons remarquer qu'il existe un très fort niveau d'autocorrélation entre les itérations successives des chaînes MCMC dans le cas de la configuration 6 par rapport à la configuration 2. Ainsi, les chaînes MCMC parcourent beaucoup plus lentement l'espace des paramètres d'où une convergence plus lente.

La corrélation linéaire positive existant entre les échantillons *a posteriori* des paramètres μ et ρ n'a globalement pas d'impact négatif sur l'estimation de la biomasse moyenne. En effet, comme vu précédemment, la médiane *a posteriori* estime correctement la biomasse moyenne dans toutes les configurations. Comme unique impact possible de cette corrélation, nous pouvons noter que les estimations de biomasse moyenne sont plus précises (i.e., CV_{100} plus faible) quand le niveau de corrélation linéaire entre μ et ρ augmente. Concernant l'estimation du couple (μ,ρ), nous pouvons remarquer que les précisions d'estimation diminuent et que les pourcentages de recouvrement et indicateurs de sur-estimation par la médiane *a posteriori* deviennent sensiblement plus proches quand le niveau de corrélation augmente.

5.4 Simulation 2 : Comparaison des performances d'estimation des modèle LOL et $\Delta\Gamma$ en fonction du nombre d'observations

La biomasse moyenne est souvent la quantité principale que souhaite pouvoir estimer le biologiste. En raison du coût élevé des relevés d'abondance, le nombre d'observations réalisées pour estimer cette biomasse moyenne est généralement réduit. Par exemple, lors des campagnes de suivi du "Centre des Pêches du Golfe", 5 traits de chalut sont effectués

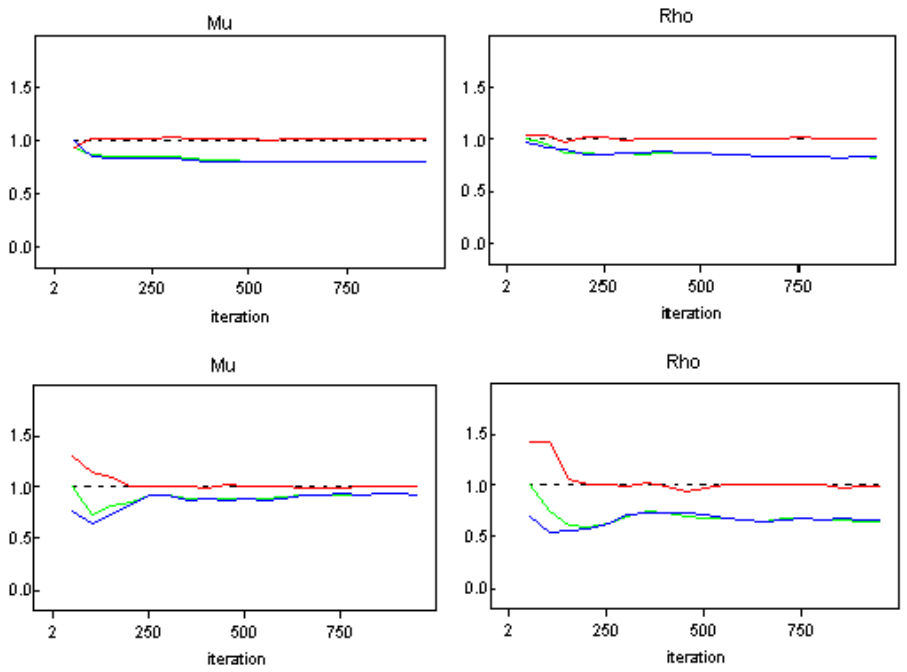


Figure 5.10 – Evolution du diagnostic de convergence de Brooks et Gelman (1998) sur les 1000 premières itérations des chaînes MCMC obtenues pour les paramètres μ et ρ (En haut) Configuration $(\mu, \rho) = (0.5, 0.1)$ avec faible corrélation linéaire (En bas) Configuration $(\mu, \rho) = (5, 0.1)$ avec forte corrélation linéaire. La courbe verte indique la largeur de l'intervalle de crédibilité inter-chaînes à 80%. La courbe bleue indique la largeur moyenne des intervalles de crédibilité intra-chaîne à 80%. La courbe rouge indique la statistique de Brooks et Gelman (i.e., le ratio des courbes vertes/bleues)

en moyenne chaque année dans chacune des 27 strates du Sud du Golfe-du-Saint-Laurent.

La deuxième simulation réalisée a deux objectifs. Tout d'abord, il s'agit de comparer les performances d'estimation des modèles LOL et $\Delta\Gamma$ en fonction du nombre d'observations. Puis, il s'agit de définir le nombre moyen d'observations nécessaires à l'obtention d'une estimation "suffisamment" précise de la biomasse moyenne.

J'ai montré que, à efforts d'échantillonnage égaux à l'effort de référence :

- Les deux modèles ont globalement les mêmes performances d'estimation au-delà de 20 observations.
- Le nombre d'observations nécessaire à l'obtention d'un estimateur suffisamment précis de la biomasse moyenne dépend du niveau de présence de l'espèce dans la région considérée.

5.4.1 Description de la simulation

7 tailles d'échantillon ont été testées : 5, 10, 20, 30, 40, 50 et 100. Afin de tenir compte de l'influence possible des configurations de paramètres (μ, ρ), j'ai considéré les 8 couples de valeurs du tableau 5.1 pour chaque nombre d'observations testé.

J'ai simulé des données de biomasse selon le modèle LOL et selon la distribution $\Delta\Gamma$. Afin de se placer sous les mêmes hypothèses que celles considérées pour le modèle LOL, les paramètres (a, b, δ) du modèle $\Delta\Gamma$ sont fixés de telle sorte que les probabilités d'occurrence d'un zéro et quantités moyennes de biomasses générées correspondent à chacun des 10

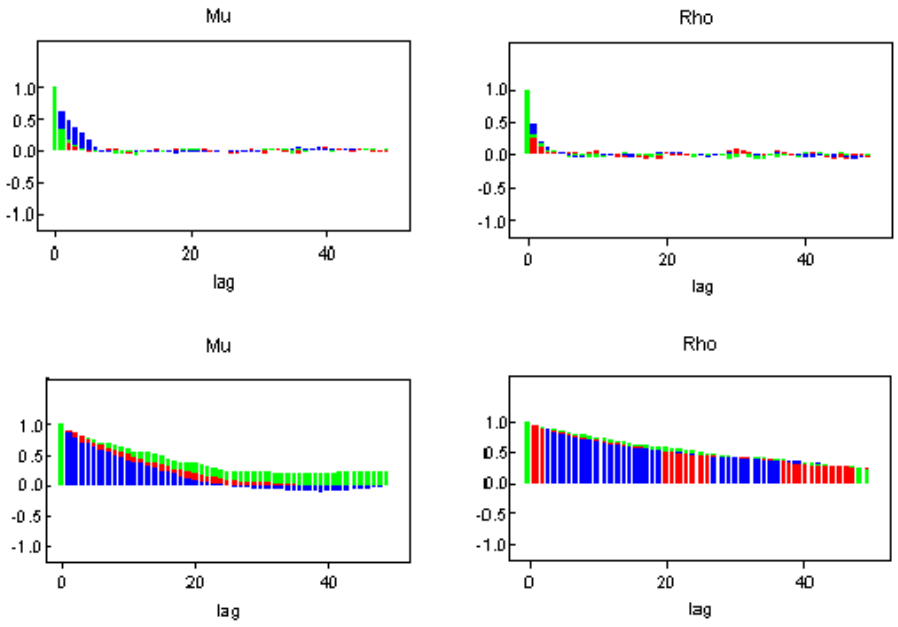


Figure 5.11 – Evolution du niveau d'autocorrélation des chaînes MCMC pour μ et ρ en fonction du nombre d'itérations séparant deux itérations d'intérêt successives (i.e., lag). (En haut) Configuration $(\mu, \rho) = (0.5, 0.1)$ avec faible corrélation linéaire. (En bas) Configuration $(\mu, \rho) = (5, 0.1)$ avec forte corrélation linéaire. Chaque couleur correspond à une chaîne MCMC.

cadres de simulation du tableau 5.1. Les paramètres (μ, ρ) du modèle LOL et (a, b, δ) du modèle $\Delta\Gamma$ sont liés par les relations suivantes :

$$\begin{cases} \delta = e^{-\mu} \\ \frac{a}{b}(1-\delta) = \frac{\mu}{\rho} \\ \frac{a}{b^2}(1-\delta) + \frac{a^2}{b^2}(1-\delta)\delta = \frac{2\mu}{\rho^2} \end{cases} = \begin{cases} \delta = e^{-\mu} \\ a = \frac{\mu}{2-\delta(2+\mu)} \\ b = \frac{\rho(1-\delta)}{2-\delta(2+\mu)} \end{cases}$$

Pour chacune des 8 configurations testées et chaque nombre d'observations, 100 simulations indépendantes ont été réalisées. J'ai également exclu toute variabilité dans l'effort d'échantillonnage.

Les échantillons *a posteriori* ont de nouveau été calculés sur la base de 40000 itérations MCMC après une période de chauffe de 20000 itérations.

J'ai calculé, pour 2 configurations particulières (μ, ρ), les médianes *a posteriori*, coûts *a posteriori*, pourcentages de recouvrement et coefficients de variation *a posteriori* moyens sur 100 simulations pour la biomasse moyenne en fonction du nombre d'observations (cf section 5.3.2). L'idée a été de comparer les performances d'estimation des modèles LOL et $\Delta\Gamma$ en fonction de l'occurrence de valeurs nulles. La configuration $(\mu, \rho) = (0.5, 10)$ a permis d'examiner le cas où cette occurrence est forte et la configuration $(\mu, \rho) = (5, 10)$ le cas où cette occurrence est faible.

5.4.2 Résultats

Comme on pouvait s'y attendre, les tableaux 5.3 et 5.4 indiquent que, lorsque le nombre d'observations augmente, les coûts *a posteriori* diminuent, les taux de recouvrement augmentent et les coefficients de variation diminuent pour les deux modèles. Cela indique que

les performances d'estimation des deux modèles augmentent avec le nombre d'observations.

Nbre obs	$\widehat{Q}_{100}$		$C_{100}(Q)$		$R(Q)$		$CV_{100}(Q)$	
	LOL	$\Delta\Gamma$	LOL	$\Delta\Gamma$	LOL	$\Delta\Gamma$	LOL	$\Delta\Gamma$
5	0.05	0.06	0.034	0.11	82	78	1.01	0.57
10	0.05	0.04	0.025	0.019	86	86	0.66	0.54
20	0.05	0.05	0.017	0.016	93	90	0.45	0.41
30	0.05	0.05	0.015	0.013	97	87	0.37	0.35
40	0.05	0.05	0.013	0.012	97	92	0.32	0.30
50	0.05	0.05	0.011	0.011	95	94	0.29	0.27
100	0.05	0.05	0.007	0.007	95	93	0.20	0.20

Tableau 5.3 – μ=0.5, ρ=10, $Q = 0.05$: Médianes *a posteriori*, coûts *a posteriori*, pourcentages de recouvrements et coefficients de variation *a posteriori* (moyens sur 100 simulations) pour l'évaluation de la biomasse moyenne Q en fonction du nombre d'observations.

Nbre obs	$\widehat{Q}_{100}$		$C_{100}(Q)$		$R(Q)$		$CV_{100}(Q)$	
	LOL	$\Delta\Gamma$	LOL	$\Delta\Gamma$	LOL	$\Delta\Gamma$	LOL	$\Delta\Gamma$
5	0.52	0.43	0.091	0.095	88	86	0.23	0.29
10	0.49	0.44	0.067	0.071	91	87	0.17	0.21
20	0.51	0.47	0.053	0.054	93	92	0.13	0.14
30	0.50	0.47	0.044	0.043	95	91	0.11	0.11
40	0.51	0.49	0.039	0.039	91	97	0.10	0.10
50	0.49	0.49	0.034	0.035	96	95	0.09	0.09
100	0.50	0.49	0.025	0.025	95	97	0.06	0.06

Tableau 5.4 – μ=5, ρ=10 $Q = 0.5$: Médianes *a posteriori*, coûts *a posteriori*, pourcentages de recouvrements et coefficients de variation *a posteriori* (moyens sur 100 simulations) pour l'évaluation de la biomasse moyenne Q en fonction du nombre d'observations.

Les tableaux 5.3 et 5.4 montrent qu'au-delà de 20 observations, les modèles LOL et $\Delta\Gamma$ ont globalement les mêmes performances d'estimation si on compare les médianes *a posteriori*, coûts *a posteriori* et coefficients de variation moyens pour les deux configurations testées. Toutefois, nous pouvons remarquer que les taux de recouvrement relatifs au modèle $\Delta\Gamma$ sont souvent inférieurs à ceux du modèle LOL lorsque les échantillons *a posteriori* sont identiquement dispersés (i.e., CV_{100} similaires). Cela indique que le modèle $\Delta\Gamma$ a généralement plus de difficultés à s'orienter vers la valeur simulée de Q par rapport au modèle LOL.

Dans le cas où le nombre d'observations est inférieur à 20 et $\mu = 0.5$, les deux modèles semblent en difficulté. Le modèle LOL donne des échantillons *a posteriori* plus dispersés que $\Delta\Gamma$ ce qui peut expliquer que les taux de recouvrement soient plus élevés. En revanche, les coûts *a posteriori* sont plus élevés sous le modèle $\Delta\Gamma$.

Dans le cas où le nombre d'observations est inférieur à 20 et $\mu = 5$, le modèle LOL semble fournir de meilleures estimations de biomasse moyenne par rapport au modèle $\Delta\Gamma$. En effet, les coûts et coefficients de variation *a posteriori* sont plus faibles. Par ailleurs,

les taux de recouvrement sont plus élevés.

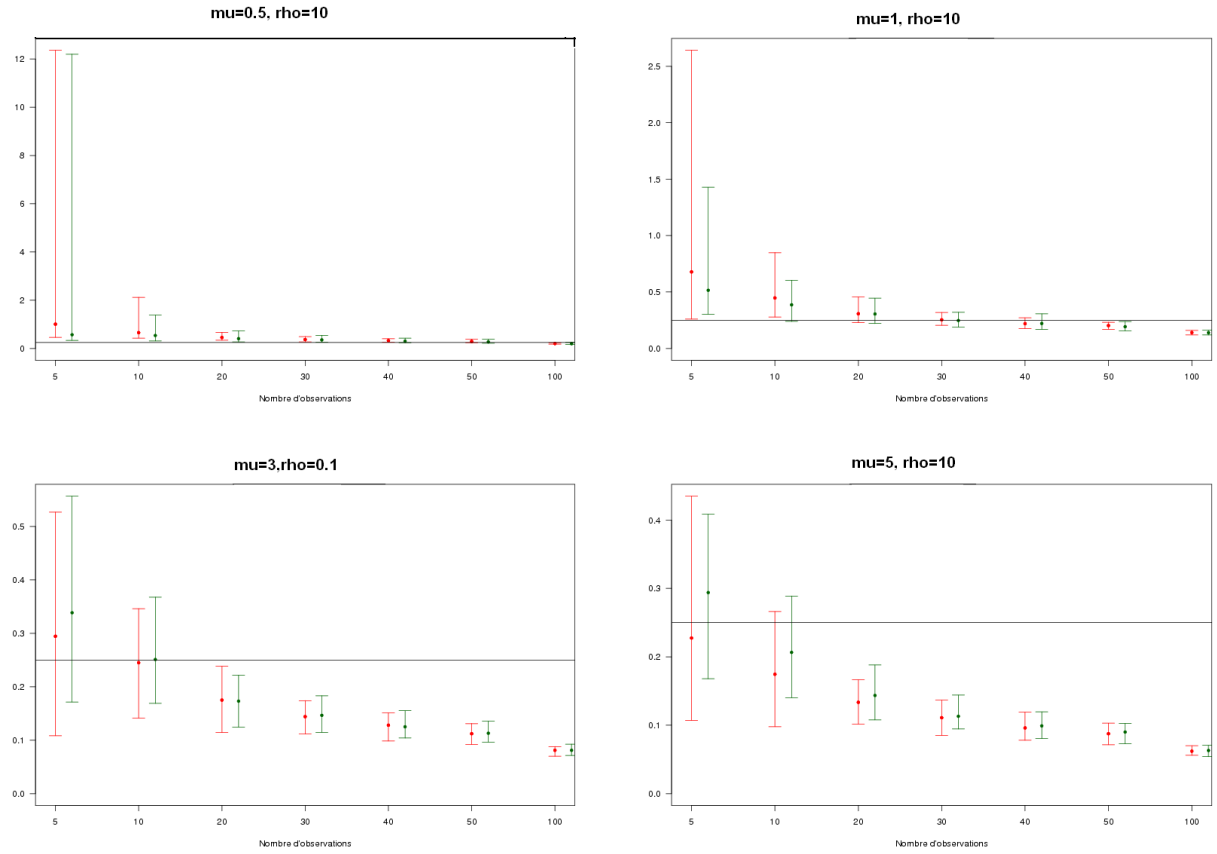


Figure 5.12 – Boxplots à 95% des coefficients de variation *a posteriori* sur l'estimation de la biomasse moyenne en fonction du nombre d'observations. Les points correspondent aux médianes *a posteriori* (En rouge) modèle LOL (En vert) modèle $\Delta\Gamma$. Le trait horizontal indique un coefficient de variation de 0.25 correspondant à une erreur d'estimation approximative de plus ou moins 50% autour de la biomasse moyenne.

La figure 5.12 indique, sous forme de boxplots à 95%, l'évolution des précisions d'estimation en fonction du nombre d'observations pour quatre configurations (μ, ρ). Ces boxplots ont été réalisés à partir de 100 simulations indépendantes. La figure montre que le nombre d'observations nécessaire à l'obtention d'un niveau donné de précision des estimations de biomasse moyenne dépend de la valeur de μ. Cela est le cas pour les modèles LOL et $\Delta\Gamma$. Considérons par exemple un coefficient de variation moyen de 0.25 qui correspond à une erreur d'estimation approximative de plus ou moins 50% autour de la biomasse moyenne. Près de 100 observations sont nécessaires sous les deux modèles pour atteindre une telle précision quand μ est "petit" (cf tableau 5.3). En revanche, moins de 10 observations suffisent dans le cas où μ est "fort" (cf tableau 5.4). Concrètement, cela signifie que, lorsque l'espèce étudiée est peu présente dans le domaine d'échantillonnage (beaucoup de valeurs de biomasses nulles), l'écologue doit réaliser près de 100 mesures de biomasses pour espérer obtenir une estimation de la biomasse moyenne précise à plus ou moins 50% que ce soit sous le modèle LOL ou le modèle $\Delta\Gamma$.

5.5 Simulation 3 : Comparaison de la robustesse des modèles LOL et $\Delta\Gamma$ pour l'analyse de relevés d'abondance à efforts d'échantillonnage variables

En théorie, le modèle $\Delta\Gamma$ n'est pas stable par addition (cf section 4.3.2). Il est donc défini par rapport à un effort d'échantillonnage de référence. Dans le cas où cet effort est variable entre les observations, une approche courante est de standardiser les données par rapport à un effort de référence et d'appliquer le modèle $\Delta\Gamma$ sur ces données standardisées. L'objectif de cette simulation est d'étudier l'impact d'une variabilité des efforts d'échantillonnage sur les modèles LOL et $\Delta\Gamma$ pour l'estimation de la probabilité d'occurrence d'un zéro et de la biomasse moyenne.

J'ai montré que :

- La précision des estimations de probabilité d'occurrence de zéros et de biomasse moyenne n'est globalement pas détériorée lors de l'analyse de relevés d'abondance à efforts d'échantillonnage variables et ce, quel que soit le modèle.
- Sous le modèle $\Delta\Gamma$, une standardisation des données de biomasse induit des biais dans l'estimation des probabilités d'occurrence de zéros. Ceci se produit tout particulièrement quand ces efforts sont dans la plupart des cas inférieurs ou supérieurs à l'effort de référence.
- Sous le modèle LOL, une variabilité des efforts d'échantillonnage n'induit aucune tendance particulière à la sur ou sous estimation.

5.5.1 Description de la simulation

J'ai simulé 1 million de quantités de biomasse sous le modèle $\Delta\Gamma$ de paramètres $\delta = 0.9995$, $a = 2$ et $b = 2$. J'ai supposé qu'elles correspondaient à un effort d'échantillonnage de référence de 0.001 unités. Par la suite, j'appelerai O cet ensemble de données.

Dans un deuxième temps, j'ai construit un jeu de 1000 quantités de biomasse par tirages aléatoires successifs (sans répétition) puis sommation de 1000 éléments de l'ensemble O. Les données obtenues formaient un échantillon associé à un effort d'échantillonnage de référence de 1 unité. Par la suite, j'appelerai e cet effort de référence. En théorie, ces données ne suivent ni la loi $\Delta\Gamma$ ni le modèle LOL. La biomasse moyenne empirique ($\widehat{Q}$) et probabilité d'occurrence d'un zéro empirique ($\widehat{\delta}$) obtenues sous l'effort e étaient respectivement de 0.48 et 0.61.

Dans un troisième temps, j'ai simulé des relevés d'abondance de 100 quantités de biomasse associées à des efforts d'échantillonnage variables par rapport à l'effort e. J'ai envisagé 26 scénarii possibles et cohérents par rapport aux relevés d'abondance du Centre des Pêches du Golfe que j'ai classé en 4 grandes catégories :

1. Efforts constants mais strictement inférieurs à 1 unité. 3 valeurs ont été considérées : 0.1, 0.5 et 0.9
2. Efforts constants mais strictement supérieurs à 1 unité. 3 valeurs ont été considérées : 1.1, 1.5 et 2
3. Efforts distribués selon une loi gaussienne centrée en 1. 10 valeurs de coefficients de variation ont été testées (entre 0.002 et 0.25).
4. un échantillon X de 100 valeurs a été simulé selon une loi exponentielle de paramètre λ. Les efforts d'échantillonnage S sont ensuite définis comme : 1-X (cf figure 5.13).

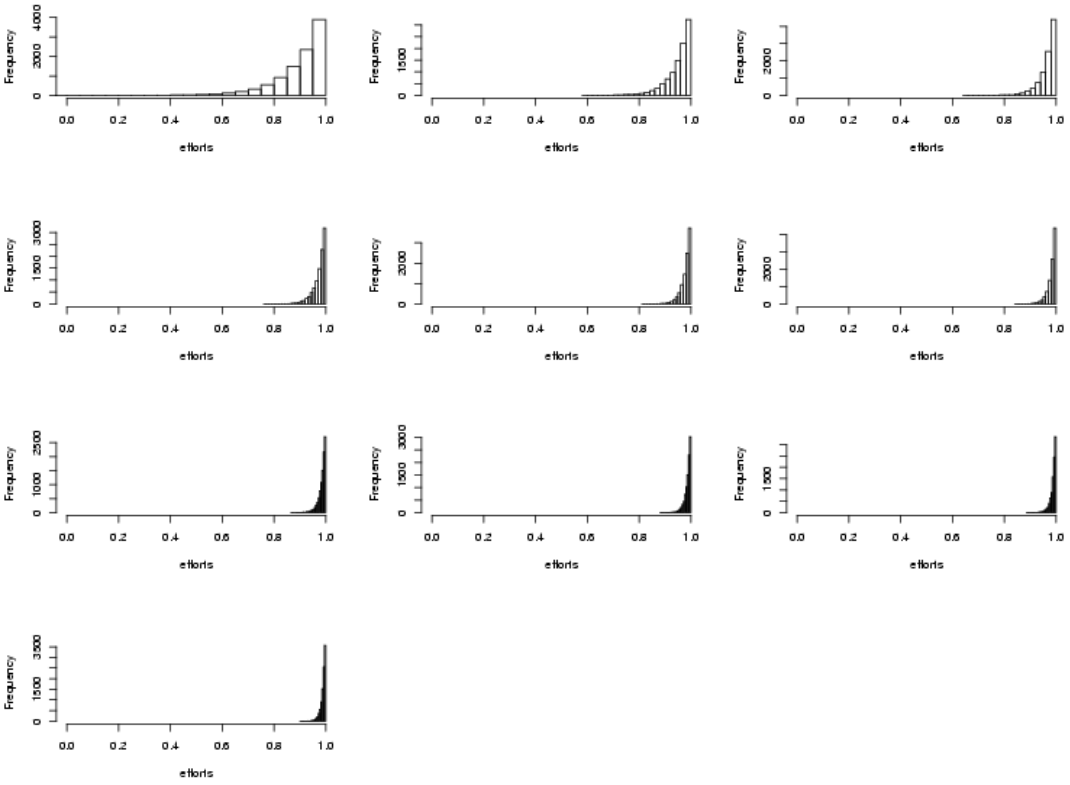


Figure 5.13 – 10 histogrammes empiriques d'efforts d'échantillonnage simulés selon le principe suivant : S=1-X avec X issu d'une loi exponentielle de paramètre λ.

> Par simplicité, je qualifierai ces efforts d'exponentiels. Ceci a permis d'obtenir une distribution empirique d'efforts d'échantillonnage variables mais inférieurs à l'effort de référence ce qui, en pratique, est souvent le cas (cf figure 4.1). En effet, l'effort d'échantillonnage a tendance à être écourté par rapport à l'effort de référence pour diverses raisons techniques (e.g., passage dans des milieux aux fonds escarpés susceptibles de déchirer le chalut). 10 valeurs de λ ont été testées correspondant à des coefficients de variation de la loi exponentielle allant de 0.09 à 0.32.

J'ai simulé 100 échantillons de 100 efforts d'échantillonnage pour chaque scénario. Puis, pour chaque échantillon, j'ai généré les quantités de biomasse associées à chaque effort d'échantillonnage par tirages aléatoires successifs (sans répétition) puis sommation de données issues de l'ensemble O. Le nombre de quantités de biomasse de l'ensemble O à sommer a été défini en fonction de l'effort d'échantillonnage associé : par exemple, un effort de 0.5 unités reviendrait à sommer 0.5/0.001 données de O.

L'inférence bayésienne des différents relevés d'abondance simulés a été réalisée sous les modèles LOL et $\Delta\Gamma$ en supposant que l'effort de référence est égal à 1 unité. Pour le modèle LOL, les efforts d'échantillonnage ont été introduits en tant que variables explicatives. Pour le modèle $\Delta\Gamma$, une standardisation préalable des quantités de biomasse a été réalisée afin de se ramener à l'effort de référence de 1 unité. Les échantillons *a posteriori* ont de nouveau été calculés sur la base de 40000 itérations MCMC après une période de chauffe de 20000 itérations.

J'ai comparé les performances d'estimation des deux modèles en fonction de la va-

riabilité des efforts d'échantillonnage. Dans cette optique, j'ai calculé certains critères de comparaison décrits dans la section 5.3.2 : médianes *a posteriori* moyennes (i.e., $\widehat{\theta}_{100}$), pourcentages de recouvrement (i.e., $R(\theta)$), indicateurs de la position des échantillons *a posteriori* par rapport aux valeurs simulées (i.e., $P_{100}(\theta)$), et coefficients de variation moyens indiquant la précision des estimations (i.e., $CV_{100}(\theta)$). J'ai comparé les performances d'estimation sur la probabilité d'occurrence de zéros δ et la biomasse moyenne Q. Leur valeur simulée respective est 0.61 et 0.48.

5.5.2 Résultats

Les figures 5.14, 5.15 et 5.16 indiquent, sous forme de boxplots à 95% (calculés sur 100 simulations), la répartition empirique des indicateurs de sur-estimation de δ et Q par la médiane *a posteriori* ainsi que les coefficients de variation *a posteriori* associés, en fonction de la variabilité des efforts d'échantillonnage.

Les performances d'estimation de la biomasse moyenne Q ne sont globalement pas affectées par une variabilité des efforts d'échantillonnage. Cela est le cas pour le modèle LOL et le modèle $\Delta\Gamma$. Ainsi, la répartition empirique des indicateurs de positionnement des distributions *a posteriori* par rapport aux valeurs simulées est globalement similaire quel que soit la variabilité des efforts d'échantillonnage. $P_{100}(Q)$ est proche de 0.5 sous les deux modèles : cela indique que la médiane *a posteriori* constitue systématiquement un estimateur ponctuel non biaisé de la biomasse moyenne. Par ailleurs, dans le cas d'une répartition Gaussienne ou exponentielle de ces efforts, la répartition empirique des coefficients de variation est également similaire quel que soit la variabilité de l'effort d'échantillonnage. Cela signifie que, sous les deux modèles, les précisions d'estimation sur Q ne dépendent pas de la variabilité des efforts d'échantillonnage. En revanche, dans le cas où ces efforts sont constants mais strictement inférieurs ou supérieurs à l'effort standard de référence, les coefficients de variation diminuent quand l'effort d'échantillonnage augmente (cf Figure 5.14). Il s'agit là d'une des hypothèses des modèles LOL et $\Delta\Gamma$ (cf section 5.2.2). A noter qu'en moyenne, les coefficients de variation *a posteriori* sur Q sont légèrement inférieurs sous le modèle $\Delta\Gamma$ confirmant les résultats obtenus dans la simulation 2.

Concernant les coefficients de variation *a posteriori* sur l'estimation de la probabilité d'occurrence de zéros δ, les mêmes commentaires que ceux réalisés pour Q s'appliquent. Une variabilité des efforts d'échantillonnage n'a globalement aucun impact sur les précisions d'estimation sur δ.

Le seul impact visible d'une variabilité de l'effort d'échantillonnage apparaît sur la capacité du modèle $\Delta\Gamma$ à retrouver la valeur simulée de la probabilité d'occurrence de zéros. Dans le cas où les efforts d'échantillonnage sont constants mais strictement inférieurs ou supérieurs à l'effort de référence, le modèle $\Delta\Gamma$ ne parvient pas à retrouver la valeur simulée de δ (cf Tableau 5.5). Ainsi, les taux de recouvrement sont très faibles lorsque qu'on s'éloigne de l'effort de référence et les médianes *a posteriori* ont systématiquement tendance à sur ou sous estimer δ. La répartition empirique des indicateurs de sur-estimation de δ est alors très proche de 1 (i.e., toutes les distributions *a posteriori* se situent au-dessus de la valeur simulée) ou proche de 0 (i.e., toutes les distributions *a posteriori* se situent en-dessous de la valeur simulée). Sous le modèle $\Delta\Gamma$, une standardisation des données ne permet pas de modifier l'occurrence de valeurs nulles en fonction de l'effort d'échantillonnage ce qui induit des biais dans les estimations de probabilité d'occurrence de zéros. Ceci montre clairement qu'il n'est pas raisonnable de comparer des relevés d'abon-

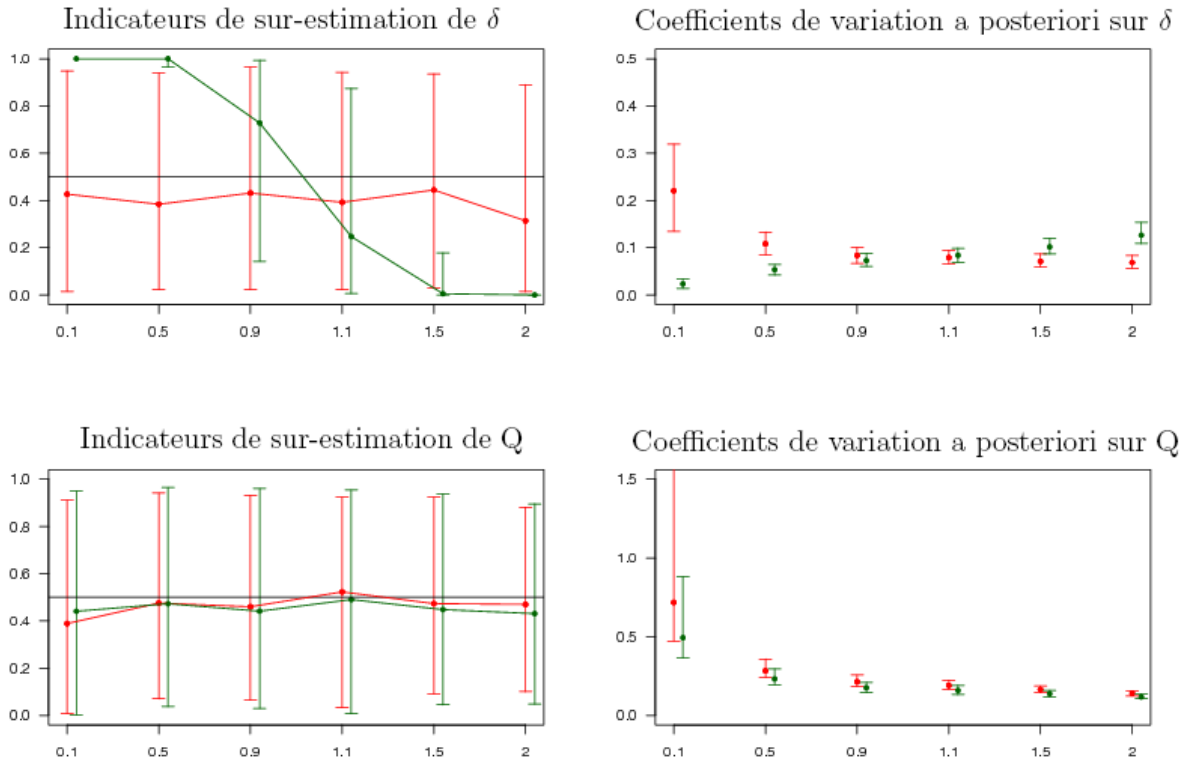


Figure 5.14 – Efforts d'échantillonnage constants : Boxplots à 95% des indicateurs de sur-estimation par la médiane *a posteriori* et des coefficients de variations *a posteriori* pour l'estimation de la probabilité d'occurrence d'un zéro δ et de la biomasse moyenne Q. L'axe des abscisses indique les efforts d'échantillonnage considérés. (Rouge) modèle LOL (Vert) modèle $\Delta\Gamma$

dance réalisés à efforts d'échantillonnage distincts avec le modèle $\Delta\Gamma$. Dans le cas où les efforts d'échantillonnage sont exponentiels, les médianes *a posteriori* ont également tendance à sur-estimer δ mais uniquement lorsque les efforts d'échantillonnage sont très variables i.e., $\lambda > 0.16$ (cf Tableau 5.5). Notons cependant que les performances d'estimation du modèle $\Delta\Gamma$ sont meilleures par rapport aux cas d'efforts systématiquement inférieurs ou supérieurs à 1. Les taux de recouvrement sont supérieurs à 85%, les coefficients de variation *a posteriori* et les indicateurs de sur-estimation par la médiane *a posteriori* sont plus faibles. Ceci peut s'expliquer par des effets de compensation de biais entre les valeurs liées à des efforts très proches de 1 et celles liées à des efforts très inférieurs. Dans le cas Gaussien, une variabilité plus ou moins importante des efforts d'échantillonnage par rapport à l'effort de référence ne semble pas avoir d'impact sur la capacité du modèle $\Delta\Gamma$ à estimer δ (cf Figure 5.16). Ceci vient confirmer l'hypothèse selon laquelle des phénomènes de compensation de biais se produisent entre les valeurs liées à des efforts inférieurs et des valeurs liées à des efforts supérieurs à 1. La symétrie de la loi normale fait disparaître complètement les possibles biais d'estimation.

Comme on pouvait s'y attendre, sous le modèle LOL, une variabilité des efforts d'échantillonnage n'induit aucune tendance particulière à la sur ou sous estimation de δ par la médiane *a posteriori*. La répartition empirique des indicateurs de sur-estimation de δ est globalement similaire quel que soit la répartition des efforts d'échantillonnage.

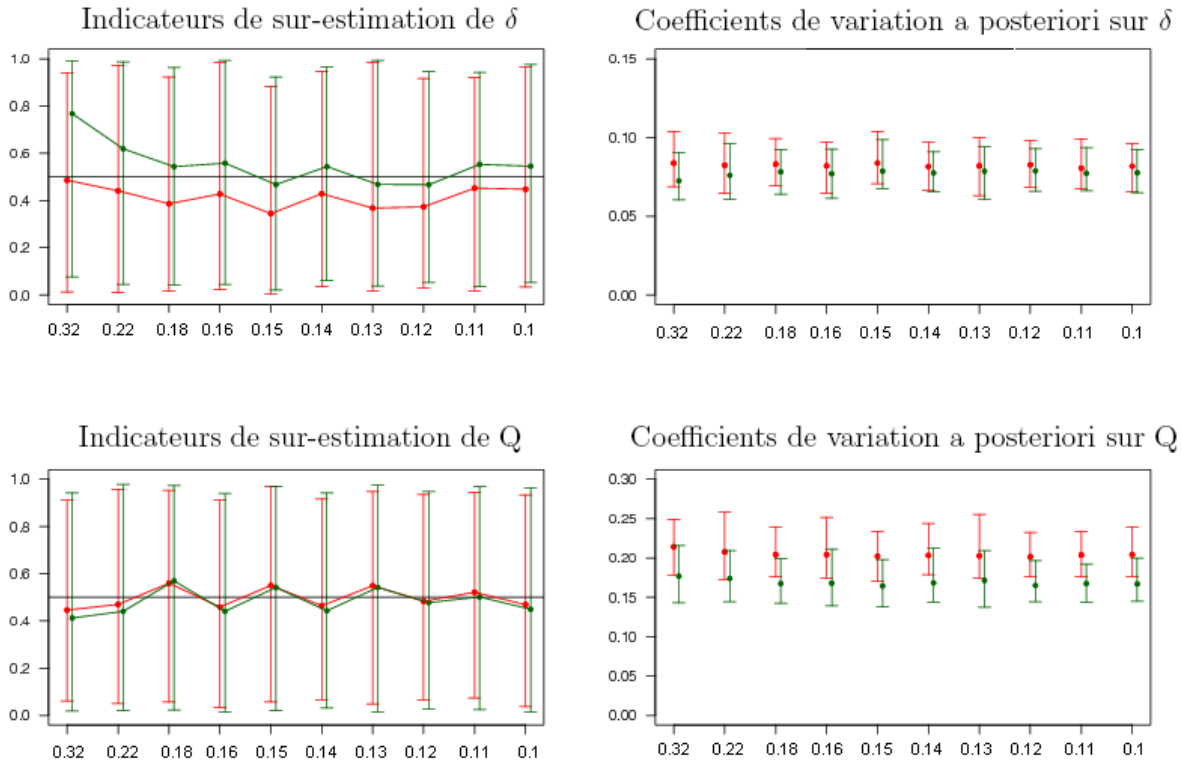


Figure 5.15 – Efforts d'échantillonnage exponentiels : Boxplots à 95% des indicateurs de sur-estimation par la médiane *a posteriori* et des coefficients de variations *a posteriori* pour l'estimation de la probabilité d'occurrence d'un zéro δ et de la biomasse moyenne Q. L'axe des abscisses indique les écarts-types des échantillons d'effort d'échantillonnage simulés. (Rouge) modèle LOL (Vert) modèle $\Delta\Gamma$

Aussi, contrairement au modèle $\Delta\Gamma$, ce modèle a l'avantage de pouvoir être utilisé pour comparer des relevés d'abondance réalisés à efforts d'échantillonnage distincts. Notons cependant que ces indicateurs sont, en moyenne, systématiquement inférieurs à 0.5 (cf Tableau 5.5). Cela confirme les résultats de la simulation 1 indiquant une sous-estimation systématique de la probabilité d'occurrence de zéros sous le modèle LOL.

5.6 Discussion

Le *modèle des fuites* est un exemple de modèle hiérarchique dans lequel les variables aléatoires latentes Z sont de deux natures différentes. Il y a les variables discrètes N_k indiquant le nombre de *gisements* collectés au site k (k=1,2,..,r) et les variables continues M_{g_k} désignant les quantités de biomasse contenues dans chaque *gisement* ramassé au site k.

L'explicitation des variables latentes N_k et M_{g_k} proposée dans ce chapitre permet de débrider la portée de la structure du modèle LOL pour l'analyse de données de biomasse *zero-inflated*. En effet, imaginons un instant que n'ait pas été donné le mode de construction conditionnelle du modèle LOL donné par les équations 5.1.1, 5.1.2 et 5.1.3. Passant directement à la vraisemblance par intégration sur les variables latentes N_k, supposons

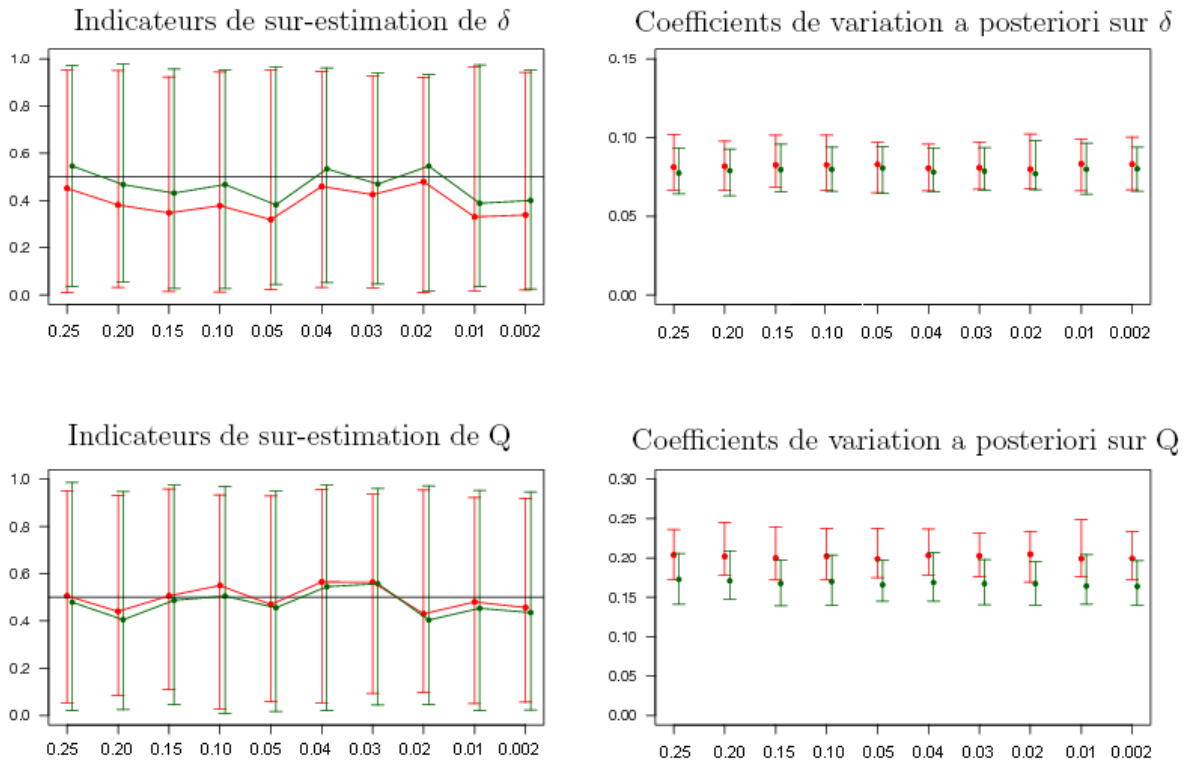


Figure 5.16 – Efforts d'échantillonnage Gaussiens : Boxplots à 95% des indicateurs de sur-estimation par la médiane *a posteriori* et des coefficients de variations *a posteriori* pour l'estimation de la probabilité d'occurrence d'un zéro δ et de la biomasse moyenne Q. L'axe des abscisses indique les écarts-types des échantillons d'effort d'échantillonnage simulés. (Rouge) modèle LOL (Vert) modèle $\Delta\Gamma$

que nous nous soyons intéressés à la variable aléatoire Y dont la distribution de probabilité s'écrit avec deux paramètres (μ, ρ) sous la forme :

$$[y_k|\mu,\rho] = 1_{y_k=0}e^{-\mu S_k} + 1_{y_k>0}\frac{\mu S_k \rho}{\sqrt{\mu S_k \rho y_k}}e^{-S_k\mu-\rho y_k}I_1(2\sqrt{\mu S_k \rho y_k})$$

où I_1 désigne la fonction de Bessel modifiée de première espèce et d'ordre 1 . Compte-tenu de l'apparente complexité de cette densité et de la difficile interprétation des paramètres μ et ρ , il y a fort à parier que le modèle LOL soit peu considéré.

L'approche hiérarchique permet de conceptualiser chaque étape du processus d'échantillonnage d'organismes vivants dans une couche différente du modèle LOL. Elle donne un sens conceptuel aux paramètres ce qui facilite le dialogue avec le non-statisticien : μ désigne le nombre moyen de *gisements* collectés pour un effort d'échantillonnage standard et $\frac{1}{\rho}$ la biomasse moyenne contenue dans un *gisement* dans un milieu homogène donné.

Quoique séduisantes, les interprétations conceptuelles attribuées aux variables latentes et aux paramètres du modèle LOL sont à prendre avec beaucoup de précautions. Elles concernent des quantités non observables dans la réalité. Tenter de quantifier le niveau d'agrégation spatiale d'une espèce à partir des paramètres μ et ρ n'est pas judicieux car il existe une corrélation linéaire positive entre ces deux paramètres qui confond leur interprétation.

		$\widehat{\delta}_{100}$		$R(\delta)$		$P_{100}(\delta)$		$CV_{100}(\delta)$	
	Efforts simulés	LOL	$\Delta\Gamma$	LOL	$\Delta\Gamma$	LOL	$\Delta\Gamma$	LOL	$\Delta\Gamma$
constant	0.1	0.59	0.94	94	0	0.45	1	0.22	0.02
	0.5	0.60	0.78	97	4	0.44	0.99	0.11	0.05
	0.9	0.61	0.64	95	91	0.47	0.70	0.08	0.07
	1.1	0.60	0.58	97	89	0.43	0.32	0.08	0.08
	1.5	0.60	0.48	98	26	0.44	0.03	0.07	0.10
	2	0.60	0.37	94	0	0.36	0	0.07	0.13
	λ	LOL	$\Delta\Gamma$	LOL	$\Delta\Gamma$	LOL	$\Delta\Gamma$	LOL	$\Delta\Gamma$
exponentiel	0.32	0.60	0.64	99	90	0.46	0.68	0.08	0.07
	0.22	0.60	0.63	93	93	0.46	0.59	0.08	0.08
	0.18	0.60	0.62	94	98	0.41	0.53	0.08	0.08
	0.16	0.60	0.62	93	85	0.44	0.54	0.08	0.08

Tableau 5.5 – Efforts d'échantillonnage constants et exponentiels : Médianes *a posteriori*, pourcentages de recouvrements , indicateurs de sur-estimation et coefficients de variation *a posteriori* associés à l'estimation de la probabilité d'occurrence de zéros δ. Valeur simulée de δ : 0.61

Plusieurs extensions et modifications du modèle LOL pourront être explorées dans le futur. Une version discrète du modèle LOL dans laquelle la loi exponentielle est remplacée par une loi géométrique pourrait constituer une alternative interessante aux traditionnels modèles ZIP et ZINB utilisés pour l'analyse de données *zero-inflated* discrètes. La construction et l'inférence fréquentiste, à l'aide d'un algorithme EM, d'un tel modèle est actuellement étudiée par Marie-Pierre Etienne, Eric Parent et Jacques Bernier. L'introduction de covariables environnementales susceptibles d'expliquer la répartition spatiale d'une espèce serait également à envisager. Elles permettraient, par exemple, de spécifier des lois *a priori* plus informatives sur les paramètres μ et ρ via des modèles linéaires généralisés.

Comme le modèle $\Delta\Gamma$, le modèle LOL ne permet pas de distinguer les deux sources de zéros observées dans les données de biomasse collectées. Ainsi, le paramètre $\delta{=}e^{-\mu}$ confond la probabilité d'occurrence d'un *vrai* zéro issu d'un échantillonnage dans une zone où l'espèce cible est absente ou très rare et la probabilité d'occurrence d'un *faux* zéro considéré comme une erreur d'observation dûe, par exemple, à la procédure d'échantillonnage ou à un faible niveau de capturabilité de l'espèce étudiée. Par conséquent, les modèles LOL et $\Delta\Gamma$ ne permettent pas d'estimer la probabilité "réelle" d'absence de l'espèce en un site i.e., uniquement dûe à l'occurrence de *vrais* zéros. Il serait donc intéressant de proposer des extensions aux modèles LOL et $\Delta\Gamma$ permettant de distinguer, dans la mesure du possible, ces deux sources de zéros afin d'obtenir une meilleure estimation de la probabilité d'absence de l'espèce dans le domaine de suivi. Par exemple, une prise en compte explicite d'effets susceptibles de gonfler "artificiellement" la probabilité d'occurrence de zéros (e.g., niveau de capturabilité de l'espèce) serait une première voie d'approche à examiner. A noter que, dans la littérature, le terme "probabilité d'absence" est souvent mal employé car cette probabilité confond généralement les *vrais* et les *faux* zéros. Cette confusion conduit à sur-estimation de la probabilité d'absence réelle de l'espèce en un site. Par ailleurs, cette probabilité est souvent mal définie car elle n'est pas rapportée à un effort d'échantillonnage. Or, une "probabilité d'absence" de 0.8 sur une distance chalutée de 1 mile n'a pas la même interprétation écologique qu'une "probabilité d'absence" de 0.8 sur

une distance chalutée de 10 miles !

Au cours de ce travail, des simulations ont été réalisées afin de comparer les performances d'estimation des modèles LOL et $\Delta\Gamma$ en fonction du nombre d'observations et de la variabilité de l'effort d'échantillonnage. J'ai montré, qu'à efforts d'échantillonnage constants, les performances d'estimation des deux modèles sont globalement similaires. Par ailleurs, j'ai montré que le modèle $\Delta\Gamma$ ne permet pas de comparer des relevés d'abondance réalisés à efforts d'échantillonnage distincts. Contrairement au modèle LOL, nous avons vu que le modèle $\Delta\Gamma$ propose en théorie un traitement séparé des zéros et non-zéros. Une étude par simulations à envisager serait d'évaluer la pertinence d'une telle hypothèse dans le cas de l'analyse de données continues *zero-inflated*.

Dans ce chapitre (et le précédent), nous nous sommes placés dans la situation simple où les données de biomasse sont recueillies dans un domaine d'échantillonnage suffisamment petit pour abriter un milieu de vie homogène. Cette hypothèse est peu vérifiée en pratique. Ainsi, par exemple, le domaine d'étude parcouru par le chalutier scientifique affrété par le Centre des Pêches du Golfe couvre $73182km^2$ et *a fortiori* les caractéristiques environnementales sont loin d'être homogènes. Dans le chapitre suivant, nous considérons la situation plus complexe où le domaine d'échantillonnage abrite des types d'habitats différents induisant une hétérogénéité dans la répartition des espèces. Dans ce cas, l'utilisation des modèles LOL ou $\Delta\Gamma$ n'est plus adaptée.

Dans sa version étendue du modèle $\Delta\Gamma$ (cf section 4.3.3), Stefansson (1996) propose d'introduire la latitude en chaque site échantillonné k comme variable explicative de la probabilité d'occurrence d'un zéro en ce site. Une fonction de lien logit permet d'écrire :

$$logit(1 - \delta_k) = \alpha_2 + \beta_2 h_k$$

où h_k désigne la latitude au site k, α_2 la probabilité d'occurrence d'un zéro moyenne (à échelle logit) de l'espèce sur l'ensemble du domaine d'étude et β_2 mesure l'effet de la covariable latitude. Une telle approche permet de quantifier l'effet de la localisation géographique sur la distribution des espèces. Toutefois, elle reste non-spatialement explicite. Dans le chapitre suivant, je propose de tirer partie de la vision hiérarchique pour introduire un modèle spatial dans la couche phénoménologique des modèles LOL et $\Delta\Gamma$. L'idée est de générer des dépendances entre points de mesure géographiquement proches afin d'expliquer la variabilité spatiale de la distribution des organismes vivants.

Chapitre 6

Modéliser l'hétérogénéité spatiale de la biomasse avec des structures sur grille latentes

Nous supposons désormais que le domaine d'échantillonnage est partitionné en I unités géographiques suffisamment petites pour abriter chacune un type d'habitat homogène. Cette hypothèse est écologiquement "plausible". En effet, les processus physiques (courants, vents, tectonique des plaques,..) créent de la structuration spatiale dans l'environnement qui peut notamment prendre la forme de régions homogènes ("patch") séparées par des discontinuités brutales des caractéristiques environnementales (Legendre & Legendre, 1998).

Une première approche possible est de mener l'inférence des modèles LOL ou $\Delta\Gamma$ séparément pour chaque unité géographique i $(1, 2, \ldots, I)$. Cela revient à supposer que ces unités sont mutuellement indépendantes i.e., qu'elles n'entretiennent aucun lien de similarité les unes avec les autres. De telles structures aléatoires ont pour principale faiblesse de ne pas être parcimonieuses : les modèles LOL et $\Delta\Gamma$ comptent alors respectivement $2 \times I$ et $3 \times I$ paramètres inconnus. On s'attend à ce que ces modèles surparamétrés aient un faible pouvoir prédictif.

Une approche très parcimonieuse consiste à agréger toutes les unités géographiques en une seule et même unité et de mener l'inférence dans le domaine d'échantillonnage global. Cela revient à supposer que les paramètres contrôlant la variabilité intra-strate dans les modèles LOL et $\Delta\Gamma$ sont tous identiques (e.g., $\mu_i = \mu$ pour tout $i = 1, 2, \ldots I$). Nous parlerons de modèle agrégé. Quoique parcimonieuse, cette approche a pour principale faiblesse de ne pas tenir compte de la structuration spatiale possible des quantités de biomasse observées. Par ailleurs, elle viole l'hypothèse d'homogénéité de chaque unité géographique.

Dans ce chapitre, je traite le cas intermédiaire où les I unités géographiques présentent des comportements similaires (mais ne sont pas identiques). D'un point de vue formel, cela revient à considérer les paramètres contrôlant la variabilité intra-unité de la biomasse non plus comme des paramètres mais comme des variables aléatoires latentes (ou effets aléatoires). L'objectif de ce chapitre est d'utiliser la démarche hiérarchique pour complexifier les structures de base LOL et $\Delta\Gamma$ afin de tenir compte de la variabilité inter-unités.

Ce chapitre s'articule en 3 parties. Dans un premier temps, je décris deux types de structures possibles pour décrire l'hétérogénéité d'une même variable aléatoire entre

différentes unités géographiques. La première introduit un lien non-spatialisé global entre toutes les unités. La seconde introduit en plus un lien de dépendance spatiale local entre des unités géographiquement voisines. Dans une deuxième partie, j'ajoute ces nouvelles "briques" aux structures de base LOL et $\Delta\Gamma$ afin de modéliser, à un niveau latent, l'hétérogénéité de leurs effets aléatoires respectifs. Dans la partie 3, les capacités d'ajustement et les capacités prédictives de 10 constructions hiérarchiques différentes basées sur les modèles LOL ou $\Delta\Gamma$ sont comparées à partir des données de biomasse en invertébrés épibenthiques du Golfe-du-Saint-Laurent (cf chapitre 1). Des facteurs de Bayes sont notamment calculés pour comparer les capacités d'ajustement des différents modèles en compétition.

6.1 Modélisation de la variabilité inter-unité

Dans cette section, Z_i désigne un effet aléatoire spécifique à l'unité géographique i. Dans le cas du modèle LOL, par exemple, $Z_i = \mu_i$ ou $Z_i = \rho_i$. Deux grands types de structures aléatoires permettent de générer des liens entre les effets aléatoires propres à différentes unités géographiques.

6.1.1 La modélisation *régionalisée*

Les effets aléatoires sont supposés indépendants mais issus d'une distribution paramétrique commune H :

$$[Z_i|\theta] \overset{i.i.d}{\sim} H(\theta)$$

Dans la suite de ce document, je qualifie H de distribution *régionale* car elle modélise l'hétérogénéité des unités géographiques à l'échelle globale de la région d'échantillonnage. La distribution régionale H assure un lien probabiliste entre les I unités considérées. Le vecteur de paramètres θ module leur niveau de similarité.

Dans le cas du modèle LOL, les nombres moyens de *gisements* $\mu_i (i = 1, \ldots, I)$ et/ou l'inverse de la biomasse moyenne contenue dans chaque *gisement* ρ_i peuvent être liés par les distributions régionales suivantes :

$$\mu_i \overset{i.i.d}{\sim} Gamma(a, b)$$
$$\rho_i \overset{i.i.d}{\sim} Gamma(c, d)$$

Le choix de la loi gamma permet de bénéficier des commodités calculatoires offertes par la conjugaison (cf Annexe B) lorsque celle-ci est combinée avec une loi de Poisson ou une loi exponentielle. Ceci est notamment le cas pour le modèle LOL. Plus la variance des lois gamma, a/b^2 et/ou c/d^2, est faible, plus les unités géographiques se ressemblent en terme d'intensité de *gisements* ou de biomasse moyenne contenue dans chaque *gisement*.

Sous le modèle $\Delta\Gamma$, 3 effets aléatoires $Z_i = (\delta_i, a_i, b_i)$ contrôlent la variabilité de la biomasse à l'intérieur de chaque unité géographique i. La probabilité d'occurrence d'un

zéro δ_i et/ou les variables latentes a_i et b_i de la loi gamma peuvent être liés comme suit :

$$\begin{aligned} logit(\delta_i) &\overset{i.i.d}{\sim} \mathcal{N}(m_1, \sigma_\delta^2) \\ a_i &\overset{i.i.d}{\sim} Gamma(a, b) \\ b_i &\overset{i.i.d}{\sim} Gamma(c, d) \end{aligned}$$

Les structures *régionalisées* permettent de combiner l'information disponible dans toutes les unités géographiques et ce, quelle que soit leur distance géographique. Les liens de similarités entre unités facilitent les estimations dans les unités pour lesquelles le nombre d'observations est faible.

6.1.2 Le modèle BYM

Au-delà de l'existence de similarités globales entre les I unités considérées, on peut s'attendre à ce que des dépendances spatiales locales lient entre-elles les unités géographiques voisines.

Dans le cadre spatial, Besag, York et Mollié ont proposé le modèle suivant, appelé modèle BYM (Besag & al., 1991), pour représenter des effets aléatoires Z_i non-gaussiens et spatialement corrélés :

$$W_i = m_0 + \Phi_i + \epsilon_i \qquad \text{avec} \qquad Z_i = g^{-1}(W_i) \tag{6.1.1}$$

Ce modèle permet de décomposer l'effet de l'unité géographique i en un effet spatialement structuré Φ_i et un effet résiduel non-spatialement structuré ϵ_i. Dans l'expression 6.1.1 :

- g désigne une fonction de lien canonique fixée.
- m_0 est un terme constant. Il représente l'effet moyen commun à toutes les unités géographiques.
- les ϵ_i sont des effets aléatoires gaussiens indépendants et identiquement distribués :

$$\epsilon_i \overset{i.i.d}{\sim} \mathcal{N}(0, \sigma_\epsilon^2) \qquad \forall i = 1, 2, \ldots, I$$

- le vecteur $\Phi = (\Phi_1, \ldots, \Phi_I)$ désigne un ensemble d'effets aléatoires gaussiens spatialement structurés selon un modèle CAR intrinsèque. Nous noterons ce modèle IAR pour "Intrinsically AutoRegressive". Il est défini par les lois conditionnelles suivantes :

$$[\Phi_i | \Phi_{-j}] \propto \exp\left\{ -\frac{(\Phi_i - \frac{1}{\omega_{i+}} \sum_{j \sim i} \Phi_j)^2}{2\frac{\sigma_{IAR}^2}{\omega_{i+}}} \right\}.$$

  où $\Phi_{-j} = (\Phi_1, \ldots, \Phi_{j-1}, \Phi_{j+1}, \ldots, \Phi_I)$, $j \sim j'$ signifie que les unités géographiques j et j' sont voisines, ω_{i+} désigne le nombre de voisins de l'unité i et σ_{IAR}^2 est un paramètre de variance locale. Ce modèle suppose que la distribution conditionnelle de l'effet Φ_i dans l'unité géographique i est une loi normale centrée en la moyenne des effets de ses unités voisines et de variance inversement proportionnelle au nombre de voisins : le niveau d'incertitude diminue quand le nombre de voisins augmente. Deux unités géographiques sont supposées voisines quand elles partagent une frontière commune.

– les effets Φ_i et ϵ_i sont indépendants.

Le modèle IAR a l'avantage d'être facilement estimable. En effet, ses distributions conditionnelles complètes ont une forme analytique connue ce qui permet de recourir à l'échantillonneur de Gibbs, décrit dans l'annexe B. En revanche, le modèle IAR est impropre : sa moyenne est non-définie et sa variance est infinie. D'après le lemme de Brook, la loi jointe des Φ_i est définie par :

$$[\phi_1, \phi_2, \ldots, \phi_I | \sigma^2_{IAR}] = (\frac{1}{\sigma^2_{IAR}})^{-I/2} exp\left(-\frac{\sigma^2_{IAR}}{2}\phi^T \Sigma^{-1} \phi\right) \tag{6.1.2}$$

où :

$$\Sigma^{-1}_{ij} = \begin{cases} \omega_{i+} & \text{si i=j} \\ -1 & \text{si i est voisin de j} \\ 0 & \text{sinon} \end{cases}$$

La distribution 6.1.2 a la forme d'une loi gaussienne multivariée. Cependant, la contrainte $\sum_i \phi_i = 0$ doit être imposée pour rendre le modèle identifiable. Par ailleurs, la matrice Σ^{-1} est singulière ce qui signifie que la matrice de variance-covariance Σ n'existe pas. Le modèle IAR est une mesure de masse infinie. Des variables observables ne peuvent donc pas être issues d'un tel mécanisme aléatoire. En revanche, ce modèle peut, comme cela est le cas ici, être utilisé comme *prior* pour modéliser des effets aléatoires. Dans ce cas, si les *posteriors* induits sont des lois propres, l'inférence bayésienne ne pose pas de difficultés particulières.

Les modèles régionalisés proposés dans la section 6.1.1 permettent d'effectuer un lissage global des effets aléatoires entre les différentes unités géographiques. Le modèle BYM a l'avantage de modéliser simultanément l'hétérogénéité globale et l'hétérogénéité locale des effets aléatoires. L'introduction de Φ permet de ne pas leur imposer la même variance pour chaque unité géographique puisque le nombre de voisins n_i est différent pour chaque unité. Les variances σ^2_ϵ et σ^2_{IAR} modulent les niveaux d'hétérogénéité globale et locale respectivement. Plus σ^2_ϵ est petit, plus les effets aléatoires ont tendance à être similaires entre toutes les unités géographiques. Plus σ^2_{IAR} est petit, plus les effets aléatoires ont tendance à être similaires entre unités géographiques voisines.

Dans le cas du modèle LOL, le modèle BYM peut être ajouté pour introduire des corrélations spatiales entre les effets aléatoires μ_i ou ρ_i. Ces deux effets aléatoires prennent des valeurs positives. Aussi, nous avons opté pour la fonction de lien log, un choix usuel qui permet de se ramener dans l'ensemble des nombres réels $\mathbb{R}$:

$$log(\mu_i) = \alpha_0 + \Phi_i + \epsilon_i \qquad \text{ou} \qquad log(\rho_i) = \alpha_0 + \Phi_i + \epsilon_i$$

Le modèle BYM pourrait également être utilisé pour modéliser l'hétérogénéité spatiale des biomasses moyennes, données par le rapport $Q_i = \frac{\mu_i}{\rho_i}$ pour chaque unité i : la biomasse moyenne dans l'unité i dépendrait alors des biomasses moyennes de ses unités voisines.

Dans le cas du modèle $\Delta\Gamma$, le modèle BYM peut être ajouté pour introduire des corrélations spatiales entre les probabilités d'occurrence d'un zéro δ_i via un lien logit (ou probit) ou entre les biomasses moyennes des quantités strictement positives $\frac{a_i}{b_i}$ via un lien log :

$$logit(\delta_i) = \alpha_0 + \Phi_i + \epsilon_i \qquad \text{ou} \qquad log\left(\frac{a_i}{b_i}\right) = \alpha_0 + \Phi_i + \epsilon_i$$

6.2 Dix modèles hiérarchiques possibles

Selon le modèle d'observations (LOL ou $\Delta\Gamma$), le type de structure ajoutée à chaque effet aléatoire (indépendante, régionale, BYM, aggrégée) et les distributions régionales supposées, de multiples constructions hiérarchiques distinctes peuvent être spécifiées pour modéliser des données de biomasse *zero-inflated* et spatialement hétérogènes. La table 6.1 résume les 10 constructions hiérarchiques, spatialisées ou non, que je propose de comparer dans la section 6.3.

Structure inter-unité / Modèle des observations	LOL $Z_i = (\mu_i, \rho_i)$	$\Delta\Gamma$ $Z_i = (\delta_i, b_i)$
Indépendante (I)	$(\text{LOL})^{\otimes I}$	$(\Delta\Gamma)^{\otimes I}$
Partiellement *régionalisée* (PR)	R_μ-LOL	R_δ-$\Delta\Gamma$
Régionalisée (R)	$\text{R}_{\mu,\rho}$-LOL	$\text{R}_{\delta,b}$-$\Delta\Gamma$
Partiellement spatialisée (PS)	BYM_μ-LOL	BYM_δ-$\Delta\Gamma$
Spatialisée et régionalisée (S+R)	$\text{BYM}_\mu R_\rho$-LOL	$\text{BYM}_\delta R_b$-$\Delta\Gamma$

Tableau 6.1 – Dix structures hiérarchiques possibles pour la modélisation de données *zero-inflated* spatialement hétérogènes. Les lettres entre parenthèses i.e., (I),(PR),(R),(PS) et (S+R) seront utilisées par la suite pour désigner les versions hiérarchiques associées.

5 variantes du modèles LOL ont été considérées : le modèle LOL appliqué indépendamment dans les I unités géographiques notée $(LOL)^{\otimes I}$, le modèle $R_{\mu,\rho}$-LOL dans lequel les effets μ_i (i=1,2,..,I) et ρ_i sont régionalisés selon une loi gamma (cf section 6.1.1), une version partiellement régionalisée R_μ-LOL dans laquelle seules les variables latentes μ_i sont régionalisées selon une loi gamma et les ρ_i sont supposés constants $\rho_i = \widetilde{\rho}$, une version spatialisée sur les μ et constante sur les ρ appelée BYM_μ-LOL et enfin une version spatialisée sur les μ et régionalisée sur les ρ appelée $BYM_{\mu,\rho}$-LOL.

5 modèles similaires basés sur la loi $\Delta\Gamma$ ont également été considérés : le modèle $\Delta\Gamma$ appliqué indépendamment dans I unités géographiques noté $(\Delta\Gamma)^{\otimes I}$, le modèle $R_{\delta,b}$-$\Delta\Gamma$ dans lequel les effets $logit(\delta_i)$ et b_i sont respectivement régionalisés selon une loi normale et gamma (cf section 6.1.1), une version partiellement régionalisée R_δ-LOL dans laquelle seules les $logit(\delta_i)$ sont régionalisés et les b_i sont supposés constants $b_i = \widetilde{b}$, une version spatialisée sur les $logit(\delta_i)$ et constante sur les b_i appelée BYM_δ-LOL et enfin une version spatialisée sur les $logit(\delta_i)$ et régionalisée sur les b_i appelée $BYM_{\delta,b}$-LOL.

Les mêmes structures sont attribuées aux effets aléatoires δ_i et μ_i car ils sont tous deux liés à une même quantité inconnue : la probabilité d'occurrence de zéros dans l'unité i. Dans tous les modèles basés sur la loi $\Delta\Gamma$, nous avons supposé que $a_i = \widetilde{a}$ pour tout $i = 1, 2, \ldots, I$. Cette hypothèse revient à fixer le coefficient de variation des lois gamma modélisant la distribution des quantités de biomasse strictement positives. D'un point de vue pratique, ce choix permet de spécifier un modèle basé sur un nombre d'effets aléatoires (=2I+1) similaire à celui induit par l'utilisation du modèle LOL (=2I). Concrètement, cette hypothèse revient à supposer que, quelle que soit la zone géographique, la variabilité des quantités de biomasse strictement positives observées est identique une fois rapportée au niveau moyen de biomasse observé.

Les lois *a priori* choisies pour mener l'inférence bayésienne des dix structures hiérarchiques considérées sont résumées dans le tableau 6.2. Pour les mêmes raisons que celles évoquées

dans la section 5.1.4 du chapitre précédent, nous avons opté pour des *priors* plats ou peu informatifs. Les lois gamma sont très souvent choisies pour les précisions (inverses des variances) des lois normales. Dans le cas du modèle BYM, nous avons utilisé les lois *a priori* gamma recommandées par Banerjee et al. (2004).

LOL		$\Delta\Gamma$	
Paramètres	Loi *a priori*	Paramètres	Loi *a priori*
a	G(0.01,0.01)	c	G(0.01,0.01)
b	G(0.01,0.01)	d	G(0.01,0.01)
c	G(0.01,0.01)	m	N(0, 10^6)
d	G(0.01,0.01)	s^2	IG(0.01,0.01)
$\widetilde{\rho}$	N(0, 10^6) tronquée	$\widetilde{a}$	N(0, 10^6) tronquée
		$\widetilde{b}$	N(0, 10^6) tronquée
α_0	N(0, 10^6)	α_0	N(0, 10^6)
σ^2_ϵ	IG(0.001,0.001)	σ^2_ϵ	IG(0.001,0.001)
σ^2_{IAR}	IG(0.1,0.1)	σ^2_{IAR}	IG(0.1,0.1)

Tableau 6.2 – Lois *a priori* des paramètres

6.3 Application aux données de relevés au chalut de fond du Centre des Pêches du Golfe

6.3.1 Objectifs et descriptif de l'analyse

Dans le chapitre précédent, les performances d'estimation des modèles LOL et $\Delta\Gamma$ ont été comparées à partir de données simulées. Je propose désormais de comparer leurs performances respectives à partir de données réelles : les relevés au chalut de fond du Centre des Pêches du Golfe (cf section 1.1.2, chapitre 1).

Le sud du Golfe du Saint-Laurent a été partitionné en 27 strates d'échantillonnage. A partir de données environnementales annexes, j'ai rafiné cette partition en 38 strates plus petites afin de s'assurer que celles-ci abritent un milieu de vie "relativement" homogène (cf Figure 6.1). Par ailleurs, j'ai choisi deux espèces caractérisées par des répartitions géographiques différentes dans la région d'étude : les oursins et les anémones de mer (cf Figure 6.2). La distribution empirique des données de biomasse en anémones de mer est beaucoup plus piquée en zéro que celle des oursins. Afin de garantir des estimations suffisamment précises, les données recoltées ont été agrégées sur trois campagnes consécutives : 1999, 2000, 2001 (i.e., 540 observations).

L'analyse de ces jeux de données réelles a trois objectifs principaux :

- Comparer les capacités d'ajustement et les capacités prédictives des cinq versions hiérarchiques du modèle LOL aux versions correspondantes du modèle $\Delta\Gamma$ (cf Tableau 6.1).
- Valider ou invalider l'hypothèse selon laquelle chacune des espèces considérées a une répartition spatialement structurée à échelle des unités géographiques définies dans le domaine d'étude.
- Déterminer des facteurs d'impact susceptibles d'expliquer l'existence et la distribution de chacune de ces espèces.

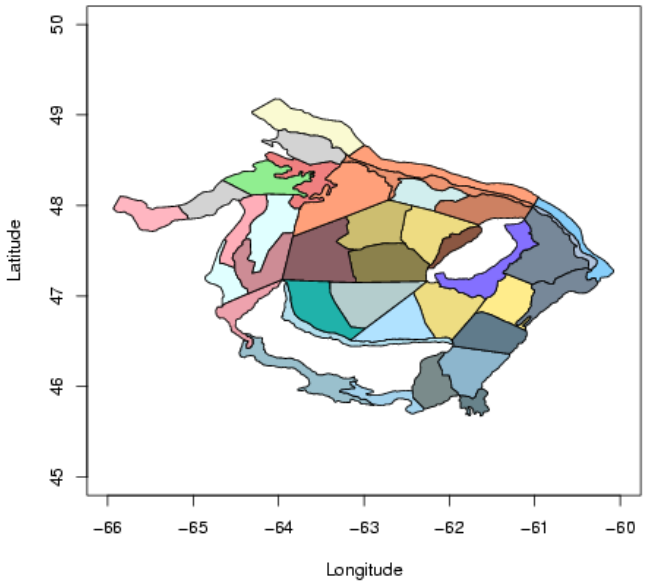


Figure 6.1 – Partition du domaine d'échantillonnage en 38 strates

J'ai réalisé l'inférence bayésienne des 10 modèles hiérarchiques définis dans le tableau 6.1. Les échantillons *a posteriori* ont été calculés sur la base de 200000 itérations MCMC après une période de chauffe de 50000 itérations. Dans le chapitre précédent, nous avons vu que, pour le modèle LOL, de fortes autocorrélations pouvaient exister entre les itérations successives des chaînes MCMC (cf section 5.3.3). Aussi, afin d'éviter ces phénomènes d'auto-corrélation, les échantillons *a posteriori* ne sont composés que des valeurs simulées toutes les 100 itérations.

Dans un premier temps, j'ai comparé les capacités d'ajustement des différents modèles. J'ai calculé deux critères bayésiens différents de sélection de modèles :

- Un critère basé sur la déviance, appelé DIC (acronyme de *Deviance Information Criterion*). Ce critère a été suggéré par Spiegelhalter et al. (2002) comme une méthode générale pour comparer les capacités d'ajustement de modèles complexes. Il peut être perçu comme une version bayésienne du critère d'Akaike et est défini par :

$$DIC = \overline{D(\theta)} + (\overline{D(\theta)} - D(\overline{\theta}))$$

  où $\overline{D(\theta)}$ est la moyenne *a posteriori* de la déviance définie par $D(\theta) = -2log([y|\theta])$ et résume l'adéquation du modèle aux données. $D(\overline{\theta})$ est la valeur de la déviance quand les paramètres sont remplacés par leurs moyennes *a posteriori*. Leur différence reflète la complexité du modèle et représente le nombre effectif de paramètres souvent noté p_D. Il s'agit en fait d'un terme de pénalisation. Le modèle dont le DIC est le plus petit est considéré comme le modèle ayant la meilleure adéquation aux données. Le DIC est facilement calculable à partir d'échantillons *a posteriori*. Toutefois, il a pour inconvénient majeur de dépendre du choix, souvent peu trivial, de la paramétrisation utilisée pour le calculer.
- Un critère appelé, facteur de Bayes, qui quantifie via un rapport probabiliste la crédibilité *a posteriori* d'un modèle par rapport à un autre. Plusieurs méthodes

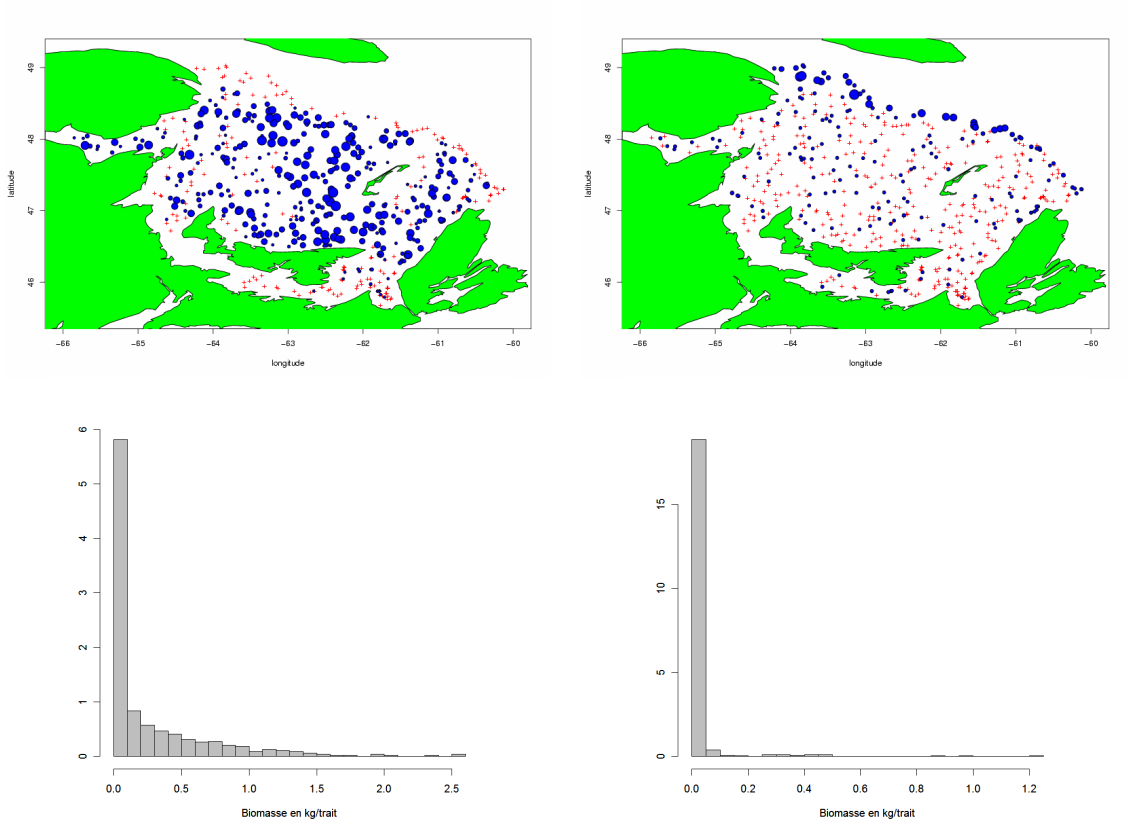


Figure 6.2 – (Haut) Répartition (Bas) Histogrammes empiriques de la biomasse en oursins (gauche) et en anémones de mer (droite) sur les années 1999-2000-2001. Les croix rouges représentent des sites où la biomasse recueillie est nulle. L'aire de chaque rond bleu est proportionnelle à la biomasse strictement positive collectée.

existent pour le calculer. Nous avons choisi la technique assymptotique approximée basée sur la formule de Raftery (Kass & Raftery, 1994) décrite en détail dans l'Annexe F. Dans les analyses, tous les facteurs de Bayes ont été calculés à partir d'un échantillon *a posteriori* de 500000 valeurs. En effet, un suivi de l'évolution du facteur de Bayes en fonction de la taille G de l'échantillon *a posteriori* utilisé a montré que 500000 valeurs étaient suffisantes pour obtenir des approximations stables du facteur de Bayes. Par ailleurs, une analyse de sensibilité des facteurs de Bayes calculés par rapport au degré d'information contenu dans les lois *a priori* a été menée.

Afin de juger de la pertinence des différents modèles proposés, j'ai également comparé les caractéristiques des estimations ponctuelles et des distributions *a posteriori* associées aux biomasses moyennes et aux probabilités d'occurrence de zéros par unité géographique.

Les modèles hiérarchiques considérés peuvent être utilisés en tant qu'outils prédictifs. Ils peuvent permettre de prédire l'évolution des stocks d'abondance ou aider à planifier les futures campagnes scientifiques. Aussi, dans un deuxième temps, j'ai comparé les modèles en compétition du point de vue de leurs capacités prédictives. Ces comparaisons sont basées sur le critère proposé par Gelfand et Ghosh (1998b) qui revient à minimiser une fonction de perte quadratique tenant compte à la fois de l'écart des prédictions aux données observées et de l'écart des prédictions à la prédiction moyenne. Ce critère, appelé critère de perte prédictive *a posteriori*, est décrit dans l'Annexe G. J'ai également comparé les

cartes des distributions prédites aux distributions observées.

Enfin, dans un troisième temps, j'ai estimé l'impact de covariables environnementales spécifiques (i.e., profondeur et type de sédiments) sur la répartition géographique de la biomasse en oursins et en anémones de mer. Pour cela, j'ai introduit une régression linéaire sur ces covariables dans la couche phénoménologique latente associée au paramètre μ du modèle $BYM_{\mu,\rho}$-LOL.

6.3.2 Analyse de la sensibilité du facteur de Bayes par rapport au degré informatif des lois *a priori*

Les facteurs de Bayes sont connus pour être sensibles aux choix des lois *a priori* (Kass & Raftery, 1994), (Sinharay & Stern, 2002). En théorie, leur calcul nécessite une loi *a priori* informative. Or, dans cette étude, toutes les lois *a priori* considérées sont non-informatives (cf tableau 6.2). Aussi, j'ai calculé des facteurs de Bayes particuliers, appelés facteurs de Bayes partiels (O'Hagan, 1991), à partir d'un niveau d'information équivalent à la spécification d'une loi *a priori* informative. Le principe général de cette méthode, détaillée dans l'Annexe F, est de partitionner l'ensemble des observations en un échantillon d'apprentissage et un échantillon de validation. Selon sa taille, l'échantillon d'apprentissage revient à spécifier une loi *a priori* plus ou moins informative pour le calcul du facteur de Bayes. Puis, l'idée est d'utiliser l'échantillon de validation pour calculer un facteur de Bayes partiel.

J'ai utilisé cette méthode pour étudier la sensibilité des facteurs de Bayes partiels, calculés sur les données de biomasse du Golfe-du-Saint-Laurent, par rapport au degré informatif des lois *a priori*. Le tableau 6.3 décrit les 7 partitions considérées du jeu de données agrégées (1999,2000,2001) en un échantillon d'apprentissage et un échantillon de validation.

- La partition 1 a un échantillon d'apprentissage vide. Le facteur de Bayes est calculé à partir des 3 années de données et sur la base d'une loi *a priori* non-informative
- Les partitions 2, 3 et 4 ont un échantillon d'apprentissage formé d'une année de données (3 possibilités : 1999 ou 2000 ou 2001). Le facteur de Bayes est alors calculé à partir des 2 années restantes et sur la base d'une loi *a priori* peu informative.
- Les partitions 5, 6 et 7 ont un échantillon d'apprentissage formé de deux années de données (3 possibilités : (2000,2001) ou (1999,2001) ou (1999,2000)). Le facteur de Bayes est calculé à partir d'une année de données (resp. 1999, 2000 ou 2001) sur la base d'une loi *a priori* informative.

Les échantillons d'apprentissage ont été définis afin de contenir des données indépendantes de celles contenues dans l'échantillon de validation.

La figure 6.3 indique, pour les oursins et les anémones de mer, la valeur des facteurs de Bayes partiels du modèle BYM$_\mu$-LOL relativement au modèle BYM$_\delta$-$\Delta\Gamma$ pour chacune des 7 partitions du tableau 6.3. Cette figure met clairement en évidence une sensibilité du facteur du Bayes :

- au degré d'information contenu dans la loi *a priori*,
- à l'échantillon de validation à partir duquel il est calculé.

Ce résultat est généralisable à tous les modèles comparés pour chaque espèce (résultats non montrés)

Dans le cas des oursins, les facteurs de Bayes partiels (échelle $2 \times log$) sont d'autant plus élevés, en faveur du modèle BYM$_\mu$-LOL, que le degré d'information *a priori*

	échantillon de validation	échantillon d'apprentissage	qualité de la loi *a priori*
1	1999-2000-2001	-	non-informative
2	2000-2001	1999	peu informative
3	1999-2001	2000	peu informative
4	1999-2000	2001	peu informative
5	1999	2000-2001	informative
6	2000	1999-2001	informative
7	2001	1999-2000	informative

Tableau 6.3 – 7 partitions possibles des quantités de biomasse observées lors des campagnes scientifiques effectuées en septembre 1999, 2000 et 2001

diminue. Cependant, les conclusions restent les mêmes : le modèle BYM$_\mu$-LOL s'ajuste mieux aux données que le modèle BYM$_\delta$-$\Delta\Gamma$ (i.e., les facteurs de Bayes sont strictement supérieurs à 2). Pour un niveau d'information *a priori* donné, notre analyse montre que le degré de présomption en faveur du modèle BYM$_\mu$-LOL varie énormément en fonction de l'échantillon de validation utilisé. Ainsi, on peut observer une très forte présomption (i.e., supérieure à 10) en faveur du modèle BYM$_\mu$-LOL pour les données observées en 2000 contre de faibles présomptions (i.e., inférieures à 6) pour celles observées en 1999 et 2001. La sensibilité aux données du facteur de Bayes partiel est très visible dans le cas des anémones de mer. Ainsi, le facteur de Bayes partiel calculé à partir d'une loi *a priori* non-informative indique une très forte présomption en faveur du modèle BYM$_\mu$-LOL alors que, dans le cas de lois *a priori* informatives, cette présomption dépend de l'échantillon de validation. Ainsi, les partitions 2, 3 et 5 conduisent à des facteurs de Bayes partiels négatifs (à échelle log) indiquant, au contraire, une forte présomption en faveur du modèle BYM$_\delta$-$\Delta\Gamma$. Ces résultats remettent en cause toute hypothèse d'homogénéité inter-annuelle des données de biomasse en oursins et en anémones de mer.

6.3.3 Comparaison des capacités d'ajustement des modèles en compétition

Pour mener l'inférence bayésienne de chaque modèle hiérarchique du tableau 6.1, 3 chaînes MCMC ont été lancées simultanément à partir de positions initiales différentes. De manière générale, la convergence des algorithmes MCMC est plus lente sous le modèle LOL par rapport à des versions similaires du modèle $\Delta\Gamma$. Cela s'explique en particulier par le fait que l'inférence sous le modèle LOL implique la mise à jour d'un nombre plus élevé de variables. En effet, elle nécessite la mise à jour du nombre de *gisements* N_k ramassé en chaque site échantillonné k=1,2,..,r. Globalement, pour chaque modèle, la convergence a été atteinte en 50000 itérations pour tous les paramètres et variables latentes. Considérons par exemple les structures hiérarchiques les plus complexes : BYM$_{\mu,\rho}$-LOL et BYM$_{\delta,b}$-$\Delta\Gamma$. La figure 6.4 représente 3 chaînes MCMC obtenues pour l'estimation des paramètres du modèle BYM$_{\mu,\rho}$-LOL à partir des données de biomasse en oursins. Pour chaque paramètre, les trois chaînes se mélangent bien indiquant une bonne convergence de l'algorithme MCMC. Ces résultats visuels ont été validés par le calcul de la diagnostic de convergence de Brooks et Gelman (cf Annexe B). En particulier, la convergence est atteinte sur les paramètres de variances σ^2_{IAR} et σ^2_ϵ ce qui, en pratique, est parfois difficile à obtenir dans le cas de structures BYM latentes. Cela signifie qu'il y a suffisamment de

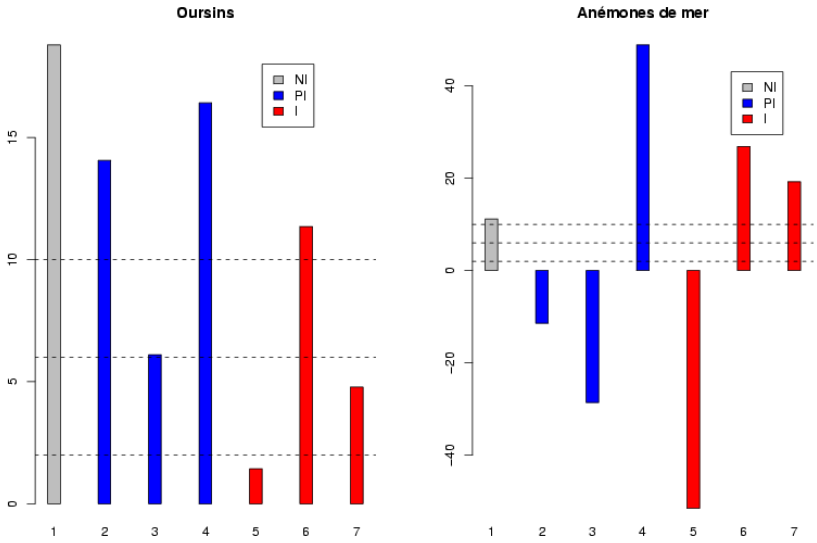


Figure 6.3 – Facteurs de Bayes à l'échelle log du modèle BYM_{μ}-LOL relativement au modèle BYM_{δ}-$\Delta\Gamma$ calculés pour chacune des 7 partitions du tableau 6.3. "NI"=*prior* non-informatif, "PI"=*prior* peu-informatif, "I"=*prior* informatif. Les lignes en pointillés représentent les seuils de significativité du facteur de Bayes donnés dans l'Annexe F.

signal dans les données pour pouvoir distinguer une part de variabilité spatialement structurée et une part de variabilité résiduelle non-structurée. Les figures 6.5 et 6.6 indiquent les densités *a posteriori* obtenues pour l'estimation de la biomasse moyenne dans 6 unités géographiques pour les modèles $\text{BYM}_{\mu,\rho}$-LOL et $\text{BYM}_{\delta,b}$-$\Delta\Gamma$ respectivement. Les densités associées à chaque chaîne MCMC se superposent indiquant une bonne convergence des algorithmes MCMC.

La figure 6.7 indique le degré de corrélation entre les chaînes MCMC des paramètres μ et ρ en fonction de la biomasse moyenne observée dans chacune des 38 unités géographiques d'intérêt. Ces corrélations sont calculées à partir des chaînes MCMC obtenues lors de l'inférence sous le modèle $\text{BYM}_{\mu,\rho}$-LOL. Dans le cas des oursins, les corrélations sont globalement fortes dans toutes les unités. Nous sommes dans la situation où les probabilités d'occurrence de zéros sont relativement faibles et les biomasses moyennes observées relativement élevées. En revanche, dans le cas des anémones de mer, cette corrélation est globalement faible car les biomasses moyennes sont plus faibles et probabilités d'occurrence de zéros plus élevées. Ces résultats sont généralisables à toutes les versions hiérarchiques du modèle LOL testées sur chaque espèce.

Les diagonales des tableaux 6.4 et 6.5 indiquent la crédibilité relative de chaque version hiérarchique du modèle LOL par rapport à la version correspondante du modèle $\Delta\Gamma$ lorsque ces modèles sont appliqués aux données de biomasse en oursins et en anémones de mer respectivement. Compte-tenu de leur sensibilité aux lois *a priori*, les facteurs de Bayes partiels indiqués dans ces tableaux sont ceux calculés sur la base de lois *a priori* informatives (partitions 5, 6, 7 du tableau 6.3). Ainsi, chaque case contient 3 valeurs correspondant aux facteurs de Bayes partiels calculés selon les partitions 5, 6 et 7 respectivement. La dernière colonne et la dernière ligne contiennent les DIC associés aux modèles basés sur LOL et $\Delta\Gamma$ respectivement. Dans le cas des oursins, les facteurs de Bayes partiels conduisent aux mêmes conclusions que les DIC du point de vue de la sélection de modèles.

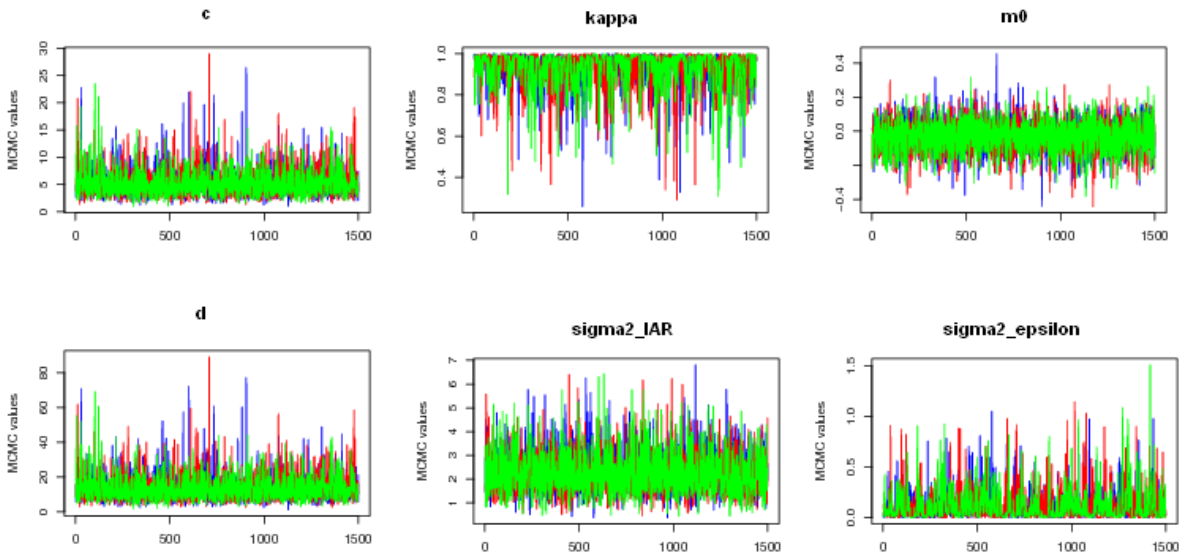


Figure 6.4 – Cas des oursins : 3 chaînes MCMC obtenues pour l'estimation de quelques paramètres du modèle $BYM_{\mu,\rho}-$LOL. c et d sont les paramètres de la loi régionale gamma posée sur les ρ_i. κ la part de variabilité expliquée par la structure IAR. m_0 la tendance sur les $log(\mu_i)$. σ^2_{IAR} et σ^2_ϵ les paramètres de variance locale et marginale. Chaque couleur désigne une chaîne MCMC. Les chaînes se mélangent bien : la convergence est atteinte.

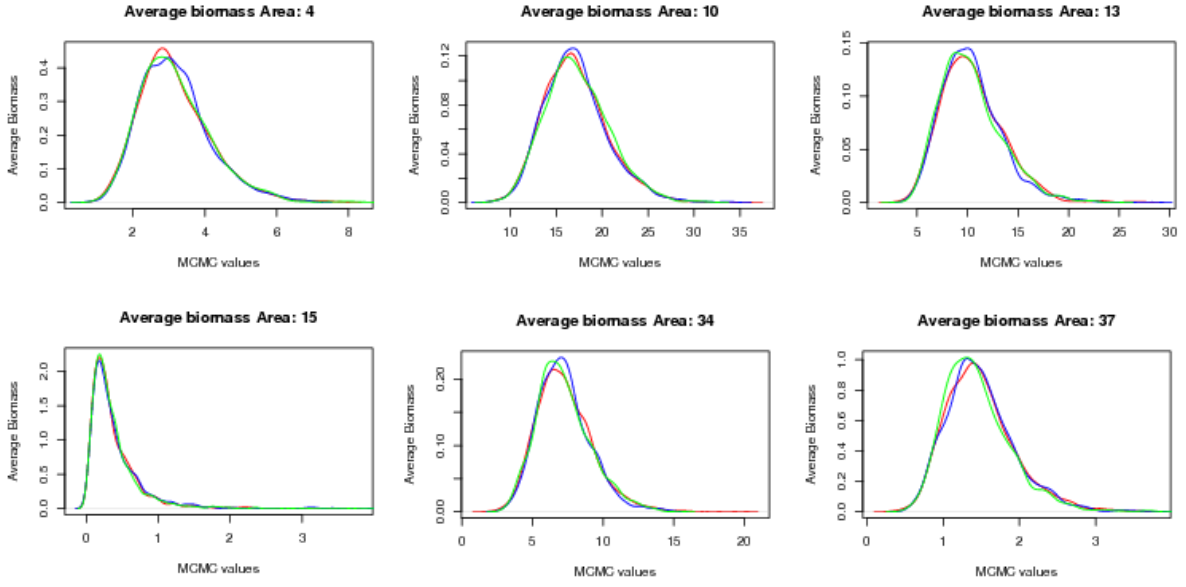


Figure 6.5 – Cas des oursins : Distributions *a posteriori* obtenues sous le modèle $BYM_{\mu,\rho}-$LOL pour la biomasse moyenne $Q = \frac{\mu}{\rho}$. Chaque couleur désigne une chaîne MCMC. Les courbes se superposent : la convergence est atteinte.

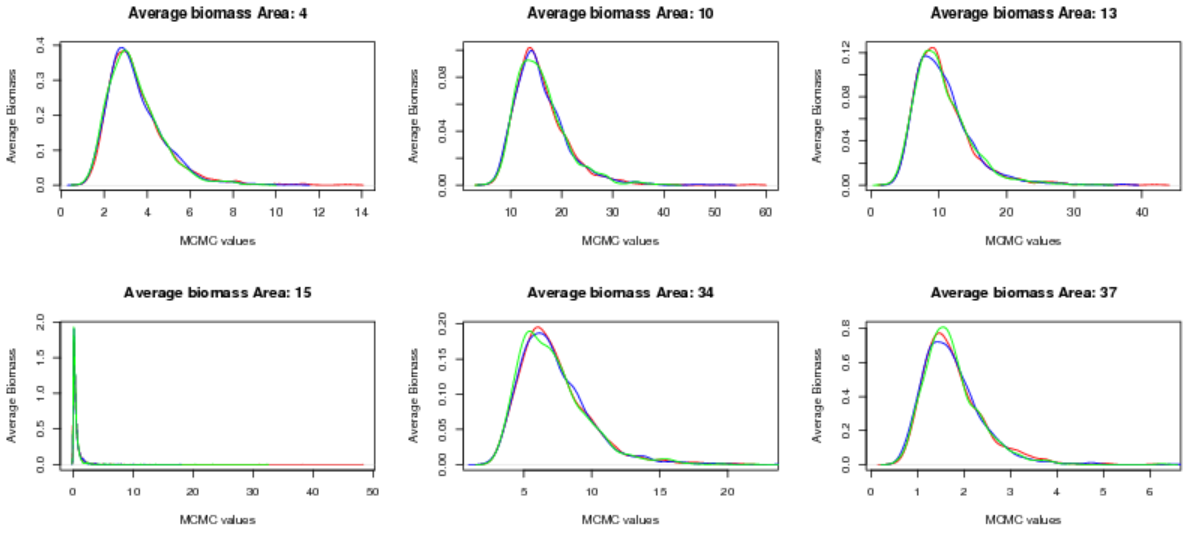


Figure 6.6 – Cas des oursins : Distributions *a posteriori* obtenues sous le modèle $BYM_{\delta,b} - \Delta\Gamma$ pour la biomasse moyenne $Q = (1-\delta)\frac{a}{b}$. Chaque couleur désigne une chaîne MCMC. Les courbes se superposent : la convergence est atteinte.

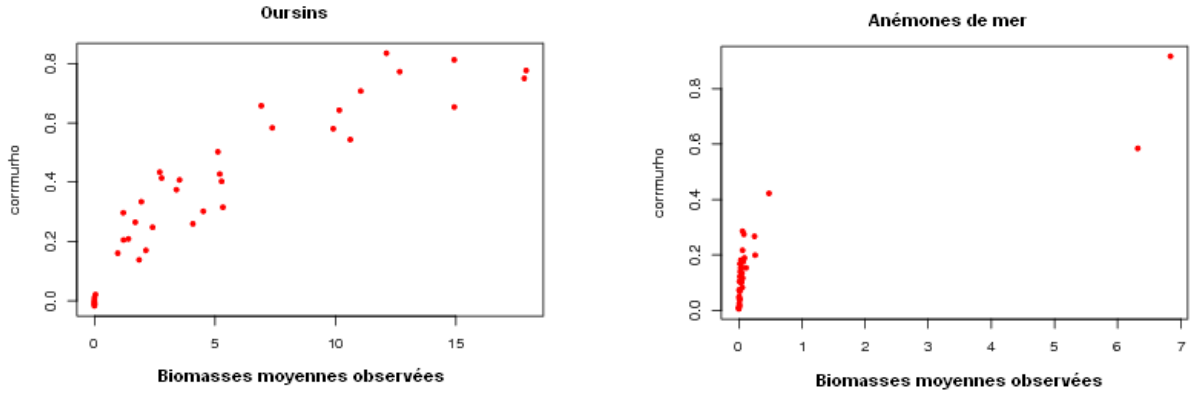


Figure 6.7 – Corrélations empiriques entre les chaînes MCMC des paramètres μ et ρ en fonction de la biomasse moyenne observée dans chaque unité géographique

Les facteurs de Bayes partiels positifs à échelle log indiquent que la version hiérarchique basée sur le modèle LOL s'ajuste mieux aux observations que la version similaire basée sur le modèle $\Delta\Gamma$. Et inversement dans les cas où ces facteurs de Bayes partiels sont négatifs. Dans le cas des anémones de mer, les facteurs de Bayes partiels sont très variables en fonction des années de données utilisées pour leur calcul : ils peuvent être fortement positifs comme fortement négatifs ce qui rend difficile toute comparaison de modèles et toute comparaison avec les conclusions induites par les calculs de DIC.

Les figures 6.8 et 6.9 permettent de comparer les médianes et écarts-types *a posteriori* obtenus, sous des versions hiérarchiques similaires des modèles LOL et $\Delta\Gamma$, pour l'estimation des biomasses moyennes Q_i et probabilités d'occurrence de zéros δ_i (i=1,2,...,38). Ces deux figures sont respectivement associées aux oursins et aux anémones de mer. Les versions hiérarchiques indépendantes ("I"), régionalisées ("R") et spatialisées et régionalisées ("S+R") sont comparées.

Enfin, les tableaux 6.6 et 6.7 contiennent des indices synthétiques calculés sur l'ensemble des 38 unités géographiques d'intérêt permettant de comparer les capacités d'adéquation aux données de versions hiérarchiques similaires des modèles LOL et $\Delta\Gamma$. Ces indices, calculés pour l'estimation de la biomasse moyenne Q et la probabilité d'occurrence de zéros δ, sont :

- La médiane *a posteriori* minimale et maximale i.e., $[\widehat{\theta}_{Min},\widehat{\theta}_{Max}]$ avec $\theta = Q$ ou $\theta = \delta$. Cet intervalle nous indique dans quelle gamme de valeurs sont estimées Q et δ.
- La médiane *a posteriori* moyenne i.e., $\widehat{\theta}_{mean}$.
- Le coût *a posteriori* moyen i.e., $C_{38}(\theta)$, associé à la fonction de coût absolu $|d-\theta|$ (où d est la médiane *a posteriori*). Il permet de quantifier le coût moyen associé à une estimation ponctuelle de la biomasse moyenne ou de la probabilité d'occurrence d'un zéro par la médiane *a posteriori* (cf Annexe C).
- L'écart-type *a posteriori* moyen i.e., $sd_{mean}(\theta)$ associé aux distributions *a posteriori* sur la biomasse moyenne ou la probabilité d'occurrence de zéros. Il nous renseigne sur le degré de précision des estimations obtenues.

Comparaison des variantes LOL et $\Delta\Gamma$ similaires

Quelle que soit l'espèce considérée, les tableaux 6.6 et 6.7 indiquent que le différentes versions hiérarchiques basées sur les modèles LOL et $\Delta\Gamma$ ont tendance à sur-estimer les biomasses moyennes observées. Ainsi, les médianes *a posteriori* moyennes $\widehat{\theta}_{mean}$ sont systématiquement supérieures aux médianes moyennes observées.

Pour les deux espèces considérées, les principales différences d'adéquation entre LOL et $\Delta\Gamma$ s'observent pour les structures hiérarchiques partiellement régionalisées ou partiellement spatialisées ie., R_μ-LOL, R_δ-$\Delta\Gamma$, BYM_μ-LOL et BYM_δ-$\Delta\Gamma$. Une comparaison des DIC indiquent que les modèles basés sur $\Delta\Gamma$ s'ajustent moins bien aux données que leur version similaire respective basée sur le modèle LOL. Les tableaux 6.6 et 6.7 montrent que les modèles basés sur $\Delta\Gamma$ conduisent à un lissage sur l'estimation de la biomasse moyenne nettement plus marqué que sous les modèles basés sur LOL. Ainsi, sous $\Delta\Gamma$, les intervalles $[\widehat{\theta}_{Min},\widehat{\theta}_{Max}]$ sont très réduits par rapport aux versions indépendantes, régionalisées et spatialisées. Sous les modèles basés sur LOL, des réductions d'intervalles apparaissent également mais elles sont moins importantes. Cela s'explique en partie par le fait que les modèles $\Delta\Gamma$ sont plus contraints par l'hypothèse $b_i = \widetilde{b}$ que les modèles LOL par l'hypothèse $\rho_i = \widetilde{\rho}$. En effet, les modèles R_δ-$\Delta\Gamma$ et BYM_δ-$\Delta\Gamma$ supposent qu'en moyenne, la quantité de biomasse non-nulle ramassée sur un effort d'échantillonnage standard, donnée

↱	$(\Delta\Gamma)^{\otimes 38}$	R_{δ}-$\Delta\Gamma$	$R_{\delta,b}$-$\Delta\Gamma$	BYM_{δ}-$\Delta\Gamma$	$BYM_{\delta,b}$-$\Delta\Gamma$	**DIC**
$(LOL)^{\otimes 38}$	3.88 -4.93 -5.53					2538.25
R_{μ}-LOL		3.49 13.05 4.64				2591.63
$R_{\mu,\rho}$-LOL			-3.66 -2.51 -7.08			2559.64
BYM_{μ}-LOL				1.42 11.36 4.78		2595.46
$BYM_{\mu,\rho}$-LOL					-4.41 -3.79 -7.01	2563.92
DIC	2524.62	2607.76	2545.57	2608.51	2545.68	

Tableau 6.4 – Cas des oursins : facteurs de Bayes à l'échelle log calculés sur 3 échantillons de validation différents : les données collectées en 1- 1999, 2- 2000, 3- 2001. La dernière colonne et la dernière ligne contiennent les DICs associés aux modèles basés sur LOL et $\Delta\Gamma$. respectivement.

↱	$(\Delta\Gamma)^{\otimes 38}$	R_{δ}-$\Delta\Gamma$	$R_{\delta,b}$-$\Delta\Gamma$	BYM_{δ}-$\Delta\Gamma$	$BYM_{\delta,b}$-$\Delta\Gamma$	**DIC**
$(LOL)^{\otimes 38}$	-11.04 3.28 -4.98					520.21
R_{μ}-LOL		-50.04 36.63 16.82				811.87
$R_{\mu,\rho}$-LOL			-0.66 3.29 0.23			509.95
BYM_{μ}-LOL				-51.51 26.85 19.22		803.89
$BYM_{\mu,\rho}$-LOL					-0.42 8.41 0.49	499.97
DIC	516.06	814.57	512.28	808.93	508.40	

Tableau 6.5 – Cas des anémones de mer : facteurs de Bayes à l'échelle log calculés sur 3 échantillons de validation différents : les données collectées en 1- 1999, 2- 2000, 3- 2001. La dernière colonne et la dernière ligne contiennent les DICs associés aux modèles basés sur LOL et $\Delta\Gamma$ respectivement.

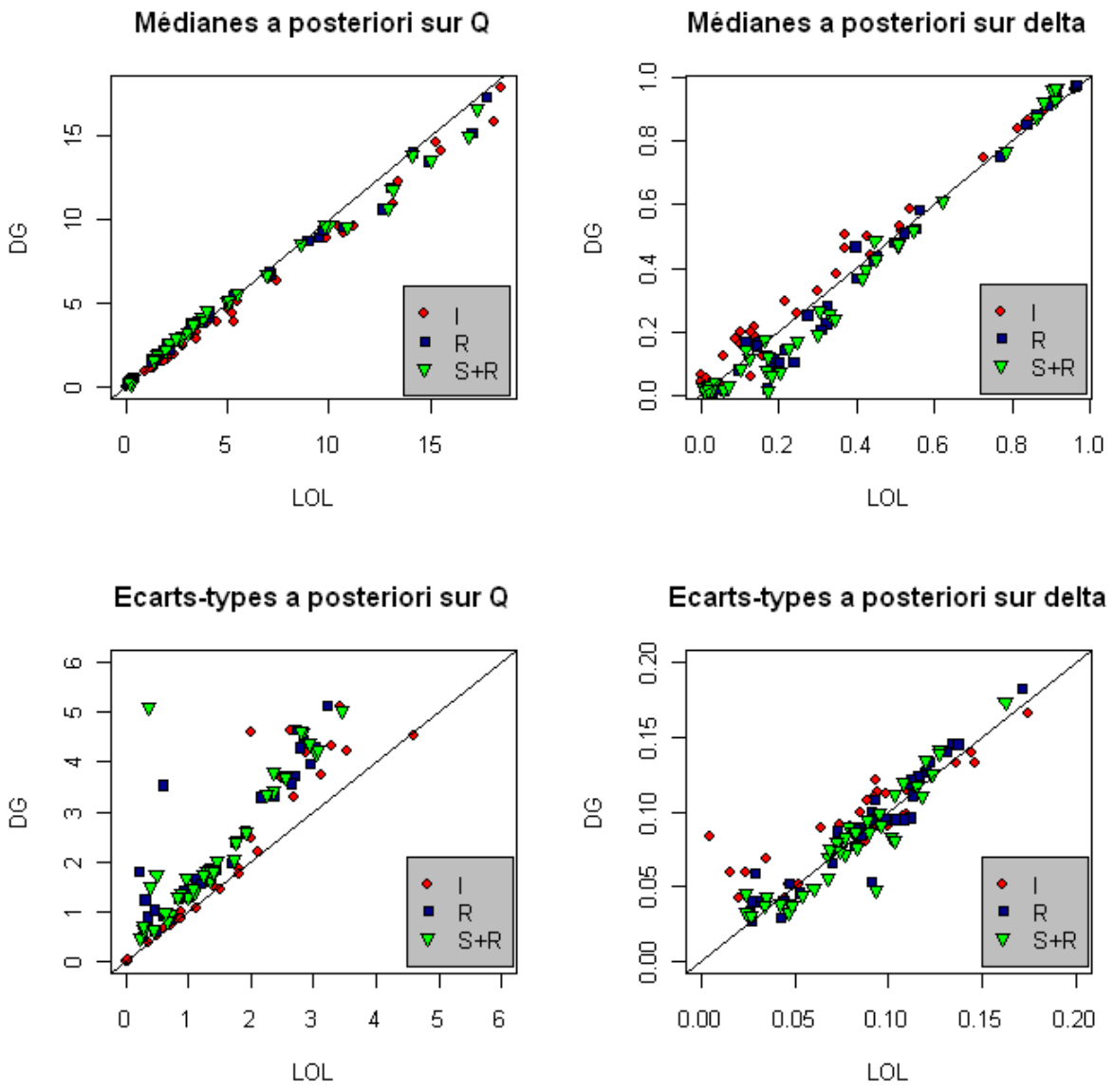


Figure 6.8 – Cas des oursins : Comparaison des médianes et écarts-types *a posteriori* pour l'estimation des biomasses moyennes Q_i (i=1,2,..,38) et des probabilités d'occurrence de zéros δ_i entre des versions hiérarchiques similaires des modèles LOL et $\Delta\Gamma$. I : versions indépendantes. R : versions régionalisées. S+R : versions spatialisées et régionalisées

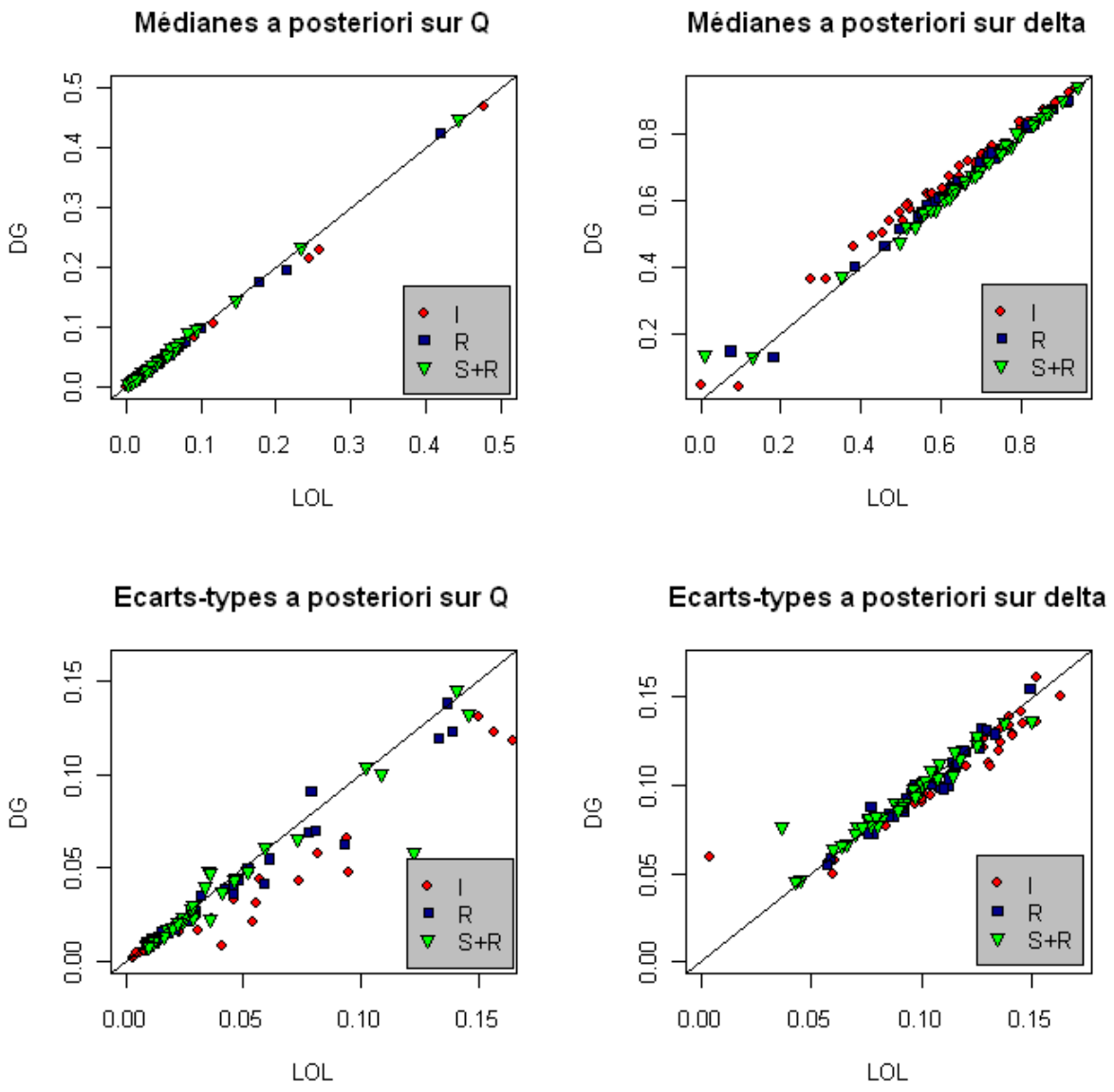


Figure 6.9 – Cas des anémones de mer : Comparaison des médianes et écarts-types *a posteriori* pour l'estimation des biomasses moyennes Q_i (i=1,2,..,38) et des probabilités d'occurrence de zéros δ_i entre des versions hiérarchiques similaires des modèles LOL et $\Delta\Gamma$. I : versions indépendantes. R : versions régionalisées. S+R : versions spatialisées et régionalisées

		$[\widehat{\theta}_{Min},\widehat{\theta}_{Max}]$			$\widehat{\theta}_{mean}$			$C_{38}(\theta)$		$sd_{mean}(\theta)$	
		obs	LOL	$\Delta\Gamma$	obs	LOL	$\Delta\Gamma$	LOL	$\Delta\Gamma$	LOL	$\Delta\Gamma$
Q	I	[0.19.63]	[0,18.48]	[0,17.89]	4.16	5.48	4.90	1.11	1.34	1.49	1.87
	PR	[0,19.63]	[0.1,16.43]	[0.22,8.51]	4.16	5.42	5.78	1.02	0.71	1.30	0.90
	R	[0,19.63]	[0.1,17.73]	[0.15,17.33]	4.16	5.32	5.22	1.11	1.53	1.47	2.25
	PS	[0,19.63]	[0.26,16.32]	[0.31,8.46]	4.16	5.41	5.80	1.02	0.70	1.30	0.86
	S+R	[0,19.63]	[0.21,17.31]	[0.20,16.60]	4.16	5.30	5.21	1.11	1.53	1.47	2.30
δ	I	[0,1]	[0.00,0.97]	[0.03,0.97]	0.33	0.32	0.35	0.07	0.07	0.08	0.09
	PR	[0,1]	[0.01,0.97]	[0.01,0.97]	0.33	0.36	0.33	0.06	0.07	0.08	0.09
	R	[0,1]	[0.01,0.97]	[0.01,0.97]	0.33	0.35	0.33	0.07	0.07	0.08	0.09
	PS	[0,1]	[0.01,0.92]	[0.01,0.96]	0.33	0.36	0.33	0.06	0.06	0.07	0.08
	S+R	[0,1]	[0.00,0.92]	[0.02,0.97]	0.33	0.36	0.33	0.06	0.06	0.08	0.08

Tableau 6.6 – Cas des oursins : Comparaison de quelques statistiques calculées sur l'ensemble des 38 médianes *a posteriori* associées à l'estimation de la biomasse moyenne Q_i et la probabilité d'occurrence de zéros δ_i pour des versions hiérarchiques similaires des modèles LOL et $\Delta\Gamma$.Les colonnes obs indiquent les statistiques empiriques calculées sur les observations.

		$[\widehat{\theta}_{Min},\widehat{\theta}_{Max}]$			$\widehat{\theta}_{mean}$			$C_{38}(\theta)$		$sd_{mean}(\theta)$	
		obs	LOL	$\Delta\Gamma$	obs	LOL	$\Delta\Gamma$	LOL	$\Delta\Gamma$	LOL	$\Delta\Gamma$
Q	I	[0,7.42]	[0,6.81]	[0,6.22]	0.27	0.40	0.37	0.07	0.07	0.11	0.10
	PR	[0,7.42]	[0,6.14]	[0.13,1.11]	0.27	0.40	0.42	0.07	0.12	0.09	0.15
	R	[0,7.42]	[0,5.68]	[0,5.44]	0.27	0.34	0.33	0.08	0.08	0.19	0.17
	PS	[0,7.42]	[0.02,6.38]	[0.07,1.11]	0.27	0.40	0.43	0.07	0.11	0.09	0.13
	S+R	[0,7.42]	[0,6.31]	[0,5.48]	0.27	0.37	0.34	0.08	0.08	0.13	0.15
δ	I	[0,1]	[0.00,0.94]	[0.04,0.94]	0.67	0.61	0.65	0.09	0.09	0.11	0.11
	PR	[0,1]	[0.00,0.98]	[0.13,0.90]	0.67	0.69	0.67	0.08	0.08	0.10	0.10
	R	[0,1]	[0.07,0.92]	[0.13,0.90]	0.67	0.66	0.67	0.08	0.08	0.10	0.10
	PS	[0,1]	[0.00,0.95]	[0.13,0.94]	0.67	0.69	0.67	0.07	0.07	0.09	0.09
	S+R	[0,1]	[0.01,0.94]	[0.13,0.94]	0.67	0.66	0.67	0.07	0.07	0.09	0.09

Tableau 6.7 – Cas des oursins : Comparaison de quelques statistiques calculées sur l'ensemble des 38 médianes *a posteriori* associées à l'estimation de la biomasse moyenne Q_i et la probabilité d'occurrence de zéros δ_i pour des versions hiérarchiques similaires des modèles LOL et $\Delta\Gamma$. Les colonnes obs indiquent les statistiques empiriques calculées sur les observations.

par $\frac{a}{b}$, est constante dans les différentes unités géographiques. En revanche, les modèles R_μ-LOL et BYM_μ-LOL supposent que la quantité de biomasse moyenne contenue dans un *gisement* i.e., $\frac{1}{\rho}$ est identique dans toutes les unités géographiques mais cela n'implique pas qu'en moyenne, les quantités moyennes de biomasse non-nulles ramassées soient constantes puisque ces quantités dépendent également du nombre de *gisements* ramassés.

Les figures 6.8 et 6.9 montrent que le modèle $(LOL)^{\otimes 38}$ a tendance à sous-estimer la probabilité d'occurrence de zéros par rapport au modèle $(\Delta\Gamma)^{\otimes 38}$. Ceci confirme les résultats précédemment obtenus par simulation indiquant que le modèle $(LOL)^{\otimes 38}$ (cf section 5.3.3) à tendance à sur-estimer le paramètre μ (et donc à sous-estimer $\delta{=}e^{-\mu}$). Contrairement au modèle LOL, le modèle $\Delta\Gamma$ propose un traitement séparé des zéros et non-zéros. C'est pourquoi il a tendance à fournir des estimations de la probabilité d'occurrence de zéros plus proches de la proportion de zéros observée que le modèle LOL (cf tableaux 6.6 et 6.7). Il en va de même pour les estimations de biomasse moyenne. Cela explique en partie pourquoi le DIC du modèle $(\Delta\Gamma)^{\otimes 38}$ est plus faible que celui du modèle $(LOL)^{\otimes 38}$ pour les deux espèces étudiées.

Dans le cas des oursins, la figure 6.8 montre que les capacités d'adéquation aux données des versions hiérarchiques régionalisées et des versions spatialisées et régionalisées sont globalement similaires entre les modèles LOL et $\Delta\Gamma$ et ce, malgré des différences très significatives en terme de DIC. Parmi les différences d'ajustements notables, nous pouvons remarquer que les modèles basés sur LOL ont tendance à fournir des estimations plus précises sur la biomasse moyenne. Ainsi, les distributions *a posteriori* sur Q sont moins dispersées comme l'indiquent les écarts-types *a posteriori* plus faibles du tableau 6.6 ainsi que les figures 6.5 et 6.6. En revanche, les variantes basées sur LOL ont tendance à sur-estimer les probabilités d'occurrence de zéros dans les strates pour lesquelles la biomasse moyenne observée est élevée. Cette sur-estimation peut en partie s'expliquer par l'existence d'une forte corrélation positive entre les paramètres μ et ρ (cf figure 6.7) qui force μ à prendre de faibles valeurs.

Dans le cas des anémones de mer, les versions hiérarchiques régionalisées $R_{\mu,\rho}$-LOL et spatialisées et régionalisées $BYM_{\mu,\rho}$-LOL ont des DIC (i.e., 509.95 et 499.97 respectivement) inférieurs à ceux des versions similaires basées sur le modèle $\Delta\Gamma$ (i.e., 512.28 et 508.40 respectivement). Toutefois, cette différence entre DIC semble s'expliquer par des différences mineures entre les capacités d'adéquation aux données respectives de ces modèles. En effet, la figure 6.9 montre que les médianes et écarts-types *a posteriori* relatifs à chaque unité géographique sont globalement identiques entre des versions LOL et $\Delta\Gamma$ similaires et ce, pour l'estimation des Q_i comme des δ_i (i=1,2,..,38). De même, le tableau 6.7 indique que les coûts et écarts-types *a posteriori* moyens sont très proches pour des versions LOL et $\Delta\Gamma$ similaires. La corrélation positive entre les paramètres μ et ρ des variantes basées sur LOL est globalement très faible pour chaque unité géographique (cf figure 6.7). Ainsi, la probabilité d'occurrence des zéros s'estime relativement indépendamment sous les variantes LOL. Cela explique pourquoi, par rapport au cas des oursins, les ajustements obtenus sont plus proches de ceux obtenus sous des versions hiérarchiques similaires basées sur $\Delta\Gamma$. Un examen plus approfondi des ajustements strate par strate montre que la seule différence importante concerne l'estimation de la probabilité d'occurrence de zéros dans une strate pour laquelle aucun zéro n'a été observé : la strate 425. Dans ce cas, la probabilité δ_{425} est largement sur-estimée dans les modèles basés sur $\Delta\Gamma$ par rapport aux observations et aux versions similaires basées sur le modèle LOL (e.g. 0.13 sous $R_{\delta,b}$-$\Delta\Gamma$ contre 0.07 sous $R_{\mu,\rho}$-LOL). Cela indique que le modèle $\Delta\Gamma$ a été plus sensible au "lien"

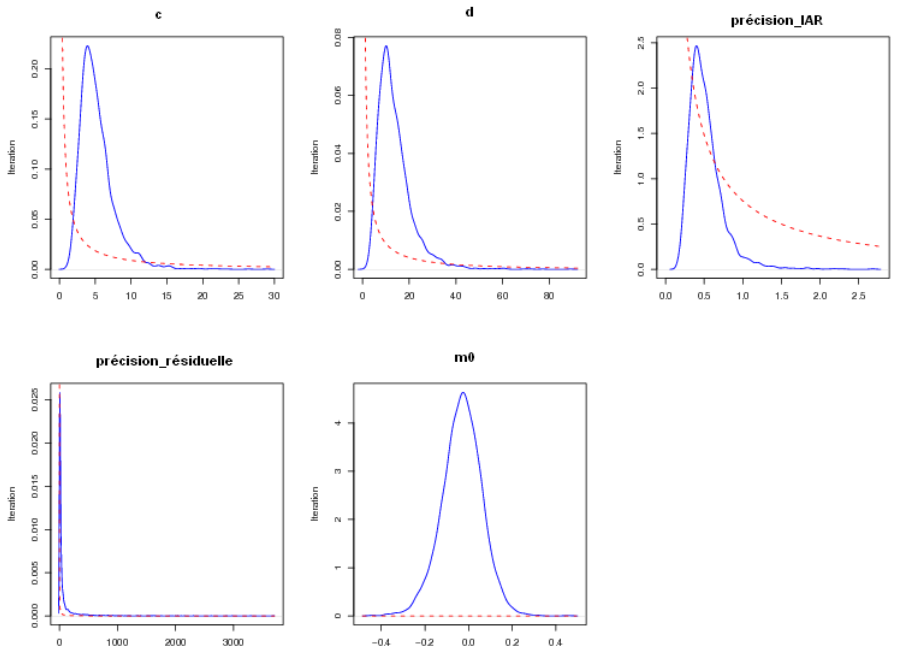


Figure 6.10 – Cas des oursins : Comparaison des distributions *a posteriori* (lignes bleues) aux distributions *a priori* (pointillés rouges) pour plusieurs paramètres du modèle $BYM_{\mu,\rho}$-LOL. Les lois *a priori* gamma sur les paramètres de précisions ont été multipliées par 10.

stochastique introduit entre les paramètres δ_i que le modèle LOL. Ainsi, les structures régionalisées et spatialisées sur les δ_i ont "poussé" δ_{425} vers la probabilité d'occurrence moyenne d'un zéro. Les modèles LOL n'imposent pas de traitement séparé des zéros et non-zéros. Aussi, ils ont utilisé l'information disponible dans les quantités de biomasse strictement positives pour estimer δ_{425} et induire un lissage moins important sur cette quantité.

Les capacités d'ajustement des versions hiérarchiques basées sur le modèle $\Delta\Gamma$ s'améliorent quand les coefficients de variation des lois gamma a_i (i=1,2,..,I) sont considérés indépendants ou liés par une distribution régionale commune. Cependant, les conclusions finales ne changent pas en terme de comparaison de modèles (facteurs de Bayes et DICs calculés mais non-montrés).

Comparaison des cinq variantes LOL

Comme nous pouvions s'y attendre, la version indépendante $(\text{LOL})^{\otimes 38}$ est sur-ajustée par rapport aux données observées. En particulier, les médianes et écarts-types *a posteriori* sur l'estimation de la biomasse moyenne sont beaucoup plus variables d'une unité géographique à l'autre par rapport à celles obtenues pour les autres structures hiérarchiques testées. En fait, la version indépendante du modèle LOL est très sensible aux valeurs extrêmes observées dans certaines unités géographiques. Prenons le cas des oursins pour lequel un certain nombre de valeurs extrêmes apparaissent dans les données observées. Le modèle $(\text{LOL})^{\otimes 38}$ va fournir des estimations plus proches des observations par rapport aux autres structures hiérarchiques proposées qui, elles, vont lisser les estimations par rapport à l'information contenue dans toutes les unités géographiques et ainsi réduire les biais induits par de possibles erreurs de mesure. Cela explique pourquoi le DIC as-

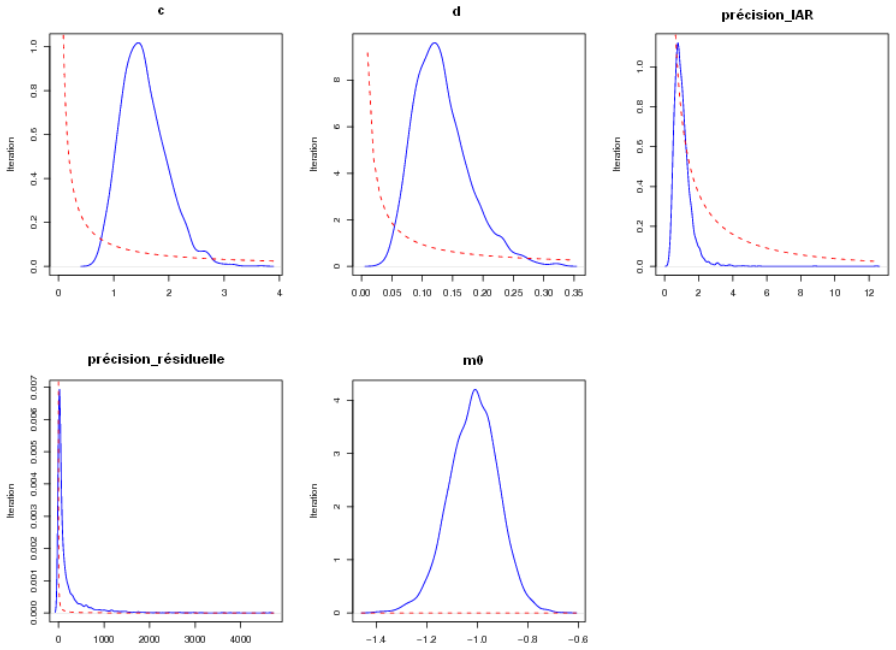


Figure 6.11 – Cas des anémones de mer : Comparaison des distributions *a posteriori* (lignes bleues) aux distributions *a priori* (pointillés rouges) pour plusieurs paramètres du modèle $BYM_{\mu,\rho}-$LOL. Les lois *a priori* gamma sur les paramètres de précisions ont été multipliées par 10.

socié à la variante indépendante est plus faible (cf tableau 6.4). En revanche, dans le cas des anémones de mer, les données contiennent très peu de valeurs extrêmes si bien que le modèle $(\text{LOL})^{\otimes 38}$ s'ajuste aux observations de façon similaire par rapport aux autres structures hiérarchiques proposées mais au prix d'un sur-paramétrage qui le pénalise en terme de DIC d'où un DIC plus élevé (cf tableau 6.5).

Les structures hiérarchiques R_μ-LOL et BYM_μ-LOL possèdent les plus mauvaises capacités d'ajustement aux données. Ainsi, nous pouvons constater que leur DIC sont nettement plus élevés que ceux obtenus sous les versions indépendantes ou les versions régionalisées et spatialisées respectivement associées ie., $\text{R}_{\mu,\rho}$-LOL et $\text{BYM}_{\mu,\rho}$-LOL. En particulier, nous pouvons observer une différence de DIC de près de 300 dans le cas des anémones de mer. De même, la crédibilité relative des versions partiellement régionalisées ou spatialisées par rapport à des versions similaires régionalisées ou spatialisées est nettement inférieure à 1 avec des facteurs de Bayes partiels calculés proches de 0.1 (résultats non montrés). Les tableaux 6.6 et 6.7 montrent que les performances d'estimation sur la probabilité d'occurrence de zéros δ sont globalement identiques entre des modèles pour lesquels $\rho_i = \widetilde{\rho}$ pour tout i=1,2,...,38 et des modèles pour lesquels ces quantités sont supposées différentes entre les unités géographiques. En revanche, les hypothèses $\rho_i = \widetilde{\rho}$ influent sur l'estimation des biomasses moyennes qui sont directement reliées aux paramètres ρ_i. Ainsi, les modèles R_μ-LOL et BYM_μ-LOL ont tendance à sur-estimer les biomasses moyennes estimées (synthétisées par $\widehat{\theta}_{mean}$) par rapport aux modèles pour lesquels les ρ_i ne sont pas supposés constants. Par ailleurs, ils conduisent à une forte réduction de la largeur des intervalles $[\widehat{\theta}_{Min},\widehat{\theta}_{Max}]$ par rapport aux versions indépendantes, régionalisées et spatialisées. Comme on pouvait s'y attendre, l'hypothèse $\rho_i = \widetilde{\rho}$ a conduit à un lissage des biomasses

moyennes estimées : les faibles valeurs ont tendance à être sur-estimées et les fortes valeurs sous-estimées. Cependant, ce lissage semble trop fort par rapport aux observations d'où de mauvaises capacités d'ajustement aux données de ce type de modèles. Ces résultats invalident l'hypothèse selon laquelle, en moyenne, les *gisements* d'espèces contiennent la même quantité de biomasse dans toutes les unités géographiques.

Dans le cas des oursins, les modèles $R_{\mu,\rho}$-LOL et $BYM_{\mu,\rho}$-LOL ont des capacités d'adéquation aux données similaires. Ainsi, les médianes, coûts et écarts-types *a posteriori* moyens répertoriés dans le tableau 6.6 sont identiques pour ces deux modèles. De même, les DIC calculés sont très proches avec une différence calculée de seulement 4.28. La structure spatialisée sur μ et régionalisée sur ρ a induit un lissage plus fort sur les estimations extrêmes comme l'indique la largeur réduite des intervalles $[\widehat{\theta}_{Min},\widehat{\theta}_{Max}]$. En particulier, nous pouvons nous attendre à ce qu'une telle structure "gomme" plus judicieusement les erreurs d'observation en fonction de l'information locale par rapport à une variante régionalisée. Dans le cas des anémones de mer, le modèle $BYM_{\mu,\rho}$-LOL semble posséder de meilleures capacités d'ajustement aux données que le modèle $R_{\mu,\rho}$-LOL. Ainsi, le DIC associé à $BYM_{\mu,\rho}$-LOL est de 499.97 contre 508.40 pour le modèle $R_{\mu,\rho}$-LOL. Le tableau 6.7 semble confirmer cette comparaison. En effet, les écarts-types *a posteriori* sur l'estimation de la biomasse moyenne sont plus élevés sous le modèle $R_{\mu,\rho}$-LOL. De plus, les estimations de biomasse moyennes sont moins proches des valeurs observées sous ce même modèle par rapport au modèle $BYM_{\mu,\rho}$-LOL. Ainsi, une prise en compte des liens de similarité entre unités géographiques voisines semble avoir permis d'améliorer la fiabilité et les précisions d'estimation sur Q et δ.

Les figures 6.10 et 6.11 permettent de comparer, au vu des données de biomasse respectives en oursins et en anémones de mer, les distributions *a posteriori* aux distributions *a priori* associées à plusieurs paramètres du modèle $BYM_{\mu,\rho}$-LOL. Nous pouvons constater que ces données de biomasse ont permis de réduire clairement l'incertitude *a priori* sur les paramètres du modèle $BYM_{\mu,\rho}$-LOL. Les précisions résiduelles sont les paramètres pour lesquels le modèle a le moins appris au vu des données. En effet, toute la variabilité des données de biomasse a été capturée par la structure IAR si bien que le signal résiduel était trop faible pour apprendre sur la variabilité résiduelle de ces données.

Enfin, la table 6.8 contient les moyennes *a posteriori* et intervalles de crédibilité à 95% estimés pour les paramètres du modèle $BYM_{\mu,\rho}$-LOL à partir des données en oursins et en anémones de mer. Dans le cas des oursins, on s'attend à ramasser en moyenne un *gisement* par trait de chalut standard quelle que soit la strate d'échantillonnage. En revanche, dans le cas des anémones, on s'attend à ne ramasser aucun *gisement* en moyenne. La part de variabilité expliquée par le modèle spatial IAR est mesurée par le ratio κ de la variance du modèle IAR sur la variabilité globale. Les fortes valeurs estimées pour κ indiquent l'existence d'une forte autocorrélation spatiale entre strates voisines pour le nombre moyen de *gisements* collectés. En effet, pour les deux espèces, la structure spatiale IAR capture en moyenne plus de 93% de l'hétérogénéité globale inter-strates. Cela explique pourquoi la variance résiduelle s^2_ϵ est plus petite que la variance de la structure IAR.

Par comparaison au modèle $R_{\mu,\rho}$-LOL, le modèle $BYM_{\mu,\rho}$-LOL a l'avantage de permettre de quantifier statistiquement la part de variabilité qui est spatialement structurée quand on s'intéresse au nombre de *gisements* susceptibles d'être ramassés pour une espèce donnée. Ses bonnes capacités d'adéquation aux données ainsi que les fortes valeurs estimées pour le paramètre de variance locale σ^2_{IAR} et le paramètre κ permettent de valider l'hypothèse selon laquelle les oursins et les anémones de mer ont une répartition géographique

Espèce	Paramètres	Post-mean	$IC_{95\%}$
Oursins	s^2_{IAR}	2.23	[0.90,4.30]
	s^2_ϵ	0.12	[0.002, 0.53]
	m_0	-0.03	[-0.22, 0.15]
	κ	0.93	[0.68, 0.99]
Anémones	s^2_{IAR}	1.15	[0.45,2.36]
	s^2_ϵ	0.05	[0.001, 0.29]
	m_0	-1.015	[-1.21, -0.83]
	κ	0.94	[0.65, 0.99]

Tableau 6.8 – Moyenne *a posteriori* et intervalles de crédibilité à 95% des paramètres du modèle BYM$_{\mu,\rho}$-LOL. κ indiques la part de variabilité expliquée par la structure IAR.

spatialement structurée à échelle des unités géographiques considérées.

6.3.4 Comparaison des capacités prédictives des modèles en compétition

Les échantillons *a posteriori* obtenus lors des précédents ajustements ont été utilisés pour générer, sous chacun des dix modèles, des échantillons prédictifs de quantités de biomasses. Dans chaque strate j (1,2,..,38), les "nouvelles" quantités de biomasses générées peuvent être vues commes des "réplicats" des quantités observées en 1999, 2000 et 2001. Elles peuvent être utilisées comme prédicteurs des quantités de biomasse présentes en des sites non-observés ou comme prédicteurs des quantités de biomasse susceptibles d'être ramassées lors de futures campagnes scientifiques si on néglige toute dynamique temporelle d'évolution de la biomasse.

Des échantillons ont été générés pour deux statistiques d'intérêt à partir des échantillons prédictifs précédents. Dans chaque strate j (j=1,2,..,38), j'ai calculé la biomasse moyenne prédite et la proportion de zéros prédite. Les tableaux 6.9 et 6.10 contiennent, pour chacun des 10 modèles hiérarchiques en compétition, les valeurs du critère de perte prédictive *a posteriori* D_k ainsi que celles de ces deux composantes P, relatif à la précision des prédictions, et G, indicateur de la fidélité des prédictions par rapport aux observations (cf Annexe G). J'ai calculé D_k pour trois valeurs du paramètre k : 1, 3 et $+\infty$. Ces valeurs reviennent à pondérer respectivement par 0.5, 0.75 et 1 la part du critère qui quantifie la fidélité des prédictions par rapport aux observations.

Notons tout d'abord que le classement des dix modèles hiérarchiques en compétition reste globalement inchangé quand k varie.

Les quantités moyennes de biomasse susceptibles d'être collectées dans chaque strate sont systématiquement mieux prédites avec le modèle LOL. Les prédictions sont nettement plus fidèles et plus précises aux observations (i.e., G et P plus petits). La figure 6.12 vient confirmer ces résultats.

Pour les proportions de zéros, les différences entre les capacités prédictives des versions LOL et celles des versions $\Delta\Gamma$ sont généralement moins significatives. Cependant, nous pouvons remarquer que le modèle LOL a tendance à mieux prédire la proportion de zéros dans le cas de distributions très piquées en zéro. Ainsi, dans le cas des anémones de mer, le critère D lié à la proportion de zéros est inférieur aux versions $\Delta\Gamma$ pour trois modèles sur cinq : les modèles R_μ-LOL, BYM$_\mu$-LOL et BYM$_{\mu,\rho}$-LOL.

Les versions indépendantes $(LOL)^{\otimes 38}$ et $\Delta\Gamma^{\otimes 38}$ sont sur-ajustées ce qui explique pour-

Statistique	Modele	G	P	D_1	D_3	$D_{+\infty}$
Biomasse moyenne	$(LOL)^{\otimes 38}$	6.04	254.59	257.61	259.12	260.63
	R_μ-LOL	63.64	185.47	217.28	233.19	249.10
	$R_{\mu,\rho}$-LOL	16.64	233.04	241.36	245.52	249.68
	BYM_μ-LOL	64.64	187.59	219.91	236.07	252.23
	$BYM_{\mu,\rho}$-LOL	19.06	225.92	235.45	240.22	244.98
	$\Delta\Gamma^{\otimes 38}$	7.45	481.69	485.42	487.28	489.14
	R_δ-$\Delta\Gamma$	633.11	327.33	643.88	802.16	960.44
	$R_{\delta,b}$-$\Delta\Gamma$	27.37	963.44	977.13	983.97	990.81
	BYM_δ-$\Delta\Gamma$	642.97	321.74	643.23	803.97	964.71
	$BYM_{\delta,b}$-$\Delta\Gamma$	27.89	542.13	556.08	563.05	570.02
Proportion de zéros	$(LOL)^{\otimes 38}$	0.11	0.72	0.78	0.80	0.83
	R_μ-LOL	0.33	0.74	0.90	0.98	1.07
	$R_{\mu,\rho}$-LOL	0.19	0.76	0.85	0.90	0.95
	BYM_μ-LOL	0.32	0.71	0.88	0.96	1.04
	$BYM_{\mu,\rho}$-LOL	0.20	0.74	0.84	0.89	0.94
	$\Delta\Gamma^{\otimes 38}$	0.10	0.80	0.85	0.88	0.90
	R_δ-$\Delta\Gamma$	0.03	0.73	0.74	0.75	0.76
	$R_{\delta,b}$-$\Delta\Gamma$	0.03	0.73	0.74	0.75	0.76
	BYM_δ-$\Delta\Gamma$	0.04	0.67	0.69	0.70	0.71
	$BYM_{\delta,b}$-$\Delta\Gamma$	0.04	0.68	0.70	0.71	0.72

Tableau 6.9 – Cas des oursins : Critères de perte prédictive *a posteriori* de Gelfand et Ghosh (1998b) calculés pour différentes valeurs de k

Statistique	Modele	G	P	D_1	D_3	$D_{+\infty}$
Biomasse moyenne	$(LOL)^{\otimes 38}$	0.05	6.51	6.54	6.55	6.56
	R_μ-LOL	6.75	1.64	5.01	6.70	8.39
	$R_{\mu,\rho}$-LOL	2.24	8.26	9.38	9.94	10.51
	BYM_μ-LOL	5.89	1.76	4.70	6.17	7.65
	$BYM_{\mu,\rho}$-LOL	0.66	7.27	7.60	7.77	7.93
	$\Delta\Gamma^{\otimes 38}$	0.30	7.55	7.70	7.77	7.84
	R_δ-$\Delta\Gamma$	67.63	7.22	41.04	57.94	74.85
	$R_{\delta,b}$-$\Delta\Gamma$	1.98	10.19	11.18	11.67	12.17
	BYM_δ-$\Delta\Gamma$	67.94	7.23	41.19	58.18	75.16
	$BYM_{\delta,b}$-$\Delta\Gamma$	2.38	13.87	15.06	15.66	16.25
Proportion de zéros	$(LOL)^{\otimes 38}$	0.17	1.08	1.17	1.21	1.26
	R_μ-LOL	0.07	0.90	0.94	0.95	0.97
	$R_{\mu,\rho}$-LOL	0.13	0.98	1.04	1.07	1.10
	BYM_μ-LOL	0.20	0.85	0.95	1.00	1.05
	$BYM_{\mu,\rho}$-LOL	0.23	0.90	1.01	1.07	1.12
	$\Delta\Gamma^{\otimes 38}$	0.06	1.02	1.05	1.07	1.08
	R_δ-$\Delta\Gamma$	0.14	0.97	1.04	1.07	1.11
	$R_{\delta,b}$-$\Delta\Gamma$	0.13	0.98	1.05	1.08	1.11
	BYM_δ-$\Delta\Gamma$	0.25	0.91	1.03	1.10	1.16
	$BYM_{\delta,b}$-$\Delta\Gamma$	0.25	0.92	1.04	1.10	1.17

Tableau 6.10 – Cas des anémones de mer : Critères de perte prédictive *a posteriori* de Gelfand et Ghosh (1998b) calculés pour différentes valeurs de k

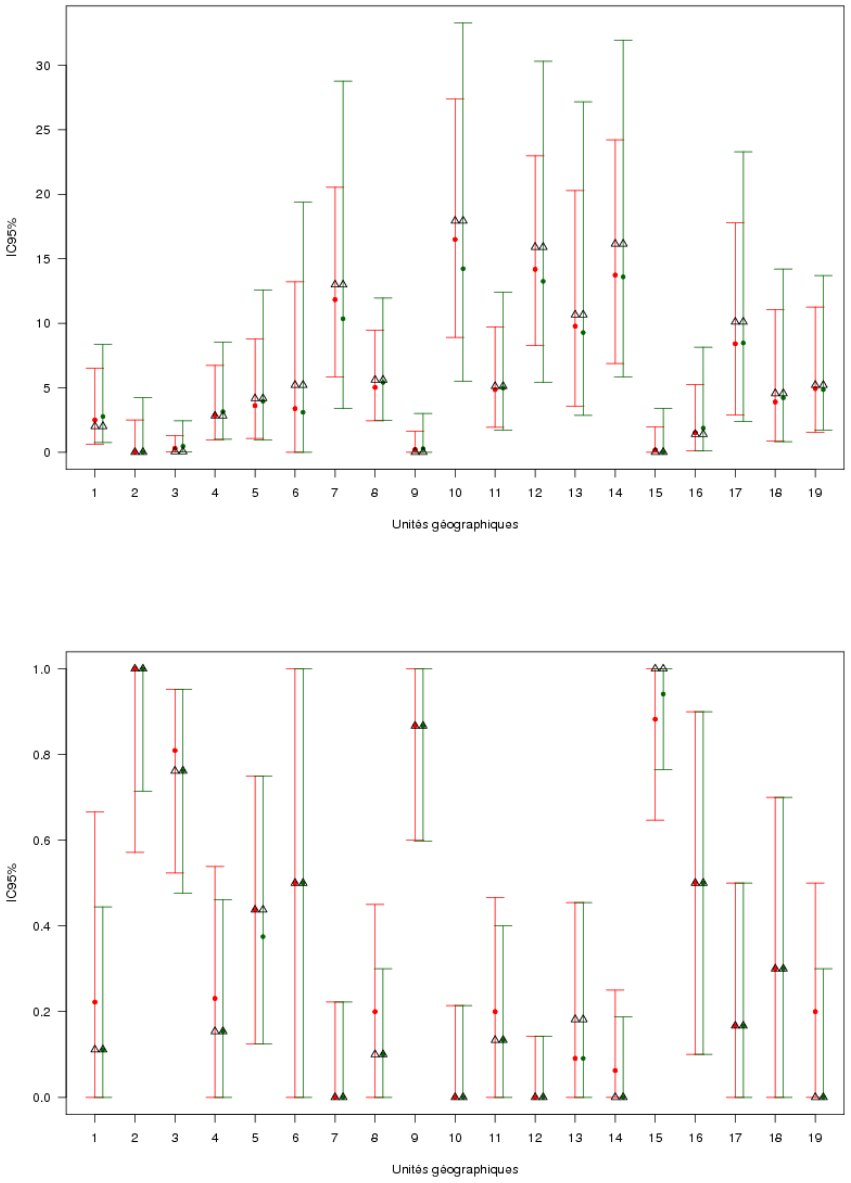


Figure 6.12 – Intervalles de crédibilité à 95% des distributions prédictives de la biomasse moyenne (gauche) et la probabilité d'occurrence d'un zéro (droite) dans 19 unités géographiques. Les triangles noirs désignent les valeurs observées. Les prédictions sous le modèle BYM$_{\mu,\rho}$-LOL (en rouge) sont plus précises que sous le modèle BYM$_{\delta,b}$-$\Delta\Gamma$ (en vert)

quoi elles fournissent généralement de mauvaises prédictions par rapport aux versions hiérarchiques régionalisées ou spatialisées. Les variances de prédiction associées (i.e., P) prennent souvent de fortes valeurs.

Les cartes contenues dans les figures 6.13 et 6.14 indiquent les quantités de biomasse moyennes observées et prédites sous le modèle BYM$_{\mu,\rho}$-LOL pour les oursins et les anémones de mer respectivement. Elles indiquent également les probabilités observées et prédites de ramasser chacune de ces espèces lors d'un trait de chalut standard de 1.75 miles. Nous pouvons remarquer que les cartes des quantités observées et prédites se ressemblent indiquant une bonne capacité prédictive du modèle BYM$_{\mu,\rho}$-LOL. Comme vu précédemment, le modèle BYM$_{\mu,\rho}$-LOL a effectué un lissage local sur la probabilité d'occurrence de zéros et la biomasse moyenne si bien que les cartes prédites sont relativement plus lisses que les cartes observées. Les patterns de répartition prédits pour les deux espèces considérées sont distincts. Les oursins sont relativement présents sur l'ensemble du domaine d'échantillonnage. De grosses prises (plus de 17 kg par trait standard) sont estimées dans une strate située au nord ouest du domaine d'étude. Cette strate offre des conditions de vie très favorables à la survie des oursins de par le type de sédiments présents (substrat de graviers ou sable) et la profondeur de ses fonds marins (entre 50 et 100 mètres). En revanche, les strates situées au nord du Golfe, dans le chenal Laurentien, sont trop profondes (>200 mètres) pour être propices à l'existence des oursins. Les anémones de mer sont essentiellement présentes dans les strates profondes du chenal Laurentien. En fait, le relevé du Centre des Pêches du Golfe capture 3 espèces d'anémones de mer qui ont des préférences d'habitat différentes. L'espèce la plus grosse mais aussi la plus abondante se trouve en eaux profondes (>200 mètres) dans le chenal Laurentien. Dans cette zone, le substrat est surtout du pélite.

A noter que les capacités prédictives des versions hiérarchiques basées sur le modèle $\Delta\Gamma$ se détériorent quand les coefficients de variation des lois gamma a_i (i=1,2,..,I) sont considérés indépendants ou liés par une distribution régionale commune. Ainsi, les prédictions de biomasse moyenne restent meilleures sous le modèle LOL que sous le modèle $\Delta\Gamma$ (D_k calculés mais non-montrés).

6.3.5 Estimation de l'impact de covariables environnementales sur la répartition géographique des espèces

Le modèle BYM$_{\mu,\rho}$-LOL suppose que le nombre moyen de *gisements* susceptibles d'être collectés en un site d'échantillonnage k dépend uniquement de l'unité géographique dans laquelle est localisé ce site. Or, en pratique, on s'attend à ce que ce nombre de *gisements* dépende également des caractéristiques environnementales propres au site d'échantillonnage. Du point de vue statistique, on s'attend à ce que, dans ce cas, la prise en compte de covariables environnementales permettent d'améliorer les performances d'ajustement et de prédiction du modèle BYM$_{\mu,\rho}$-LOL. Dans cette optique, le modèle spatial BYM, décrit dans le paragraphe 6.1.2, a l'avantage de faciliter l'insertion de covariables environnementales. Il suffit d'inclure les covariables sous la forme :

$$log(\mu_{ik}) = m_0 + \sum_{q=1}^{Q} \beta_q X_{qik} + \Phi_i + \epsilon_i \tag{6.3.3}$$

où

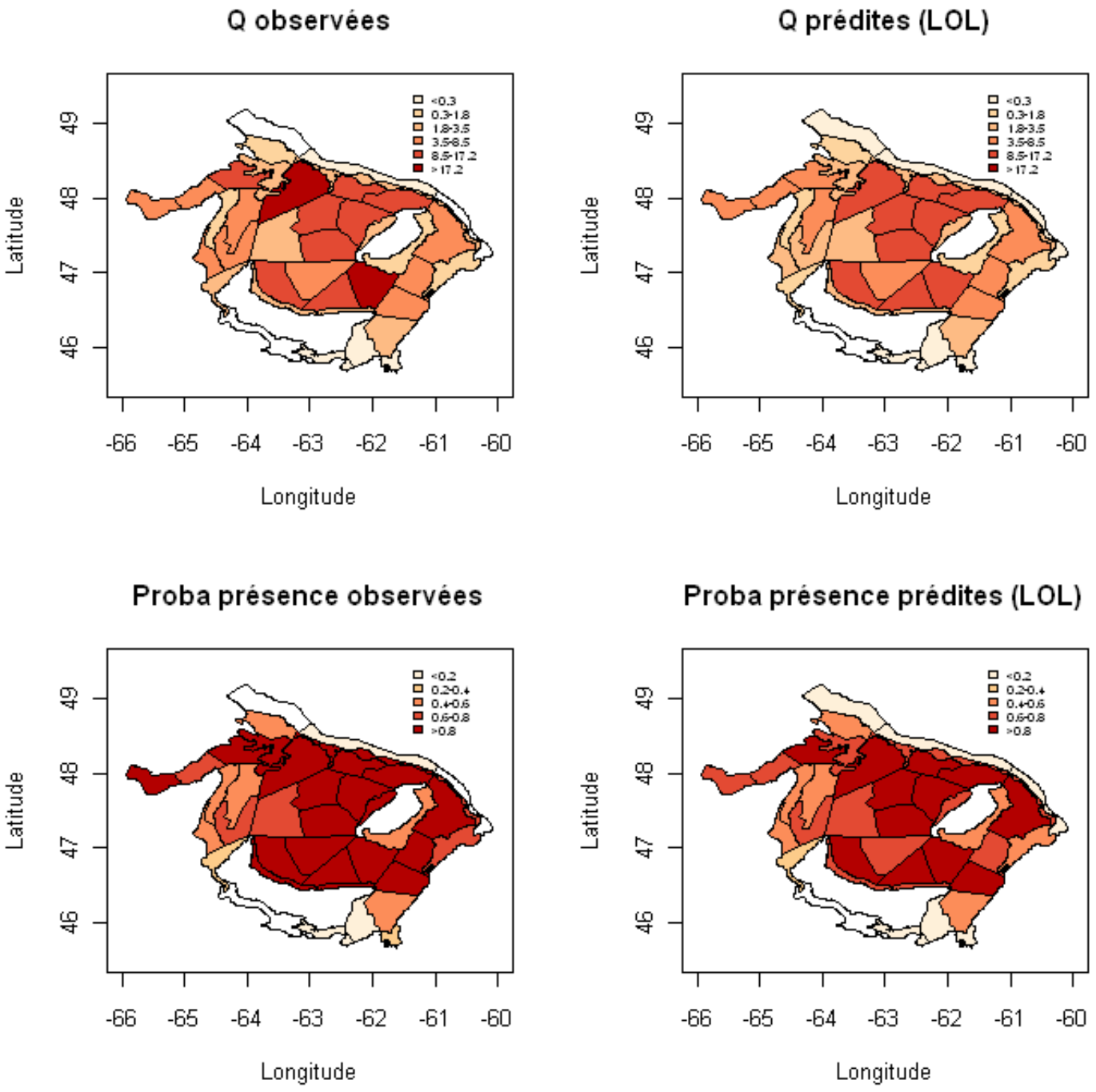


Figure 6.13 – Cas des oursins : médianes *a posteriori* observées et prédites sous le modèle $\text{BYM}_{\mu,\rho}$-LOL pour la biomasse moyenne Q (à gauche) et la probabilité de ramasser l'espèce d'intérêt $1-\delta$ (à droite) sur une distance chalutée fixée de 1.75 miles

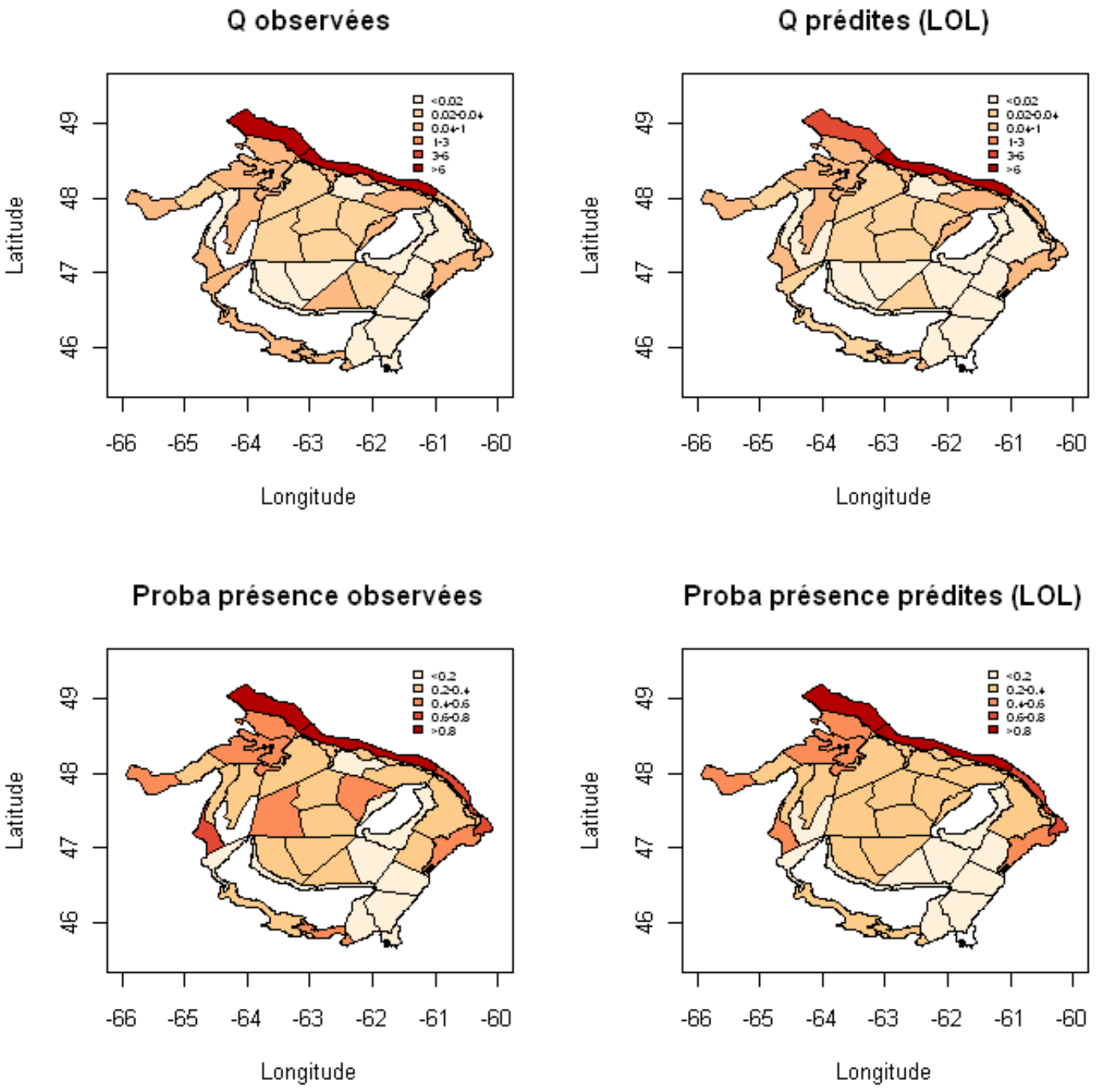


Figure 6.14 – Cas des anémones de mer : médianes *a posteriori* observées et prédites sous le modèle $\text{BYM}_{\mu,\rho}$-LOL pour la biomasse moyenne Q (à gauche) et la probabilité de ramasser l'espèce d'intérêt $1-\delta$ sur une distance chalutée fixée de 1.75 miles (à droite)

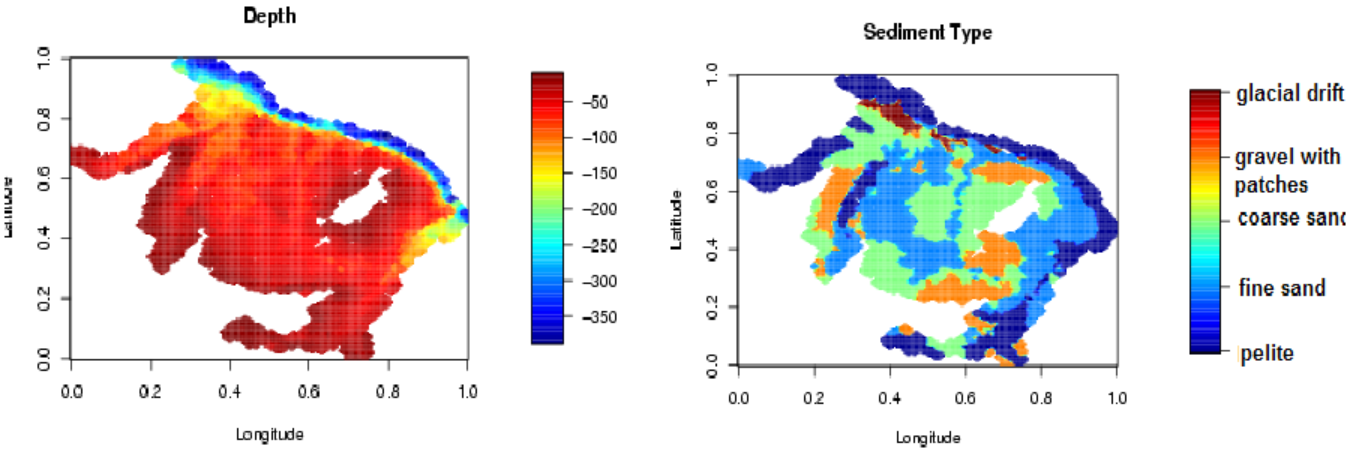


Figure 6.15 – Cartes de la profondeur (en mètres) des fonds marins et des types de sédiments dans le Sud du Golfe-du-Saint-Laurent

- μ_{ik} désigne le nombre moyen de *gisements* susceptibles d'être collectés au site k situé dans l'unité géographique i
- X_{qik} représente la valeur d'une covariable X_q relative au site k situé dans l'unité i
- β_q est le coefficient de régression associé à la q-ième covariable ($q = 1, 2, \ldots, Q$).

La variabilité résiduelle du nombre moyen de *gisements* collectés au site k, c'est-à-dire non expliquée par les covariables, est de nouveau expliquée par deux effets aléatoires spécifiques à l'unité géographique i, Φ_i et ϵ_i. Dans la régression 6.3.3, $exp(\beta_q)$ représente l'effet de la q-ième covariable sur le nombre moyen de *gisements* collectés quand celle-ci augmente d'1 unité. En effet, le nombre moyen de *gisements* susceptibles d'être collectés lors d'un trait standard est alors multiplié par $exp(\beta_q)$.

J'ai appliqué ce modèle aux données d'abondance en oursins et en anémones de mer en considérant les deux covariables explicatives suivantes :

- une covariable continue : la profondeur du site, mesurée lors de chaque trait de chalut. Elle prend ses valeurs entre 17 et 351 mètres.
- une covariable catégorielle : le type de sédiments, interpolé en chaque site à partir d'une carte géologique du domaine d'étude réalisée en 1973 (Loring & Nota, 1973). 5 classes de sédiments sont considérés (pélite, sable fin, sable dur, graviers, dérive glaciaire).

La distribution de ces deux covariables dans le Sud du Golfe-du-Saint-Laurent est indiquée dans la figure 6.15. Nous pouvons remarquer que ces deux covariables sont spatialement structurées. Nous pouvons donc nous attendre à ce qu'elles expliquent en partie la variabilité spatiale de la distribution des oursins et des anémones de mer.

Les mêmes lois *a priori* que celles du modèle BYM$_{\mu,\rho}$-LOL ont été attribuées aux paramètres m_0, s^2_{IAR} et s^2_ϵ (cf tableau 6.2). Par ailleurs, une loi *a priori* normale plate a été attribuée aux coefficients de régression b_q. Une contrainte d'identifiabilité a été posée pour permettre l'estimation de l'effet des sédiments. J'ai considéré la pélite comme le sédiment de référence à partir duquel des effets relatifs ont été estimés (i.e, β_1=0 où β_1 désigne l'effet pélite). J'ai effectué l'inférence bayésienne puis des prédictions selon les deux modèles suivants :

- le BYM$_{\mu,\rho}$-LOL incluant l'effet de la profondeur des sites. Par la suite, j'appelerai

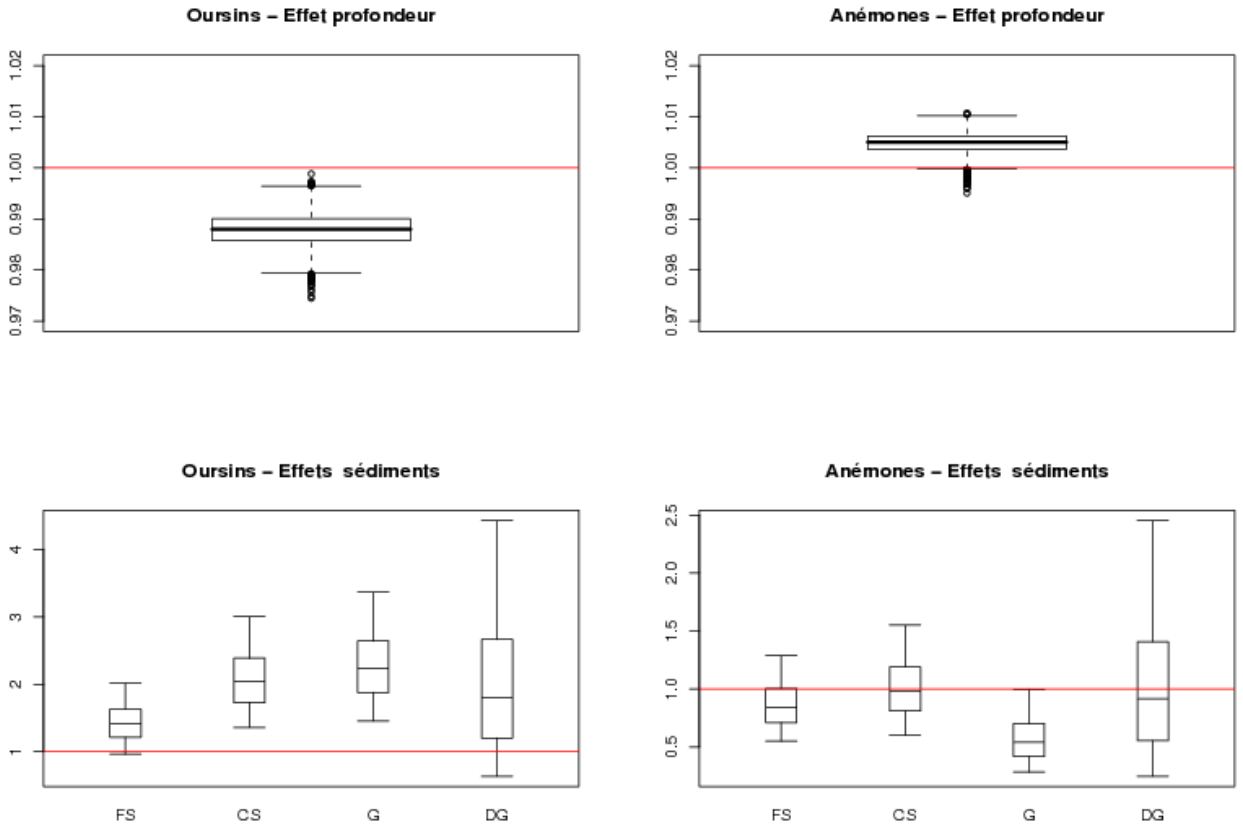


Figure 6.16 – Intervalles de crédibilité à 95% pour l'effet ($exp(b_q)$) de la profondeur et l'effet relatif de chaque type de sédiments (effet de base : pélite) sur le nombre de *gisements* collectés en oursins (gauche) et en anémones de mer (droite)

ce modèle BYM$_{\mu,\rho}$-LOL-P.

– BYM$_{\mu,\rho}$-LOL incluant l'effet de la profondeur et le type de sédiments en chaque site. Par la suite, j'appelerai ce modèle BYM$_{\mu,\rho}$-LOL-PS.

La figure 6.16 indique les distributions *a posteriori* obtenues pour l'estimation de l'effet de la profondeur et des effets relatifs de chaque type de sédiments (effet de base : pélite). Ces effets sont indiqués à l'échelle exponentielle. Dans le cas des oursins, la profondeur a un léger effet négatif. Ainsi, l'intervalle de crédibilité à 95% associé est strictement inférieur à 1 (i.e., $IC_{90\%} = [0.98, 0.99]$). En moyenne, si la profondeur augmente d'un mètre, l'intensité du nombre de *gisements* diminue : elle est multipliée par 0.99. Cela est cohérent avec les cartes de probabilité de présence observées (cf figure 6.13). En effet, les strates profondes situées au Nord du Chenal Laurentien sont peu propices à l'existence d'oursins. Dans le cas des anémones de mer, la profondeur a au contraire un très léger effet positif avec un intervalle de crédibilité à 95% supérieur à 1 ($IC_{95\%} = [1.001, 1.01]$) . En moyenne, si la profondeur augmente d'un mètre, l'intensité du nombre de *gisements* augmente : elle est multipliée par 1.005. Une fois encore, ce résultat est cohérent avec les cartes observées de répartition de l'espèce. Concernant l'effet des sédiments, la présence de graviers ou de sable dur a un effet relatif significativement positif sur le nombre moyen de *gisements* en oursins. Ainsi, les intervalles de crédibilité à 95% associés à ces deux types de sédiments sont strictement positifs. En moyenne, le nombre de *gisements* d'oursins que

	Oursins		Anémones de mer	
	Post-mean	$IC_{95\%}$	Post-mean	$IC_{95\%}$
Profondeur	0.99	[0.98,0.99]	1.005	[1.001,1.01]
Pélite (P)	1		1	
Sable fin (FS)	1.44	[0.96,2.02]	0.87	[0.55,1.29]
Sable dur (CS)	2.09	[1.36,3.00]	1.02	[0.60,1.55]
Graviers (G)	2.30	[1.46,3.37]	0.58	[0.28,0.99]
Dérive glaciaire (DG)	2.09	[0.64,4.43]	1.08	[0.25,2.45]

Tableau 6.11 – Moyennes *a posteriori* et intervalles de crédibilité à 95% pour l'estimation de l'effet de la profondeur et de l'effet relatif de chaque type de sédiments (effet de base : pélite) sur le nombre de *gisements* collectés en oursins et en anémones de mer. Les effets sont estimés à l'échelle exponentielle sous le modèle BYM$_{\mu,\rho}$-LOL-PS.

l'on peut espérer ramasser est deux fois plus élevé sur des graviers ou du sable dur que sur du pélite. En revanche, la présence de sable fin ou de dérive glaciaire n'a pas d'effet significatif sur le nombre de *gisements* en oursins. Concernant les anémones de mer, seule la présence de graviers a un effet significativement négatif sur le nombre de *gisements* ramassés ($IC_{95\%} = [0.28, 0.99]$). En moyenne, le nombre de *gisements* en anémones de mer est deux fois plus faible sur des graviers que sur du pélite.

Les cartes de la figure 6.17 indiquent la répartition géographique prédite de la biomasse moyenne et de la probabilité de ramasser au moins un *gisement* une fois l'effet des covariables profondeurs et types de sédiment enlevé. Nous pouvons constater que des lissages apparaissent dans certaines strates par rapport aux cartes obtenues avec le modèle sans covariable $BYM_{\mu,\rho}$-LOL.

La table 6.13 contient les DIC ainsi que les composantes du critère prédictif de Gelfand et Ghosh (1998b) pour les modèles BYM$_{\mu,\rho}$-LOL, BYM$_{\mu,\rho}$-LOL-P et BYM$_{\mu,\rho}$-LOL-PS. Nous pouvons constater que la prise en compte des covariables n'a pas le même effet sur les capacités d'ajustement du modèle BYM$_{\mu,\rho}$-LOL selon les espèces. Dans le cas des oursins, la prise en compte de la profondeur des sites d'échantillonnage améliore nettement les capacités d'ajustement : le DIC du modèle BYM$_{\mu,\rho}$-LOL est réduit de 26 unités. De même, la prise en compte des sédiments améliore légèrement les capacités d'ajustement du modèle BYM$_{\mu,\rho}$-LOL-P. En revanche, dans le cas des anémones de mer, la prise en compte des covariables profondeur et sédiments ne permet pas d'améliorer les capacités d'ajustement du modèle BYM$_{\mu,\rho}$-LOL. En effet, les différences entre DIC sont faibles ce qui signifie que les modèles BYM$_{\mu,\rho}$-LOL-P et BYM$_{\mu,\rho}$-LOL-PS ont des capacités d'ajustement comparables au modèle BYM$_{\mu,\rho}$-LOL. Le tableau 6.12 montre que les covariables profondeur et sédiments permettent d'expliquer en partie la variabilité spatiale locale de la répartition de la biomasse. En effet, nous pouvons constater que la moyenne *a posteriori* du paramètre s^2_{IAR} diminue avec l'ajout successif de ces deux covariables. Par ailleurs, ce paramètre est estimé plus précisément puisque la largeur des intervalles de crédibilité à 95% diminue. Ces résultats indiquent que la structuration spatiale de la distribution des oursins et des anémones de mer dans le sud du Golfe du Saint-Laurent s'explique en partie par la structuration spatiale des covariables environnementales (cf figure 6.15).

Du point de vue prédictif, la prise en compte de la profondeur et des sédiments a plutôt tendance à améliorer la fidélité des prédictions par rapport aux observations (i.e., G diminue) mais à dégrader la précisions des prédictions (i.e., P augmente). De ce fait, les

Espèce	Paramètre	Modèle	Post-mean	$IC_{95\%}$
Oursins	s^2_{IAR}	BYM$_{\mu,\rho}$-LOL	2.23	[0.90,4.30]
		BYM$_{\mu,\rho}$-LOL-P	1.67	[0.68, 3.2]
		BYM$_{\mu,\rho}$-LOL-PS	1.58	[0.56, 3.11]
	s^2_{ϵ}	BYM$_{\mu,\rho}$-LOL	0.12	[0.002,0.53]
		BYM$_{\mu,\rho}$-LOL-P	0.10	[0.001, 0.44]
		BYM$_{\mu,\rho}$-LOL-PS	0.10	[0.001,0.47]
Anémones	s^2_{IAR}	BYM$_{\mu,\rho}$-LOL	1.15	[0.45,2.36]
		BYM$_{\mu,\rho}$-LOL-P	0.42	[0.08, 1.28]
		BYM$_{\mu,\rho}$-LOL-PS	0.43	[0.08,1.24]
	s^2_{ϵ}	BYM$_{\mu,\rho}$-LOL	0.05	[0.001,0.29]
		BYM$_{\mu,\rho}$-LOL-P	0.03	[0.001, 0.17]
		BYM$_{\mu,\rho}$-LOL-PS	0.03	[0.001,0.19]

Tableau 6.12 – Moyennes *a posteriori* et intervalles de crédibilité à 95% pour l'estimation de la variabilité spatiale locale s^2_{IAR} et de la variabilité résiduelle non-spatialisée s^2_{ϵ}

Espèce	Modèle	DIC	G	P	D_1
Oursins	BYM$_{\mu,\rho}$-LOL	2563.92	19.06	225.92	235.45
	BYM$_{\mu,\rho}$-LOL-P	2537.51	18.74	219.24	228.61
	BYM$_{\mu,\rho}$-LOL-PS	2531.98	22.79	204.94	216.34
Anémones	BYM$_{\mu,\rho}$-LOL	499.97	0.66	7.27	7.60
	BYM$_{\mu,\rho}$-LOL-P	504.49	0.71	9.17	9.52
	BYM$_{\mu,\rho}$-LOL-PS	508.29	29.17	4.03	18.61

Tableau 6.13 – DIC et composantes du critère prédictif de Gelfand et Ghosh (1998b) pour les modèles BYM$_{\mu,\rho}$-LOL, BYM$_{\mu,\rho}$-LOL-P et BYM$_{\mu,\rho}$-LOL-PS.

capacités de prédiction sont généralement moins bonnes pour les modèles BYM$_{\mu,\rho}$-LOL-P et BYM$_{\mu,\rho}$-LOL-PS.

6.4 Conclusions et discussion

Dans ce chapitre, j'ai montré que la modélisation hiérarchique est adaptée à la spécification de modèles décrivant l'hétérogénéité d'une même variable, entre différentes unités spatiales. J'ai montré que, lorsqu'il existe une forme de cohérence, les modèles hiérarchiques incluant une structure spatiale ou régionale latente offrent une alternative plus performante, notamment du point de vue prédictif, à l'hypothèse d'indépendance mutuelle des unités géographiques. En effet, une telle hypothèse conduit à des modèles sur-ajustés dont les capacités de prédiction sont souvent très faibles.

J'ai comparé cinq versions hiérarchiques possibles basées sur le modèle LOL ou le modèle $\Delta\Gamma$. J'ai pu en déduire que le modèle LOL a globalement de meilleures capacités de prédiction que le modèle $\Delta\Gamma$. En particulier, les prédictions de biomasses moyennes sont plus précises et plus fidèles aux observations. En revanche, du point de vue des capacités d'adéquations aux données, les modèles basés sur LOL ont globalement des performances similaires par rapport aux modèles basés sur $\Delta\Gamma$. De légères améliorations semblent toutefois être apportées par les modèles basés sur LOL lorsque les données sont

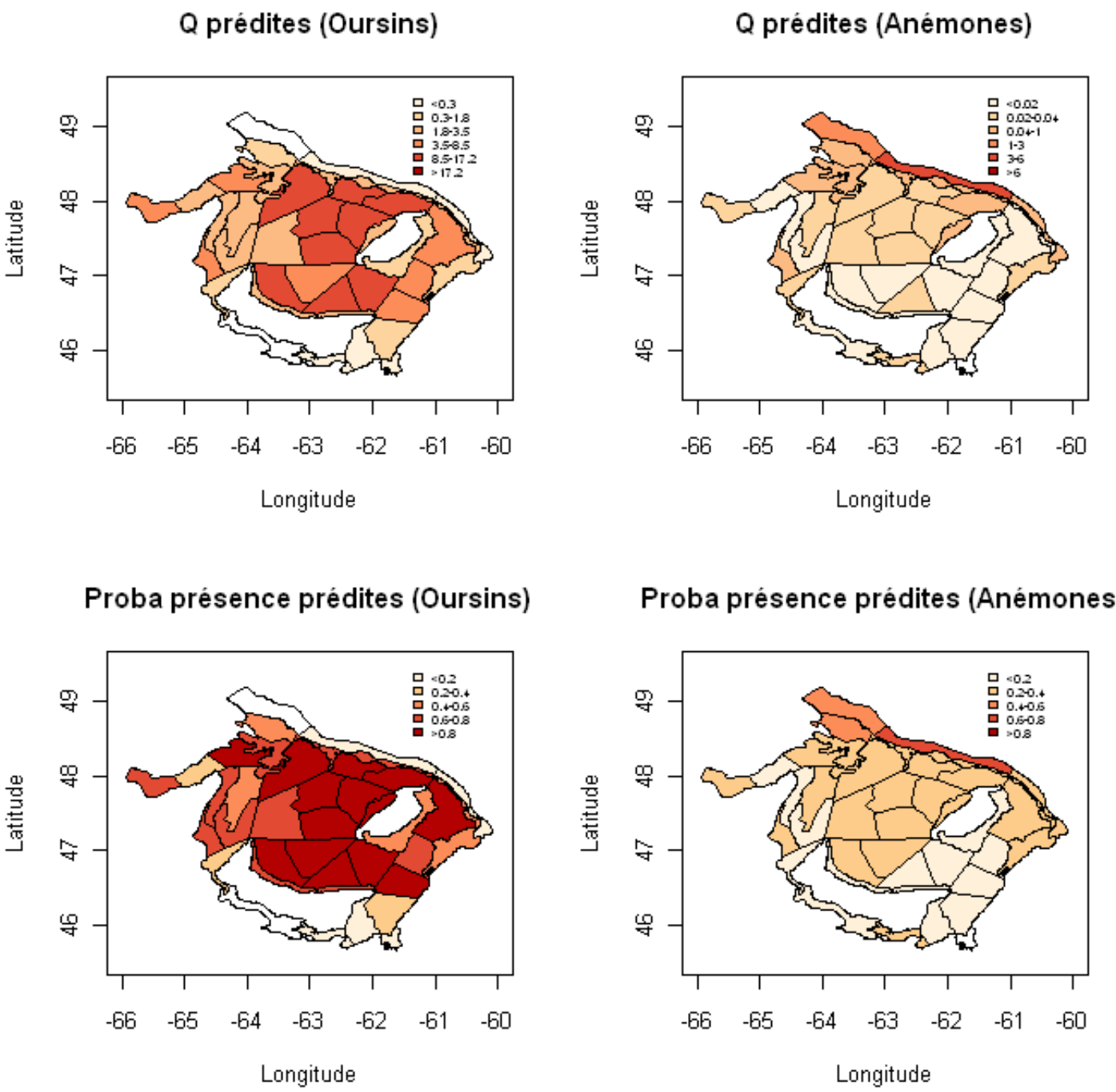


Figure 6.17 – Biomasses moyennes et probabilités de ramasser l'espèce, prédites sous le modèle $\text{BYM}_{\mu,\rho}$-LOL-PS pour les oursins (à gauche) et les anémones de mer (à droite)

caractérisées par une très forte occurrence de zéros et des biomasses observées faibles. Typiquement, cela est le cas du jeu de données sur les anémones de mer. Dans ce cas, la corrélation entre les paramètres μ et ρ est faible si bien que l'estimation des probabilités d'occurrence de zéros est similaire entre les modèles LOL et $\Delta\Gamma$. En revanche, le modèle LOL a tendance à mieux se comporter dans des unités géographiques particulières où l'occurrence de zéros est nulle et/ou les quantités de biomasse observées très élevées. Cela peut s'expliquer par le fait que le modèle LOL n'impose pas de traitement séparé entre zéros et non-zéros ce qui le rend un peu moins sensible aux liens stochastiques introduits entre unités géographiques pour l'estimation des δ_i (i=1,2,..,I).

La structure $BYM_{\mu,\rho}$-LOL apparaît comme la structure hiérarchique la mieux adaptée pour l'estimation et la prédiction des données de biomasse en oursins et en anémones de mer. En particulier, elle permet de quantifier statistiquement la part de variabilité spatialement structurée dans la distribution des organismes. Elle a permis de valider l'hypothèse selon laquelle les oursins et les anémones de mer ont une répartition spatialement structurée à échelle des unités géographiques définies dans le sud du Golfe du Saint-Laurent.

La prise en compte de covariables environnementales, facilitée par l'utilisation d'une structure BYM, a permis d'améliorer les capacités d'ajustement du modèle $BYM_{\mu,\rho}$-LOL et de quantifier les impacts respectifs de ces covariables sur la répartition des espèces. Les résultats obtenus indiquent que les oursins ont tendance à se répartir majoritairement dans les zones peu profondes avec des substrats de type gravillonneux ou sableux durs. Quant aux anémones de mer, elles se répartissent majoritairement dans les zones profondes et sont peu présentes sur les substrats gravillonneux (par rapport aux substrats de type pélite).

Les différentes structures régionalisées et BYM attribuées aux effets aléatoires b_i du modèle $\Delta\Gamma$ (cf section 6.2) ont également été testées sur la moyenne des quantités de biomasses strictement positives i.e., $\frac{a_i}{b_i}$ (i=1,2,..,I). Cependant, dans le cas des données de biomasse du Golfe-du-Saint-Laurent, les qualités d'ajustement et les qualités de prédiction des modèles sont systématiquement moins bonnes lorsque les structures régionalisées et BYM sont attribuées aux rapports $\frac{a_i}{b_i}$ (résultats non montrés). C'est pourquoi, afin de ne pas désavantager le modèle $\Delta\Gamma$, j'ai choisi de donner les résultats obtenus lorsque sont modélisés les effets aléatoires b_i.

Le choix de la structure de voisinage choisie pour le modèle IAR est critiquable. En effet, lorsque le domaine d'échantillonnage est constitué d'unités de taille et de forme très différentes -ce qui est notamment le cas pour le Golfe-du-Saint-Laurent- définir comme voisins les unités qui ont une frontière commune n'est pas toujours le choix le mieux adapté. Une autre possibilité pourrait être de définir le voisinage en fonction de la distance entre unités géographiques. L'avantage de cette structure est qu'elle est facilement adaptable pour des domaines irréguliers et discontinus.

L'utilisation d'un modèle CAR intrinsèque pour lier les unités géographiques est également critiquable. Ce choix est uniquement cohérent dans la mesure où on accepte de regarder l'espace comme une partition intangible en sous-domaines. Toutefois, à partir du moment où des partitions emboîtées du domaine sont considérées, ce choix induit une perte de cohérence distributionnelle des observations de biomasse à grande échelle. En effet, le modèle LOL assure la stabilité par addition des observations uniquement à l'intérieur d'un même domaine homogène. Une approche alternative serait de modéliser des dépendances non plus entre unités géographiques mais directement au niveau des observations afin de conserver une propriété de cohérence distributionnelle à grande échelle.

Ce chapitre met clairement en évidence la grande flexibilité offerte par la modélisation hiérarchique. De nombreuses constructions hiérarchiques peuvent être imaginées à partir d'un même modèle d'observations. Dans ce chapitre, nous proposons cinq versions hiérarchiques possibles basées sur le modèle LOL ou le modèle $\Delta\Gamma$. De nombreuses autres structures pourraient être testées pour représenter l'hétérogénéité inter-unités des effets aléatoires propres à chaque modèle. D'autres distributions régionales pourraient être attribuées. Par exemple, nous aurions pu poser :

$$log(\frac{\mu_i}{\rho_i}) \overset{i.i.d}{\sim} \mathcal{N}(m_0, \sigma_0^2)$$
$$\mu_i \overset{i.i.d}{\sim} \text{Gamma}(a, b)$$

ou encore introduire de la covariation entre les effets aléatoires μ_i et ρ_i (i=1,2,..,38) étant donné qu'ils semblent être liés par une corrélation positive (cf figure 5.8) :

$$log \begin{pmatrix} \mu_i \\ \rho_i \end{pmatrix} \overset{i.i.d}{\sim} \mathcal{N} \begin{pmatrix} \mu_0 \\ \rho_0 \end{pmatrix}, \Sigma_0$$

avec Σ_0 une matrice de variance-covariance non-diagonale. De même, dans le cas du modèle LOL, la structure BYM pourrait être utilisée pour représenter la variabilité des ρ_i ou des biomasses moyennes μ_i/ρ_i plutôt que celle des μ_i (i=1,2,..,I). De même, pour le modèle $\Delta\Gamma$, elle pourrait être utilisée pour représenter la variabilité des b_i ou des biomasses moyennes strictement positives a_i/b_i (i=1,2,..,I). Enfin, la flexibilité des structures hiérarchiques permet d'introduire facilement des covariables environnementales, susceptibles d'expliquer la répartition spatiale des espèces, dans la couche phénoménologique latente des modèles. Ainsi, nous avons pu quantifier les effets respectifs de la profondeur des fonds marins et du type de sédiments présents dans le Golfe du Saint-Laurent sur la distribution des oursins et des anémones de mer. Nous avons supposé que la profondeur a un effet linéaire sur l'intensité du nombre de *gisements*. Une alternative intéressante serait d'inclure un effet non-linéaire de la profondeur plutôt qu'un effet linéaire. En effet, cette hypothèse semble en théorie plus réaliste qu'un effet linéaire. Nous pouvons imaginer qu'une espèce soit de plus en plus présente jusqu'à un certain seuil de profondeur puis de plus en plus rare au-delà de ce seuil.

Les relations inter-espèces jouent un rôle non-négligeable sur la structure et le fonctionnement des écosystèmes. En particulier, elles peuvent expliquer la répartition spatiale de la biodiversité. Certaines espèces peuvent co-exister dans un même milieu ou au contraire vivre dans des milieux différents. Les écologues peuvent ainsi être interessés par les patterns spatiaux joints de plusieurs espèces. Jusqu'à présent, nous n'avons pas tenu compte de ces relations inter-espèces comme, par exemple, entre les oursins et les anémones de mer. Nous nous sommes limités à la modélisation de structures spatiales latentes univariées. Ainsi, tous les modèles proposés permettent uniquement de représenter la répartition spatiale de la biomasse espèce par espèce. Par ailleurs, les modèles construits supposent que la distribution des organismes vivants est spatialement structurée mais uniquement à large échelle, celles des unités géographiques. Or, les écologues savent que des processus biologiques comme la reproduction et la compétition peuvent induire des structures spatiales à échelle plus fine. Aussi, le chapitre suivant propose une version continue du modèle LOL, basée sur un champ géostatistique multivarié latent construit par convolution de noyaux de lissage.

6.5 Article 3 soumis à Environmental and Ecological Statistics

Environmental and Ecological Statistics manuscript No.
(will be inserted by the editor)

Modelling spatial zero-inflated continuous data with an exponentially compound Poisson process

Sophie Ancelet · Marie-Pierre Etienne · Hugues Benoît · Eric Parent

Received: date / Accepted: date

Abstract A parsimonious model is presented as an alternative to *delta* approaches to modelling zero-inflated continuous data. The data model relies on an exponentially compound Poisson process, also called the law of leaks (LOL). It represents the process of sampling resources that are spatially distributed as Poisson distributed patches, each containing a certain quantity of biomass drawn from an exponential distribution. In an application of the LOL, two latent structures are proposed to account for spatial dependencies between zero values at different scales within a hierarchical Bayesian framework. The LOL is compared to the *delta*-gamma distribution using bottom-trawl survey data. Results of this case study emphasize that the LOL not only provides better fits to learning samples but also offers better predictions of validation samples.

Keywords Bayes factor · *delta* models · hierarchical modelling · Intrinsic AutoRegressive spatial model · law of leaks · MCMC algorithms

1 Introduction

Abundance survey data (e.g., counts, biomass) typically contain a large proportion of zeros accompanied by a skewed distribution of the remaining values, including extremes. Martin et al. (2005) recently argued that the analysis of such zero-inflated data [8] should begin by distinguishing the sources of "excess zeros". Three general sources can affect the choice of a model. The most trivial is the inclusion in a data set of observations from areas where there are no chances of making a non-zero observation, such as areas outside the possible environmental range of a species. These cases are easily addressed by excluding them from analysis. Remaining true zero

Sophie Ancelet
AgroParisTech-INRA UMR518, Equipe MORSE,19 Avenue du Maine, 75732 Paris cedex 15
Tel.: +0033-145498928
Fax: +0033-145498827
E-mail: sophie.ancelet@orange.fr

Marie-Pierre Etienne
AgroParisTech-INRA UMR518, Equipe MORSE

Hugues Benoît
Fisheries and Oceans Canada, Gulf Fisheries Centre, P.O.Box5030, Moncton, New Brunswick

Eric Parent
AgroParisTech-INRA UMR518, Equipe MORSE

values can occur as a direct result of the effect under study (e.g., suitability of a given intertidal habitat) or as a stochastic result of sampling from areas of low density. On the other hand, false zeros can occur as a result of detection limits, observer effects (e.g., hiding behavior in the presence of observers) or a mistiming of observation (e.g., sampling a site that is normally occupied by a species, just not at the time of observation).

Zero-inflated data create many problems for statistical analysis. First, the sample mean computed from overdispersed data can be an imprecise indicator of stock abundance [16]. Second, designing realistic models for such data remains a challenge. For instance, the Poisson and the negative binomial distributions (if the data are discrete) or the lognormal and gamma distributions (in the continuous case) are common models proposed for marine surveys samples. But the spike at zero in the empirical histograms often contains many more zeros than would be expected from these standard distributions. Consequently, these common sense approaches often lead to poor fits [21] since the underlying distributional assumptions (e.g. variance-mean relationship for the Poisson distribution) often are violated.

Model properties are commonly exploited to define efficient indicators of stock abundance useful to monitor species over time and guide management decisions. Consequently, a fair amount of statistical effort has recently been devoted to dealing with zero-inflated data sets (Martin et al, 2005). Counts with extra zeros can be modelled with the well-known Zero-Inflated Poisson and Negative Binomial regression models (ZIP and ZINB respectively) [11] or with their respective Random-Effects version (RE-ZIP and RE-ZINB)[7]. In the continuous case, *Delta* models (i.e., conditional or two part-models) are routinely used by ecologists to analyze abundance data with many zero values. These models assume that the presence-absence of a species at a site and its abundance when present result from separate ecological mechanisms. Formally, these mixture models separately define the occurrence of a zero value as a Bernoulli random trial and the positive abundances using either a gamma [19] or a lognormal [1] random variable. They have the attractive advantage of an orthogonal parametrization, which makes them easy to fit and interpret. They offer a powerful framework to analyze and predict the spatial distribution of organisms as they can incorporate a parametric regression to relate the sources of zero observations and species abundance when present to environmental characteristics. However, the break between zero and non-zero values presents a particularly unnatural discontinuity in species abundance or density data, where many zeros are an important component of decreasing gradients. In addition, there is no change of parameters of *delta* models that coherently matches a change of sampling effort.

In this paper, we propose an alternative model for the analysis of continuous zero-inflated data. Its parsimonious stochastic structure is based on a mixture of distributions : a Poisson random sum of exponential variables. This compound Poisson process, originally coined *la loi des fuites* (*the law of leaks* - LOL) by Bernier and Fandeux (1970) represents the process of sampling resources that have a latent patchy spatial (or temporal) distribution. Abundance data result from an appealing representation of data collected from a hidden Poisson sampling process, ensuring spatial coherence with regards to a change of sampling effort. If covariates were known to explain the latent data model variability, the occurrence of zeros could then be interpreted as a natural endpoint of a progression from high to low intensities.

For a fair and realistic comparison, the LOL or the competing *delta* models were embedded as a data submodel within a general hierarchical model. Hierarchical Bayesian modelling distinguishes between the three different stages of a model's hierarchy : a data submodel describing how the data y are collected, a process submodel accounting the spatial covariations (or the dynamics) of the phenomenon through latent variables z, and a top structure that quantifies the partial knowledge about the unknown parameters θ. No temporal dependency was considered in

the process submodel, but a spatial effect representation (via a IAR structure) was compared to a random effect model.

We computed Bayes factors from a bottom-trawl survey dataset to compare the fitting abilities of competing models. Bayes factors are known to be sensitive to prior choice, and informative priors are recommended to get a meaningful comparison between models. We adapted the frequentist approach of split test sample analysis to easily check Bayes factor sensitivity to priors in our case study.

In this paper, we argue that :

1. The LOL conveniently accommodates the presence of a large number of zeros and a skewed distribution of non-zero values.
2. In our case study, the LOL fits and predicts the data much better than the commonly used *delta* models.
3. Thanks to Markov Chain Monte-Carlo inferential techniques (MCMC), only few additional computational costs are required when dealing with the two additional latent layers of the LOL (The LOL can itself be viewed as a hierarchical data model).
4. The introduction of a spatial structure on the top of the hierarchy is straightforward. It can improve the robustness of marine resource survey analyses which are frequently characterized by relatively small sample sizes due to the high cost of sampling at sea.

In the following section, we detail the available bottom-trawl survey dataset then, the conceptual and mathematical hypotheses underpinning the LOL model used to represent zero-inflated data. Two possible underlying spatial patterns of abundance data for the process submodel are proposed. In the third section, we briefly describe the Bayesian inference techniques used. In the fourth section, we propose an ingenious method to perform quickly a Bayes factor sensitivity analysis to priors. The fifth section compares the fitting and predictive abilities of the LOL model with a *delta* model using our case study. Finally, we discuss the advantages and limits of the LOL.

2 Hierarchical modelling of spatially structured zero-inflated data

2.1 Data description

Scientific bottom-trawl surveys have been conducted by Fisheries and Oceans Canada in the southern Gulf of St.Lawrence (sGSL) (NW Atlantic) each September since 1971 [9]. The main objective is to quantify the abundance and the distribution of various marine species.

The survey follows a stratified random design, with stratification based on depth and geographic area. Each year, the number of sampling sites chosen is generally proportional to the size of each stratum. The total number of sites sampled annually has varied from about 65-75 during the 1970s to 140-200 in all but one year since 1989.

The target fishing procedure is a 30-min straight-line tow at 3.5 knots (i.e., 3.21 km trawled distance). However, the actual distance trawled can vary because some tows are shortened to avoid tearing the net on rough bottom and because of variations in vessel speed resulting from prevailing winds and currents. Consequently, it is recorded after the catch as the difference between starting and ending positions of tow.

Collected species are identified, sorted and weighted (in kilogrammes per tow). While the survey was initially targetted at fish, data on the abundance of epibenthic invertebrates such as urchins, starfishes, whelks and anemones have also been collected since 1988.

Environmental covariables are measured at each sampling site : depth, water temperature. Moreover, the type of sediments is interpolated at each sampling site from a geological map of the

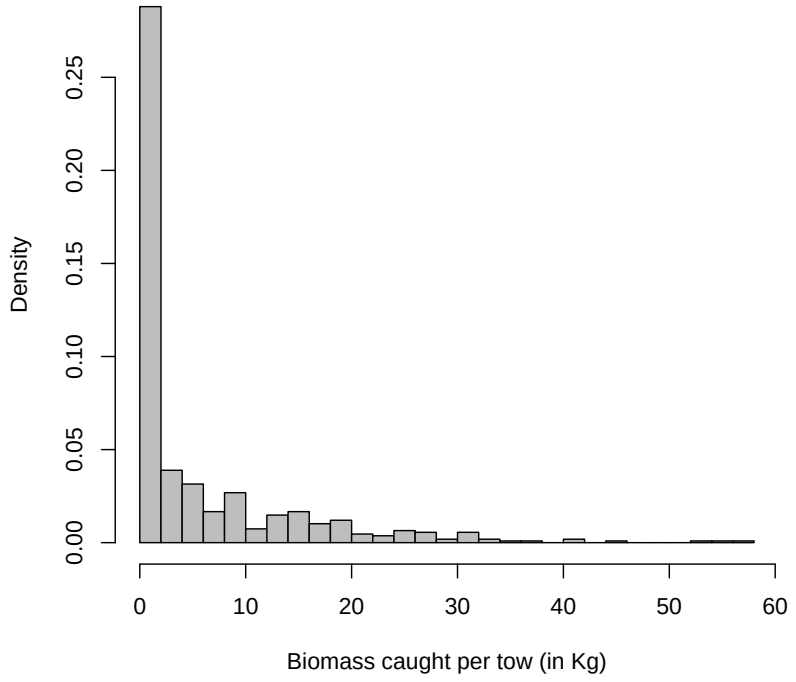


Fig. 1 An example of zero-inflated dataset: the biomass of sea urchins (kg/tow) from individual tows in the sGSL bottom-trawl survey, 1999-2000-2001

sGSL, made in 1973 [12]. These covariables will not be considered in this study since they have already been used to refine the sampling stratification and to partition the sGSL into 38 areal units with approximately uniform habitats. The spatial repartition of species will consequently be studied in this paper from a lattice data perspective.

We used data on the abundance of sea urchins (*Strongylocentrotus sp.*) collected from 1999 to 2001, which corresponds to 540 bottom-trawl surveys (Fig 1). The data can be obtained on request from the third author (email: BenoitH@dfo-mpo.gc.ca). This species was chosen as it is relatively sessile, with very negligible interannual movements on the scale of the survey. Furthermore, the time period was chosen so as to minimize the impact of interannual changes in abundance. Fig 2 clearly emphasizes that the sea urchins in the sGSL are distributed, like the majority of marine organisms, in patches of localized variable abundance, interspersed by numerous and relatively large areas where the species is absent.

2.2 The traditional *delta* models

Let $\{y_k; k = 1, 2, 3, ..., r\}$ denote quantities of biomass measured in r sampling sites and $\{S_k; k = 1, 2, 3, ..., r\}$ the corresponding sampling effort (e.g., swept area as in our case study but could also be volume filtered, observation time). The records of r independent sampling events contain strictly positive continuous values but zero values can also occur. Zero-inflated continuous data

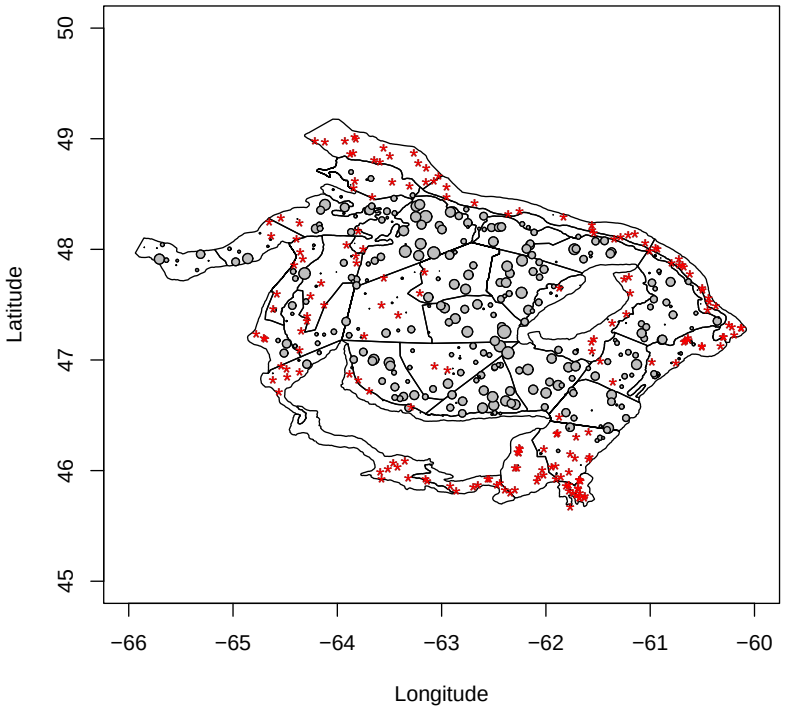


Fig. 2 The spatial repartition of sea urchins biomass from individual tows in the sGSL bottom-trawl survey, 1999-2000-2001. The "*" denotes the sites where no urchins were caught. The radii of the circles are proportional to the biomass (in kg/taw) caught.

sets are often modelled using *delta* models. *Delta* models are named after the Dirac function at zero, modelling the occurrence of a zero value with a Bernoulli random variable. The other component gives the positive abundances using either a gamma or a lognormal distribution after rescaling the data y_k (k=1,2,..,r) by the catch effort S_k. Let G be the distribution of the biomass per unit of area. Thus, the cumulative distribution function (c.d.f) of the abundance Y_k is:

$$\begin{aligned} [Y_k = 0] &= \delta \\ [Y_k < y_k] &= \delta + (1-\delta)G\left(\frac{y_k}{S_k}\right) \text{ with } y_k > 0 \end{aligned}$$

where δ denotes the probability that the studied species is absent ($0 \leq \delta \leq 1$) and G is a continuous c.d.f (gamma or lognormal) describing the abundance of strictly positive values. In this paper, the notation [ ] means either a distribution function for discrete variable or a density for continuous ones.

In what follows, G denotes a gamma cumulative distribution function. Actually, we followed the results of Myers and Pepin (1990) which suggested that the use of the gamma density is preferable to the use of a lognormal density for fisheries data especially when there is a considerable probability of small observations as this is the case for the urchins of the sGSL. A

shape parameter $\alpha > 0$ and a rate parameter $\beta > 0$ are added by the gamma density G:

$$[Y_k = 0] = \delta$$
$$[Y_k = y_k] = (1-\delta)\frac{\beta^\alpha}{\Gamma(\alpha)}\left(\frac{y_k}{S_k}\right)^{\alpha-1}\exp\left(-\beta\left(\frac{y_k}{S_k}\right)\right) \text{ with } y_k > 0$$

Consequently, three unknown quantities (α, β, δ) characterize the random mecanism of data occurrence for this particular *delta* model called *delta*-gamma model ($\Delta\Gamma$ model).

A property of towing a net is that the same overall distribution would be expected to hold for a long tow as for a short tow. Unfortunately, it is not the case with the $\Delta\Gamma$ distribution. This can be simply seen by considering the characteristic function of the $\Delta\Gamma$ distribution, given by:

$$\varphi^{\Delta\Gamma}(t) = \mathbb{E}(e^{itY_k}|\alpha,\beta,\delta) = \delta + (1-\delta)\frac{1}{(1-it\frac{S_k}{\beta})^\alpha}$$

It is clear that the probability distribution of the sum of two $\Delta\Gamma$ variables is not a $\Delta\Gamma$ distribution but a quite complex one for which analytical results are not trivial to obtain. The lack of such additivity property is a major draw-back when working of the $\Delta\Gamma$ distribution [19]. A preliminary standardization of data is needed to consider biomass abundances rescaled to the same sampling effort. Therefore, when tow durations are very dispersed, this standardization produces an artificial increase of the number of zero values and indicators of stock abundance are then biased.

2.3 A new competing data model: the LOL

The LOL belongs to the class of compound Poisson processes until now used to model continuous-time stochastic processes. It was initially designed to model losses from French gas pipelines [3]. Exponential gas intensities would flow out from Poisson distributed holes all along a pipe and only the sum of the local contributions was measured. It is also well known in the insurance domain industry if one assumes exponential damage events occurring as a time Poisson process. In what follows, we point out a spatial and ecological analogy : a latent Poisson sampling process collecting patches of organisms that each have an associated exponential mass. Formally, the LOL can be presented under a hierarchical setting.

2.3.1 The underlying process submodel with latent patches

Conceptually, the LOL describes the process involved in sampling many of the living organisms that are the focus of ecological analysis. For example, imagine one bottom-trawl survey that consists in sweeping the sea floor with a large fishing net to collect organisms of a given species. The model assumes there are patches of organisms to be collected. Patches are drawn from an homogeneous Poisson process so that the random number N_k of patches collected during the sampling event k is given by:

$$N_k \sim Poisson(S_k\mu)$$

where μ is the expected number of patches for a unit of sampling effort. The latent N_k are independent but not identically distributed Poisson variables depending on the corresponding sampling effort S_k.

Each patch $p_k = 1, .., N_k$ contains a certain biomass M_{p_k}. The simplest version of the LOL assumes that the M_{p_k}, $p_k = 1, .., N_k$, are $iid-$exponentially distributed with parameter ρ such that $\mathbb{E}(M_{p_k}) = \frac{1}{\rho}$:

$$M_{p_k} \overset{iid}{\sim} Exp(\rho), \quad p_k = 1, .., N_k$$

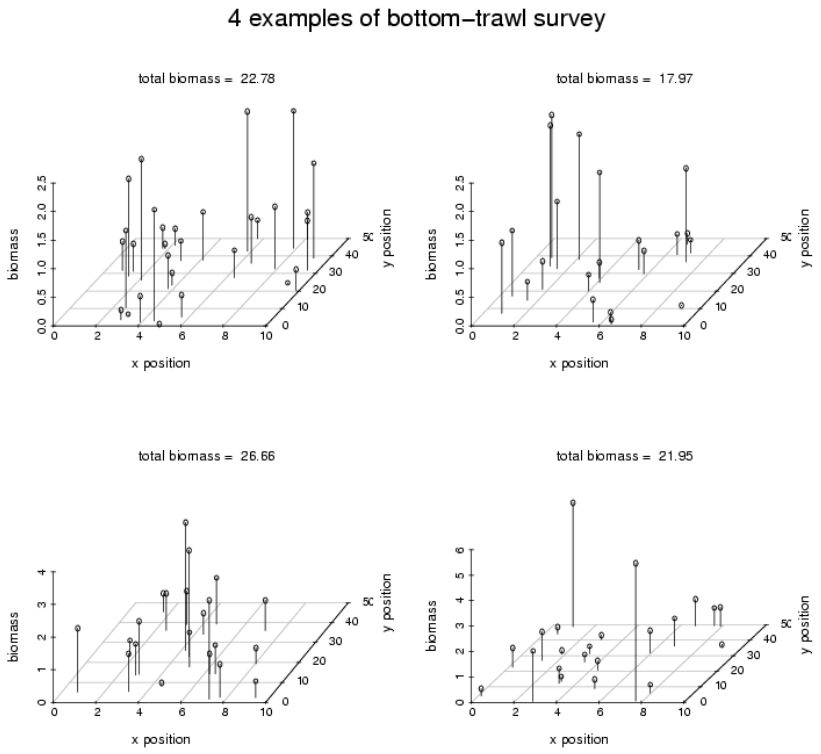


Fig. 3 The LOL is obtained by harvesting a marked Poisson point process

2.3.2 The data submodel

The sum of the individual patches captured by the trawl yields the total observed sample biomass $Y_k = \sum_{p_k=1}^{N_k} M_{p_k}$ (see Fig 3). By definition, an absence of patches (i.e., $N_k = 0$) offers a zero value and the occurrence of at least one patch (i.e., $N_k \geq 1$) produces a strictly positive outcome, distributed as the random sum of independent exponential variables (i.e., gamma pdf).

Conditionally on the unknown number of patches N_k $(k = 1, 2, .., r)$, we see that:

$$Y_k \sim \begin{cases} \Gamma(N_k, \rho) & \text{if} \quad N_k > 0 \\ 0 & \text{if} \quad N_k = 0 \end{cases}$$

where the scale parameter ρ is interpreted as the inverse of average biomass in each patch.

Working conditionally on μ and ρ only, we integrate out the latent N_k to obtain the pdf for the LOL:

$$[Y_k = y_k|\mu, \rho] = \begin{cases} \sum_{n=1}^{\infty} \left(e^{-S_k\mu} \frac{(\mu S_k)^n}{n!} \right) \frac{\rho^n}{\Gamma(n)} y_k^{n-1} e^{-\rho y_k} & \text{if} \quad y_k > 0 \\ e^{-S_k\mu} & \text{if} \quad y_k = 0 \end{cases} \tag{1}$$

The zero occurrence (no patch) is only a function of the unknown parameter μ. In contrast to the $\Delta\Gamma$ model, only two unknown quantities (ρ, μ) are necessary to describe the random mecanism of data occurrence.

2.3.3 Model properties

In this subsection, we briefly sum up some interesting properties of the LOL model. The characteristic function of the LOL is :

$$\varphi^{LOL}(t) = \mathbb{E}(e^{itY_k}|\mu,\rho) = e^{\frac{it\mu S_k}{\rho - it}} \tag{2}$$

Common quantities of interest for ecologists are easily derived. The probability of getting a zero value at site k is given by $e^{-S_k\mu}$ and the pointwise mean and variance of the abundance linearly depends on the sampling effort S_k :

$$\mathbb{E}(Y_k|\rho,\mu) = \frac{\mu S_k}{\rho}$$
$$\mathbb{V}ar(Y_k|\rho,\mu) = \frac{2\mu S_k}{\rho^2}$$

The coefficient of variation $\sqrt{\frac{2}{\mu S_k}}$ will desirably increase towards infinity as the expected number of patches tends to zero.

To represent the variability of each capture event, the LOL assumes that, within a survey stratum, the occurrence of individuals of a species follows an homogeneous Poisson point process (with exponential marks). Sampling events are considered as interchangeable random experiments. This strong hypothesis means that the survey stratum must be small enough to encompass a rather uniform habitat, suitable for the studied species and characterized by a non-structured spatial repartition of patches at the (micro) scale of a tow.

As a consequence, the model inherits a Poisson spatial coherence with regards to a change of sampling effort. Imagine that the sampling effort goes from S_1 to $\widetilde{S} = S_1 \cup S_2$, with $S_1 \cap S_2 = \varnothing$. The events Y_1 and Y_2 made in a same stratum occur independently on S_1 and S_2, and equation 2 shows the total catch event $Y' = Y_1 + Y_2$ also follows a LOL distribution. Therefore, contrary to the $\Delta\Gamma$ distribution, the LOL enables to work with raw data directly. Each survey set corresponds to a sampling event according to an homogeneous Poisson point process and the greater is the sampling effort, the greater is the number of collected patches.

Moreover, the LOL offers a natural representation of data collected along a gradient, where zeros are often a natural endpoint of a progression from high to low abundances. Contrary to *delta* models with a lognormal or a gamma non zero component, the probability density function of equation 1 tends to a strictly positive value at zero (namely $\mu\rho e^{-\mu}$). Hence, some interplay between parameters μ (controlling the probability of absence) and ρ ensures a much smoother link between the zero and non-zero values than the *delta* models characterized by an abrupt shift modelled by a Dirac function.

2.4 Accounting for spatial dependencies

The sGSL has been divided into 38 homogeneous strata (Sect. 2.1). Consequently, a first basic idea consists in leading inference of the model on each stratum i ($i = 1, 2, ..., 38$) separately. The weakness of this method is obvious: in small strata with few samples only, the quality of the estimation will be very poor.

But all strata share a common feature : they are located in the same part of the ecosystem. Then, we expect that the properties of those strata present generally similar behavior. Borrowing information from neighboring sites via a hierarchical model is developed in the following sections through two proposed spatial structure.

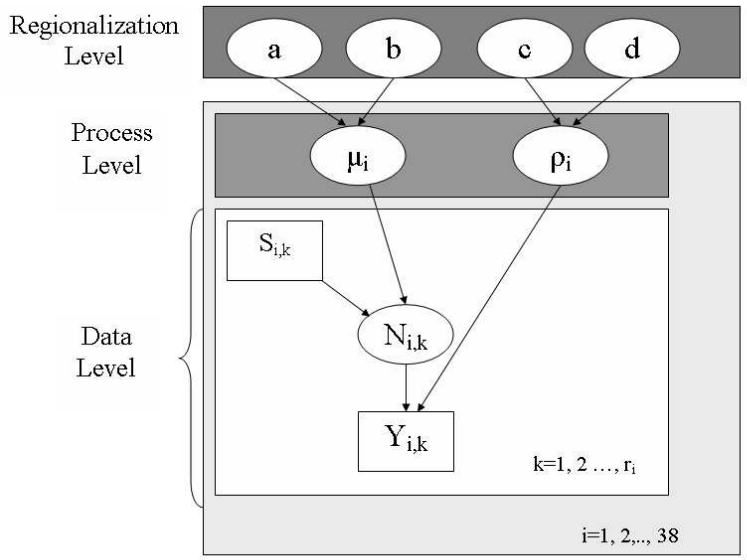


Fig. 4 Directed Acyclic Graph for the $R_{\mu,\rho}$-LOL model with regionalization structure. The square nodes represent observed data and the circles refer to latent variables or parameters. Arrows depict stochastic dependence between nodes.

2.4.1 A regionalized structure to borrow strength from exchangeable areal units

The LOL and $\Delta\Gamma$ submodels yield the total amount of biomass collected $Y_{i,k}$ in stratum i. Both the LOL and $\Delta\Gamma$ distribution are defined by area specific sets of parameters: (μ_i, ρ_i) and $(\delta_i, \alpha_i, \beta_i)$ respectively.

A simple structure to take into account a similar behavior between strata is a regionalized version of the data submodels in which an upper level is added to the basic models. The parameters of the data submodels called z_i (i=1,2,...,38) are assumed to stem from an iid common *regional* distribution H describing the various degrees of ressemblance between sites. We obtain a non linear mixed model:

$$[z_i|\theta] \sim^{iid} H(\theta)$$

For instance, the LOL with $z_i = (\mu_i, \rho_i)$ can be regionalized (model called $R_{\mu,\rho}$-LOL) via a gamma structure as a convenient regional distribution H with $\theta = (a, b, c, d)$. For stratum i taken at random in the sGSL, the expected number of patches μ_i and the expected biomass contained in each patch $1/\rho_i$ are random effects following a gamma distribution of parameters (a, b) and (c, d) respectively. The four letters are first level parameters. The Directed Acyclic Graph in Fig 4 suggests a simple representation of the $R_{\mu,\rho}$-LOL model. For each stratum i, the expected number of patches to be collected is a/b and the inverse of the expected biomass contained in one patch is c/d. Similarity between all strata is determined by the dispersion among the z_i. As the variances a/b^2 or c/d^2 are decreased, the strata will become more similar.

For the $\Delta\Gamma$ distribution of section (2.2), a regional distribution has to be put on a three dimensional vector $z_i = (\alpha_i, \beta_i, \delta_i)$. We chose to draw $logit(\delta_i) = \log(\frac{\delta_i}{1-\delta_i})$ from a normal distribution and β_i from a Gamma distribution (model called $R_{\delta,\beta}$-$\Delta\Gamma$). For a fair comparison among models with the same number of parameters, we decided in what follows to set the

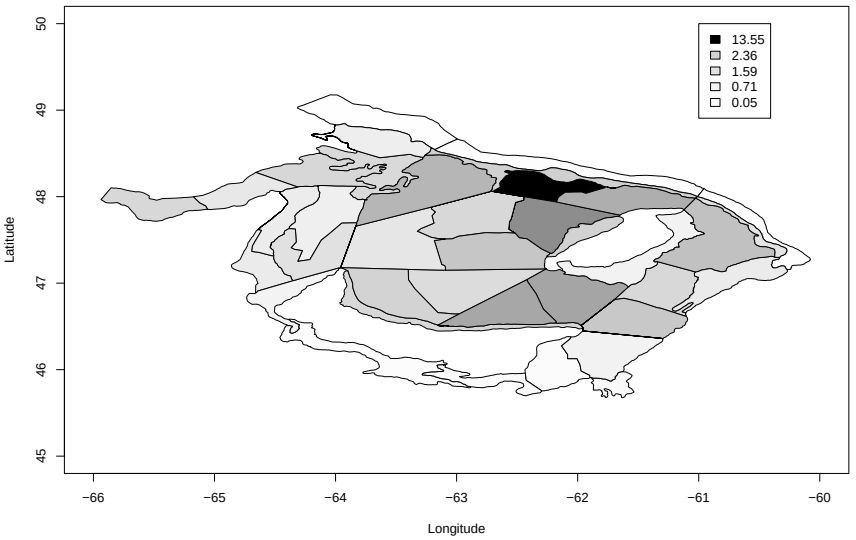


Fig. 5 Spatial structure of the expected number of patches μ obtained by fitting LOL independently in each stratum

coefficients of variation of the gamma distributions to a constant $\alpha_i = \tilde{\alpha}$, which is not an unrealistic assumption from an ecological point of view.

Regionalized model structures take advantage of similarity between survey strata regardless of the distance that separates them. However, one could argue that spatially more proximate strata should have a more similar behavior. The idea that vicinity matters is developed in the following section using an intrinsic Gaussian conditional autoregressive model (IAR model) to impose a spatial structure.

2.4.2 An IAR to link neighboring areal units

One may wish to assume that the expected numbers of patches in stratum i ressembles those in neighboring strata (see Fig 5). Big catches (about 14 kg of urchins per tow) were observed in the stratum located in the northern portion of the sGSL. This stratum provides a suitable habitat for urchins with gravelly coarse sand as the dominant sediment type and favorable water temperature conditions. One expects that the number of patches μ (if the LOL model is used as data model) will also be great in the neighboring strata which may also benefit from rather favorable physiographical conditions.

We assume that some univariate link function g of the latent vector $z = (z_1, z_2, ...z_{38})$ satisfies:

$$g(z_i) = m_0 + \varepsilon_i + \Phi_i$$

where:

- m_0 is a constant term.
- ε_i captures local heterogeneity using a normal pdf : $(\varepsilon_i)_{1\leq i\leq 38} \overset{i.i.d}{\sim} \mathcal{N}(0, s_\varepsilon^2)$.

– Φ_i is the spatially structured section of the model. We suppose two strata sharing a common boundary are neighbors. Here, we use a IAR model [2] on $\Phi = (\phi_1, \phi_2, .., \phi_{38})$:

$$[\Phi_i | \Phi_j \, , j \sim i] \propto \exp \left\{ - \frac{(\Phi_i - \frac{1}{n_i} \sum_{j \sim i} \Phi_j)^2}{2 \frac{s^2_{IAR}}{n_i}} \right\}.$$

where $i \sim j$ means stratum i and stratum j are neighboring, n_i denotes the number of adjacent strata to stratum i and s^2_{IAR} is the local standard deviation.

The IAR structure is improper (the sum of the Φ_i must be centered so that the conditional distributions make sense to define a joint pdf on $\mathbb{R}^{n-1}$), but we will see in the next section that, since it is used only as a top level structure (prior) in a hierarchical Bayesian setting, the inference doesn't present any difficulty and the posterior will be proper. Each time a component of z is modelled by a IAR structure through a univariate link function g, three top-level parameters $(m_0, s^2_\varepsilon, s^2_{IAR})$ have to be inferred. It is worth noting that when $s^2_{IAR} = 0$, the IAR structure yields a regionalized model with the normal pdf as a regional distribution H on the $g(z_i)$.

Modelling ρ with a spatial structure instead of μ or modelling both ρ and μ as a multivariate IAR is discussed later.

A IAR layer can also be added to a *delta* model. For the $\Delta\Gamma$ distribution with the three dimensional vector $z_i = (\alpha_i, \beta_i, \delta_i)$ (section 2.2), spatial structure can be added on the probability of a zero e.g., using the simple link function $g(z_i) = logit(\delta_i)$ or it could be added on the expected non zero biomass in stratum i e.g., writing $log\left(\frac{\alpha_i}{\beta_i}\right) = m_0 + \varepsilon_i + \Phi_i$. Additionally, more sophisticated models could also rely on a multivariate IAR structure.

2.5 Competing hierarchical constructions

Depending on the number of parameters in each data submodel to be *spatialized*, the chosen link functions and the hypothetized regional distribution, many hierarchical constructions can be specified. Table 1 sums up the eight combinations for the LOL and $\Delta\Gamma$ submodels, with or without spatial structure, which are compared in section 5.

Four variants of the LOL were considered (Table 1): the LOL independently applied to each stratum (called $(\text{LOL})^{\otimes 38}$), the $\text{R}_{\mu,\rho}$-LOL model (see Sect. 2.4.1) , a partially regionalized version of LOL (called R_μ-LOL) in which the latent variables μ_i (i=1,....,38) are *iid* following a gamma distribution and $\rho_i = \widetilde{\rho}$, and the IAR_μ-LOL model (Sect. 2.4.2).

Four models based on similar structures have been considered for the $\Delta\Gamma$ distribution (Table 1): the $\Delta\Gamma$ distribution independently applied to each stratum (called $(\Delta G)^{\otimes 38}$), the $R_{\delta,\beta}$-$\Delta\Gamma$ model (Sect. 2.4.1), a partially regionalized version (called R_δ-$\Delta\Gamma$) in which $\text{logit}(\delta_i)$ are iid following a normal distribution and both $\alpha_i = \tilde{\alpha}$ and $\beta_i = \tilde{\beta}$ for all i from 1 to 38 and a spatialized version (called IAR_δ-$\Delta\Gamma$) in which the $\text{logit}(\delta_i)$ follows a IAR model and both $\alpha_i = \tilde{\alpha}$ and $\beta_i = \tilde{\beta}$ for all i from 1 to 38.

3 Bayesian inference

The inference for the family of models relying on the LOL or on the $\Delta\Gamma$ distribution has been developed under the Bayesian paradigm. This setting offers the possibility of accounting for external qualitative information through the prior, the commonsense intuitive interpretation of

Table 1 Eight competing models based on the LOL and the $\Delta\Gamma$ distribution

Spatial structure/ Data submodel	LOL $z_i = (\mu_i, \rho_i)$	$\Delta\Gamma$ $z_i = (\delta_i, \alpha_i, \beta_i)$
No spatial effect	$(\text{LOL})^{\otimes 38}$	$(\Delta\Gamma)^{\otimes 38}$
Neighbors are exchangeable	$\text{R}_{\mu,\rho}$-LOL	$\text{R}_{\delta,\beta}$-$\Delta\Gamma$
Neighbors are partially exchangeable	R_{μ}-LOL	R_{δ}-$\Delta\Gamma$
Vicinity only matters	IAR_{μ}-LOL	IAR_{δ}-$\Delta\Gamma$

posterior statements in terms of probabilistic bets and the coherence with probability theory when deriving predictive judgments by integrating out nuisance parameters. Thanks to MCMC techniques, the Bayesian framework is also particularly appropriate for working with hierarchical models such as the ones we presented here.

Consider for instance the $R_{\mu,\rho}$-LOL model with observations y, two layers of latent variables $z = (N, \mu, \rho)$ and parameters $\theta = (a, b, c, d)$. The posterior distribution of the unknowns (i.e., both the parameter θ and the latent variable z), given the data y is simply obtained by Bayes theorem as:

$$[\theta, z|y] \propto [y|z, \theta][z|\theta][\theta]$$

In this expression:

– the data occurrence submodel $[y|z, \theta]$ is:

$$[y|z, \theta] = [y|N, \rho] = \prod_{i=1}^{38} \prod_{k=1}^{r_i} [y_{i,k}|\rho_i, N_{i,k}]$$

$$= \prod_{i=1}^{38} \left(\prod_{\{k;N_{i,k}=0\}} \delta_0(y_{i,k}) \prod_{\{k;N_{i,k}>0\}} \widetilde{G}(N_{i,k}, \rho_i)(y_{i,k}) \right)$$

where $\widetilde{G}(N_{i,k}, \rho_i)$ denotes the Gamma p.d.f with $n_{i,k}$ as shape parameter and ρ_i as rate parameter

– the distribution of the latent variable z given the parameter θ, is :

$$[z|\theta] = [N|\mu][\mu|a, b][\rho|c, d]$$

– the prior $[\theta] = [a, b, c, d]$.

Inferences on (θ, z) are carried out by simulating the posterior distribution $[\theta, z|y]$ through a Markov Chain Monte Carlo (MCMC) sampling algorithm - the Gibbs sampler and/or the Metropolis Hastings algorithm (e.g., Metropolis et al., 1953; Tanner, 1992).

The Bayesian model specification requires prior distributions. In our case, it is difficult to produce informative priors for *theoretical* quantities such as an expected number of collected patches during a sampling event given that the definition of what consists of a patch depends on scale. Consequently, a simple approach was to consider flat priors. When the LOL was applied independently to each stratum, μ_i and ρ_i (i=1,2,...,38) were independent with a flat normal prior truncated at 0. For the regionalized versions of the LOL (Sect. 2.4.1), $\overline{\mu} = \frac{a}{b}$, $\overline{\rho} = \frac{c}{d}$, $\tau_\mu = \frac{b^2}{a}$ and $\tau_\rho = \frac{d^2}{c}$ all followed a gamma prior with 0.001 as shape and rate parameters. Finally, for the

IAR$_\mu$-LOL model, we used the priors recommended in [2]: the precision $\frac{1}{s_\varepsilon^2}$ of the non-spatialized residual component (Sect. 2.4.2) with a gamma prior with 0.001 as shape and rate parameters and the local precision of the IAR model $\tau_{IAR} = \frac{1}{s_{IAR}^2}$ following a gamma prior with 0.1 as shape and rate parameters.

Models fitting was performed using the software OpenBUGS [17] and the BRugs package from R Project for Statistical Computing. OpenBUGS appears as a nice and well documented tool [4] for Bayesian analysis of complex statistical models using Markov chain Monte Carlo (MCMC) techniques. We ran each model of Table 1 for 200,000 cycles with a burn-in period of 50,000 cycles and a thinning of 100 cycles. We checked the convergence of MCMC algorithms by computing Gelman-Rubin Statistics [6] thanks to the R package Coda.

4 Testing Bayes factor sensitivity to priors

The Bayes factor $BF_{i,j}$ is a measure of the relative credibility of the model M_i compared to the model M_j given the data y [10] and appears as the ratio of the posterior probability of the competing models when the prior probabilities are equal. Consequently, it is simply the ratio of the marginal likelihood of the data under the model M_i to the marginal likelihood under the model M_j :

$$BF_{i,j} = \frac{\frac{[M_i|y]}{[M_j|y]}}{\frac{[M_i]}{[M_j]}} = \frac{[y\,|M_i\,]}{[y\,|M_j\,]} \tag{3}$$

Computing the marginal likelihood can be done by the standard decomposition for a given model M

$$[y\,|M\,] = \int [y\,|M,\phi][\phi\,|M\,]d\phi \tag{4}$$

In this section, ϕ is a generic notation employed for *any* subset of the unknowns (z, θ). To emphasize that prior, likelihood and predictive pdfs are model specific, we write in this section $[\phi\,|M\,], [y\,|M,\phi], [y\,|M\,]$ instead of $[\phi], [y\,|\phi], [y]$.

Equation (4) shows that the marginal likelihood is not defined when an improper distribution $[\theta\,|M\,]$ is chosen as a prior. More generally, Bayes factors are known to be sensitive to prior choice, and informative priors are recommended to get a meaningful comparison between models relying on Bayes factors.

Now consider a split sample test analysis. Suppose that the data y can be divided into two independent samples blocks $y = (y_{(-t)}, y_{(t)})$. t is a partition index such that:

$$[y\,|M,\theta] = [y_{(-t)}\,|M,\theta] \times [y_{(t)}\,|M,\theta] \tag{5}$$

The subsample $y_{(-t)}$ is a *learning* sample : starting with the non informative prior $[\theta\,|M\,]$, even improper, information conveyed in $y_{(-t)}$ is processed into an informative pdf $[\theta\,\big|M, y_{(-t)}\,]$. This latter posterior pdf is used as a proper prior for the study with the *test* sample $y_{(t)}$, and using definition (3), predictive Bayes factor
$BF_{ij,(t)} = \frac{[y_{(t)}|M_i\,,y_{(-t)}]}{[y_{(t)}|M_j\,,y_{(-t)}]}$ can be straightforwardly derived as a measure of the relative credibility of the model M_i compared to the model M_j for the test data $y_{(t)}$ after *assimilating* $y_{(-t)}$ during a learning stage. The learning sample $y_{(-t)}$ must be made large enough such that $[\theta\,\big|M, y_{(-t)}\,]$ is proper and sufficiently informative, and small enough so that $y_{(t)}$ remains a representative sample of the whole dataset. A robustness analysis to prior specification can be easily conducted by varying the partition index t, possibly after rearrangering the ordering of the data.

Appendix A gives some insights about the computation of the Bayes factor via MCMC techniques and its possible numerical instabilities. In the present application, a 500000 MCMC replicates were used to compute Bayes factors.

5 Case study: bottom-trawl survey data

In this section, we report results from an intensive survey of the abundance data concerning the urchins and collected in the sGSL from 1999 to 2002 (Sect. 2.1). For the purpose of our case study, we used four years (1999-2002) of data from the bottom-trawl survey (Sect. 2.1). Data from the first three years were used to predict urchin abundance in 2002.

The goals of our analysis are:

1. to select a model by comparing the fitting and predictive abilities of the $(LOL)^{\otimes 38}$, $R_{\mu,\rho}$-LOL, R_{μ}-LOL and IAR_{μ}-LOL models with the corresponding versions of the $\Delta\Gamma$ model (see Table 1).
2. to analyse more precisely the results obtained with the best structure for predictions, the IAR_{μ}-LOL model.

5.1 Comparing the fitting abilities

Generally speaking, it is difficult to define meaningful criteria to compare hierarchical bayesian models. Spiegelhalter et al. (2002) proposed a Deviance Information Criterion (DIC) as a measure of complexity and fit:

$$DIC = \overline{Dev(\theta)} + pD$$

where $\overline{Dev(\theta)}$ measures the goodness of fit and $pD = \overline{Dev(\theta)} - Dev(\overline{\theta})$ acts as a penalty term depending on the model complexity. Even if MCMC sampling techniques make it easy to compute, the DIC has no clear statistical interpretation (except for normal linear models) and depends on the parametrization choice. That's why, we reinforced our comparative analysis by computing Bayes Factors.

One can be irrespectful towards chronology since we ignore dynamic evolution. Consequently, there are seven ways to split the 3 years sample, so as to define a learning sample to get a proper pdf for the unknowns, that will be used as a prior on the remaining sample:

- empty learning sample (non informative priors as defined in the previous section),
- one year of data to learn (3 possibilities : either 1999 or 2000 or 2001), yielding an informative state of knowledge for the unknows and two years to compare models,
- two years of data to learn (3 possibilities : 2 years to be taken among 3) and the remaining year as a validation data set. This case corresponds to very informative priors.

In order to obtain meaningful predictive Bayes factors, we used two years of data (about 66%) to learn and define the prior distributions. The diagonal line of Table 2 shows the relative credibility of similar versions of each model, based on validation samples consisting of single years (either 1999 or 2000 or 2001).

With the exception of the regionalized version $R_{\mu,\rho}$-LOL,

The partially regionalized version R_{μ}-LOL and the spatial version IAR_{μ}-LOL of the LOL fit to the data better than the analogous $\Delta\Gamma$ model (Table 2). The DIC confirm these results. The independent version of the LOL clearly provides a superior fit compared to the independent version of the $\Delta\Gamma$.

Table 2 Predictive Bayes factors computed from the validation samples consisting of single years (1999, 2000 or 2001). The last column and last line contain the DIC related to models based on the LOL and $\Delta\Gamma$ data submodels respectively.

↱	$(\Delta\Gamma)^{\otimes 38}$	$R_{\delta,\beta}$-$\Delta\Gamma$	R_{δ}-$\Delta\Gamma$	IAR_{δ}-$\Delta\Gamma$	**DIC**
$(LOL)^{\otimes 38}$	3.88 -4.93 -5.53				2538.25
$R_{\mu,\rho}$-LOL		-3.66 -2.51 -7.08			2559.64
R_{μ}-LOL			3.49 13.05 4.64		2591.63
IAR_{μ}-LOL				1.42 11.36 4.78	2595.46
DIC	2524.62	2545.57	2607.76	2608.51	

The fitting capabilities of models based on the $\Delta\Gamma$ distribution improve when independent or regionalized coefficients of variations of the gamma distributions are considered. However, the fitting of the different versions of the LOL remain better than the $\Delta\Gamma$ distribution even in that case (Bayes factors computed but not shown).

To check Bayes factors sensitivity to priors, we performed a partial robustness analysis to prior specification. In order to smooth the numerical instability due to the computation of harmonic means (see Appendix A eq(10)), we monitored the evolution of the Bayes factors from an initial sample of 200,000 iterations to a sample of 500,000 iterations by successively increasing the former by 1,500 additional iterations. The sensitivity of Bayes factors both to prior choice and to the corresponding validation sample is clearly observed in Fig 6. This is particularly true for all the models compared. Although the evidence in favor of the IAR_{μ}-LOL model decreases when the level of prior information increases, the same conclusion remains: the IAR_{μ}-LOL fits the data better than the IAR_{δ}-$\Delta\Gamma$ (i.e., Bayes factors always are strictly higher than 1). Moreover, for a given level of prior information, this analysis shows the degree of evidence in favor of the IAR_{μ}-LOL model can vary enormously depending on the data used to compute Bayes factors: a very strong evidence in favor of the LOL versions based on the data collected in 2000 but much less so for the ones collected in 1999.

5.2 Comparing the predictive abilities

Such statistical models can be used as predictive tools, for instance, to predict the evolution of stock abundance indicators or to help to plan future surveys. Therefore, comparing models can also be performed within a predictive setting with a loss function to evaluate the discrepancy between the test data and the predictive structure.

By dividing again the data y into a learning sample $y_{(-t)}$ and a test sample $y_{(t)}$ of quantities we want to re-evaluate, we computed the *posterior Predictive Loss Criterion* (PPLC) [5] for each model M of Table 1:

$$C^{\omega}_{(y_{(-t)},y_{(t)})}(M) = \sum_{l=t}^{n} \hat{\sigma}_l^2 + \frac{\omega}{\omega+1} \sum_{l=t}^{n} (\hat{\mu}_l - y_l)^2 \tag{6}$$

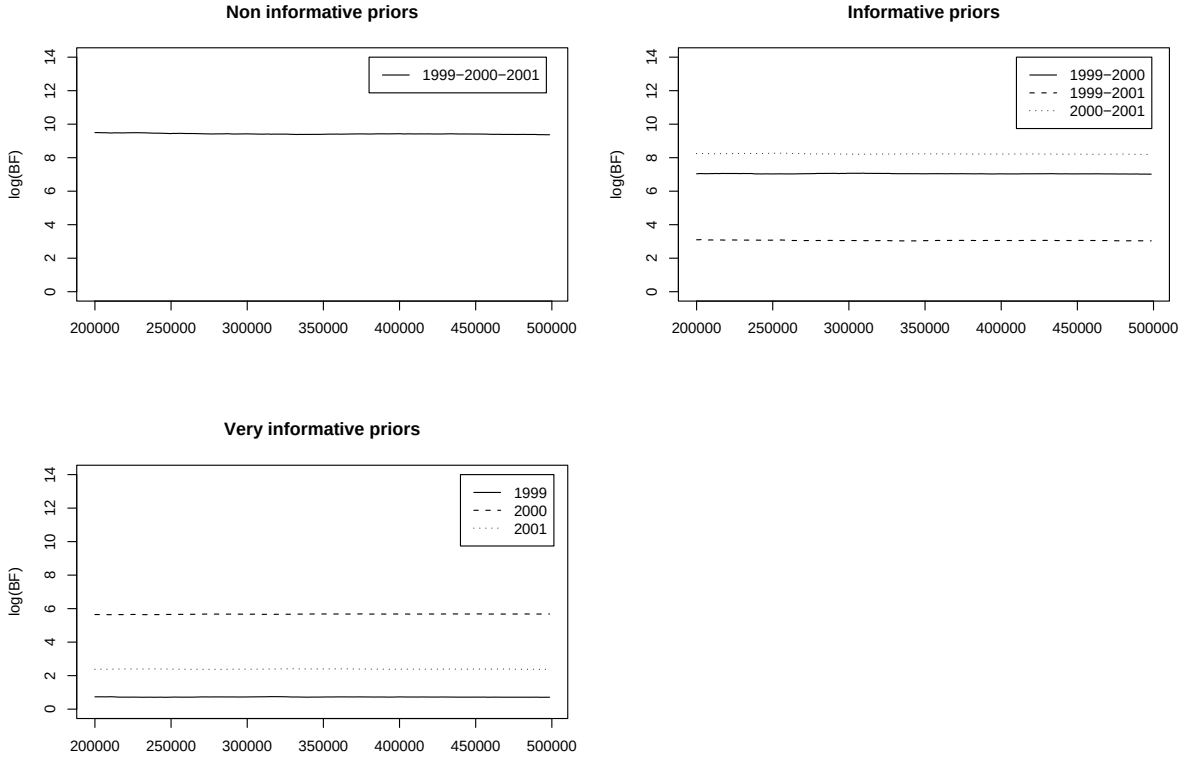


Fig. 6 Evolution of the predictive Bayes factors (log scale) comparing the IAR$_\mu$-LOL model to the IAR$_\delta$-$\Delta\Gamma$ model according different priors specifications

where $\hat{\mu}_l = \mathbb{E}(\hat{Y}_l)$ and $\hat{\sigma}_l^2 = \mathbb{V}ar(\hat{Y}_l)$, i.e., the mean and variance of the predictive distribution of $\hat{Y}_l$ (the predictive variable for y_l) given the learning sample $Y_{(-t)}$.

Starting from a non-informative prior, we processed the abundance data from 1999 to 2001 and then tried to predict replicates of these data. For each tow made from 1999 to 2001, we generate a 2000-sample of abundance values according to the related predictive distribution. Summary statistics are calculated from this. For instance, Table 3 summarizes the PPLC computed when $\omega = 1$ (see eq 6) for the proportion of null catches (no organisms caught), the average biomass expected to be collected and the degree of variability of strictly positive measured abundances.

The LOL provides better predictions of the average expected biomass collected and the coefficient of variation of strictly positive abundances relative to the analogous $\Delta\Gamma$ models (Table 3). Indeed, the precision of predictions is clearly improved (Figure 7).

The largest improvements were obtained for the predictions of the average biomass with PPLC at least twice as small as the ones related to the versions of the $\Delta\Gamma$ model. Independent versions $(\text{LOL})^{\otimes 38}$ and $(\Delta\Gamma)^{\otimes 38}$ are overfitted models which explains why they generally provide the worst predictions for the average biomass collected. For the proportion of zero values, the evidence in favor of the LOL is less obvious except for the independent versions.

The predictive capabilities of models based on the $\Delta\Gamma$ distribution are worse when independent or regionalized coefficients of variations of the gamma distributions are considered.

Table 3 PPLC computed from abundance data collected in 1999-2000-2001 for k=1

	Proportion of zero values	Average biomass (in Kg)	CV of positive quantities
$(\text{LOL})^{\otimes 38}$	0.78	257.61	10.45
$\text{R}_{\mu,\rho}$-LOL	0.85	241.36	10.43
R_{μ}-LOL	0.90	217.28	10.47
IAR_{μ}-LOL	0.88	219.91	10.84
$(\Delta\Gamma)^{\otimes 38}$	0.85	485.42	10.13
$\text{R}_{\delta,\beta}$-$\Delta\Gamma$	0.74	977.13	10.02
R_{δ}-$\Delta\Gamma$	0.74	643.88	10.58
IAR_{δ}-$\Delta\Gamma$	0.69	643.23	10.57

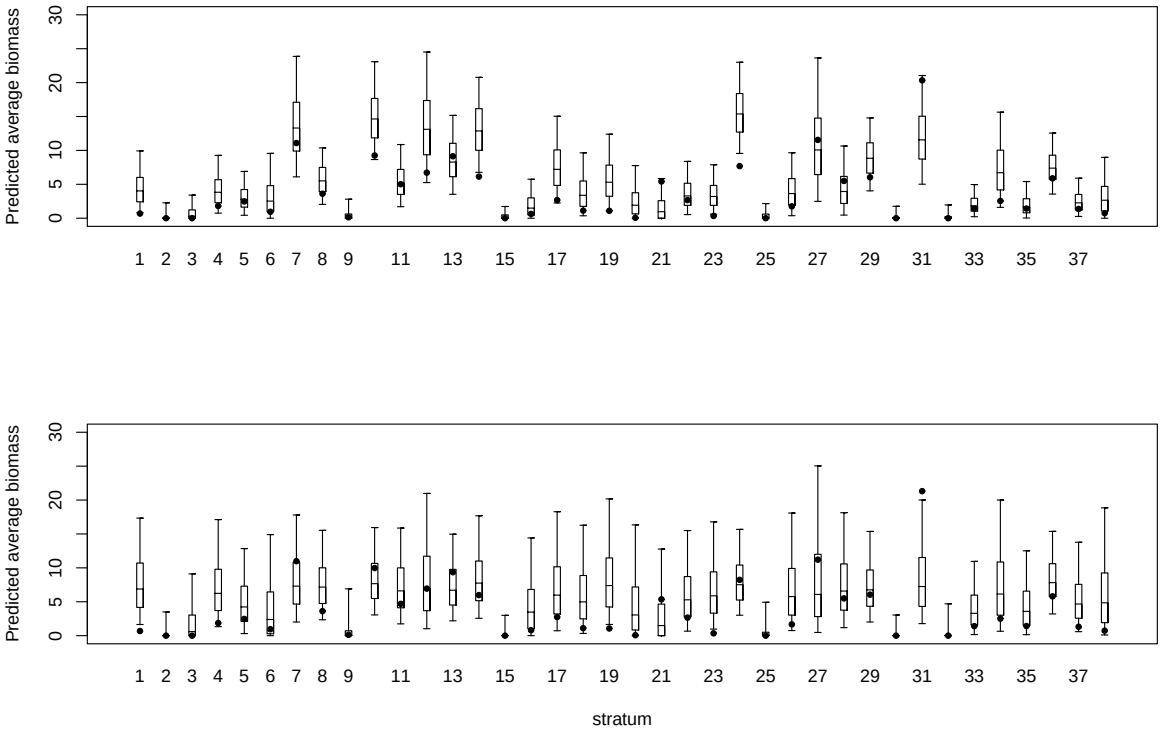


Fig. 7 95% credibility intervals computed from the predictive distributions for the average biomass expected to be caught in each stratum in 2002 (Top: IAR_{μ}-LOL model, Bottom: IAR_{δ}-$\Delta\Gamma$ model). The points correspond to the empirical values observed in 2002.

Consequently, the predictions of the different versions of the LOL remain better than the $\Delta\Gamma$ distribution even in that case (PPLC computed but not shown).

5.3 The IAR_{μ}-LOL model

Although it is the poorest member of the LOL family in terms of fitting ability, the spatial version of the LOL (IAR_{μ}-LOL) is clearly superior to the non-spatial version $(\text{LOL})^{\otimes 38}$ and the

Table 4 posterior mean, coefficient of variation and 95% credibility intervals for the IAR_μ-LOL model parameters (see 2.4.2). κ indicates the part of variability explained by the IAR structure.

parameters	posterior mean	posterior CV	$IC_{95\%}$
s^2_{IAR}	2.44	0.99	[0.87,4.75]
s^2_ϵ	0.19	0.21	[0.002, 0.74]
m_0	-0.04	0.10	[-0.26, 0.15]
κ	0.89	0.12	[0.57, 0.99]

regionalized versions ($R_{\mu,\rho}$-LOL, R_μ-LOL) for prediction purposes (Table 3). That is why, we favor this model and we give details on the results obtained for this model.

Table 4 contains the posterior means, coefficients of variation and 95% credibility intervals for the parameters of IAR_μ-LOL model from MCMC runs performed on abundance data collected in 1999, 2000 and 2001. The proportion of variability explained by the IAR model is measured by the ratio κ of the IAR model and the global variability.

In the absence of spatial effects, the expected number of patches is rather variable around 1 (0 is contained within the 905% posterior credible interval for m_0, the global mean for the $log(\mu_i)$ i=1,2,..38). Strong spatial autocorrelation between neighboring strata for the average number of patches collected is estimated (Table 4). Indeed, the IAR structure accounts for about 89% of the global heterogeneity. The variance of the residual noise is therefore smaller than that of the IAR structure.

6 Discussion & Perspectives

The LOL has considerable potential for the general analysis of zero-inflated abundance survey data. First, its latent variables μ and ρ can easily incorporate several relevant properties of organisms distribution. For example, environmental covariables can be used in determining the prior distributions for μ and ρ via generalized linear models. Another approach could be to explain the average biomass $\frac{\mu}{\rho}$ by environmental covariables instead of μ and ρ. Secondly, many hierarchical constructions can be imagined from this conceptual data submodel. In this paper, we proposed four possible hierarchical variants. We excluded multivariate $IAR_{\mu,\rho}$-LOL to favor parsimony and to avoid mathematical intricacy. We did not consider spatial effects on ρ (IAR_ρ-LOL or R_ρ-LOL) on the basis of a preliminary descriptive analysis of the sGSL dataset. We computed Moran's statistics for $\rho = (\rho_1, \rho_2, .., \rho_{38})$ and $\mu = (\mu_1, \mu_2, .., \mu_{38})$ respectively. In our case, as it must refer to latent variables z, we used the MCMC samples obtained from the independent case $(LOL)^{\otimes 38}$to compute a posterior sample of Moran's statistics for ρ and μ respectively. These are plotted in Figure 8. Moran [2] showed that if the μ_i (respectively ρ_i) are independent and identically distributed, Moran's index is asymptotically distributed with mean $-\frac{1}{\text{Number of strata}-1} = -\frac{1}{37}$ and an explicit variance. Figure 8 shows a larger departure from the independence hypothesis for the average number of patches collected during a standard survey μ_i (p_{value}=0.0005) than for the inverse of the average quantities of biomass in one patch ρ_i (p_{value}=0.0049). Consequently, we put a IAR structure on $\log(\mu_i)$ and assumed no spatial effect on ρ_i.

The analytical p.d.f for the LOL is:

$$[Y = y|\mu, \rho] = 1_{y=0}e^{-S\mu} + 1_{y>0}\frac{\mu S\rho}{\sqrt{\mu S\rho y}}\exp(-S\mu - \rho y)I_1(2\sqrt{\mu S\rho y})) \quad (7)$$

where S denotes the catch effort, and I_1 the modified Bessel function of the first kind linked with the occurrence of a strictly positive event. We did not present eq (7) earlier as we preferred

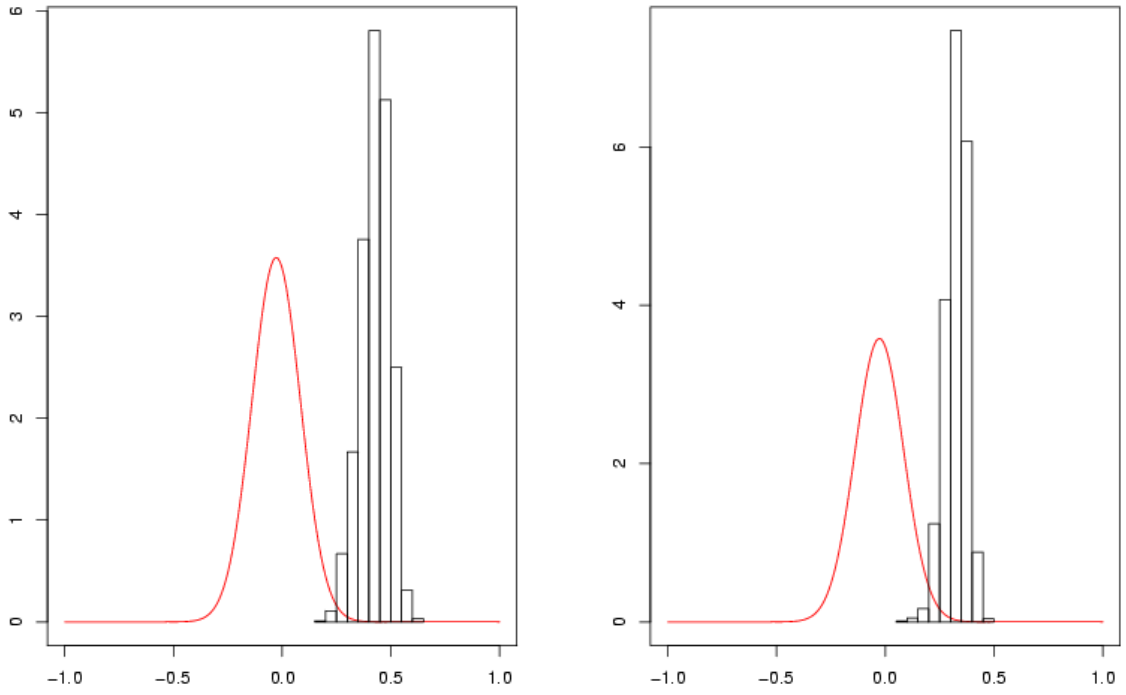


Fig. 8 Posterior distributions of Moran's index for the latent variables μ_i (left) and ρ_i (right). The solid line represents the distribution of Moran's index under the independence hypothesis.

to highlight its conceptual properties of compound Poisson process. The hierarchical Bayesian approach makes clear the role of different processes by specifying them at different stages of the model's hierarchy and confers a conceptual sense on parameters: μ, the expected number of patches to be collected during a standard sampling event and $1/\rho$ the average biomass contained in any patch in a given habitat.

Although the LOL is easily interpretable in its hierarchical form (see eq(1)), great care must be given not to over-interpret the significance of the LOL parameters. These unknowns are not directly observed when sweeping a net through water to capture organisms. Furthermore, the definition of patchiness depends on the scale at which it is measured. While it might be tempting to use the two variables to make statements about the patchiness of organism distribution, the high degree of correlation between the two confounds their interpretation (Figure 9). Indeed, this linear correlation increases with the average biomass collected in a given habitat. In other words, when a lot of biomass is caught, the LOL model does not properly resolve the differences between sampling a single large patch or numerous small single-individual patches. This correlation should not however be viewed necessarily as a weakness of the model. Indeed, when the LOL is used to estimate or predict the biomass in a given habitat, the correlation between μ and ρ doesn't matter since the quantity of interest only is $\frac{\mu S}{\rho}$ (eq (3)). It is only jointly however that these two parameters can be used to describe the distribution of observed biomass.

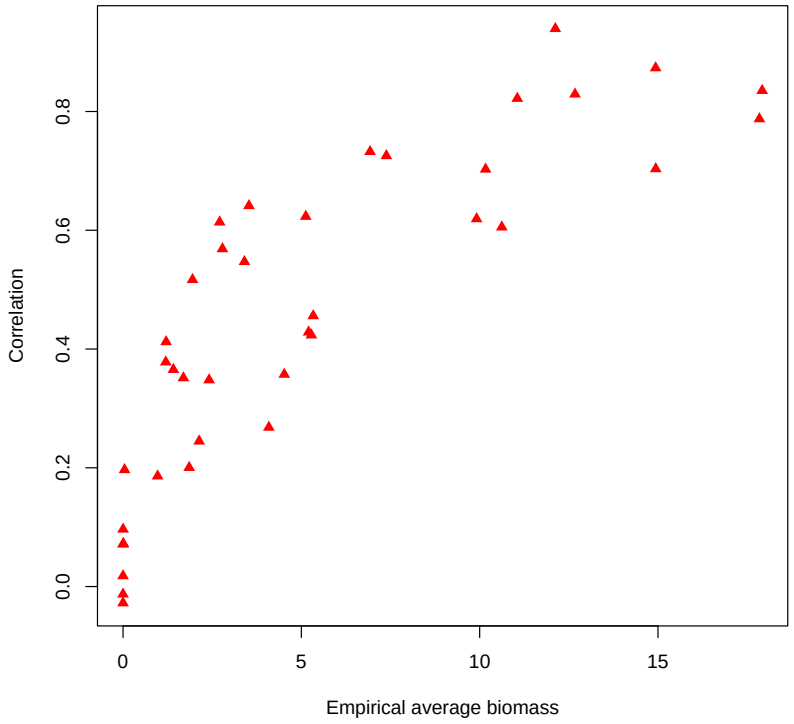


Fig. 9 Posterior correlations between ρ_i and μ_i (i=1,2,...,38) according to the empirical average biomass collected in September from 1999 to 2001. They were computed from MCMC samples obtained by fitting the $(\mathrm{LOL})^{\otimes 38}$ model. Each red point corresponds to one stratum.

Compared to the $\Delta\Gamma$ distribution, the LOL is more difficult to implement within OpenBUGS and unfortunately its deviance is not available from this software. Consequently, we made recourse to R to compute DIC and Bayes factors. It is worth mentioning that the introduction of an additional layer in the definition of the LOL and of a spatial prior for zero values has not to be paid much when inferring the unknowns. In particular, the IAR prior can be easily implemented thanks to GeoBUGS, an add-on to OpenBUGS that fits spatial models.

Many future extensions and modifications of the LOL are to be explored in possible works. For example, a multivariate LOL structure could be used to describe interspecies relationships in an analysis of ecological communities. Also, a discrete version of the LOL, obtained by changing the exponential into a geometric p.d.f, may provide an alternate model to the traditional ZIP and ZINB models used for counts. We also bet that a geostatistical version of the LOL would be most appropriate for abundance data with zeros at a thinner spatial scale.

Appendix A : MCMC Computation of Bayes factor

A traditional stochastic approximation by drawing a sample $\left(\phi^{(1)}, \phi^{(2)}, ...\phi^{(k-1)}, \phi^{(k)}\right)$ from the "prior" distribution $[\phi\,|M\,]$ and performing the prior arithmetic mean of the $[y\,\big|M, \phi^{(k)}\,]'s$ is often not very efficient since probable values from the sampling prior distribution and from the

likelihood function may lie far apart. We used the method proposed by Kass and Raftery (1994) :

$$[y\,|M]^{-1} = \int [y\,|M,\phi]^{-1}[\phi\,|M,y]d\phi$$

A more robust approximation of the marginal likelihood of model M can be computed by the posterior harmonic mean :

$$[y\,|M\,] \approx \left(\frac{1}{H}\sum_{h=1}^{H}[y\,\Big|\phi^{(h)},M\,]^{-1}\right)^{-1} \tag{8}$$

where $\phi^{(h)}$ is the g^{th} set of unknowns of the posterior sample (from $[\phi\,|y,M\,]$), $[y\,|\phi,M\,]$ is the value of the sampling density of the observed data in M calculated at $\phi^{(h)}$, and H is the sample size. This estimate of the marginal likelihood is known to be unstable, because the harmonic mean of the likelihood is highly sensitive to the very small values that may appear during the sampling. This may hinder the comparison of two models when their difference in credibility is low. However, in practice, if the contrast between the credibility of the different models is high, the approximation from (8) gives results which are accurate enough for identifying the most credible model(s). An additional trick to increase numerical stability of (4) or (8) is to ensure ϕ as the smallest subset of (z,θ) such that $[y\,|M,\phi]$ is explicitly known. As hierarchical models often exhibit partial conjugacy between their successive layers, dimension reduction can be obtained by performing integration whenever possible to shrink $[y\,|M,z,\theta]$ into $[y\,|M,\phi]$.

When using an informative pdf $[\theta\,\big|M,y_{(-t)}\,]$ as a prior for the study with the validation sample $y_{(t)}$, provided that the likelihood can be decomposed using eq(5), eq(8) gives the marginal likelihood:

$$[y_{(t)}\,\big|M,y_{(-t)}\,]^{-1} = \int [y_{(t)}\,|M,\phi]^{-1}[\phi\,\big|M,y_{(-t)},y_{(t)}\,]d\phi = \int [y_{(t)}\,|M,\phi]^{-1}[\phi\,|M,y]d\phi \tag{9}$$

Varying the partition index t in the computation of the Bayes factor $BF_{ij,(t)} = \frac{[y_{(t)}|M_i\,,y_{(-t)}]}{[y_{(t)}|M_j\,,y_{(-t)}]}$ (see definition (3)) does not induce any additional numerical cost : indeed eq (9) can be approximated for various t, by the following expression analogous to (8), but with the same replicates $\phi^{(h)}$ drawn once from the complete posterior $[\phi\,|M,y]$:

$$[y_{(t)}\,|M\,] \approx \left(\frac{1}{H}\sum_{h=1}^{H}[y_{(t)}\,\Big|\phi^{(h)},M\,]^{-1}\right)^{-1} \tag{10}$$

Acknowledgements We are grateful to Etienne Rivot and Liliane Bel for their constructive suggestions and to Jacques Bernier for pointing out the possible ecological interpretation of the LOL. Financial support was provided by AgroParisTech for Sophie Ancelet to work overseas with ecologists from Fisheries and Oceans Canada in Moncton, N.B.

Biographical Sketches

Sophie Ancelet is a PhD student in applied statistics at the academic institution AgroParisTech from the french Ministry of Agriculture (http://www.engref.fr). She belongs to the research team for Environmental Statistics named "Modelisation et Risque en Statistique Environnementale" (MORSE) of the UMR AgroParisTech/INRA 518 "Mathématiques et Informatique Appliquées". Her thesis, supervised by Eric Parent, deals with hierarchical spatial modelling and Bayesian inference applied to environmental management and population biology. She also supervises

students during tutorial works in statistics and probability at the University of Paris 12, Créteil, France.

Marie-Pierre Etienne is a Senior Lecturer in applied statistics for ecology, biodiversity and environment in AgroParisTech. She belongs to the research team MORSE. Her main research subjects are environmental spatial statistics and genomics with a particular interest for hierarchical modelling and Bayesian inference.

Hugues Benoît is a fisheries ecologist working for the federal department of Fisheries and Oceans Canada at the Gulf Fisheries Centre in Moncton, NB. His research interests include trying to understand the direct and indirect impacts of harvesting and environmental change on marine communities.

Eric Parent is a Professor in applied statistics and probabilistic modelling for environmental engineering in AgroParisTech. He is the director of the research group MORSE. His broader interests include " Bayesian Statistics at work", statistical modelling for environmental sciences with a particular interest for spatial statistics, extremes and their decisional aspects. He co-authored two books (in French), one on Bayesian statistics for environmental engineering and the other on theoretical and algorithmic aspects of Bayesian theory.

References

1. Aitchison J. , Brown J.A.C. (1957), The Lognormal Distribution, Cambridge University Press
2. Banerjee S., Carlin B., Gelfand A. (2004), Hierarchical Modeling and Analysis for Spatial Data, Chapman and Hall/ CRC, In: Chapter 4:98-120
3. Bernier J., Fandeux D. (1970), Théorie du renouvellement- Application à l'étude statistique des précipitations mensuelles, Revue de statistique appliquée, 18:75-87
4. Congdon P. (2001), Bayesian Statistical Modelling, Wiley, Collection Wiley Series in Probability and Statistics
5. Gelfand A.E., Ghosh S. (1998), Model choice: A minimum posterior predictive loss approach, Biometrika, 85:1-11
6. Gelman A., Rubin D.B. (1992), Inference from Iterative Simulation Using Multiple Sequences, Statistical Science, 7:457-72
7. Hall D.B (2000), Zero-Inflated Poisson and binomial regression with Random Effects: A Case Study, Biometrics, 56:1030-1039
8. Heilbron D.C (1994), Zero-altered and other regression models for count data with added zeros, Biometrical Journal, 36:531-547
9. Hurlbut T., Clay D. (1990), Protocols for research vessel cruises within the Gulf Region (demersal fish)(1970-1987), Canadian Manuscript Report of Fisheries and Aquatic Sciences, 2082:2143
10. Kass R.E., Raftery A.E. (1994), Bayes factors, Journal of the American Statistical Association, 90:773-795
11. Lambert D. (1992), Zero-inflated Poisson regression, with an application to defects in manufacturing., Technometrics, 34:1-4
12. Loring D.H., Nota D.J.G. (1973), Morphology and sediments of the Gulf of St.Lawrence., Bulletin of the Fisheries Research Board of Canada, 182:147
13. Martin T.G. et al. (2005), Zero tolerance ecology: improving ecological inference by modelling the source of zero observations, Ecology Letters, 8:1235-1246
14. Metropolis N. et al. (1953), Equations of State Calculations by Fast Computing Machines, Journal of Chemical Physics, 21:1087-1091
15. Myers R.A., Pepin P. (1990), The robustness of lognormal based estimators of abundance, Biometrics, 46:1185-1192
16. Pennington M. (1996), Estimating the mean and variance from highly skewed marine data, Cambridge University Press, 94:498-505
17. Spiegelhalter D.J. et al. (2000), WinBUGS Version 1.3. User Manual, MRC Biostatistics Unit
18. Spiegelhalter D.J. et al. (2002), Bayesian measures of model complexity and fit, Journal of the Royal Statistical Society Serie B, 64:1-34
19. Stefansson G. (1996), Analysis of groundfish survey abundance data: combining the GLM and delta approaches, ICES Journal of Marine Science, 53:577-588
20. Tanner M.H. (1992), Tools for Statistical Inference : Observed Data and Data Augmentation Methods, Springer-Verlag,New York
21. Welsh A.H. et al. (1996), Modelling the abundance of rare species: statistical counts with extra zeros, Ecological Modelling, 88:297-308

Chapitre 7

Tenir compte de relations inter-espèces avec un champ géostatistique bivarié latent défini par convolution de fonctions noyaux

Le travail présenté dans ce chapitre a été réalisé en collaboration avec Vincent Garreta, doctorant au CEREGE d'Aix-en-Provence, dans le cadre du réseau "Modèles hiérarchiques spatiaux" organisé par le département INRA MIA.

Tous les modèles hiérarchiques spatiaux spécifiés dans les chapitres précédents ont, dans leur couche phénoménologique latente, un champ de Markov caché induisant, pour un effet aléatoire donné, des dépendances spatiales entre sites (e.g., modèle de Potts) ou unités géographiques (e.g., modèle IAR) liés par une relation de voisinage. Bien évidemment, nous sommes loin d'avoir parcouru toutes les possibilités offertes par l'approche hiérarchique pour la modélisation de structures spatiales latentes. Les modèles géostatistiques (cf chapitre 1) et autres structures spatiales multivariées peuvent également être introduits dans la couche cachée des modèles hiérarchiques (Banerjee & al., 2004). L'intérêt de ce chapitre est avant tout méthodologique. Il s'agit d'élargir la panoplie des structures spatiales possibles pour modéliser, à un niveau caché, la distribution d'organismes vivants.

Pour cela, je propose de franchir un pas supplémentaire dans la modélisation de quantités de biomasse *zero-inflated* et spatialement structurées. L'objectif du chapitre est de présenter un modèle hiérarchique permettant de décrire la répartition spatiale conjointe de deux espèces et ce, à une échelle spatiale plus fine que celle des unités géographiques. Par rapport aux modèles hiérarchiques construits dans les chapitres précédents, ce modèle a pour particularité de posséder, dans sa couche latente, un champ bivarié gaussien construit par convolution de bruits blancs gaussiens avec des noyaux de lissage.

L'utilisation d'un modèle bivarié permet de répondre aux hypothèses écologiques selon lesquelles la répartition des organismes vivants est régie par des relations inter-espèces (e.g., compétition, relation proie-prédateur,..) et des processus biologiques induisant de la structuration spatiale à petite échelle. D'un point de vue statistique, l'un des intérêts majeur à l'utilisation d'une structure bivariée est de pouvoir réaliser des transferts d'information entre des données relatives à deux espèces afin d'améliorer la performance globale des estimations de biomasse par rapport à un traitement séparé de ces deux espèces.

Ce chapitre s'articule en quatre parties. Dans la première partie, je décris brièvement

le principe de la modélisation spatiale bivariée et souligne les principales faiblesses de deux classes de modèles usuels : les modèles séparables et le modèle linéaire de corégionalisation. La deuxième partie se concentre autour d'une méthode peu utilisée pour la contruction de champs gaussiens : l'approche par convolution initialement développée par Thiébaux et Pedder (1987). Nous verrons que cette méthode gagne à être connue de par sa grande flexibilité pour la modélisation de dépendances spatiales et sa capacité à s'accommoder de "gros" jeux de données. La troisième partie est consacrée à la présentation d'un modèle hiérarchique pour la modélisation de données de biomasse *zero-inflated*, spatialisées et bivariées. Ce modèle est le fruit de la combinaison d'une version continue de la loi des fuites avec un champ gaussien bivarié latent construit par convolution. L'inférence bayésienne de ce modèle est actuellement en cours d'implémentation. Aucune simulation ni analyse sur données réelles n'ont donc été réalisées pour le moment. Dans la partie 4, je liste quelques pistes de recherche envisagées pour la poursuite de ce projet.

7.1 Le point sur la modélisation spatiale bivariée

7.1.1 Principe général

Soit $\psi(s) = (\psi_1(s), \psi_2(s))^T$ un vecteur à 2 composantes aléatoires spatialisées définies au site s d'un domaine d'étude D. Dans notre cas, 2 désigne le nombre d'espèces d'intérêt. Une étape fondamentale dans la modélisation du processus spatial bivarié $\psi = \{\psi(s); s \in D\}$ consiste à spécifier la matrice de covariance croisée C(s,s') de dimension 2×2 et définie par $C_{i,i'}(s, s') = cov(\psi_i(s), \psi_{i'}(s'))$ $(i = 1, 2,\ i' = 1, 2)$. Celle-ci doit permettre de générer deux formes de dépendances : l'une entre les variables aléatoires associées à un même site et l'autre entre les variables associées à plusieurs sites distincts.

Comme dans le cas univarié (cf section 1.2.2), les processus gaussiens sont parcimonieux et simples à manipuler et, par conséquent, largement utilisés pour décrire des processus spatiaux bivariés (Banerjee & al., 2004). Ils sont entièrement définis par leur moyenne et la matrice de covariance croisée $C(s, s')$. Ainsi, dans ce chapitre, je considère uniquement des processus spatiaux bivariés gaussiens. Par ailleurs, je me limite au cas de processus spatiaux isotropes : C(s,s') dépend de s et s' via la distance géographique séparant ces deux sites $\|s - s'\|$.

En pratique, il est extrêmement difficile de spécifier directement une matrice de covariance croisée $C(s, s')$ valide. En effet, cela nécessite que, pour tout entier naturel n et tout choix de n sites dans D, la matrice de covariance globale $np \times np$ induite soit définie-positive (Banerjee & al., 2004). Plusieurs approches ont ainsi été développées pour construire des fonctions de covariance croisée valides à partir de fonctions de covariance standards. Les deux paragraphes suivants présentent deux approches usuelles.

7.1.2 Les modèles séparables

Soit ρ une fonction de corrélation valide pour un processus spatial univarié stationnaire. Soit T une matrice de covariance 2×2 définie-positive décrivant les dépendances entre les 2 processus évalués au site s. Dans ce cas, une matrice de covariance croisée de la forme :

$$C(s, s') = \rho(s - s')T \tag{7.1.1}$$

est dite de forme séparable : les dépendances spatiales entre sites, contrôlées par ρ, sont modélisées séparément des dépendances entre les p processus d'intérêt, contrôlées par T.

La matrice de covariance du processus bivarié ψ issu de 2 types de processus évalués en n sites se déduit facilement de 7.1.1. Elle s'écrit :

$$\Sigma_\psi = H \otimes T \tag{7.1.2}$$

où $(H)_{ij} = \rho(s_i - s_j)$ et $\otimes$ désigne le produit de Kronecker. Σ_ψ est définie positive car le produit de Kronecker de deux matrices définies positives est une matrice définie positive. Celle-ci est donc une matrice de covariance valide.

L'inférence des modèles séparables est facilitée par la relation 7.1.2. En effet, $\Sigma_\psi^{-1} = H^{-1} \otimes T^{-1}$ et $|\Sigma_\psi| = |H|^2|T|^n$ ce qui signifie que l'étape d'inférence nécessite d'inverser une matrice 2×2 et une matrice $n \times n$ plutôt qu'une matrice $2n \times 2n$. Néanmoins, ces modèles présentent deux faiblesses :

- La matrice de covariance croisée est symétrique, i.e., $C_{i,i'}(s, s') = C_{i',i}(s, s')$. Cette symétrie signifie que la covariance entre le processus i au site s et le processus i' au site s' est la même que la covariance entre le processus i' au site s et le processus i au site s'. Selon les applications, une telle hypothèse n'est pas toujours valide.
- Les 2 processus spatiaux mesurés en chaque site ont la même matrice de covariance H. Dans notre cas, une telle hypothèse reviendrait à supposer que les deux espèces d'intérêt ont un pattern spatial identique ce qui est difficilement justifiable en pratique.

7.1.3 Le modèle linéaire de corégionalisation

Le modèle linéaire de corégionalisation (LMC), à l'origine développé par Grzebyk et Wackernagel (1994), permet de combler les lacunes des modèles séparables. Le terme "corégionalisation" signifie que plusieurs variables aléatoires covarient ensemble sur un même domaine d'étude.

Soient $\omega_1, \omega_2, \ldots, \omega_I$ I processus spatiaux gaussiens univariés et indépendants de moyennes respectives $\mu_1, \mu_2, \ldots, \mu_I$ et de fonctions de covariance respectives $\rho_1, \rho_2, \ldots, \rho_I$. La version la plus simple du modèle LMC suppose qu'en un site s donné, le processus spatial bivarié $\psi(s)$ peut s'écrire comme la combinaison linéaire suivante :

$$\psi(s) = A\omega(s) \tag{7.1.3}$$

où $\omega(s) = (\omega_1(s), \omega_2(s), \ldots, \omega_I(s))^T$ et A est une matrice $2 \times I$ triangulaire supérieure. En pratique, pour réduire la dimension du modèle, le nombre I de processus gaussiens univariés est choisi tel que : $I \leq 2$.

Dans notre cas, l'utilisation d'un tel modèle revient à supposer que les répartitions spatiales des deux espèces d'intérêt sont contrôlées par des champs gaussiens communs. Ceci permet de générer des ressemblances ou dissemblances entre les patterns de distribution des deux espèces. La contribution de chaque champ gaussien ω_i (i=1,..,I) sur la répartition spatiale de chaque espèce est prise en compte avec la matrice A qui contient les poids intervenant dans la combinaison linéaire 7.1.3. Ces poids permettent de définir des répartitions spatiales différentes d'une espèce à l'autre.

L'espérance de $\psi(s)$ est donnée par $\mathbb{E}(\psi(s)) = A\mu$ avec $\mu = (\mu_1, \mu_2, \ldots, \mu_I)^T$ et la

matrice de covariance croisée 2×2 associée à $\psi(s)$ est définie par :

$$C(s, s') = \sum_{i=1}^{I} \rho_i(s - s')T_i \tag{7.1.4}$$

où $T_i = a_i a_i^T$ et a_i désigne la i-ième colonne de la matrice A. Ainsi, la combinaison linéaire 7.1.3 définit des processus spatiaux stationnaires d'ordre 2.

Le modèle LMC définit une classe de modèles riche et flexible pour la modélisation de processus spatiaux bivariés. En particulier, il permet de combiner des processus spatiaux caractérisés par des fonctions de covariance différentes ce qui permet de créer des dépendances entre des processus de portée différentes. A noter que lorsque $\rho_1 = \rho_2 = \ldots = \rho_I$, la matrice de covariance croisée a la forme séparable donnée par l'expression 7.1.1 avec $T = \sum_{i=1}^{I} T_i$. Ainsi, les modèles séparables forment un cas particulier des modèles LMC.

La loi jointe du processus spatial bivarié ψ est définie par :

$$[\psi|\{\rho_j\}, T, \mu] = \mathcal{N}(\mu, \sum_{i=1}^{I}(H_i \otimes T_i)) \tag{7.1.5}$$

où H_i est la matrice $n \times n$ telle que $(H_i)_{k,k'} = \rho_i(s_k, s_{k'})$.

L'inférence de la structure jointe 7.1.5 pose un certain nombre de difficultés techniques. En particulier, plus le nombre d'observations est important plus les techniques d'inférence (classiques et bayésiennes) sont coûteuses en temps de calcul (Banerjee & al., 2004). Dans la section suivante, je décris une approche alternative pour la construction de champs spatiaux bivariés qui, nous le verrons, a l'avantage de faciliter l'inférence de "gros" jeux de données.

7.2 L'approche par convolution

Une troisième approche a récemment été proposée par Higdon (2002) pour modéliser des processus spatiaux bivariés. Il propose de construire un champ gaussien bivarié à 2 composantes en convolant des bruits blancs gaussiens avec des noyaux de lissage spécifiques à chacun des 2 processus d'intérêt. Dans cette section, je décris l'approche par convolution pour la construction de champs gaussiens dans le cas univarié. Les concepts présentés serviront ensuite de bases pour modéliser, à un niveau latent, la répartition spatiale conjointe de deux espèces.

7.2.1 Construire un champ gaussien univarié par moyennes mobiles

La fonction de covariance d'un champ gaussien univarié ϕ défini sur un domaine spatial $D \in \mathbb{R}^2$, est généralement choisie directement dans la panoplie des fonctions de covariance valides (cf section 1.2.2).

Une approche constructive alternative, initialement développée par Barry et Ver Hoef (1996), est de construire ce champ gaussien en convolant un bruit blanc gaussien continu x défini sur D avec un noyau de lissage κ :

$$\phi(s|x) = \kappa * x(s) = \int_D \kappa(\omega - s)x(\omega)d\omega \qquad \text{pour } s \in D \tag{7.2.6}$$

où $*$ désigne l'opérateur de convolution. Le champ ϕ est construit par moyennes mobiles : en chaque site s, il peut être vu comme la moyenne des valeurs de x sur D pondérées par le noyau κ. En pratique, κ définit une fonction décroissante de la distance géographique séparant deux sites. Ainsi, en chaque site s, les valeurs de x associées à des sites proches de s auront une influence plus forte sur la valeur de $\phi(s|x)$ par rapport aux valeurs associées à des sites géographiquement éloignés.

La fonction de covariance du champ gaussien ϕ se déduit facilement de l'expression 7.2.6. Elle dépend uniquement du noyau de lissage κ et est donnée par :

$$\begin{aligned} cov(\phi(s|x), \phi(s'|x)) &= \int_D \kappa(\omega - s)\kappa(\omega - s')d\omega \\ &= \int_D \kappa((\omega - s + s') - s')\kappa(\omega - s')d\omega \\ &= \int_D \kappa(u - (s - s'))\kappa(u)du \qquad \text{changement de variable } u = \omega - s' \\ &= \kappa * \kappa(d) \qquad \text{avec d=s-s'} \\ &= c(d) \end{aligned}$$

La fonction de covariance entre deux sites s et s' dépend uniquement de la distance séparant ces deux sites : tout processus gaussien construit par convolution d'un bruit blanc avec un noyau de lissage est donc stationnaire d'ordre 2. Par ailleurs, si le noyau κ est isotrope, intégrable et de carré intégrable alors la fonction de covariance induite est également isotrope (Higdon, 2002).

La figure 7.1 représente trois noyaux isotropes classiquement utilisés pour lisser un bruit blanc :

- le noyau gaussien, régulier à l'origine, pour lequel les dépendances spatiales entre sites diminuent lentement avec la distance :

$$\kappa(d) \propto \exp\left(-\frac{d^2}{2\alpha^2}\right) \tag{7.2.7}$$

- le noyau exponentiel, très piqué à l'origine, pour lequel les dépendances spatiales entre sites diminuent exponentiellement avec la distance :

$$\kappa(d) \propto \exp\left(-\frac{d}{\alpha}\right) \tag{7.2.8}$$

- le noyau disque, défini comme une fonction constante sur un disque de rayon α centré en un point d'intérêt :

$$\kappa(d) \propto \begin{cases} \frac{1}{\sqrt{\alpha^2\pi}} & \text{si} \quad d \leq \alpha \\ 0 & \text{si} \quad d > \alpha \end{cases} \tag{7.2.9}$$

Dans tous ces noyaux de lissage, le paramètre α s'interprète comme un paramètre de portée gérant la distance d'influence spatiale du noyau sur les sites d'observation.

En pratique, il est intéressant de pouvoir relier les paramètres d'un noyau de lissage aux paramètres de covariance classiques : palier, portée (cf section 1.2.2). En effet, ces derniers sont en général plus facilement interprétables. La portée désigne la distance à partir de laquelle il n'y a presque plus de dépendance entre deux sites et le palier peut être vu

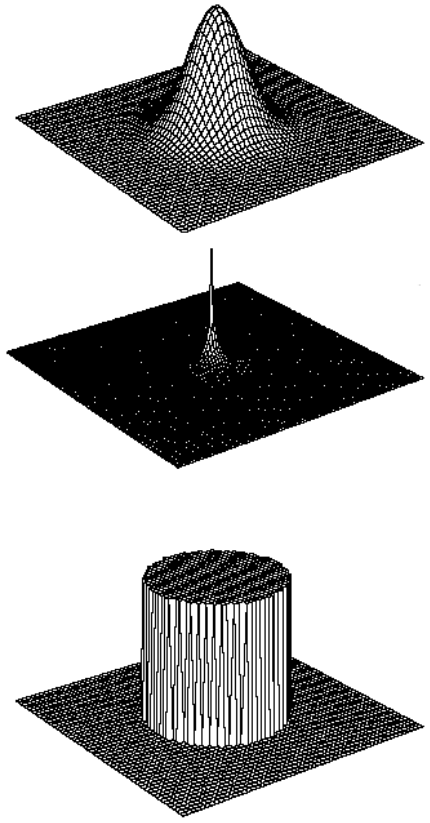


Figure 7.1 – Trois noyaux de lissage spatiaux isotropes classiques. De haut en bas : un noyau gaussien, un noyau exponentiel et un noyau disque

comme la variance de la structure spatiale. La relation entre κ et la fonction de covariance c(d) est donnée par le théorème de convolution des transformées de Fourier * (Pinkus & Zafrany, 1997). La transformée de Fourier d'un produit de convolution est donnée par le produit des transformées de Fourier. Ainsi, c(d) est la transformée de Fourier inverse † du carré de la transformée de Fourier de $\kappa(d)$. La construction par moyenne mobile du champ gaussien ϕ garantit que sa fonction de covariance soit définie positive.

Dans le cas isotrope, chaque noyau $\kappa(d)$ correspond à une unique fonction de covariance c(d) et vice versa.

– le noyau gaussien induit une fonction de covariance gaussienne :

$$c(d) \propto e^{-\frac{d^2}{4\alpha^2}}$$

– le noyau exponentiel induit une fonction de covariance exponentielle :

$$c(d) \propto e^{-\frac{d}{\alpha}}$$

– le noyau disque induit une fonction de covariance de la forme :

$$c(d) = \frac{I_{(d\leq 2\alpha)}}{\alpha^2.\pi}\left(2.\alpha^2.\cos^{-1}(d/2\alpha) - d.\sqrt{\alpha^2-(d/2)^2}\right)$$

∗. La transformée de Fourier $\mathcal{F}$ d'une fonction intégrable f est : $\mathcal{F}(f) : \xi \mapsto \widehat{f}(\xi) = \int_{-\infty}^{+\infty} f(x)e^{-i\xi x}dx$

†. La transformée de Fourier inverse $\mathcal{F}^{-1}$ est définie par : $\mathcal{F}^{-1}(f) : x \mapsto f(x) = \frac{1}{2\pi}\int_{-\infty}^{+\infty} \widehat{f}(\xi)e^{i\xi x}d\xi$

Cette fonction de covariance est semblable à la covariance sphérique avec un comportement linéaire à l'origine et une forme "aplatie" lorsqu'on s'approche de la portée.

Cette relation n'est plus vraie si κ est anisotrope : dans ce cas, différents noyaux peuvent conduire à la même fonction de covariance (Higdon, 2002). La relation entre le noyau de lissage κ et la fonction de covariance est détaillée dans (Barry & VerHoef, 1996) et (Kern, 2000).

7.2.2 Approximation d'un champ gaussien par convolution discrétisée sur grille latente

Higdon (1998) propose d'approcher un champ gaussien en utilisant une version discrète de l'approche par convolution. Au lieu de définir le processus x sous-jacent comme un bruit blanc continu, celui-ci peut être défini sur une grille latente recouvrant le domaine d'étude D. Ce bruit blanc discret peut être convolé avec un noyau de lissage afin de générer une approximation du champ gaussien d'intérêt. Nous verrons dans les sections suivantes que cette approche discrétisée a de multiples avantages pratiques.

Considérons un domaine d'échantillonnage D recouvert par une grille régulière dont les noeuds $\{\omega_1, \omega_2, \ldots, \omega_M\}$ définissent des sites géographiques *conceptuels* où se manifeste un bruit blanc discret latent. Une approximation du champ gaussien ψ au site s s'obtient en discrétisant l'écriture 7.2.6 sous la forme :

$$\phi(s|x) = \sum_{i=1}^{M} \kappa(\omega_i - s)x(\omega_i) + \sigma\epsilon(s) \tag{7.2.10}$$

avec M le nombre de noeuds de la grille, $x(\omega_i) \overset{i.i.d}{\sim} \mathcal{N}(0,1)$, σ la variabilité résiduelle dûe à l'erreur d'approximation sur grille et $\kappa(\omega_i - .)$ un noyau de lissage centré en ω_i. Cette approche permet de construire une approximation d'un processus gaussien continu à partir d'un processus discret latent (cf figure 7.2).

Le choix de la taille de la grille latente est une étape importante dans la spécification du modèle. En effet, une grille trop grossière aura tendance à induire un lissage trop fort. Une grille trop fine empêche de bénéficier de la réduction de dimension offert par la discrétisation.

L'écriture 7.2.10 indique que le champ gaussien d'intérêt $\phi = \{\phi(s), s \in D\}$ est entièrement contrôlé par les valeurs du bruit blanc discrétisé x se manifestant aux M noeuds d'une grille régulière. Ainsi, cette approche permet de modéliser un champ gaussien avec un nombre réduit de paramètres. Ceci permet de diminuer considérablement les temps d'inférence et de faciliter les inversions matricielles induites lors du traitement de champs gaussiens (e.g., krigeage) en un nombre important de sites. Les problèmes de dimension étant fréquents en statistique spatiale multivariée, l'approche par convolution sur grille discrète est une solution possible pour palier à ces difficultés. C'est pourquoi nous avons choisi cette approche pour modéliser la répartition spatiale conjointe de deux espèces.

De nombreux travaux ont déjà été réalisés autour de l'approche par convolution : (Barry & VerHoef, 1996) et (Kern, 2000) discutent du choix du noyau de lissage. (Ickstadt & Wolpert, 1999) proposent d'utiliser cette méthode pour définir des processus spatiaux non-gaussiens par la convolution de processus de Lévy.

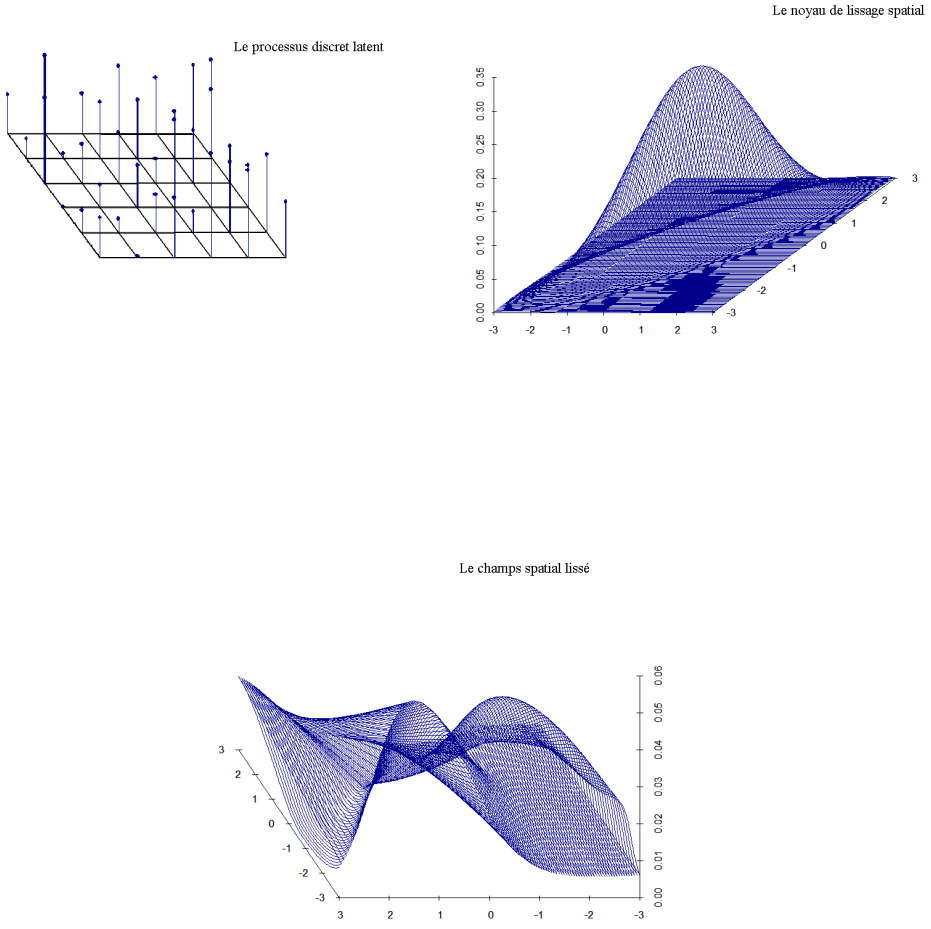


Figure 7.2 – Approcher un processus gaussien continu en convolant un noyau de lissage et un bruit blanc discrétisé latent

7.3 Une version hiérarchique continue et bivariée du modèle LOL

7.3.1 Notations et idée générale

Soit D un domaine d'échantillonnage. Nous notons $Y(s) = (Y_1(s), Y_2(s))$ le vecteur aléatoire des quantités de biomasse évaluées en un site s de D pour deux espèces d'intérêt. L'objectif est de modéliser le champ bivarié $Y = \{Y(s), s \in D\}$ de telle sorte que chacun des processus $Y_1 = \{Y_1(s), s \in D\}$ et $Y_2 = \{Y_2(s), s \in D\}$ induise des dépendances spatiales entre sites géographiquement proches et qu'en chaque site s, les quantités $Y_1(s)$ et $Y_2(s)$ covarient. Nous supposons disposer d'une réalisation partielle du champ bivarié Y en n sites d'échantillonnage $s_1, s_2, .., s_n$.

Nous nous plaçons de nouveau dans la situation où les quantités de biomasse recueillies sont *zero-inflated*. Pour mettre en oeuvre l'approche par convolution décrite dans la section 7.2, nous définissons $\mu_l(s)$ comme le nombre moyen de *gisements* que l'on peut espérer collecter pour l'espèce l (l=1,2) en un site s. Puis, nous proposons de lier les $\mu_1(s)$ et $\mu_2(s)$ relatifs à chaque espèce via un champ gaussien bivarié contruit par convolution sur grille discrète.

7.3.2 Le processus interne : un champ gaussien bivarié régit la répartition spatiale conjointe des espèces

Nous supposons qu'un champ gaussien bivarié latent $Z = \{(Z_1(s), Z_2(s)); s \in D\}$ régit la répartition spatiale conjointe des deux espèces d'intérêt dans D. Bien que plusieurs approches existent pour modéliser Z, nous avons choisi d'utiliser l'approche par convolution car :

- cela permet d'exploiter une méthode peu connue mais prometteuse pour la modélisation de structures spatiales bivariées
- grâce à sa forme discrétisée, cette approche permet de palier aux fréquents problèmes d'inférence liés à la grande dimension des modèles spatiaux bivariés

Soient $X_1 = \{X_1(s), s \in D\}$ et $X_2 = \{X_2(s), s \in D\}$ deux bruits blancs gaussiens continus et indépendants. En chaque site s, le champ gaussien Z est construit par convolution et par mélange à partir de ces 2 champs générateurs :

$$Z_1(s) = \int_D \kappa_{\theta_1}(s-w)x_1(w)dw \tag{7.3.11}$$

$$Z_2(s) = \frac{\zeta}{\sqrt{\zeta^2 + (1-|\zeta|)^2}} \int_D \kappa_{\theta_2}(s-w)x_1(w)dw + \frac{(1-|\zeta|)}{\sqrt{\zeta^2 + (1-|\zeta|)^2}} \int_D \kappa_{\theta_2}(s-w)x_2(w)dw) \tag{7.3.12}$$

Dans cette expression :

- ζ caractérise la force de liaison entre les champs Z_1 et Z_2 et est à valeurs dans $[-1, 1]$. Selon son signe, elle crée une dépendance positive ou négative entre les deux espèces. Si ζ vaut 1 ou -1, toute la variabilité spatiale du champ Z_2 est expliquée par X_1. Les champs Z_1 et Z_2 auront tendance à présenter de fortes ressemblances si $\zeta = 1$ et de fortes dissemblances si $\zeta = -1$. Si $\zeta = 0$, toute la variablité spatiale de Z_2 est expliquée par X_2 : les champs Z_1 et Z_2 sont alors indépendants.

– κ_{θ_1} et κ_{θ_2} désignent deux noyaux de convolution paramétriques, spécifiques à chaque espèce.

La structure spatiale bivariée ci-dessus, définie par les équations 7.3.11 et 7.3.12, s'inspire de l'approche par corégionalisation (cf section 7.1.3). Ainsi, Z_2 s'écrit comme la combinaison linéaire de deux champs gaussiens : l'un permet de modéliser la part de variabilité expliquée par la structure spatiale de l'espèce 1 et l'autre permet de modéliser la part de variabilité spatiale résiduelle spécifique à l'espèce 2.

La différence entre l'approche par corégionalisation et l'approche par convolution apparaît dans l'origine de la dépendance entre les deux champs Z_1 et Z_2. Dans l'approche par corégionalisation, cette dépendance provient du fait que la variabilité spatiale des deux champs est expliquée par les mêmes processus spatiaux. Dans l'approche par convolution, la dépendance entre Z_1 et Z_2 est induite par le bruit blanc gaussien X_1 qui contrôle simultanément ces deux processus (cf figure 7.3). La variabilité de chaque champ est ainsi expliquée par des processus spatiaux différents.

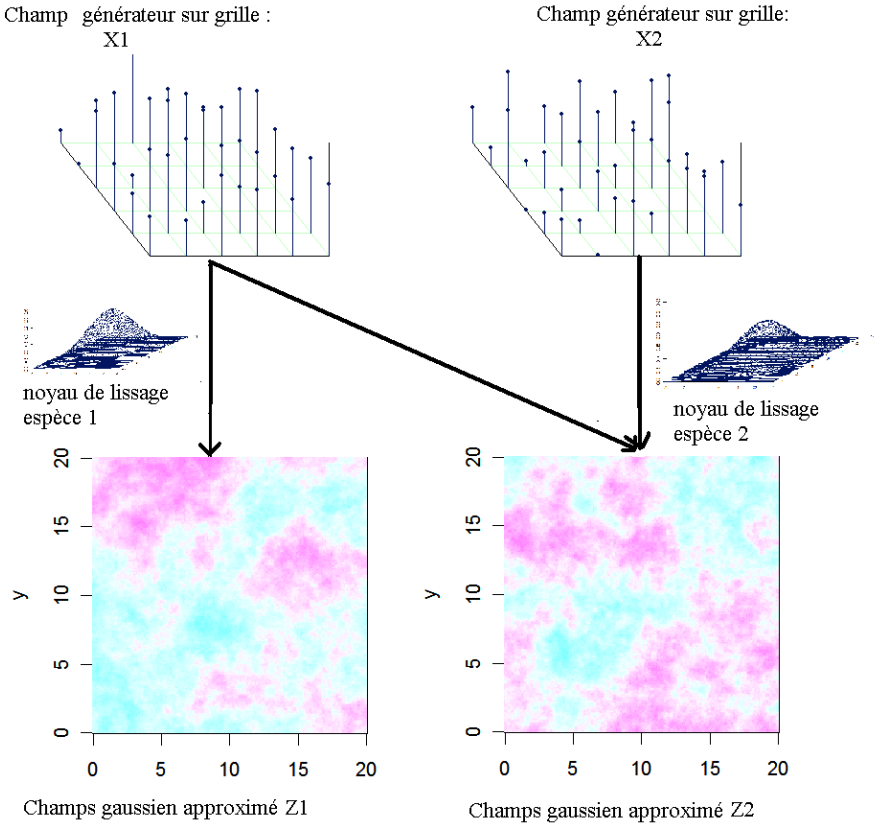


Figure 7.3 – Le champ générateur sur grille X_1, commun à l'expression des champs Z_1 et Z_2, assure la corrélation entre ces deux processus.

Les matrices de covariance intra et inter champs associées aux équations (7.3.11) et (7.3.12) sont calculables analytiquement :

$$C_{ii}(s,s') = \int_{R^2} \kappa_{\theta_i}(\omega)\kappa_{\theta_i}(\omega - d(s,s'))d\omega \tag{7.3.13}$$

$$C_{ij}(s,s') = \frac{\zeta}{\sqrt{\zeta^2 + (1-|\zeta|)^2}} \int_{R^2} \kappa_{\theta_i}(\omega)\kappa_{\theta_j}(\omega - d(s,s'))d\omega \tag{7.3.14}$$

avec d(s,s') désignant la distance entre les sites s et s'.

Malheureusement, l'inférence d'une telle structure est souvent très coûteuse en temps de calcul car la dimension de la couche latente est égale à 2n où n désigne le nombre de sites d'échantillonnage. Aussi, afin de spécifier un modèle opérationnel pour le traitement de "gros" jeux de données, nous proposons d'approximer les convolutions des bruits blancs gaussiens générateurs X_1 et X_2 sur une grille latente régulière (cf section 7.2.2).

Soient $\omega_1, \omega_2, ..., \omega_m$ les m noeuds d'une grille régulière définissant un support commun sur D pour les bruits blancs discrétisés X_1 et X_2. Avec un léger abus de notation, nous noterons : $X_l = (X_l(\omega_1), X_l(\omega_2), ..., X_l(\omega_j), ..., X_l(\omega_m)) = (X_l^1, ..., X_l^j, ..., X_l^m)$ pour l=1,2.

Les équations (7.3.11) et (7.3.12) définissant les champs gaussiens latents Z_1 et Z_2 se discrétisent sous la forme :

$$Z_1^i = \sum_{j=1}^{m} \kappa_{\theta_1}(i,j) X_1^j + \sigma_1 \varepsilon_1^i \tag{7.3.15}$$

$$Z_2^i = \frac{\zeta}{\sqrt{\zeta^2 + (1-|\zeta|)^2}} \sum_{j=1}^{m} \kappa_{\theta_2}(i,j) X_1^j + \frac{(1-|\zeta|)}{\sqrt{\zeta^2 + (1-|\zeta|)^2}} \sum_{j=1}^{m} \kappa_{\theta_2}(i,j) X_2^j + \sigma_2 \varepsilon_2^i \tag{7.3.16}$$

avec i=1,2,....n où n désigne le nombre de sites d'observation.

La discrétisation de la convolution en une somme finie est complétée par un aléa d'erreur d'approximation $\sigma_l \varepsilon_l$ propre à chaque espèce l=1,2 :

$$\varepsilon_l^i \overset{i.i.d}{\sim} \mathcal{N}(0,1)$$

Si $\sigma_1 = \sigma_2$ alors cela signifie que les deux erreurs d'approximation ont la même amplitude. Les vecteurs aléatoires ε_1 et ε_2 sont supposés indépendants.

Les covariances intra et inter champs ainsi que les variances respectives des champs Z_1 et Z_2 ont finalement la forme simplifiée suivante :

$$C_{12}(i,j) = \frac{\zeta}{\sqrt{\zeta^2 + (1-|\zeta|)^2}} \sum_{k=1}^{m} \kappa_{\theta_1}(i,k) \kappa_{\theta_2}(j,k)$$

$$C_{11}(i,j) = \sum_{k=1}^{m} \kappa_{\theta_1}(i,k) \kappa_{\theta_1}(j,k) \qquad \text{si i} \neq \text{j}$$

$$C_{22}(i,j) = \sum_{k=1}^{m} \kappa_{\theta_2}(i,k) \kappa_{\theta_2}(j,k) \qquad \text{si i} \neq \text{j}$$

$$Var(Z_1(i)) = \sum_{k=1}^{m} \kappa_{\theta_1}(i,k)^2 + \sigma_1^2$$

$$Var(Z_2(i)) = \sum_{k=1}^{m} \kappa_{\theta_2}(i,k)^2 + \sigma_2^2$$

Construire les champs gaussiens Z_1 et Z_2 par convolution discrète implique de définir des versions discrétisées des noyaux gaussien (cf eq.7.2.7), exponentiel (cf eq. 7.2.8) et disque (cf eq.7.2.9). Nous proposons les formes discrétisées suivantes :

– le noyau gaussien

$$\kappa_\theta(i,j) = \beta exp\left(-\frac{d(i,j)^2}{2\alpha^2}\right) \tag{7.3.17}$$

– le noyau exponentiel

$$\kappa_\theta(i,j) = \beta exp\left(-\frac{d(i,j)}{\alpha}\right) \tag{7.3.18}$$

– le noyau disque

$$\kappa_\theta(i,j) \quad = \quad \begin{cases} \frac{\beta}{\sqrt{n_{i+}}} \text{ si } d(i,j) < \alpha \\ 0 \text{ si } d(i,j) \geq \alpha \end{cases} \tag{7.3.19}$$

où i désigne un site d'observation (i=1,2,..,n), j désigne un noeud de la grille (j=1,2,..,m) et n_{i+} désigne le nombre de noeuds de la grille latente situés à une distance inférieure à α du site d'observation i. Les paramètres β caractérisent la force de contribution des champs générateurs sur grille sur les champs continus approchés par convolution.

7.3.3 Le modèle des observations

En chaque site s de D, nous supposons que les quantités de biomasse relatives à chaque espèce sont issues d'une loi des fuites dont les paramètres sont à la fois spécifiques au site et à l'espèce considérée :

$$Y_1(s) \sim \mathcal{F}uite(\mu_1(s), \rho_1(s))$$
$$Y_2(s) \sim \mathcal{F}uite(\mu_2(s), \rho_2(s))$$

Dans l'expression ci-dessus, $\mu_l(s)$ et $\rho_l(s)$ (l=1,2) désignent respectivement le nombre moyen de *gisements* et l'inverse de la quantité moyenne de biomasse par *gisement* que l'on peut espérer collecter pour l'espèce l au cours d'un effort d'échantillonnage standard réalisé autour du site s. Contrairement au chapitre précédent où les effets aléatoires μ et ρ étaient supposés varier sur une grille, ceci sont désormais supposés varier continûment dans le domaine d'étude D.

Nous proposons d'introduire les dépendances spatiales intra et inter-espèces au niveau de l'intensité du nombre de *gisements* via la fonction de lien log :

$$log(\mu_1(s)) = m_1(s) + Z_1(s)$$
$$log(\mu_2(s)) = m_2(s) + Z_2(s)$$

Dans cette expression :
- $m_l(s)$ désigne la tendance pour l'espèce l au site s
- $Z = \{Z_1(s), Z_2(s), s \in D\}$ est un champ gaussien bivarié centré approché par convolution sur grille discrète (cf eq. 7.3.15 et 7.3.16).

Les intensités moyennes de *gisements* sont supposées continûment structurées spatialement dans D. Par ailleurs, elles sont supposées covarier d'une espèce à l'autre.

Dans la suite, les tendances $m_l(s)$ (l=1,2) seront supposées constantes : $m_l(s) = m_l$. A noter qu'il ne serait pas difficile de lever cette hypothèse en expliquant les tendances en chaque site s par des covariables environnementales (e.g., profondeur, température, sédiments...).

Pour gagner en parcimonie, nous faisons l'hypothèse simple selon laquelle, pour chaque espèce, il n'y a pas de dépendances spatiales entre les quantités de biomasse moyennes

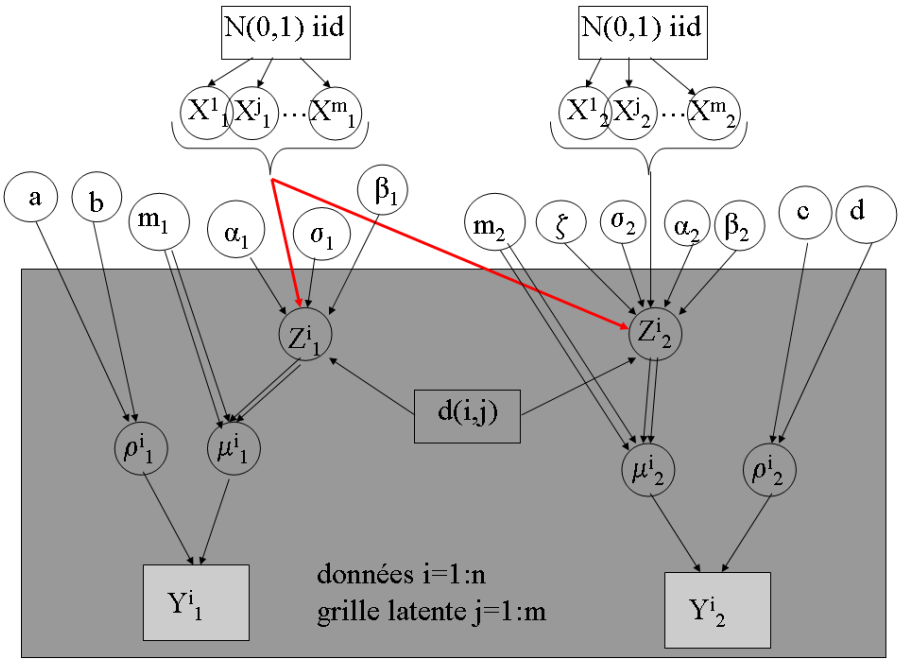


Figure 7.4 – DAG d'une version hiérarchique continue du modèle LOL basée sur un champ gaussien bivarié construit par approximation de convolutions de bruits blancs gaussiens sur une grille latente

contenues dans chaque *gisement*. Nous supposons que les champs $\rho_1 = \{\rho_1(s), s \in D\}$ et $\rho_2 = \{\rho_2(s), s \in D\}$ sont indépendants et distribués comme suit :

$$\rho_1(s) \overset{i.i.d}{\sim} Gamma(a, b)$$
$$\rho_2(s) \overset{i.i.d}{\sim} Gamma(c, d)$$

$\frac{a}{b}$ (resp. $\frac{c}{d}$) et $\frac{a}{b^2}$ (resp. $\frac{c}{d^2}$) représentent les tendances et variances respectives du champ ρ_1 (resp. ρ_2) sur le domaine D.

Le graphe acyclique direct de la Figure 7.4 fait apparaître les relations de dépendances entre les diverses grandeurs du modèle hiérarchique ainsi construit.

7.3.4 Inférence bayésienne

Nous envisageons de mener l'inférence du modèle hiérarchique de la figure 7.4 sous le paradigme bayésien. Les paramètres inconnus sont $\sigma_1, \sigma_2, \alpha_1, \alpha_2, \beta_1, \beta_2, m_1, m_2, \zeta, a, b, c, d$. Les variables latentes sont les bruits blancs gaussiens discrétisés X_1 et X_2, les champs gaussiens approximés Z_1 et Z_2 ainsi que les champs Gamma indépendants ρ_1 et ρ_2.

Compte-tenu des nombreuses propriétés d'indépendances conditionnelles mises en évidence dans le DAG de la figure 7.4, la loi jointe du modèle hiérarchique construit s'écrit :

$$\begin{aligned}
&[Y_1, Y_2, \rho_1, \rho_2, Z_1, Z_2, X_1, X_2, a, b, c, d, m_1, m_2, \sigma_1, \sigma_2, \alpha_1, \alpha_2, \beta_1, \beta_2, \zeta] \\
&= [Y_1 \,|Z_1, m_1, \rho_1] \times [Y_2 \,|Z_2, m_2, \rho_2] \times [\rho_1 \,|a, b] \times [\rho_2 \,|c, d] \times [X_1] \times [X_2] \\
&\times [Z_1 \,|X_1, \sigma_1, \alpha_1, \beta_1] \times [Z_2 \,|X_2, X_1, \sigma_2, \alpha_2, \beta_2, \zeta] \\
&\times [\sigma_1] \times [\sigma_2] \times [\alpha_1] \times [\alpha_2] \times [\beta_1] \times [\beta_2] \times [m_1] \times [m_2] \times [\zeta] \times [a] \times [b] \times [c] \times [d]
\end{aligned} \tag{7.3.20}$$

Par commodité, nous proposons :
– Des *priors* conjugués inverse-gamma pour les paramètres de variance :

$$\frac{1}{\sigma_l^2} \sim G(0.001, 0.001) \qquad l = 1, 2$$

– des *priors* uniformes sur un intervalle $I = [\theta_{min}, \theta_{max}]$ pour les paramètres de portée θ_1 et θ_2. Le choix des bornes de l'intervalle I est important, notamment dans le cas d'un noyau de lissage disque. En effet, il existe un $\theta_{\min}$ en dessous duquel tout point d'observation n'est plus voisin d'un point de la grille : le modèle approché n'a alors plus de sens. De même il faut limiter les θ à un $\theta_{\max}$ pour ne pas introduire une correlation spatiale excessive irréaliste entre des observations proches. Nous choisissons $\theta_{min} = \underset{i=1,..,nj=1,2,..,m}{max\ min}\ d(i,j)$ et $\theta_{max} = \underset{i=1,..,nj=1,..,m}{min\ max}\ d(i,j)$ où d(i,j) désigne la distance entre un site d'échantillonnage i et le jème noeud de la grille.
– les paramètres β_1 et β_2 de force de contribution des champs sur grille suivent des normales plates tronquées en zéro
– ζ suit une loi uniforme sur $[-1, 1]$.
– m_1 et m_2 suivent des lois normales plates
– a,b,c,d suivent des lois Gamma(0.01,0.01)

La conditionnelle complète associée à chaque variable (ou vecteur) du modèle se *lit* directement sur l'expression de la loi jointe 7.3.20 en cherchant tour à tour les composantes du membre de droite de cette équation qui dépendent de cette variable (ou vecteur). Pour simplifier les notations de conditionnement, on appelle $\langle - \langle \phi \rangle \rangle$, la partie complémentaire de ϕ dans l'ensemble des inconnues du problème. Les conditionnelles complètes des inconnues relatives à la première espèce sont :

$$\begin{aligned}
[m_1 \,|Y_1, Y_2, \langle - \langle m_1 \rangle \rangle] &\propto [Y_1 \,|Z_1, m_1, \rho_1] \times [m_1] \\
[\sigma_1 \,|Y_1, Y_2, \langle - \langle \sigma_1 \rangle \rangle] &\propto [Z_1 \,|X_1, \sigma_1, \alpha_1, \beta_1]] \times [\sigma_1] \\
[X_1 \,|Y_1, Y_2, \langle - \langle X_1 \rangle \rangle] &\propto [Z_1 \,|X_1, \sigma_1, \alpha_1, \beta_1] \times [X_1] \\
[\beta_1 \,|Y_1, Y_2, \langle - \langle \beta_1 \rangle \rangle] &\propto [Z_1 \,|X_1, \sigma_1, \alpha_1, \beta_1] \times [\beta_1] \\
[\alpha_1 \,|Y_1, Y_2, \langle - \langle \alpha_1 \rangle \rangle] &\propto [Z_1 \,|X_1, \sigma_1, \alpha_1, \beta_1] \times [\alpha_1] \\
[Z_1 \,|Y_1, Y_2, \langle - \langle Z_1 \rangle \rangle] &\propto [Y_1 \,|Z_1, m_1, \rho_1] \times [Z_1 \,|X_1, \sigma_1, \alpha_1, \beta_1] \\
[a \,|Y_1, Y_2, \langle - \langle a \rangle \rangle] &\propto [\rho_1 \,|a, b] \times [a] \\
[b \,|Y_1, Y_2, \langle - \langle b \rangle \rangle] &\propto [\rho_1 \,|a, b] \times [b]
\end{aligned}$$

Les équations pour les grandeurs inconnues associées à la deuxième espèce s'écrivent de façon symétrique à celles de la première en remplaçant $[Z_1|X_1, \sigma_1, \alpha_1, \beta_1]$ par $[Z_2|X_2, \sigma_2, \alpha_2, \beta_2, \zeta]$ et en considérant une équation de plus :

$$[\zeta \,|Y_1, Y_2, \langle - \langle \zeta \rangle \rangle] \propto [Z_2 \,|X_2, X_1, \zeta, \sigma_2, \alpha_2, \beta_2] \times [\zeta] \tag{7.3.21}$$

Nos choix des *priors* permettent d'hériter de nombreuses propriétés de conjugaison. Ainsi, la plupart des lois conditionnelles complètes ont une forme analytique connue :
– normale pour les inconnues m_1, m_2, X_1 et X_2
– normale tronquée pour les paramètres β_1 et β_2.

– gamma inverse pour les paramètres de variance σ_1 et σ_2
– gamma pour les paramètres b et d

L'inférence sur ces paramètres peut donc être menée avec un échantillonneur de Gibbs.

En ce qui concerne les paramètres α_1, α_2, a et c ainsi que les variables latentes Z_1 et Z_2, l'expression des lois conditionnelles complètes n'est connue qu'à une constante près. L'inférence sur ces paramètres peut donc être menée avec un algorithme de Metropolis-Hastings.

7.4 Perspectives

L'implémentation d'un algorithme MCMC hybride Metropolis-within-Gibbs est actuellement en cours. Il faudrait ensuite mener une étude par simulations afin de tester les performances du modèle sous différents scenarii :

– avec des noyaux de lissage différents
– avec des grilles régulières latentes plus ou moins fines
– avec des paramètres de simulation différents : portée et variance des champs Z_1 et Z_2, niveau d'intéraction ζ entre les deux champs

En appliquant ce modèles aux données de biomasse en invertébrés épibenthiques du Golfe-du-Saint-Laurent (cf section 1.1.2), l'objectif serait à terme de comparer les performances du modèle bivarié par rapport à un traitement séparé des deux espèces et d'estimer le degré de corrélation entre espèces.

Enfin, ce modèle hiérarchique spatial est généralisable pour modéliser la répartition spatiale jointe de plus de deux espèces.

Quatrième partie

Conclusion générale et perspectives

Chapitre 8

Conclusion générale et perspectives

Dans ce dernier chapitre, je propose de revenir sur les principales contributions des applications présentées dans ce travail et de tracer mes perspectives de recherche.

8.1 Principales contributions

Coupler un modèle statistique décrivant l'hétérogénéité spatiale d'un processus aléatoire latent et un modèle décrivant l'occurrence d'observations de terrain m'a permis d'apporter des solutions opérationnelles :

- au problème de détection et de localisation de lignes de discontinuités génétiques
- au problème de modélisation de la distribution d'une (ou plusieurs) espèces à partir de données de biomasse continues *zero-inflated*.

Dans la première partie de ce travail, j'ai couplé le modèle de Potts-Dirichlet avec un modèle bivarié d'occurrence de génotypes. Ce modèle permet de classer des individus en populations génétiquement homogènes en tenant compte de l'hypothèse biologique selon laquelle la variabilité génétique varie continûment dans un domaine d'étude et est seulement caractérisée par de petites discontinuités génétiques. J'ai montré que, lorsque cette hypothèse est effectivement vérifiée, modéliser la structure spatiale de la variabilité génétique peut permettre de diminuer le nombre d'erreurs de classification. Ceci est vrai quelque soit le niveau de différentiation génétique entre individus. Par ailleurs, le modèle hiérarchique spatial Geneclust conduit à un nombre d'erreurs de classification d'autant plus faible par rapport au modèle non spatial Structure que ce niveau de différentiation génétique est faible et le niveau de dépendances spatiales élevé. Dans le cas des données génotypiques d'ours bruns de Scandinavie et HGDP-CEPH, Geneclust permet d'obtenir des classifications similaires au modèle non-spatial Structure mais à partir d'un nombre plus petit de marqueurs génétiques.

Dans la seconde partie de ce travail, j'ai présenté le modèle hiérarchique LOL comme alternative aux modèles $\Delta\Gamma$ pour la modélisation de données de biomasse continues *zero-inflated*. Ce processus de Poisson composé peut être décrit comme l'association d'un processus de Poisson homogène marqué avec un modèle conditionnel d'occurrence exponentielle de biomasse. Le modèle LOL a l'avantage d'être stable par addition ce qui permet de travailler directement avec des données brutes, d'éviter les problèmes de biais induits par une standardisation préalable des données et de comparer des relevés effectués à efforts d'échantillonnage différents. Dans cette partie, j'ai également proposé de complexifier les structures LOL et $\Delta\Gamma$ en les couplant à des modèles spatiaux sur grille puis

à un champ géostatistique bivarié afin de modéliser la structure spatiale de la biomasse en une (ou deux) espèce(s) dans un domaine de suivi. Plus les données sont *zero-inflated*, plus le modèle LOL possède de meilleures capacités d'adéquation aux données. Selon le critère de perte prédictive *a posteriori*, les différentes versions du modèle LOL possèdent généralement de meilleures performances prédictives.

Les différentes applications proposées dans ce document illustrent les potentialités de l'approche hiérarchique spatiale pour la modélisation statistique en écologie des populations. Une telle approche permet de répondre aux hypothèses écologiques selon lesquelles la distribution des organismes vivants est spatialement structurée et controlée par de multiples facteurs en intéraction (cf chapitre 1). D'un point de vue technique, elle permet de bénéficier simultanément des avantages offerts par la statistique spatiale et la démarche de construction hiérarchique. Comme illustré par les deux applications traitées au cours de cette thèse, les modèles développés en écologie doivent souvent décrire l'hétérogénéité d'une (ou plusieurs) variable(s) observable(s) entre différents sites d'échantillonnage ou différentes unités géographiques (e.g., différents habitats locaux). La modélisation hiérarchique permet de relier ces observations de terrain souvent non standards (e.g., non-gaussiennes et/ou multivariées) à des variables latentes pour lesquelles il est possible de recourir aux modèles spatiaux usuels. Quant aux modèles spatiaux, ils organisent un transfert d'informations entre unités ou sites géographiques voisins. Ces relations de dépendances se propagent entre les différentes couches de l'édifice hiérarchique permettant de générer des dépendances entre observations. Par ailleurs, ce transfert d'information permet d'améliorer les performances d'estimation et/ou de prédiction des modèles par rapport à une hypothèse d'indépendance mutuelle des effets aléatoires. Ceci est notamment le cas quand le niveau d'information apporté par les données est faible. Ainsi, le modèle Geneclust tire profit des dépendances génotypiques entre individus géographiquement proches quand l'information génétique est trop faible pour classer les individus en populations génétiquement homogènes.

L'explicitation de variables aléatoires latentes permet de débrider l'interprétation des modèles statistiques : elle permet de renforcer leur sens conceptuel, facilite leur compréhension et *a fortiori* la discussion avec le praticien : des structures aléatoires complexes peuvent être imaginées étape par étape à l'aide de relations de conditionnements probabilistes. Chaque couche du modèle est interprétée par rapport au phénomène modélisé et par rapport au niveau de variabilité considéré. Cette propriété est illustrée avec le modèle LOL dont la vraisemblance a une forme complexe peu explicite mais qui, décrit avec un raisonnement conditionnel, permet de conceptualiser le processus d'échantillonnage d'organismes vivants. L'explicitation de variables aléatoires latentes indiquant le nombre aléatoire de *gisements* collectés en chaque site permet de donner un sens conceptuel aux paramètres μ et ρ de ce modèle : μ désigne le nombre moyen de *gisements* que l'on peut espérer collecter pour un effort d'échantillonnage standard et ρ l'inverse de la quantité moyenne de biomasse contenue dans un *gisement*.

La modélisation hiérarchique spatiale permet d'élargir considérablement la panoplie des modèles statistiques possibles pour le traitement de questions écologiques. Dans les chapitres 3 et 6, je donne un aperçu de la grande flexibilité offerte par une telle approche : à partir de quelques "briques" élémentaires choisies judicieusement dans la panoplie des modèles standards et de quelques modèles spatiaux, une multitude de constructions hiérarchiques spatiales différentes peut être spécifiée. En théorie, il est possible de combiner autant de modèles que l'on souhaite. En pratique, il faut veiller à rester parci-

monieux dans ces choix de modélisation afin d'éviter de tomber dans le piège des "usines à gaz" i.e., des structures hiérarchiques caractérisées par un nombre trop important de paramètres et variables aléatoires latentes, inapproprié par rapport au nombre d'observations disponibles.

Dans tous les modèles hiérarchiques construits au cours de ce travail, des lois *a priori* non-informatives ont été attribuées aux différents paramètres inconnus. En pratique, cette approche simple est la plus utilisée car elle évite une étape d'élicitation de *priors* par un (ou plusieurs) expert(s) ou la recherche d'un jeu de données susceptible d'apporter de l'information *a priori* sur les paramètres inconnus. Dans le cas de structures hiérarchiques, une deuxième raison a été mise en évidence au cours de ce travail : les paramètres inconnus n'ont généralement pas un sens physique si bien qu'ils ne sont pas estimables *a priori*. Cela est notamment le cas pour le modèle LOL : la notion de *gisements* est purement conceptuelle si bien qu'il est impossible d'inclure de l'information dans les lois *a priori* attribuées aux paramètres μ et ρ. Une des difficultés posées par le choix de *priors* non-informatifs intervient lors du calcul d'un facteur de Bayes qui nécessite un *prior* informatif (cf chapitre 6 section 6.3.1). Dans ce cas, nous proposons d'appliquer la méthode décrite dans l'Annexe F qui permet de calculer le facteur de Bayes à partir d'un niveau d'information équivalent à la spécification d'un *prior* informatif. Ceci m'a notamment permis de comparer les performances d'ajustement de plusieurs versions hiérarchiques des modèles LOL et $\Delta\Gamma$.

8.2 Perspectives

Les perspectives qui découlent de ce travail sont variées. Je retiendrai essentiellement un axe de recherche qui me paraît prometteur : exploiter l'approche hiérarchique pour la modélisation de structures spatio-temporelles. Cette perspective s'inscrit dans la continuité logique de ce travail de thèse. Les dynamiques temporelles sont omniprésentes en écologie des populations et peuvent être introduites dans des constructions hiérarchiques. L'objectif est désormais de tirer profit de la grande flexibilité offerte par la démarche de construction hiérarchique pour coupler des processus aléatoires spatialement structurés à des processus temporels.

La modélisation hiérarchique spatio-temporelle a déjà fait l'objet de quelques travaux, notamment en sciences environnementales (Wickle & al., 2001), (Calder, 2003) et en épidémiologie (Richardson & al., 2006). Un tour d'horizon dans la littérature montre que cette approche est rarement envisagée pour la modélisation en écologie des populations. Pourtant, en théorie, la plupart des phénomènes écologiques sont la manifestation d'un ou de plusieurs processus spatio-temporels covariant à diverses échelles (Legendre & Legendre, 1998). En fait, la complexité de tels processus ainsi que leur grande dimension rendent souvent extrêmement difficile l'utilisation des méthodes statistiques classiques basées sur la spécification directe d'une structure de covariance spatio-temporelle. L'approche hiérarchique propose des solutions pour palier à ces difficultés. Certaines d'entre elles sont notamment détaillées dans le chapitre 8 de l'ouvrage de Banerjee et al. (2004).

Une première application intéressante serait de développer des outils pertinents permettant de décrire la dynamique spatio-temporelle de la biomasse en une ou plusieurs espèces. Comme cas d'étude, je pourrai exploiter de nouveau la richesse des données de biomasse en invertébrés épibenthiques du Golfe-du-Saint-Laurent fournies par le Centre des Pêches du Golfe. En effet, je dispose des quantités de biomasse en 14 espèces mesurées

sur 17 mois de Septembre consécutifs et avec en moyenne 180 observations par année. La figure 8.1 montre que ces espèces connaissent une dynamique temporelle assez marquée depuis 1989.

Nous avons vu dans le chapitre 7 que l'approche par convolution sur grille discrète semble offrir un cadre de modélisation simple et flexible pour la modélisation de structures bivariées. Dans sa thèse, Catherine Calder (2003) montre, à partir de données de pollution atmosphérique, que cette approche a également de nombreuses potentialités pour la modélisation de structures spatio-temporelles. Une idée prometteuse serait donc d'adapter cette approche pour modéliser la dynamique spatio-temporelle d'évolution de la biomasse. Le principe en est simple et prometteur : il s'agit de spécifier un processus temporel latent sur une grille régulière puis d'obtenir une approximation d'un champ spatio-temporel par convolution de ce processus discrétisé avec un noyau de lissage purement spatial. La dynamique temporelle de la biomasse pourrait se manifester en chaque noeud de la grille selon un modèle de production de biomasse, pioché dans la panoplie des modèles classiquement utilisés par les écologues pour représenter la dynamique d'évolution d'une population.

Une idée supplémentaire serait d'introduire dans ce modèle hiérarchique des covariables susceptibles d'expliquer la dynamique temporelle des espèces. Une application intéressante, notamment pour les écologues du Centre des Pêches du Golfe, serait de définir un modèle permettant de dissocier et quantifier les impacts respectifs du changement climatique (via des mesures de température de l'eau de fond), de la prédation (via un indice potentiel de prédation calculé par le Centre des Pêches du Golfe) et de la pêche commerciale par engins de pêche mobiles (chalut, senne) sur la dynamique spatio-temporelle des invertébrés épibenthiques du Golfe du Saint-Laurent. Des données issues de journaux de bord de pêcheurs sont disponibles pour évaluer l'impact de la pêche commerciale sur l'écosystème marin du Golfe du Saint-Laurent. Dans le cas des invertébrés épibenthiques, les écologues savent que les engins de pêche mobiles peuvent détruire ces organismes (effet direct) ainsi que leur milieu de vie (effet indirect). Des passages fréquents et nombreux dans une même région fragmentent les habitats benthiques qui perdent de plus en plus leur diversité structurelle (Kaiser & al., 2002). A mesure que le raclage du fond s'intensifie, les communautés épibenthiques pourraient diverger de leur état naturel à l'échelle de la zone où se concentrent les activités de pêche. Jusqu'à présent, de nombreuses études ont été réalisées dans le but d'évaluer l'impact de la pêche commerciale sur les communautés marines (Auster & al., 1996), (Thrush & al., 1998), (Collie & al., 2000). Le point faible essentiel de ces études est de ne pas tenir compte de l'impact de la prédation ou encore des anomalies climatiques qui, eux aussi, peuvent influencer la survie et la répartition spatiale des espèces (Stenseth & al., 2005), (Savenkoff & al., 2007). Aussi, dans ce contexte où l'importance relative de ces différents facteurs est difficile à établir, les effets directs et indirects de la pêche commerciale sur l'écosystème marin sont souvent mal-évalués et le débat se poursuit entre une conservation accrue de la biodiversité par la limitation des activités de pêche et une exploitation efficace des ressources halieutiques canadiennes. L'objectif serait donc de contribuer positivement à ce débat par la construction d'un modèle hiérarchique spatio-temporel permettant de dissocier l'impact respectif de ces différents facteurs sur la communauté épibenthique.

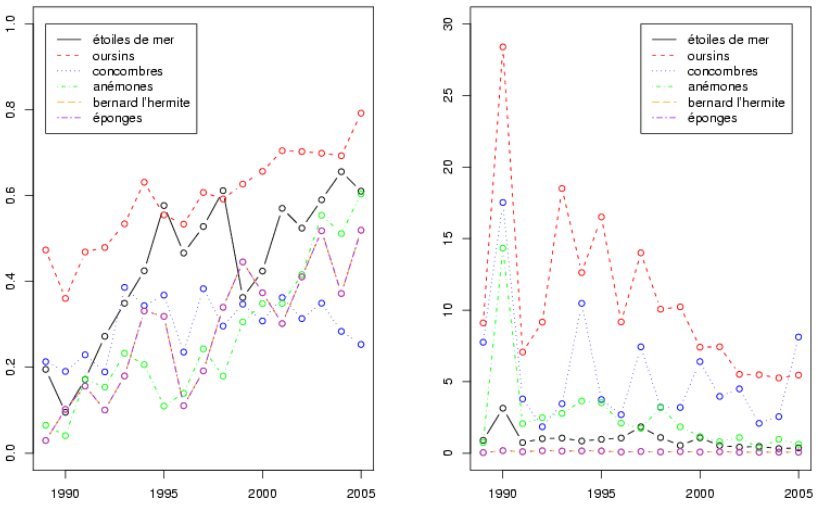


Figure 8.1 – Evolution de la probabilité de présence (à gauche) et de la biomasse moyenne (à droite) en 6 espèces d'invertébrés épibenthiques évaluée à partir des campagnes du *Centre des Pêches du Golfe*.

Cinquième partie

Annexes

Annexe A

Quelques rappels de génétique

Dans cette annexe sont rappelés les concepts génétiques de base, nécessaires à la compréhension du problème de localisation de lignes de discontinuités génétiques.

Le noyau d'une cellule est « la bibliothèque » qui renferme tout le patrimoine génétique héréditaire de chaque individu. Ce patrimoine est encodé dans une molécule appelée Acide DésoxyriboNucléique (ADN). Un chromosome est « un livre » de cette bibliothèque. Il s'agit d'un élément microscopique constitué de molécules d'ADN. Les chromosomes sont en nombre variable suivant chaque espèce. Chez les organismes dits *diploïdes*, ils sont présents en double exemplaire, l'un d'origine paternelle et l'autre d'origine maternelle. Ainsi, l'espèce humaine compte 46 chromosomes : 23 paires, dont 22 sont des chromosomes homologues, la dernière paire correspondant aux deux chromosomes sexuels. L'ensemble des chromosomes est représenté sur un caryotype, ou carte de chromosomes (cf figure A.1).

Les marqueurs génétiques sont des séquences d'ADN repérables spécifiquement. Ils sont situés à un endroit bien précis sur un chromosome appelé locus (cf figure A.2). Dans ce travail, nous utilisons un type de marqueurs particuliers : les microsatellites. Il s'agit de séquences d'ADN formées par une répétition continue de mêmes motifs nucléotidiques. Ces marqueurs génétiques sont particulièrement adaptés pour mener des analyses génétiques car ils sont le siège de variations à l'origine de polymorphismes multialléliques très informatifs. Ils permettent ainsi d'établir l'empreinte génétique d'un individu, c'est-à-dire de décrire et définir des individus en vue de leur protection et/ou de leur classification.

En tant que fragments d'ADN, les marqueurs génétiques peuvent posséder différentes versions, appelées allèles, qui diffèrent de par leur séquence nucléotidique. Il faut savoir que, chez un individu diploïde, chaque marqueur génétique est représenté par deux allèles, situés au même locus et que ces deux allèles sont soit identiques dans leur composition nucléotidique -l'individu est alors homozygote pour ce marqueur- soit différents dans leur composition -l'individu est alors hétérozygote pour ce marqueur.

Un couple de marqueurs en un locus donné s'appelle un génotype. Par exemple, s'il existe deux formes du marqueur génétique X au locus l : l'allèle Xa et l'allèle Xb, alors le génotype d'un individu pour le marqueur X peut-être soit homozygote (Xa/Xa ou Xb/Xb) soit hétérozygote (Xa/Xb).

Une fréquence allélique est la fréquence d'occurrence d'un allèle pour un marqueur génétique donné, par rapport à l'ensemble des allèles existant dans une population d'intérêt.

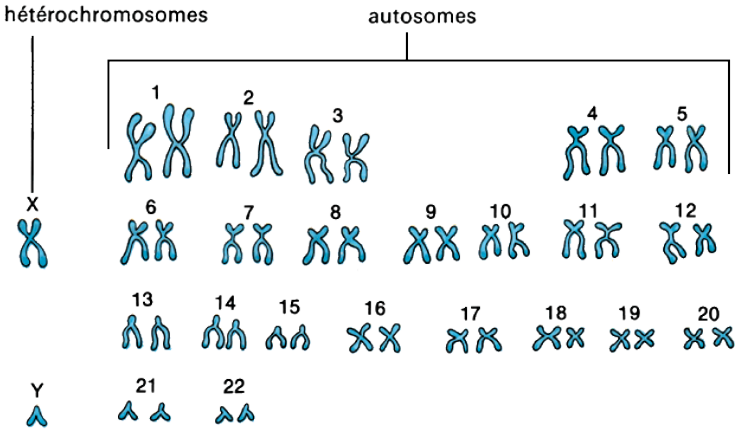


Figure A.1 – Caryotype d'un individu diploïde. Les 46 chromosomes sont présentés sous forme condensée et dupliquée

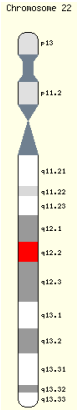


Figure A.2 – Un chromosome sous forme condensée non dupliquée (i.e., batônnet). La portion rouge représente un locus.

Annexe B

Le point sur l'inférence bayésienne dans le cadre hiérarchique

Tous les modèles construits au cours de cette thèse ont été analysés sous le paradigme bayésien. C'est pourquoi, dans cette annexe, je propose de faire un point rapide sur l'inférence bayésienne dans le cadre de la modélisation hiérarchique (objectifs, principes, algorithmes MCMC). Pour plus de détails, les aspects conceptuels de l'analyse bayésienne sont largement discutés dans des ouvrages de référence comme Gelman et al. (1995), Robert (2006) et Parent et Bernier (2007).

B.1 Les formules clés de l'inférence bayésienne

La démarche de modélisation hiérarchique consiste à spécifier successivement la loi *a priori* $[\theta]$, le modèle du processus latent $[Z|\theta]$ et le modèle des observations $[Y|Z,\theta]$. Construire un modèle hiérarchique consiste donc à définir la distribution de probabilité jointe des observables, des variables latentes et des paramètres donnée par :

$$[Y, Z, \theta] = [\theta][Z|\theta][Y|Z,\theta] \tag{B.1.1}$$

La formule des probabilités composées permet de renverser le sens du conditionnement de l'équation B.1.1 pour obtenir l'écriture alternative suivante :

$$[Y, Z, \theta] = [\theta, Z|Y][Y] \tag{B.1.2}$$

Dans cette expression :

- $[Y]$ désigne la loi prédictive *a priori*. Cette distribution marginale décrit le degré de crédibilité que l'on peut accorder à chaque valeur y prise par les observables en intégrant l'incertitude sur le vecteur des paramètres et des variables latentes :

$$[y] = \int [y|z,\theta][\theta, z] d\theta dz \tag{B.1.3}$$

- $[\theta, Z|Y]$ est la distribution *a posteriori* jointe des paramètres et des variables latentes.

Des relations stochastiques B.1.1 et B.1.2 se déduit la fameuse *formule de Bayes* qui permet d'écrire la probabilité *a posteriori* jointe des grandeurs inconnues (θ, Z) du modèle hiérarchique sous la forme :

$$[\theta, z|y] = \frac{[\theta][z|\theta][y|z,\theta]}{[y]} \tag{B.1.4}$$

Dans cette expression, la probabilité marginale [y], indépendante de θ et Z, est uniquement une constante de normalisation : elle assure que la distribution *a posteriori* $[\theta, Z|Y]$ soit bien une loi de probabilité. La distribution *a posteriori* est souvent écrite sous sa forme non normalisée :

$$[\theta, Z|y] \propto [\theta][Z|\theta][Y|Z, \theta] \tag{B.1.5}$$

Le passage de la distribution *a priori* $[\theta, Z] = [\theta][Z|\theta]$ à la distribution *a posteriori* $[\theta, Z|Y]$ des inconnues du modèle, exprimé par la *formule de Bayes*, peut-être interprété comme une mise à jour de la connaissance *a priori* sur le phénomène en une connaissance *a posteriori* sur la base des observations Y introduites à travers le modèle des observations $[Y|Z, \theta]$.

La *formule de Bayes* est centrale. Elle montre comment se combinent le modèle d'expertise, le modèle du processus interne et le modèle des observations pour exprimer la distribution *a posteriori* des inconnues du modèle. Par ailleurs, la *formule de Bayes* assure la cohérence entre les démarches de modélisation et d'inférence. Cette idée est illustrée dans la figure B.1. Le membre de droite fait apparaître le conditionnement probabiliste dans le sens suivi lors de l'élaboration d'un modèle hiérarchique. Connaissant les paramètres du modèle, on peut générer les variables latentes par leur distribution conditionnelle $[Z|\theta]$. Puis, connaissant les états et les paramètres, on peut générer des observations par le modèle d'observations $[Y|Z, \theta]$. Le membre de gauche fait apparaître le conditionnement probabiliste dans le sens inverse : celui de l'inférence. Connaissant les observations et conditionnellement à une structure de modèle, la distribution jointe *a priori* des variables latentes et des paramètres est "mise à jour" pour obtenir la distribution *a posteriori*.

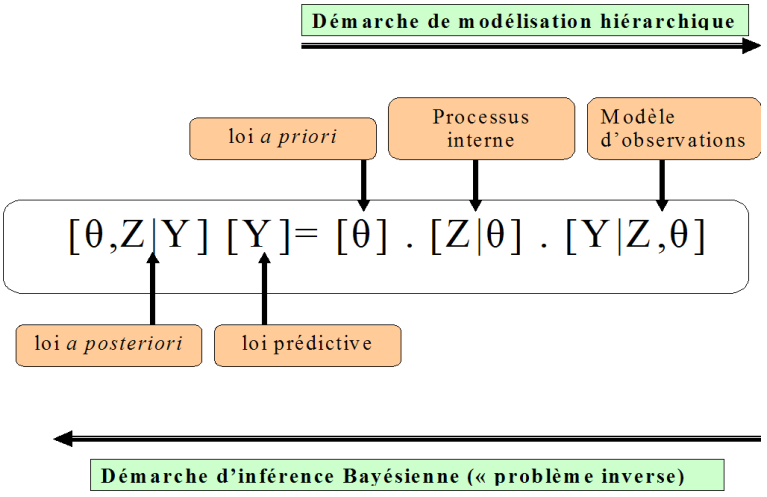


Figure B.1 – Inférence bayésienne d'un modèle hiérarchique. Le traitement bayésien réunit d'un même cadre cohérent la modélisation et l'inférence statistique

B.2 Une circonstance heureuse : la conjugaison

Dans le cas de modèles unidimensionnels, le calcul du numérateur de la *formule de Bayes* peut, pour certaines combinaisons de lois, aboutir à l'expression d'une loi de probabilité connue. Le calcul de la constante de normalisation $[y]$ peut alors se faire analytiquement. Cette méthode analytique "exacte" de calcul de la distribution *a posteriori* s'appelle la conjugaison.

Soit ϕ un vecteur de grandeurs inconnues : il peut s'agir de paramètres et/ou de variables aléatoires latentes. On dit que la loi *a priori* $[\phi]$ est conjuguée à la loi conditionnelle $[y|\phi]$ si la loi *a posteriori* issue de la combinaison $[\phi][y|\phi]$ appartient à la même famille de loi que la loi *a priori*. Les distributions conjuguées naturelles les plus utilisées sont indiquées dans le tableau B.1 pour un n-échantillon indépendant. Dans ce tableau, τ désigne la précision de la loi normale, $\overline{y}$ la moyenne empirique de l'échantillon et $s^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - m)^2$

Vraisemblance $[y\|\phi]$	*prior* $[\phi]$	Posterior $[\phi\|y]$
Binomiale(N,p) avec N connu	Beta(α,β)	Beta($\alpha + n\overline{y}$, $\beta + nN - n\overline{y}$)
Poisson(λ)	Gamma(a,b)	Gamma(a+n$\overline{y}$, b+n)
Exponentielle(ρ)	Gamma(a,b)	Gamma(a+n, b+n$\overline{y}$)
Normal(m,τ) avec τ connu	Normal(μ,ν)	Normal($\frac{\overline{y}n\tau+\mu\nu}{n\tau+\nu}$, $n\tau + \nu$)
Normal(m,τ) avec m connu	Gamma(a,b)	Gamma($a + \frac{n}{2}$, $b + \frac{ns^2}{2}$)
Multinomiale($p_1, p_2, \ldots, p_k$)	Dirichlet($\alpha_1, \alpha_2, \ldots, \alpha_k$)	

Tableau B.1 – Tableau des conjuguées naturelles

Dans le cas de structures multidimensionnelles, le calcul analytique complet des distributions *a posteriori* par conjugaison n'est possible que pour la loi multinormale ou la loi multinomiale dont les conjuguées respectives sont la loi multinormale-Wishart et la loi de Dirichlet.

En pratique, les propriétés de conjugaison sont rares et le calcul analytique "exact" de l'intégrale multidimensionnelle B.1.3 et, par conséquent, de la loi *a posteriori* B.1.4 sont généralement impossibles. Prenons l'exemple du modèle LOL décrit dans le chapitre 5. La formule de Bayes permet d'écrire facilement la loi *a posteriori* jointe des grandeurs inconnues (i.e., paramètres $\theta = (\mu, \rho)$ et variables latentes Z=N) du modèle sachant les données y :

$$[\theta, z|y] \propto [y|z, \theta][z|\theta][\theta]$$

Dans cette expression :

– Le sous-modèle des observations $[y|z, \theta]$ est :

$$\begin{aligned}[y|z,\theta] = [y|N,\rho] &= \prod_{k=1}^{r}[y_k|\rho, N_k] \\ &= \prod_{\{k;N_k=0\}}\delta_0(y_k)\prod_{\{k;N_k>0\}}\Gamma(N_k,\rho)(y_k)\end{aligned}$$

– La loi des variables latentes Z sachant les paramètres inconnus θ est :

$$[z|\theta] = [N = n|\mu] = e^{-\mu}\frac{\mu^n}{n!}$$

– la loi *a priori* $[\theta] = [\mu, \rho] = [\mu][\rho]$.

Malheureusement, la loi *a posteriori* jointe $[\theta, Z|Y]$ n'a pas une forme analytique connue.

Pour contourner ce problème d'intégration, le recours aux méthodes de simulation de Monte-Carlo par chaînes de Markov (notées MCMC) est aujourd'hui l'approche la plus utilisée.

B.3 Principe général des méthodes MCMC

Le but des méthodes dites MCMC (Markov Chain Monte Carlo) est d'obtenir des tirages aléatoires dans une loi de probabilité $f(x)$ dont l'expression est connue à une constante près c :

$$f(x) = c \times h(x)$$

Ce problème se présente typiquement dans les calculs d'inférence bayésienne où les distributions *a posteriori* $[\phi|y]$ ont une expression mathématique généralement inconnue. Dans ce cas, les méthodes MCMC consistent à construire une chaîne de Markov dont la distribution stationnaire est $[\phi|y]$. En pratique, elles permettent d'obtenir un échantillon aléatoire $\{\phi^{(1)}, \phi^{(2)}, \ldots, \phi^{(G)}\}$ de la loi *a posteriori* $[\phi|y]$ à partir duquel peuvent être empiriquement calculées toutes les quantités d'intérêt sur cette loi : moyenne, variance, percentiles etc. Si l'échantillon est "suffisamment grand", l'histogramme des tirages obtenus est lisse et fournit une bonne approximation de la distribution *a posteriori*. L'avantage que procure ces méthodes est qu'il n'est pas nécessaire de connaître l'expression analytique de la loi *a posteriori* de ϕ pour réaliser les simulations de Monte-Carlo.

Les principaux algorithmes MCMC actuellement utilisés sont : l'algorithme de Metropolis-Hastings et l'algorithme de Gibbs, un cas particulier de l'algorithme de Metropolis-Hastings (Robert & Casella, 1998).

B.4 L'algorithme de Metropolis-Hastings

L'algorithme de Metropolis-Hastings est historiquement la première des méthodes MCMC (Metropolis & al., 1953). Il est basé sur un algorithme en deux temps. A partir du tirage en cours ϕ^t, il propose une valeur "candidate" ϕ^* tirée aléatoirement d'une distribution conditionnelle $[\phi^*|\phi^t]$ choisie, appelée *loi instrumentale* ou *fonction de saut*. Cette distribution permet à la chaîne de Markov de "bouger" dans l'espace des variables inconnues à partir d'un point donné. Puis, il utilise un algorithme "d'acceptation-rejet" de ce candidat. Les étapes sont décrites ci-dessous :

Initialisation : Choisir une valeur initiale ϕ^0
Transition Markovienne $\phi^{(t)} \longrightarrow \phi^{(t+1)}$ pour $t \geq 0$:

1. Générer un candidat ϕ^* selon la *fonction de saut* $[\phi^*|\phi^{(t)}]$.
2. Calculer le ratio
$$r = \frac{[\phi^*|y] * [\phi^{(t)}|\phi^*]}{[\phi^{(t)}|y] * [\phi^*|\phi^{(t)}]}$$
3. Tirer une réalisation B issue d'une loi de Bernoulli de paramètre ρ où $\rho = min(1, r)$.
Si B=1, le candidat ϕ^* est accepté et $\phi^{(t+1)} = \phi^*$.
Si B=0, on reconduit la valeur précédente à l'étape suivante : $\phi^{(t+1)} = \phi^{(t)}$.

La convergence théorique de l'algorithme de Metropolis-Hastings est notamment prouvée dans Roberts et Smith (1993). Le choix des lois *instrumentales* est un moment décisif dans la mise en place de la méthode car elle conditionne la rapidité de convergence de l'algorithme. En effet, des lois *instrumentales* trop dispersées qui génèrent des candidats souvent refusés peuvent laisser la chaîne longtemps "bloquée" sur une valeur et demander énormément d'itérations avant d'atteindre la convergence. Inversement, avec des fonctions de saut peu dispersées, la chaîne peu bouger trop lentement. L'effet est le même : trop d'itérations sont nécessaires pour atteindre la convergence. Pour parer à ces difficultés, des méthodes de "réglage" de l'algorithme de Metropolis-Hastings, et plus précisément de la variance de la *fonction de saut* sont possibles (Parent & Bernier, 2007).

Il est nécessaire de pouvoir calculer le ratio r

$$r = \frac{[\phi^*|y] * [\phi^{(t)}|\phi^*]}{[\phi^{(t)}|y] * [\phi^*|\phi^{(t)}]}$$

pour tout couple $(\phi^*, \phi^{(t)})$. Il est important de noter qu'il n'est pas nécessaire de connaître la distribution normalisée *a posteriori* pour calculer le ratio r car la constante de normalisation $[y] = \int [\phi][y|\phi]d\phi$ se simplifie dans le calcul. L'expression non-normalisée de la distribution *a posteriori* s'obtient aisément en multipliant la loi *a priori* par la fonction de vraisemblance.

B.5 L'algorithme de Gibbs

Le second groupe de méthodes MCMC est encore appelé "échantillonneurs de Gibbs".

Soit k la dimension du vecteur ϕ et $(\phi_1, \phi_2, .., \phi_k)$ les k composantes de ϕ. L'algorithme de Gibbs génère un échantillon dans la distribution *a posteriori* de ϕ en effectuant, de façon cyclique, des tirages successifs dans chacune des lois conditionnelles complètes. Celles-ci s'obtiennent en extrayant la composante ϕ_j de la loi jointe $[\phi, y]$ et en conditionnant par les données et par toutes les autres composantes du vecteur ϕ (i.e., $(\phi_1, ..., \phi_{j-1}, \phi_{j+1}, ..., \phi_k)$) supposées constantes. L'algorithme de Gibbs fonctionne ainsi :

Initialisation : Générer un vecteur de valeurs initiales $\phi^{(0)}$
(ex : générer les paramètres initiaux selon leur loi *a priori* puis conditionnellement aux paramètres initiaux, générer des valeurs initiales pour les variables latentes selon le modèle du processus interne)
Transition Markovienne $\phi^{(t)} \longrightarrow \phi^{(t+1)}$ pour $t \geq 0$:
Générer $\phi_1^{(t+1)}$ en simulant selon la loi $[\phi_1|\phi_2^{(t)} \dots \phi_j^{(t)} \dots \phi_k^{(t)}, y]$
Générer $\phi_2^{(t+1)}$ en simulant selon la loi $[\phi_2|\phi_1^{(t+1)}, \phi_3^{(t)} \dots \phi_j^{(t)} \dots \phi_k^{(t)}, y]$
...
Générer $\phi_j^{(t+1)}$ en simulant selon la loi $[\phi_j|\phi_1^{(t+1)}, \phi_2^{(t+1)} \dots \phi_{j-1}^{(t+1)}, \phi_{j+1}^{(t)} \dots \phi_k^{(t)}, y]$
...
Générer $\phi_k^{(t+1)}$ en simulant selon la loi $[\phi_k|\phi_1^{(t+1)}, \phi_2^{(t+1)} \dots \phi_{k-1}^{(t+1)}, y]$

On montre que cette procédure itérative construit une chaîne de Markov qui admet comme distribution stationnaire la loi *a posteriori* du paramètre ϕ. La convergence est assurée sous des conditions de régularité très peu contraignantes (Parent & Bernier, 2007).

Grâce aux propriétés d'*indépendance conditionnelle* induites dans une structure hiérarchique (cf chapitre 2), le calcul des *lois conditionnelles complètes* est souvent considérablement simplifié. En effet, l'expression de la loi *conditionnelle complète* de chaque grandeur inconnue se simplifie sous la forme du produit des lois unidimensionnelles des seules variables et données qui lui sont directement reliées. Autrement dit, le conditionnement en apparence important induit lors du calcul des conditionnelles complètes est finalement réduit à quelques variables.

L'implémentation pratique de cet algorithme est particulièrement simple : elle ne demande pas de mettre en place une fonction de saut et, d'un point de vue calculatoire, ne requiert pas d'étape d'acceptation-rejet. Cependant, elle nécessite de connaître et de pouvoir simuler les distributions conditionnelles complètes des éléments de ϕ. Dans le cas contraire, il faut recourir à l'algorithme de Metropolis-Hastings qui nécessite seulement de connaître la loi visée à une constante multiplicative près. A noter que l'algorithme de Gibbs présente parfois le défaut d'une convergence très lente.

Les deux techniques Gibbs et Metropolis-Hastings ne sont pas exclusives l'une de l'autre et on peut construire des algorithmes hybrides. Si par exemple, certaines conditionnelles complètes sont difficiles à échantillonner, on peut utiliser pour la génération de la variable liée correspondante une sous-étape de Metropolis-Hastings. Robert (1998) montre l'intérêt et la validité de ces algorithmes imbriqués.

B.6 Diagnostics de convergence

Dans la pratique, une des difficultés des algorithmes MCMC est la vérification de la convergence. Avant de procéder aux inférences, il est crucial de vérifier si l'échantillon MCMC peut être considéré comme représentatif de la loi *a posteriori*.

En pratique, la méthode la plus courante consiste à appliquer des diagnostics directement sur les sorties de l'algorithme i.e., sur la (les) chaîne(s) simulée(s). Plusieurs auteurs (Gelman & Rubin, 1992), (Brooks & Gelman, 1998) ont ainsi proposé des diagnostics de convergence basés sur la comparaison entre plusieurs chaînes MCMC indépendantes, obtenues en lançant le processus d'échantillonnage à partir d'états initiaux différents, dispersés dans l'espace des paramètres. Parmi ces méthodes, les diagnostics de mélange sont les plus répandus. Ils consistent à vérifier que différentes chaînes, initiées à différents endroits, vont effectivement "perdre la mémoire" de leur état initial et explorer de la même façon le domaine des paramètres. Ces méthodes s'adaptent à tous les types d'algorithme MCMC et fournissent des diagnostics reconnus comme fiables (Kass & al., 1998).

Au cours de ce travail, la convergence des algorithmes MCMC a été vérifiée par un examen visuel des chaînes MCMC complété par un calcul de la statistique de convergence de Gelman-Rubin (1992), modifiée par Brooks et Gelman (1998). En pratique, un recours à ce diagnostic de convergence est souvent privilégié car il est simple à réaliser à partir des sorties MCMC. En particulier, le calcul de cette statistique est proposé en routine par le logiciel WinBUGS (disponible gratuitement sur http ://www.mrc-bsu.cam.ac.uk/bugs) et le package R "coda" développé parallèlement à la famille de logiciels BUGS. Par ailleurs, cette statistique a l'avantage d'être facilement interprétable car elle peut être vue comme une version non paramétrique d'un ratio de variances. En outre, elle donne une représentation visuelle et dynamique qui renseigne sur la vitesse de convergence.

Le principe de calcul est le suivant. Soient m chaînes MCMC lancées en parallèle (typiquement, $m > 2$) et G la longueur de ces chaînes. Soit θ_{kj} la j ième composante

unidimensionnelle de la chaîne k (k=1,2,..,m) et $\theta_{kj}(t)$ la chaîne composée des t premiers tirages de θ_{kj}. Soit α un seuil de probabilité préalablement choisi ($\alpha = 0.20$ dans WinBUGS). Alors, pour toute composante j et tout t :

1. On calcule $W_{kj}^{\text{within}}(t)$ la largeur de l'intervalle de crédibilité de niveau $(1-\alpha)$ calculée pour l'échantillon $\theta_{kj}(t)$ de taille t.
2. On calcule $\overline{W_j}^{\text{within}}(t)$ la moyenne des $W_{kj}^{\text{within}}(t)$ pour toutes les chaînes MCMC k (k=1,2,...,m).
3. On calcule $W_j^{\text{pooled}}(t)$ la largeur de l'intervalle de crédibilité $(1-\alpha)$ calculée pour l'échantillon de longueur $(t \times m)$ formé par l'aggrégation des m échantillons $\theta_{kj}(t)$ (k=1,2,..,m).
4. On calcule le ratio $R_j(t) = \frac{W_j^{\text{pooled}}(t)}{\overline{W_j}^{\text{within}}(t)}$.

Ce diagnostic de convergence est un critère de mélange entre plusieurs chaînes : il consiste à monitorer la convergence d'un algorithme MCMC en comparant des longueurs d'intervalles de crédibilité à $(1-\alpha)\%$, intra et inter chaînes, calculés à partir d'échantillons de valeurs de plus en plus grands. La convergence des chaînes vers la loi stationnaire se traduit, pour chaque composante j, par : $\lim_{t\to\infty} \overline{W_j}^{\text{within}}(t) = \lim_{t\to\infty} W_j^{\text{pooled}}(t) =$ constante. En pratique, pour une chaîne de longueur G, le diagnostic de convergence proposé par Brooks et Gelman (1998) consiste à vérifier que, pour toutes les composantes j, $\overline{W_j}^{\text{within}}(t)$ et $W_j^{\text{pooled}}(t)$ se stabilisent et que $R_j(t)$ se stabilise vers 1 quant t tend vers G. Kaas et al. (1998) suggèrent que le diagnostic de convergence est satisfaisant lorsque R_j est stabilisé et inférieur à 1.1 pour toutes les composantes j.

Annexe C

Coût *a posteriori* d'une estimation bayésienne ponctuelle

Pour plus de détails, les aspects théoriques liés au thème "risque et aide bayésienne à la décision" sont largement discutés dans des ouvrages de référence comme Berger (1985) et Parent et Bernier (2007).

Dans l'approche bayésienne, la distribution *a posteriori* $[\theta|y]$ décrit les incertitudes sur les paramètres inconnus θ compte-tenu de toutes les connaissances disponibles : les données expérimentales y et le savoir *a priori* modélisé par la loi *a priori* $[\theta]$. En pratique, cette distribution est souvent "résumée" à l'aide de valeurs ponctuelles telles que la moyenne, la médiane, le (ou les) mode(s) de la distribution, l'écart-type, les inter-quartiles et autres quantiles.

Un estimateur d est dit bayésien lorsque, en particulier, il minimise l'espérance *a posteriori* d'une fonction de coût, notée c. Ce coût *a posteriori* (en anglais, *posterior expected loss*) donne une indication des conséquences induites par le choix d'un estimateur ponctuel d devant notre niveau d'incertitude sur θ quantifié par $[\theta|y]$. Il est défini par :

$$C(d) = \int_{[\theta|y]} c(d,\theta)[\theta|y]d\theta \tag{C.0.1}$$

où $c(d,\theta)$ désigne le coût "ponctuel" de l'estimateur d pour une valeur possible θ. A chaque estimateur d correspond un coût aléatoire calculé à partir de la répartition des écarts à d pour chaque valeur possible *a posteriori* de θ, pondérés par la probabilité de cet écart évaluée par la probabilité *a posteriori* $[\theta|y]$.

Limitons-nous au cas simple où le paramètre θ est monodimensionnel. Le choix de c définit une mesure de l'écart entre l'estimateur d et le paramètre inconnu θ. Parmi les fonctions de coût classiques peuvent être citées :

1. les fonctions de coût quadratique définies par :

$$c_1(d,\theta) = a(d-\theta)^2$$

   a est une constante. Elles pénalisent de façon symétrique les sur- et les sous- estimations, mais pénalisent comparativement beaucoup plus les fortes déviations que les écarts-mineurs.

Fonction de coût	Estimateur ponctuel optimal (i.e., d^*)
c_1	$\mathbb{E}(\theta\|y)$
c_2	$F(d^*\|y) = \frac{1}{2}$
c_3	$F(d^*\|y) = \frac{a}{a+b}$

Tableau C.1 – Fonctions de coût classiques et décisions d'estimations ponctuelles optimales associées. $F(.|y)$ désigne la fonction de répartition *a posteriori*

.

2. les fonctions de coût "linéaire" définies par :

$$c_2(d, \theta) = a|d - \theta|$$

a est une constante. Elles pénalisent de façon symétrique les sur- et les sous- estimations ainsi que les forts et faibles écarts.

3. les fonctions de coût "linéaire" définies par :

$$c_3(d, \theta) = \begin{cases} a|d - \theta|, & d - \theta \geq 0 \\ b|d - \theta|, & d - \theta < 0 \end{cases}$$

a et b sont des constantes. Elles permettent d'affecter des poids différents aux sur- et aux sous- estimations de θ.

Le tableau C.1 indique, pour chacune de ces fonctions de coût, la meilleure décision d'estimation ponctuelle solution de la minimisation :

$$d^* = argmin_d C(d)$$

En pratique, le calcul de l'intégrale C.0.1 s'effectue par approximations numériques. Il s'agit de calculer la moyenne empirique des coûts ponctuels à partir d'un échantillon de valeurs *a posteriori* pour θ :

$$C(d) \simeq \frac{1}{G} \sum_{g=1}^{G} c(d, \theta_g)$$

où d désigne l'estimateur ponctuel choisi, G la taille de l'échantillon *a posteriori* et θ_g la g-ième valeur générée de l'échantillon.

Annexe D

Codes WinBUGS des modèles LOL et $BYM_{\mu,\rho}$-LOL

A l'exception des modèles présentés dans la partie II, toutes les estimations réalisées dans cette thèse, ont été menées sous la version 2.10 du logiciel OpenBUGS, dérivé direct du logiciel WinBUGS. Comme son préfixe "Open" l'indique, sa particularité par rapport à WinBUGS est de pouvoir être utilisée à partir du logiciel externe "R Project for Statistical Computing" (cf "http ://www.r-project.org/"). L'avantage est que cela permet de bénéficier des facilités d'inférence bayésienne offertes par WinBUGS et des nombreuses potentialités offertes par R (graphiques, manipulations des chaînes MCMC, stockage de résultats...). Pour appeler OpenBUGS sous R (version 2.4.1), j'ai utilisé la librairie R nommée *BRugs* qui utilise le même langage que WinBUGS pour la spécification des modèles et le même format de stockage pour les données et les valeurs initiales des chaînes MCMC (cf "http ://cran.r-project.org/doc/packages/BRugs.pdf" pour plus de détails)

Les deux programmes ci-après correspondent aux modèles LOL et $BYM_{\mu,\rho} - LOL$ décrits dans les chapitres 5 et 6. Ces programmes sont écrits dans le langage du logiciel WinBUGS 1.4. Ils permettent de donner au lecteur un aperçu de la simplicité avec laquelle ces modèles peuvent être implémentés sous WinBUGS. Par ailleurs, ils peuvent être utilisés par le statisticien ou l'écologue désireux d'appliquer ces modèles à leur propres jeux de données continues *zero-inflated*.

```
model{

#Calcul des intensités Poissonniennes en chaque site en fonction de l'effort
d'échantillonnage
        for(j in 1:nsampling){
                mupres[j]<- mu*(eff [j]/dstandard)
        }

        #Cas où l'espèce est présente
        for(k in 1:npres){
                #Le nombre de gisements ramassés est strictement positif
                ngis[k]~ dpois(mupres[pres[k]])I(1,)
                #La donnée observée est issue d'une loi gamma
                Y[pres[k]]~ dgamma(ngis[k], rho)
        }

        #Cas où l'espèce est absente
        for(j in 1:nabs){
                #Probabilité de présence évaluée au site où absence de l'espèce
                proba[j]<- 1-exp(-mupres[abs[j]])
                Y[abs[j]]~ dbern(proba[j])
        }

#Priors plats
        rho~ dnorm(0.0,1.0E-6)I(0.0001,)
        mu~ dnorm(0.0,1.0E-6)I(0.0001,)


#Notations :
nsampling= nombre d'observations dans le domaine D
dstandard= effort d'échantillonnage standard (e.g., distance chalutée en miles)
npres= nombre de sites où l'espèce est présente
nabs= nombre de sites où l'espèce est absente
Y= vecteur des données de biomasse
eff = vecteur des efforts d'échantillonnage
pres = indices des observations strictement positives
abs = indices des observations nulles
```

Figure D.1 – Code WinBUGS du modèle LOL. Les lignes précédées du signe ♯ correspondent à des commentaires destinés à faciliter la lecture du programme.

```
model{

#Modélisation des effets aléatoires spécifiques aux I unites géographiques
                for(i in 1:nunite){
                    log(mu[i])<- alpha0+phi[i]+epsilon[i]
                    epsilon[i]~ dnorm(0.0,tau.epsilon)
                    rho[i]~ dgamma(c,d)
                }

#Spécification d'une structure CAR Intrinsèque sur les effets aléatoires phi
        phi[1:nunite] ~ car.normal(adj[],weights[],num[],tau.IAR)
        for(l in 1:sumNumNeigh) {
                        weights[l] <- 1
        }

#Calcul des intensités Poissonniennes en chaque site en fonction de l'effort d'échantillonnage
        for(j in 1:nsampling){
                 mupres[j]<- mu[X[j]]*(eff [j]/dstandard)
        }

#Cas où l'espèce est présente
        for(k in 1:npres){
                #Le nombre de gisements ramassés est strictement positif
                ngis[k]~ dpois(mupres[pres[k]])I(1,)
                #La donnée observée est issue d'une loi gamma
                Y[pres[k]]~ dgamma(ngis[k], rho[X[pres[k]]])
        }

#Cas où l'espèce est absente
        for(j in 1:nabs){
                #Probabilité de présence évaluée au site où absence de l'espèce
                proba[j]<- 1-exp(-mupres[abs[j]])
                Y[abs[j]]~ dbern(proba[j])
        }

 #Les priors
            c~ dgamma(0.01,0.01)
            d~ dgamma(0.01,0.01)
            alpha0~ dflat()
            tau.epsilon~ dgamma(0.001,0.001)
            tau.IAR~  dgamma(0.1,0.1)
            sd.epsilon<- sd(epsilon[])
            sd.IAR<- sd(phi[])

#Part de variabilité inter-unité expliquée par le modèle IAR
            alpha<- sd.IAR/(sd.epsilon+sd.IAR)
 }
```

#Notations :
nsampling= nombre d'observations dans le domaine D
nunite= nombre d'unités géographiques
dstandard= effort d'échantillonnage standard (e.g., distance chalutée en miles)
npres= nombre de sites où l'espèce est présente
nabs= nombre de sites où l'espèce est absente
Y= vecteur des données de biomasse
X= vecteur indiquant l'indice de l'unité géographique dans laquelle a été réalisée chaque observation
eff = vecteur des efforts d'échantillonnage
pres = indices des observations strictement positives
abs = indices des observations nulles
sumNumNeigh = nombre de paires d'unités voisines
adj = vecteur listant les indices des voisins de chaque unité géographique
weight= vecteur des poids non-normalisés pour chaque paire d'unités voisines
num = vecteur indiquant le nombre de voisins de chaque unité géographique

Figure D.2 – Code WinBUGS du modèle $BYM_{\mu,\rho}$-LOL. Les lignes précédées du signe ♯ correspondent à des commentaires destinés à faciliter la lecture du programme.

Annexe E

Détails sur l'inférence MCMC du modèle Geneclust

La loi jointe du modèle Geneclust, décrit dans le chapitre 3, s'écrit :

$$[Y, Z, \theta] = [Y, c, \phi, \psi, f] = [Y|f, c, \phi][c|\psi][f|c][\phi][\psi]$$

Aucune propriété de conjugaison ne permet de calculer analytiquement les lois conditionnelles complètes respectives de c, f, ϕ et ψ. C'est pourquoi j'ai implémenté la mise à jour successive de chacune de ces grandeurs inconnues avec un pas de Metropolis-Hastings.

E.1 Mise à jour des fréquences alléliques $f_{k,l,j}$

Pour chaque population k et chaque locus l, les fréquences alléliques suivantes sont proposées pour deux allèles j et j' choisis au hasard dans le vecteur $f_{k,l,.}$ de dimension J_l :

$$f^*_{k,l,j} = a \times B_f \qquad f^*_{k,l,j'} = a - f^*_{k,l,j}$$

avec $B_f \sim \mathcal{B}(2,2)$ et $a = 1 - \sum_{i \neq j, i \neq j'} f_{k,l,i}$. Ce proposal permet de générer aléatoirement une nouvelle valeur pour $f^*_{k,l,j}$ et de définir $f^*_{k,l,j'}$ pour que la condition $\sum_{i=1}^{J_l} f_{k,l,i} = 1$ soit vérifiée. La valeur candidate proposée pour $f^*_{k,l,j}$ est bien comprise entre 0 et 1 et vaut en moyenne a. Soit f^* l'ensemble des fréquences alléliques comportant les valeurs candidates $f^*_{k,l,j}$ et $f^*_{k,l,j'}$. La mise à jour proposée pour $f_{k,l,j}$ et $f_{k,l,j*}$ est acceptée avec une probabilité :

$$p = min\left(1, \frac{[y|f^*, c, \psi, \phi]}{[y|f, c, \psi, \phi]} \frac{f_{k,l,j}(1 - f_{k,l,j})}{f^*_{k,l,j}(1 - f^*_{k,l,j})}\right)$$

E.2 Mise à jour du vecteur des coefficients de consanguinité $\phi = (\phi_1, \phi_2, ..., \phi_K)$

Pour chaque coefficient ϕ_k, une valeur candidate est générée selon une loi instrumentale uniforme sur $[0, 1]$:

$$\phi^*_k \sim \mathcal{U}[0, 1]$$

Soit ϕ^* le vecteur des coefficients de consanguinité contenant la valeur candidate ϕ_k^*. Comme les ϕ_k suivent *a priori* une loi Béta $\mathcal{B}(4,40)$ (cf section 3.5.2), cette mise à jour est acceptée avec la probabilité :

$$p = min\left(1, \frac{[y|\phi^*, \psi, f, c]}{[y|\phi, \psi, f, c]} \frac{\phi_k^{*3}(1-\phi_k^*)^{39}}{\phi_k^3(1-\phi_k)^{39}}\right)$$

E.3 Mise à jour du vecteur latent $c = (c_1, c_2, ..., c_n)$

La mise à jour du vecteur c est réalisée composante par composante en respectant l'ordre des sites $i = \{1, 2, ..., n\}$. Pour chaque label de classe c_i, une valeur candidate c_i^* est tirée dans une loi instrumentale uniforme sur tous les labels de classe possibles $\{1, 2, ..., K\}$. Soit c^* le vecteur des labels de classe contenant la valeur candidate c_i^*. La valeur c_i^* est acceptée avec une probabilité :

$$p = min\left(1, \frac{[y|c^*, f, \psi, \phi]}{[y|c, f, \psi, \phi]} exp(\psi \Delta U_i(c))\right) \qquad \text{avec} \qquad \Delta U_i(c) = \sum_{j \sim i} (\delta_{c_j, c_i^*} - \delta_{c_j, c_i})$$

E.4 Mise à jour du paramètre d'intéraction spatiale ψ

La difficulté principale à l'utilisation d'un modèle de Potts pour représenter des dépendances spatiales locales entre individus intervient au niveau de l'inférence sur le paramètre d'intéraction spatial ψ. C'est pourquoi l'inférence sur ce paramètre a retenu tout particulièrement mon attention.

Implémenter un pas de Metropolis-Hastings pour l'inférence sur le paramètre ψ nécessite de pouvoir calculer le ratio suivant :

$$\frac{[c|\psi^*]}{[c|\psi]} = \frac{Z(\psi, K)}{Z(\psi^*, K)} e^{(\psi^* - \psi)U(c)} \qquad \text{avec} \qquad Z(\psi, K) = \sum_{c \in \{1,2,..,K\}^n} exp(\psi U(c))$$

Ce ratio implique le calcul de la fonction de partition $Z(\psi, K)$ qui, en général, n'est pas calculable analytiquement. La difficulté de ce calcul repose sur le fait que la somme porte sur K^n termes. Dans le cas des ours bruns de Scandinavie par exemple, en supposant que K vaut 4, il faut calculer la valeur $exp(\psi U(c))$ sur plus de 2×10^{220} configurations de classes possibles !

L'évaluation de la constante de normalisation d'un champ aléatoire markovien est un problème délicat en statistique. C'est pourquoi, le paramètre d'intéraction spatial est souvent supposé connu dans les modèles markoviens. Récemment, des méthodes à l'origine développées en physique statistique ont été utilisées dans le cadre de processus stochastiques spatiaux pour approcher ces constantes de normalisation (Gelman & Meng, 1998). Jugée plus simple à implémenter, j'ai finalement choisi d'appliquer l'une d'entre elles, connue sous le nom d'intégration thermodynamique et notamment utilisée en épidémiologie (Green & Richardson, 2002).

Considérons la fonction :

$$\begin{aligned} \mu_K : [0, \psi_{\max}] &\mapsto]0, +\infty[\\ \psi &\to log(Z(K, \psi)) \end{aligned}$$

En particulier, en $\psi = 0$, on a : $\mu_K(0) = nlog(K)$.

Soit une configuration $c = (c_1, c_2, ..., c_n)$ d'un modèle de Potts de paramètres ψ et K. La dérivée première de μ_K vaut :

$$\frac{\partial}{\partial\psi}\mu_K(\psi) = E(U(c)|\psi, K)$$

D'après la définition de la primitive de μ_K qui s'annule en 0, on a la relation suivante :

$$\mu_K(\psi) = nlog(K) + \int_0^{\psi} \mathbb{E}(U(c)|x, K)dx$$

Le paramètre ψ est supposé prendre ses valeurs dans l'ensemble discrétisé $\{0, 0.1, 0.2, ..., 1\}$. Il suffit donc de calculer $\mu_K(\psi)$ pour chacune de ces valeurs. L'ensemble des valeurs calculées forme la table des constantes.

Pour chaque valeur de ψ, la première étape consiste à estimer l'espérance conditionnelle $\mathbb{E}(U(c)|\psi, K)$ par une méthode de simulation Monte-Carlo car celle-ci n'est pas calculable analytiquement. Pour cela, je commence par générer un échantillon de G configurations de classes selon le modèle de Potts-Dirichlet à l'aide d'un échantillonneur de Gibbs (cf section 3.2.3, figure 3.6). A chaque configuration $c^{(g)}$ (où g=1,2,..,G) simulée est associée une énergie potentielle $U(c^{(g)})$. D'après la loi des grands nombres, un estimateur convergent de $\mathbb{E}(U(c)|x, K)$ est donné par la moyenne empirique :

$$\widehat{\mathbb{E}(U(c)|x, K)} = \frac{1}{G}\sum_{g=1}^{G} U(c^{(g)})$$

où $c^{(g)}$ désigne la g-ième configuration de classes générée. 50000 itérations dont une période-de-chauffe de 20000 itérations ont été réalisées pour la simulation de configurations de Potts-Dirichlet. Afin de limiter les dépendances au sein de l'échantillon MCMC généré, un pas de 100 itérations a été choisi entre le stockage de deux configurations successives.

La deuxième étape consiste à approcher numériquement l'intégrale $\int_0^{\psi} \mathbb{E}(U(c)|x, K)dx$ à partir des 11 estimations de $\mathbb{E}(U(c)|\psi, K)$ précédemment calculées. Pour cela, j'ai utilisé l'algorithme proposé par la routine fortran *avint*. Celui-est basé sur une interpolation préalable de la fonction à intégrer à l'aide de fonctions paraboles.

Annexe F

Le facteur de Bayes

Dans cette annexe, ϕ désigne un vecteur de grandeurs inconnues pour un modèle donné. Cela peut donc désigner aussi bien des paramètres que des variables aléatoires latentes ou les deux.

F.1 Définition générale

Soient M_1 et M_2 deux modèles en compétition pour expliquer des données y. Plaçons-nous dans le cas où M_1 et M_2 sont des modèles hiérarchiques, paramétrés par ϕ_1 et ϕ_2.

Soient $[M_1]$ et $[M_2] = 1 - M_1$ les probabilités *a priori* des modèles M_1 et M_2. Elles quantifient un pari *a priori* de l'analyste en faveur des deux modèles en compétition. Etant données les observations y et les probabilités *a priori* $[M_1]$ et $[M_2]$, les probabilités *a posteriori* de chaque modèle M_j (j=1,2) sont données par la *formule de Bayes* :

$$[M_j|y] = \frac{[y|M_j][M_j]}{[y|M_1][M_1] + [y|M_2][M_2]} \tag{F.1.1}$$

Dans l'expression F.1.1, la quantité $[y|M_j]$ est la loi prédictive du modèle M_j. Cette loi marginale, aussi appelée *vraisemblance intégrée*, s'obtient par intégration de la densité jointe $[y, \phi_j]$ sur les valeurs possibles des inconnues ϕ_j :

$$[y|M_j] = \int [y|\phi_j][\phi_j] d\phi_j$$

La validité *a posteriori* du modèle M_1 par rapport au modèle M_2 est quantifiée par le rapport des probabilités *a posteriori* respectives des deux modèles :

$$\frac{[M_1|y]}{[M_2|y]} = \frac{[y|M_1]}{[y|M_2]} \frac{[M_1]}{[M_2]}$$

Dans cette formule, le ratio :

$$B_{M_1,M_2} = \frac{[y|M_1]}{[y|M_2]} \tag{F.1.2}$$

est dit *facteur de Bayes* du modèle M_1 relativement au modèle M_2.

Si les paris *a priori* sur les deux modèles en compétition sont identiques i.e., $[M_1] = [M_2]$ alors le facteur de Bayes s'écrit comme le ratio des probabilités *a posteriori* des deux modèles.

Kaas et Raftery (1994) suggèrent le barème des facteurs de Bayes indiqué dans le tableau F.1. Ils proposent de s'intéresser au double du logarithme du facteur de Bayes car cette quantité se situe à la même échelle que la deviance, notamment utilisée pour le calcul du DIC (cf chapitre 6 section 6.3.1). Dans les applications proposées dans cette thèse, je me suis basée sur ce barème indicatif pour comparer les modèles en compétition. Cela dit, d'autres barèmes, légèrement différents, ont été proposés par Jeffrey (1961) et Congdon (2001).

B_{M_1,M_2}	$2ln(B_{M_1,M_2})$	Evidence en faveur de M_1
1 à 3	0 à 2	Aucune
3 à 20	2 à 6	Positive
20 à 150	6 à 10	Forte
>150	>10	Très forte

Tableau F.1 – Interprétation de B_{M_1,M_2} (Kaas et Raftery, 1994)

F.2 Calcul du facteur de Bayes

Le calcul du facteur de Bayes pose les mêmes difficultés calculatoires que le calcul des lois *a posteriori*. En effet, à l'exception de quelques cas élémentaires, le calcul analytique de l'intégrale multidimensionnelle F.1 est impossible. De nombreuses méthodes numériques, basées sur des techniques de simulations par méthode MCMC (Cong & Bradley, 2001) ou des techniques assymptotiques approximées (Kass & Raftery, 1994) existent.

Au cours de cette thèse, nous nous sommes intéressés aux techniques d'approximation de l'intégrale F.1 par simulations de Monte-Carlo. La méthode la plus intuitive consiste à réaliser des tirages aléatoires des inconnues ϕ_j dans la loi *a priori* $[\phi_j]$ puis d'approcher l'intégrale F.1 par la moyenne arithmétique suivante :

$$[y|M_j] \simeq \frac{1}{G}\sum_{g=1}^{G}[y|\phi_j^{(g)}, M_j]$$

où G est la taille de l'échantillon des tirages, $\phi_j^{(g)}$ le g-ième tirage de ϕ_j dans la loi *a priori* (i.e., $[\phi_j|M]$) et $[y|\phi_j^{(g)}, M_j]$ la valeur de la densité des données observées sous M_j calculée en $\phi_j^{(g)}$. L'inconvénient de cette méthode est qu'en général la loi *a priori* est "plate" et par conséquent, la plupart des valeurs tirées de cette loi correspondent à des valeurs faibles de $[y|\phi_j^{(g)}]$.

Dans le chapitre 6, les facteurs de Bayes ont été calculés à partir d'une méthode jugée plus robuste : celle dite de *Raftery*. Elle prévoit des tirages de ϕ_j dans la loi *a posteriori* $[\phi_j|y]$ (Kass & Raftery, 1994). La technique se fonde directement sur la formule de Bayes dont on peut déduire que :

$$\frac{[\phi_j|M_j]}{[y|M_j]} = \frac{[\phi_j|y, M_j]}{[y|\phi_j, M_j]} \tag{F.2.3}$$

L'intégration de la formule F.2.3 sur l'espace des inconnues ϕ_j donne :

$$\frac{1}{[y|M_j]} = \int \frac{1}{[y|\phi_j, M_j]}[\phi_j|y, M_j]d\phi_j \tag{F.2.4}$$

L'intégrale F.2.4 peut être approchée par la moyenne harmonique des tirages de ϕ_j dans sa loi *a posteriori* $[\phi_j|y, M_j]$:

$$\frac{1}{[y|M_j]} \simeq \frac{1}{G}\sum_{g=1}^{G}\frac{1}{[y|\phi_j^{(g)}, M_j]} \tag{F.2.5}$$

d'où la *formule de Raftery* :

$$[y|M_j] \simeq \left(\frac{1}{G}\sum_{g=1}^{G}\frac{1}{[y|\phi_j^{(g)}, M_j]}\right)^{-1} \tag{F.2.6}$$

où est $\phi_j^{(g)}$ le g-ième tirage de ϕ_j dans la loi *a posteriori* (i.e., $[\phi_j|y, M]$).

Le calcul pratique de cette moyenne harmonique est assez simple. Il suffit, dans les calculs d'inférence par méthode MCMC, de calculer à chaque itération g la valeur de la densité $[y|\phi_j^{(g)}, M_j]$. Une fois la convergence atteinte, ces valeurs peuvent être employées pour le calcul exprimé par la formule F.2.6.

Contrairement au DIC, le choix de la paramétrisation ϕ_j n'a pas d'influence théorique sur le calcul du facteur de Bayes. En effet, celui-ci est basé sur un rapport de quantités intégrées (i.e., $[y|M1]$ et $[y|M_2]$) et donc indépendantes d'une paramétrisation ϕ_j.

L'approximation de la vraisemblance marginale par la formule de Raftery est connue pour être instable. En effet, la moyenne harmonique de la vraisemblance est très sensible aux très petites valeurs de $[y|\phi_j^{(g)}, M_j]$ susceptibles d'apparaître lors de l'échantillonnage MCMC. Cela peut notamment affecter la comparaison de modèles pour lesquels la différence de crédibilité est faible. En pratique, si le rapport entre la crédibilité des deux modèles en compétition est élevé, l'approximation F.2.6 est suffisamment précise pour identifier le modèle le plus crédible.

Pour augmenter la stabilité numérique de l'intégrale F.2.4 et de la moyenne harmonique F.2.6, il est judicieux de choisir le plus petit sous-ensemble ϕ de quantités inconnues (Z, θ) pour lequel $[y|M_j, \phi_j]$ (j=1,2) est calculable analytiquement. En effet, dans le cas de modèles hiérarchiques, il existe souvent des propriétés de conjugaison partielles entre les couches de variabilité successives qui permettent de réduire considérablement la dimension de l'intégrale F.2.4. Dans le chapitre 6, les quantités inconnues ϕ_j choisies pour définir la vraisemblance marginale $[y|M_j]$ sont : (μ, ρ) pour le modèle LOL et (δ, a, b) pour le modèle $\Delta\Gamma$.

F.3 Analyse de sensibilité du facteur de Bayes au degré d'information *a priori*

Les facteurs de Bayes sont reconnus pour être sensibles aux choix des lois *a priori* (Kass & Raftery, 1994), (Sinharay & Stern, 2002). En pratique, il est important d'étudier la sensibilité du facteur de Bayes par rapport au degré d'information contenu dans les lois *a priori* attribuées aux paramètres des modèles en compétition. Une approche usuelle simple est d'évaluer le facteur de Bayes pour différentes distributions *a priori*. Les limites de cette méthode sont claires : elle est très côuteuse en temps de calculs et surtout, la gamme des lois *a priori* possibles est limitée puisque le recours à des lois informatives est souvent difficile, en particulier dans le cas hiérarchique où les paramètres inconnus

ont souvent un sens purement conceptuel ne permettant pas de les quantifier *a priori* (cf section 5.1.4).

Les facteurs de Bayes ne sont réellement significatifs que lorsqu'ils sont calculés à partir de lois *a priori* informatives (O'Hagan, 1995). Or, pour les raisons précédemment évoquées, le recours à des lois *a priori* non-informatives est fréquent dans le cas hiérarchique. En particulier, toutes les lois *a priori* utilisées dans ce document sont non-informatives (cf tableau 6.2 pour exemple).

Pour calculer le facteur de Bayes à partir d'un niveau d'information équivalent à la spécification d'une loi *a priori* informative et vérifier la sensibilité du facteur de Bayes au degré d'information *a priori*, nous proposons la méthode suivante. Soient deux modèles statistiques en compétition M_1 et M_2 paramétrés par ϕ_1 et ϕ_2 respectivement. Considérons une partition $(y_{(t)}, y_{(-t)})$ de l'ensemble des observations y.

- $y_{(-t)}$ est l'échantillon d'apprentissage. Il permet d'apprendre sur les paramètres inconnus ϕ_j (j=1,2) pour définir une loi *a priori* informative $[\phi_j|M_j, y_{(-t)}]$. Plus la taille de l'échantillon $y_{(-t)}$ est grande, plus la loi *a priori* induite est informative.
- L'échantillon restant, $y_{(t)}$, est l'échantillon de validation. Il permet de définir le facteur de Bayes du modèle M_1 relativement au modèle M_2, appelé *facteur de Bayes partiel* (O'Hagan, 1995), comme suit :

$$BF_{1,2,(t)} = \frac{[y_{(t)}|M_1, y_{(-t)}]}{[y_{(t)}|M_2, y_{(-t)}]}$$

La formule de Raftery F.2.6 nécessite le calcul de l'inverse des lois marginales $[y_{(t)}|M_j, y_{(-t)}]$ (j=1,2). Celles-ci sont définies par :

$$[y_{(t)}|M_j, y_{(-t)}]^{-1} = \int [y_{(t)}|M_j, \phi_j]^{-1}[\phi_j|M_j, y_{(-t)}, y_{(t)}]d_{\phi_j} = \int [y_{(t)}|M_j, \phi_j]^{-1}[\phi_j|M_j, y]d\phi_j \tag{F.3.7}$$

Cette méthode s'inspire d'une idée à l'origine proposée par Lempers (1971). Son principal avantage est qu'elle permet de faire varier t, c'est-à-dire le niveau d'information *a priori*, sans coût calculatoire supplémentaire. En effet, l'intégrale F.3.7 peut être approximée pour différentes partitions $(y_{(t)}, y_{(-t)})$ et ce, à partir du même échantillon *a posteriori* $\phi_j^{(g)}$ (g=1,2,..,G) issu de la distribution *a posteriori* complète $[\phi_j|M_j, y]$:

$$[y_{(t)}|M_j, y_{(-t)}] \approx \left(\frac{1}{G} \sum_{g=1}^{G} [y_{(t)}|\phi_j^{(g)}, M_j]^{-1} \right)^{-1} \tag{F.3.8}$$

D'autres extensions de cette idée ont été plus récemment proposées sous le nom de *facteur de Bayes intrinsèque* (Berger & Perrichi, 1996) ou encore *facteur de Bayes fractionnel* (O'Hagan, 1995). Leur principale différence concerne la méthode de sélection et le choix de la taille de l'échantillon d'apprentissage $y_{(-t)}$. Dans ce travail, ces deux aspects n'ont pas fait l'objet d'études particulières. Les échantillons d'apprentissage ont simplement été définis de telle sorte qu'ils contiennent des données indépendantes de celles contenues dans l'échantillon de validation (cf chapitre 6).

L'apport principal de notre méthode concerne la méthode de calcul du facteur de Bayes qui, par rapport aux différentes méthodes précédemment citées, est basée sur la formule de Raftery F.2.6.

Annexe G

Prédictions bayésiennes et critères de perte prédictive *a posteriori*

G.1 Principe de la prédiction bayésienne

Etant donné un vecteur d'observations y_{obs} et un modèle statistique M_j, de nouvelles observations y_{new} peuvent être prédites en tirant aléatoirement dans leur distribution *prédictive*, notée $[y_{new}|y_{obs}, M_j]$. Cette distribution s'obtient par intégration de la densité conditionnelle jointe $[y_{new}, \phi_j|y_{obs}, M_j]$ sur les valeurs possibles d'un vecteur ϕ_j de grandeurs inconnues du modèle M_j :

$$[y_{new}|y_{obs}, M_j] = \int [y_{new}|\phi_j, M_j][\phi_j|y_{obs}, M_j]d\phi_j \tag{G.1.1}$$

Le calcul analytique de l'intégrale multidimensionnelle G.1.1 est généralement impossible si bien que la loi prédictive n'est pas explicitement connue. L'approche classique consiste à tirer aléatoirement dans la prédictive $[y_{new}|y_{obs}, M_j]$ par simulations Monte-Carlo. y_{new} est supposé suivre la même distribution que y_{obs}. Il suffit donc de générer un échantillon de G réplicats $(y_{new}^{(1)}, \ldots, y_{new}^{(G)})$ en tirant itérativement dans la distribution $[y|\phi_j, M_j]$ à partir d'un échantillon MCMC de valeurs de ϕ_j.

G.2 Le critère de perte prédictive *a posteriori*

Soient m modèles en compétition $(M_1, \ldots, M_m)$. Gelfand et Ghosh (1998) ont proposé un critère prédictif qui permet de sélectionner le modèle fournissant les échantillons prédictifs les plus précis et les plus fidèles aux observations y_{obs} disponibles. Soient $y_{new,l}$ et $y_{obs,l}$ les l-ième composants des vecteur y_{new} et y_{obs} respectivement. Le "meilleur" modèle est celui qui permet de minimiser :

$$\min_{a_l} \qquad \mathbb{E}_{[y_{new,l}|y_{obs,l},M_j]} L(y_{new,l}, a_l; y_{obs}, M_j) \tag{G.2.2}$$

où $L(y_{new,l}, a_l; y_{obs}, M_j)$ désigne la fonction de perte quadratique suivante :

$$L(y_{new,l}, a_l; y_{obs}, M_j) = (y_{new,l} - a_l)^2 + k(y_{obs,l} - a_l)^2 \qquad k \geq 0 \tag{G.2.3}$$

$L(y_{new,l}, a_l; y_{obs}, M_j)$ tient compte du fait que $y_{new,l}$ et $y_{obs,l}$ sont issus d'une même distribution. La quantité a_l assure le compromis entre un biais par rapport aux observations

et la variance des prédictions. Le poids k indique la perte relative engendrée par un écart par rapport à $y_{obs,l}$ (i.e., fidélité aux observations) comparé à un écart par rapport à μ_l^j (i.e., variance de prédiction).

Compte-tenu de la fonction de perte choisie, l'expression G.2.2 s'écrit :

$$\min_{a_l} \qquad \sigma_l^{2(j)} + (a_l - \mu_l^{(j)})^2 + k(a_l - y_{obs,l})^2 \tag{G.2.4}$$

où μ_l^j et $\sigma_l^{2(j)}$ désignent respectivement la moyenne et la variance de la distribution prédictive de $y_{new,l}$ sous le modèle M_j.

L'expression G.2.4 est minimisée pour $a_l = \frac{1}{k+1}(ky_{l,obs} + \mu_l^{(j)})$. Le critère de Gelfand et Ghosh (1998), nommé critère de perte prédictive *a posteriori*, est ainsi défini par :

$$D_k(M_j) = \sum_{l=1}^{n} \sigma_l^{2(j)} + \frac{k}{k+1} \sum_{l=1}^{n} (\mu_l^j - y_{obs,l})^2 \tag{G.2.5}$$

Le "meilleur" modèle est celui pour lequel le critère D_k est minimal. Dans l'expression G.2.5, posons $G(M_j) = \sum_{l=1}^{n} (\mu_l^j - y_{obs,l})^2$ et $P(M_j) = \sum_{l=1}^{n} \sigma_l^{2(j)}$. $G(M_j)$ mesure la fidélité des prédictions du modèle M_j par rapport aux observations et $P(M_j)$ désigne un terme de pénalité. Pour des modèles sous ou sur-ajustés, on s'attend à ce que les variances de prédictions soient grandes. Ainsi, pour des modèles trop simples, $D_k(M_j)$ sera élevé. Au contraire, quand les modèles M_j deviennent complexes, un compromis est généralement observé : $G(M_j)$ a tendance à décroître et $P(M_j)$ à augmenter. Les modèles complexes sont pénalisés et un choix parcimonieux est recommandé.

L'approche prédictive ne nécessite pas de spécifier la dimension du modèle utilisé. Ceci est un avantage, notamment dans le cas de modèles hiérarchiques dont la notion de dimension n'est pas bien définie.

Annexe H

Comparaison de l'autocorrélation spatiale des effets aléatoires μ et ρ du modèle LOL

La notion d'autocorrélation spatiale mesure essentiellement la ressemblance entre unités géographiques voisines. Il existe plusieurs indices pour mesurer l'autocorrélation spatiale (indice de Moran, coefficient de Geary) basés sur la notion de voisinage (contiguïté, distance). Dans ce travail, l'indice de Moran a été utilisé comme mesure exploratoire de la ressemblance entre les unités géographiques voisines du Golfe-du-Saint-Laurent.

Le modèle LOL est caractérisé par deux effets aléatoires spécifiques pour chaque unité géographique : μ_i et ρ_i (i=1,2,..,38). L'ajout d'une structure spatiale latente dans le modèle suppose de choisir l'effet le plus pertinent à spatialiser. L'annexe suivante décrit l'analyse exploratoire réalisée.

H.1 L'indice de Moran

L'indice de Moran se présente comme le rapport de l'auto-covariance sur la variance :

$$I = \frac{N \sum_{i,j} \omega_{ij}(x_i - \overline{x})(x_j - \overline{x})}{s_0 \sum_i (x_i - \overline{x})^2} \tag{H.1.1}$$

Dans cette expression :

- x_i désigne la valeur de la variable dans l'unité i et de moyenne $\overline{x}$
- N désigne le nombre total d'unités géographiques.
- W désigne la matrice $N \times N$ des pondérations basées sur la notion de voisinage choisie.
- $s_0 = \sum_{i,j} \omega_{ij}$ désigne le nombre total de paires de voisins.

Le résultat du calcul du I de Moran s'interprète comme un coefficient de corrélation linéaire. Une valeur proche de -1 s'interprète comme une autocorrélation spatiale négative (les voisins ont des valeurs opposées) et proche de +1 comme une autocorrélation spatiale positive (les voisins ont des valeurs semblables). On notera que I n'est pas strictement borné entre -1 et +1.

La matrice de voisinage choisie est la matrice basée sur la contiguïté W où $\omega_{ij} = 1$ si les unités géographiques i et j ont une frontière commune et $\omega_{ij} = 0$ sinon. La standardisation

de W consiste à diviser chaque élément ω_{ij} par le nombre de voisins de l'unité i (la somme de chaque ligne de W est alors égale à 1).

En pratique, l'indice de Moran est calculé à partir de données $x = (x_1, \ldots, x_n)$ observées en chacune des n unités géographiques considérées. Dans le cas du modèle LOL, j'ai souhaité le calculer pour chaque effet aléatoire latent spécifique aux unités géographiques i.e., $\mu = (\mu_1, \ldots, \mu_{38})$ et $\rho = (\rho_1, \ldots, \rho_{38})$. Dans la suite, je décris la méthode suivie pour le vecteur μ. La même démarche a été suivie pour le vecteur ρ. Je suis partie de r configurations MCMC obtenues pour chaque vecteur μ lors de l'inférence du modèle $LOL^{\otimes 38}$. A chaque configuration $\mu^l = (\mu_1^l, \mu_2^l, \ldots, \mu_{38}^l)$ $(l = 1, 2, \ldots, r)$ est associée une valeur de l'indice de Moran. Les r configurations MCMC ont ainsi permis d'obtenir un échantillon *a posteriori* de l'indice de Moran.

L'indice de Moran se base sur la moyenne. Il est donc sensible aux valeurs aberrantes. De plus, il est moins fragile que le coefficient de Geary car il mesure les écarts à la moyenne et non pas les écarts entre les voisins. Ces deux indices sont relativement similaires mais l'indice de Moran a l'avantage supplémentaire d'être implémenté dans beaucoup de logiciels (ex : fonction *Moran.I* du package *ape* de R).

H.2 Test d'autocorrélation spatiale

Pour tester l'existence d'une autocorrélation spatiale en utilisant l'indice de Moran, il faut connaître la distribution de I sous l'hypothèse nulle H_0 d'absence d'auto-corrélation. La distribution de I sous H_0 a été étudiée pour les variables latentes μ dans le cadre d'un modèle probabiliste de variables gaussiennes indépendantes. L'hypothèse nulle H_0 testée est :

$$log(\mu_i) \underset{i.i.d}{\sim} \mathcal{N}(m_1, \sigma_1^2)$$

Sous cette hypothèse, les moments d'ordre 1 et 2 de I sont :

$$\mathbb{E}(I) = \frac{-1}{N-1} \qquad \mathbb{V}ar(I) = \frac{n^2(n-1)S_1 - n(n-1)S_2 - 2S_0^2}{(n+1)(n-1)S_0^2} \tag{H.2.2}$$

avec :

- $S_0 = \sum_{i \neq j} w_{ij}$
- $S_1 = \frac{1}{2} \sum_{i \neq j} (\omega_{ij} + \omega_{ji})^2$
- $S_2 = \sum_k (\sum_j \omega_{kj} + \sum_i \omega_{ik})^2$

La matrice W étant standardisée, il vient $s_0 = N$. Sous H_0, la distribution de I est asymptotiquement normale (si chaque unité géographique a un nombre fini de liens avec les autres).

En pratique, l'indice de Moran I_{obs} est calculé à partir de données observées en chacune des n unités géographiques considérées puis le test d'écart à l'indépendance passe par le calcul d'une p-valeur : $\mathbb{P}_{H_0}[I \geq I_{obs}]$. Dans ce travail, j'ai calculé, pour chaque vecteur d'effets aléatoires μ et ρ, un échantillon de r valeurs de l'indice de Moran à partir des échantillons *a posteriori* obtenus. Puis, pour définir l'effet aléatoire (μ ou ρ) pour lequel il était le plus pertinent de modéliser une structuration spatiale, j'ai calculé deux indices : la p-valeur moyenne et la distance de Kullback-Leibler empirique entre la distribution théorique de I sous H_0 et l'échantillon *a posteriori* d'indices de Moran.

Espèce	Effet aléatoire	p-valeur$_{moy}$	KL$_{emp}$
Anémones	μ	0.050	0.694
	ρ	0.121	0.523
Oursins	μ	0.0005	0.818
	ρ	0.0049	0.743
Soleils de mer	μ	0.0004	0.881
	ρ	0.469	0.0015

Tableau H.1 – P-valeurs moyennes et distances de Kullback-Leibler empiriques entre la distribution de I sous H_0 et l'échantillon *a posteriori* obtenu

H.3 Application aux données d'abondance du Golfe-du-Saint-Laurent

Le tableau H.1 indique, pour chaque effet aléatoire et pour 3 espèces particulières (anémones, oursins, soleils de mer), les p-valeurs moyennes et distances de Kullback-Leibler empiriques calculées à partir d'un échantillon de 2000 valeurs de l'indice de Moran.

Dans le cadre d'un test à 5%, l'hypothèse d'indépendance est rejetée pour l'effet aléatoire μ pour les 3 espèces considérées. Les p-valeurs moyennes sont inférieures à 0.05. En revanche, pour l'effet aléatoire ρ, l'hypothèse d'indépendance est acceptée pour les anémones et les soleils de mer soit, dans 2 cas sur 3. Dans le cas des oursins, les deux effets aléatoires semblent auto-corrélés spatialement bien que cette auto-corrélation soit plus forte pour le paramètre μ avec une p-valeur moyenne plus faible. Les distances de Kullback-Leibler confirment ces résultats. En effet, les distances calculées pour l'effet aléatoire ρ sont systématiquement inférieures à celles obtenues pour l'effet aléatoire μ. Cela signifie que la distribution empirique de l'indice de Moran obtenue est plus "proche" de sa distribution théorique sous l'hypothèse d'indépendance pour l'effet aléatoire ρ que pour l'effet aléatoire μ. Le tableau H.1 vient illustrer ces résultats.

Les résultats de cette analyse exploratoire nous ont conduit à modéliser les effets aléatoires μ_i avec le modèle BYM plutôt que les ρ_i. Plus de détails sur l'indice de Moran sont donnés dans des ouvrages de référence comme (Banerjee & al., 2004), (Cressie, 1993).

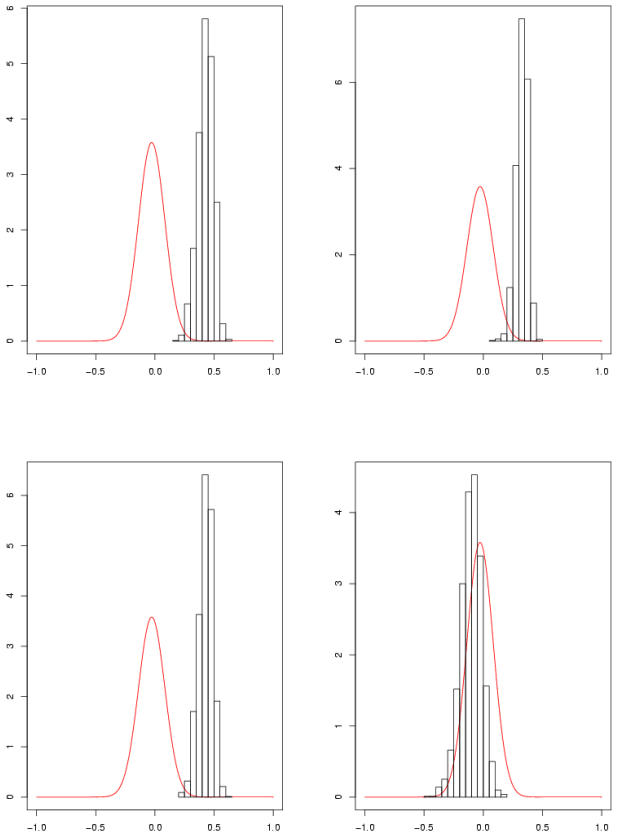


Figure H.1 – Distributions *a posteriori* de l'indice de Moran pour les effets aléatoires μ (gauche) et ρ (droite). La densité rouge représente la distribution de l'indice de Moran sous l'hypothèse d'indépendance. (Haut) : cas des oursins (Bas) : cas des soleils de mer

Bibliographie

Aitchison, J. (1955). On the distribution of a positive random variable having a discrete probability mass at the origin. *Journal of the American Statistical Association*, *50*(271), 901-908.

Auster, P. J., & al. (1996). The impacts of mobile fishing gear on seafloor habitats in the gulf of maine (northwest atlantic) : Implications for conservation of fish populations. *Reviews in Fisheries Science*, *4*(2), 185-202.

Bamshad, M., & al. (2003). Human population genetic structure and inference of group membership. *American Journal of Human Genetics*, *72*(3), 578-589.

Banerjee, S., & al. (2004). *Hierarchical modeling and analysis for spatial data*. Chapman and Hall/ CRC.

Barry, R. P., & VerHoef, J. M. (1996). Blackbox kriging : Spatial prediction without specifying variogram models. *Journal of Agricultural, Biological, and Environmental Statistics*, *1*(3), 297-322.

Berger, J. O. (1985). *Statistical decision theory and bayesian analysis*. Springer-Verlag.

Berger, J. O., & Perrichi, L. R. (1996). The intrinsic bayes factor for model selection and prediction. *Journal of the American Statistical Association*, *91*(433), 109-122.

Bernier, J., & Fandeux, D. (1970). Théorie du renouvellement- application à l'étude statistique des précipitations mensuelles. *Revue de statistique appliquée*, *18*(2), 75-87.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society. Series B.*, *36*(2), 192-236.

Besag, J., & al. (1991). Bayesian image restoration with two applications in spatial statistics. *Annals of the Institute of Statistics and Mathematics*, *43*(1), 1-51.

Blum, M., & al. (2004). Brownian models and coalescent structures. *Theoretical Population Biology*, *65*(3), 249-261.

Brooks, S. P., & Gelman, A. (1998). Alternative methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, *7*(4), 434-455.

Calder, A. C. (2003). *Exploring latent structure in spatial temporal processes using process convolutions*. Unpublished doctoral dissertation, Institute of Statistics and Decision Sciences, Duke University.

Cann, H., & al. (2002). A human genome diversity cell line panel. *Science*, *296*(5566), 261-262.

Cipra, B. A. (1987). An introduction to the ising model. *The American Mathematical Monthly*, *94*(10), 937-959.

Clark, J. S. (2007). *Models for ecological data : An introduction*. Princeton university press.

Clark, J. S., & Gelfand, A. E. (2006). A future for models and data in environmental science. *Trends in Ecology & Evolution*, *21*(7), 375-380.

Coe, R., & Stern, R. D. (1982). Fitting models to daily rainfall data. *Journal of Applied Meteorology*, *21*(7), 1024-1031.

Collie, J. S., & al. (2000). A quantitative analysis of fishing impacts on shelf-sea benthos. *Journal of Animal Ecology*, *69*(5), 785-798.

Cong, H., & Bradley, P. C. (2001). Markov chain monte-carlo methods for computing bayes factors : A comparative review. *Journal of the American Statistical Association*, *96*(455), 1122-1132.

Cressie, N. A. C. (1993). *Statistics for spatial data*. John Wiley & Sons, Inc.

Dawson, K. J., & Belkhir, K. (2001). A bayesian approach to the identification of panmictic populations and the assignment of individuals. *Genetical Research*, *78*(1), 59-77.

Falush, D., & al. (2003). Inference of population structure using multilocus genotype data : linked loci and correlated allele frequencies. *Genetics*, *164*, 1567-1587.

Feller, W. (1968). *An introduction to probability theory and its applications, volume 1, 3rd edition*. John Wiley.

Fletcher, D., & al. (2005). Modelling skewed data with many zeros : a simple approach combining ordinary and logistic regression. *Environmental and ecological statistics*, *12*(1), 45-54.

Gelfand, A. E., & al. (1998). Spatio-temporal modeling of residential sales data. *Journal of Business & Economic Statistics*, *16*(3), 312-321.

Gelfand, A. E., & Ghosh, S. K. (1998). Model choice : a minimum posterior predictive loss approach. *Biometrika*, *85*(1), 1-11.

Gelman, A., & Meng, X. L. (1998). Simulating normalizing constants : From importance sampling to bridge sampling to path sampling. *Statistical Science*, *13*(2), 163-185.

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, *7*(4), 457-472.

Geman, S., & Geman, D. (1984). Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *6*(6), 721-741.

Green, P., & Richardson, S. (2002). Hidden markov models and disease mapping. *Journal of the American Statistical Association*, *97*(460), 1055-1070.

Grzebyk, M., & Wackernagel, H. (1994). Multivariate analysis and spatial/temporal scales : real and complex models. *Proceedings of the XVIIth International Biometrics Conference*, *1*, 19-33.

Guillot, G., & al. (2005). A spatial statistical model for landscape genetics. *Genetics*, *170*, 1261-1280.

Hall, D. B. (2000). Zero-inflated poisson and binomial regression with random effects : A case study. *Biometrics*, *56*(4), 1030-1039.

Heilbron, D. C. (1994). Zero-altered and other regression models for count data with added zeros. *Biometrical Journal*, *36*(5), 531-547.

Higdon, D. (2002). Space and space-time modeling using process convolutions. In *Quantitative methods for current environmental issues* (p. 37-56). Springer.

Higdon, D., & al. (1998). Non-stationary spatial modeling. *Bayesian Statistics*, *6*, 761-768.

Hurlbut, T., & Clay, D. (1990). *Protocols for research vessel cruises within the gulf region*

(demersal fish)(1970-1987) (Tech. Rep.).
Ickstadt, K., & Wolpert, R. L. (1999). Spatial regression for marked point processes. *Bayesian Statistics*, *6*, 323-341.
Johnson, N. L., & Kotz, S. (1969). *Distributions in statistics : Discrete distributions.* Haughton Mifflin, Boston, MA.
Jonsen, I. D., & al. (2003). Meta-analysis of animal movement using state-space models. *Ecology*, *84*(11), 3055-3063.
Kaiser, M. J., & al. (2002). Modification of marine habitats by trawling activities : prognosis and solutions. *Fish and Fisheries*, *3*(2), 114-136.
Kass, R. E., & al. (1998). Markov chain monte carlo in practice : a roundtable discussion. *American Statistical Association*, *52*(2), 93-100.
Kass, R. E., & Raftery, A. E. (1994). Bayes factors. *Journal of the American Statistical Association*, *90*(430), 773-795.
Kern, J. C. (2000). *Bayesian process-convolution approaches to specifying spatial dependence structure.* Unpublished doctoral dissertation, Duke University.
Legendre, P., & Legendre, L. (1998). *Numerical ecology.* Elsevier Science, 2nd english edition.
Lempers, F. B. (1971). *Posterior probabilities of alternative linear models.* Rotterdam : University Press.
Lo, N. C. H., & al. (1992). Indices of relative abundance from fish spotter data based on delta-lognormal models. *Canadian Journal of Fisheries and Aquatic Science*, *49*, 2515-2526.
Loring, D. H., & Nota, D. J. G. (1973). Morphology and sediments of the gulf of st.lawrence. *Bulletin of the Fisheries Research Board of Canada*, *182*.
Manel, S., & al. (2003). Landscape genetics : combining landscape ecology and population genetics. *Trends in Ecology and Evolution*, *18*(4), 189-197.
Martin, T. G., & al. (2005). Zero tolerance ecology : improving ecological inference by modelling the source of zero observations. *Ecology Letters*, *8*, 1235-1246.
McCarthy, M. A. (2007). *Bayesian methods for ecology.* Cambridge University Press.
Metropolis, N., & al. (1953). Equations of state calculations by fast computing machines. *Journal of Chemical Physics*, *21*(6), 1087-1091.
Morlat, G. (n.d.). *La distribution des débits de fuite sur une conduite de gaz* (Tech. Rep.). EDF - Etudes Economiques Générales.
Myers, R. A., & Pepin, P. (1990). The robustness of lognormal based estimators of abundance. *Biometrics*, *46*(4), 1185-1192.
O'Hagan, A. (1991). Contribution to the discussion of posterior bayes factors. *Journal of the Royal Statistical Society. Series B.*, *53*(136), 132-133.
O'Hagan, A. (1995). Fractional bayes factors for model comparison. *Journal of the Royal Statistical Society. Series B.*, *57*(1), 99-138.
Parent, E., & Bernier, J. (2007). *Le raisonnement bayésien. modélisation et inférence.* Springer.
Pennington, M. (1983). Efficient estimators of abundance for fish and plankton surveys. *Biometrics*, *39*(1), 281-286.
Pinkus, A., & Zafrany, S. (1997). *Fourier series and integral transforms.* Cambridge University Press.
Pritchard, J. K., & al. (2000). Inference of population structure using multilocus genotype data. *Genetics*, *155*, 945-959.

Richardson, S., & al. (2006). Bayesian spatio-temporal analysis of joint patterns of male and female lung cancer risks in yorkshire (uk). *Statistical Methods in Medical Research*, *15*(4), 385-407.

Ridout, M., & al. (1998). Models for count data with many zeros. *International Biometric Conference*.

Ripley, B. (1996). *Pattern recognition and neural networks*. Oxford University Press.

Robert, C. P. (2006). *Le choix bayésien*. Springer.

Robert, C. P., & Casella, G. (1998). *Monte-carlo statistical methods*. Springer.

Roberts, G. O., & Smith, A. F. M. (1993). Simple conditions for the convergence of the gibbs sampler and the metropolis-hastings algorithms. *Stochastic processes and their applications*, *49*, 207-216.

Rosenberg, N., & al. (2005). Clines, clusters and the effect of study design on the influence of human population structure. *PLoS Genetics*, *1*(6), 660-671.

Savenkoff, C., & al. (2007). Effects of fishing and predation in a heavily exploited ecosystem : Comparing periods before and after the collapse of groundfish in the southern gulf of st. lawrence (canada). *Ecological modelling*, *204*(1), 115-128.

Serre, D., & Paabo, S. (2004). Evidence for gradients of human genetic diversity within and among continents. *Genome Research*, *14*(9), 1679-1685.

Sinharay, S., & Stern, H. S. (2002). On the sensitivity of bayes factors to the prior distributions. *The American Statistician*, *56*(6), 196-201.

Spiegelhalter, D. J., & al. (n.d.). Computation on bayesian graphical models.

Spiegelhalter, D. J., & al. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society Serie B*, *64*(3), 1-34.

Stefansson, G. (1996). Analysis of groundfish survey abundance data : combining the glm and delta approaches. *ICES Journal of Marine Science*, *53*(3), 577-588.

Stenseth, N. C., & al. (2005). *Marine ecosystems and climate variation. the north atlantic. a comparative perspective*. Oxford University Press.

Stern, C. (1943). The hardy weinberg law. *Science*, *97*, 137-138.

Swenson, J., & al. (1994). Size, trend, distribution and conservation of the brown bear, ursus arctos, population in sweden. *Biological Conservation*, *70*, 9-17.

Thiébaux, H. J., & Pedder, M. A. (1987). *Spatial objective analysis : With applications in atmospheric science.* Academic Press.

Thrush, S. F., & al. (1998). Disturbance of the marine benthic habitat by commercial fishing : impacts at the scale of the fishery. *Ecological Applications*, *8*(3), 866-879.

Waits, L., & al. (2000). Nuclear dna microsatellite analysis of genetic diversity and gene flow in the scandinavian brown bear *Ursus Arctos*. *Molecular Ecology*, *9*(4), 610-621.

Wickle, C. K. (2003). Hierarchical models in environmental science. *International Statistical Review*, *71*(2), 181-220.

Wickle, C. K., & al. (2001). Spatio-temporal hierarchical bayesian modeling : Tropical ocean surface winds. *Journal of the American Statistical Association*, *96*(454), 382-397.

Wu, F. (1982). The potts model. *Reviews of Modern Physics*, *54*, 235-268.