Modélisation statistique de distribution spatiale et temporelle d'espèces d'invertébrés dans le Golfe du Saint Laurent.

Jean-Baptiste Lecomte

ParisTech
INSTITUT DES SCIENCES ET TECHNOLOGIES
PARIS INSTITUTE OF TECHNOLOGY

AgroParisTech
INSTITUT DES SCIENCES ET INDUSTRIES DU VIVANT ET DE L'ENVIRONNEMENT
PARIS INSTITUTE OF TECHNOLOGY FOR LIFE, FOOD AND ENVIRONMENTAL SCIENCES

Ecole Doctorale ABIES
Agriculture
Agriculture
Alimentation
Food
Biologie
Biology
Environnement
Environment
Santé
Health

Doctorat ParisTech

T H È S E

pour obtenir le grade de docteur délivré par

L'Institut des Sciences et Industries du Vivant et de l'Environnement

(AgroParisTech)

Spécialité : Statistiques

présentée et soutenue publiquement par

Jean-Baptiste LECOMTE

le 16 avril 2014

Modélisation statistique de distribution spatiale et temporelle d'espèces d'invertébrés dans le Golfe du Saint Laurent.

Directeur de thèse : **Eric PARENT**
Co-encadrement de la thèse : **Liliane BEL**

Jury

M. Nicolas BEZ, Directeur de recherche, IRD	Rapporteur
M. Aurélien Besnard, Enseignant-Chercheur, CEFE/CNRS	Rapporteur
M. David Makowski, Directeur de recherche, INRA	Examinateur
Mme. Anick BRIND'AMOUR, Chargé de recherche, IFREMER	Examinateur
M. Samuel SOUBEYRAND, Chargé de recherche, INRA	Examinateur
M. Christophe LAPLANCHE, Enseignant-Chercheur, ENSAT/ECOLAB	Examinateur
Mme. Liliane BEL, Professeur, UMR 518 AgroParisTech/INRA,	Directeur de thèse
M. Eric PARENT, Ingénieur ENGREF, UMR 518 AgroParisTech/INRA	Directeur de thèse

AgroParisTech
UMR 518 AgroParisTech/INRA
Math. Info. Appli., F-75005 Paris, France

Le doute n'est pas une condition agréable, mais la certitude est absurde.

Voltaire

REMERCIEMENTS

Je m'estime chanceux d'avoir pu effectuer mes travaux de thèse aussi bien entouré tant professionnellement, que personnellement. Les quelques mots qui suivent me permettent de remercier ces personnes qui ont fait que mes années de thèse ont été très agréables et enrichissantes.

Je souhaitai remercier Éric Parent et Liliane Bel de m'avoir accompagné et encadré pendant cette thèse. Merci de m'avoir fait confiance pour mener à bien ces travaux, de votre disponibilité et écoute ainsi que pour vos conseils avisés. Votre duo de choc m'a beaucoup apporté pendant ces trois années. J'espère pouvoir continuer à travailler avec vous malgré la grande distance qui nous sépare.

Un grand merci à Hugues Benoît qui a suivi mes travaux depuis Moncton, et avec qui j'ai partagé de très bons moments lors de la campagne de pêches 2012 dans le golfe du Saint-Laurent. Merci de m'avoir accueilli dans ta famille et de m'avoir fait découvrir le Nouveau-Brunswick. Je ne sais pas si mon futur passage au Canada me permettra de revenir à Moncton, mais je repasserai très volontiers quelques jours avec toi et toute ta famille!

Je remercie également Marie-Pierre Étienne avec qui j'ai pu discuter sans retenue et qui a toujours été de bons conseils. Je suis vraiment très heureux d'avoir pu travailler avec toi. Tu as été pour moi plus qu'une collaboratrice lors de ces trois années et j'espère que notre amitié nous permettra de nous retrouver à Nanaimo ou ailleurs.

Je remercie également Étienne Rivot et Sophie Ancelet pour m'avoir donné leurs avis et leurs impressions chaque année de thèse lors de mon comité de suivi, mais également à chacune de nos rencontres plus informelles. Merci pour leurs conseils qui m'ont permis quelques recadrages des travaux en cours. Merci également à Nicolas Bez et Aurélien Besnard d' avoir accepté de rapporter mes travaux et pour leurs remarques constructives. Je souhaitai également remercier les membres du jury présents lors de ma soutenance : Anick Brind'amour, Samuel Soubeyrand, David Makowski et Christophe Laplanche, qui

avec leurs questions très pertinentes m'ont permis de prendre du recul sur mes travaux.

Merci à l'ensemble de l'UMR 518, j'ai passé de très bonnes années parmi vous, ponctuées notamment de pauses café mémorables! J'ai également vécu de très bons moments avec Caroline Berard et Stevenn Volant avec qui j'ai partagé mon bureau à mon arrivée dans le laboratoire. Vous avez été de très bons conseils pour le jeune doctorant que j'étais et je vous en remercie. Après votre départ, j'ai malheureusement perdu ma place dans notre beau bureau pour l'échanger contre un bureau plus petit et quelque peu isolé... Cependant, j'ai trouvé dans ce nouveau bureau deux amis. Merci à Frédéric Fer pour sa bonne humeur en toutes circonstances, son humour et ses débats toujours à la pointe de l'actualité. Merci à Aurélien Bechler pour les discussions passionnées (sportives et cinématographiques), les vidéos ludiques et surtout les jeux, toujours drôles, organisés par tes soins. Merci également pour ton soutien vis-à-vis des arguments ravageurs de Frédéric. De cette belle rencontre est né le Bureau Des Doctorants Du Bas plus connu sous le nom de BDDDB organisme à but non lucratif et festif, toujours présent lors des moments importants, j'espère que vous le ferez prospérer malgré mon absence.

Merci à ma famille, tout particulièrement à mon père et ma mère pour leur soutien inconditionnel. Vous m'avez accompagné lors de ces longues études et je vous en remercie. Merci également à Héloïse et Quentin, ma soeur et mon frère chéris, j'ai eu la chance de vivre avec vous à Paris et j'en garderai de très bons souvenirs. Merci également aux Deleys, famille adoptive, avec qui je partage de très bons moments depuis que je vous ai rencontré.

Je remercie également les copains, les bicouches lipidiques, pour les soirées, les apéros, les barbecs, les concerts, les voyages! Dans le désordre merci aux Toulousains, aux Nantais, aux Parisiens, Rennais et Néoquébécois et non-affiliés.

Merci à Noémie, mon Amour, d' avoir toujours été à mes côtés pendant les moments de joie et les périodes plus compliquées. Je te remercie pour ton soutien quotidien et pour les moments d'évasions partagés ensemble qui m'ont permis de m'échapper de mes problèmes de doctorant.

Enfin, je dédie cette thèse à mes deux grands-pères, Armand et Yannick, qui, je pense, auraient apprécié chacun à leur manière le travail accompli.

TABLE DES MATIÈRES

1

INTRODUCTION

Comprendre la distribution spatiale et temporelle des organismes a longtemps été un des objectifs majeurs de l'écologie (Andrewartha et Birch, 1954; Krebs, 1978; MacKenzie, 2006). Les premières études s'intéressant à la répartition d'espèces ont d'abord été qualitatives (Grinnell, 1904) et ont notamment permis de mettre en évidence le rôle important du climat (von Humboldt et Bonpland, 1807; Candolle, 1855), mais pourquoi aujourd'hui s'intéresser à la distribution des espèces ? La première raison est de comprendre les relations entre les espèces et leur environnement biotique et abiotique. Dans ce cas, l'étude de la répartition spatiale de l'espèce n'est pas l'objectif principal de l'étude, mais cela permet de tester des hypothèses écologiques ou phylogénétiques de relation entre l'espèce et son habitat (Peterson et Holt, 2003; Jones *et al.*, 2007). Cependant, les écologues utilisent majoritairement ces études pour interpoler ou extrapoler, à partir d'observations, la distribution spatiale d'une espèce. Ainsi, ces études fournissent des cartes de prédictions de répartition qui peuvent être utilisées pour mesurer et prédire la future distribution d'une espèce en réponse aux changements climatiques (Thuiller *et al.*, 2005; Fitzpatrick *et al.*, 2007; Araújo et New, 2007), identifier et gérer des zones de protections naturelles (Perry et Smith, 1994; Williams et Bax, 2001; Hobday et Hartmann, 2006; Hartog *et al.*, 2011), ou encore prédire la dispersion d'une espèce invasive (Cook *et al.*, 2007). Toutes ces applications, en plein essor depuis ces dernières décennies, sont possibles grâce à l'amélioration des techniques d'acquisition de données. En effet, le développement des systèmes d'information géographique (SIG) a permis de stocker et de manipuler facilement des données environnementales comme des données géoréférencées d'observations d'espèces (Foody, 2008). Ces études sont également possibles grâce au développement d'outils statistiques permettant de mettre

en évidence les relations entre espèces et environnement. Une grande variété de modèles est ainsi utilisée pour étudier la distribution d'une espèce ou d'une population. Ces modèles sont regroupés dans une famille nommée «species distribution models» (SDMs) qui comprend des modèles bioclimatiques, également appelés enveloppes climatiques, des modèles de niche écologique, des modèles d'habitats ou encore des fonctions de sélection des ressources (Elith et Leathwick, 2009).

Les modèles bioclimatiques ont été développés pour définir les limites spatiales de la distribution d'une espèce. Cette distribution spatiale est obtenue à l'aide de variables climatiques généralement corrélées à la présence (ou l'absence) de l'espèce considérée. Ces modèles sont ensuite utilisés pour extrapoler les changements de distribution spatiale de l'espèce différents scénarii climatiques. Cependant, les prédictions obtenues avec ces modèles sont à considérer avec précaution (Guisan et Thuiller, 2005). En effet, les facteurs climatiques ne sont pas les seuls à influencer significativement la distribution d'une espèce (Hampe, 2004), l'évolution de la fragmentation de l'habitat et les limites de dispersion des espèces (Iverson et Prasad, 2002), l'impact de l'augmentation de la concentration atmosphérique en dioxyde de carbone ou encore les différences génétiques au sein d'une même espèce sont autant de facteurs à prendre en considération lors de prédictions de distribution d'espèces à long terme (Heikkinen *et al.*, 2006).

Les modèles de niche écologique tendent à ajouter ces autres facteurs tout aussi importants dans la répartition d'une espèce (Hutchinson, 1957). Ces modèles de niche associent donc variables climatiques, comme les modèles bioclimatiques, mais également des facteurs comme les facteurs génétiques. Dans le contexte des SDMs, le terme de niche est souvent considéré comme les besoins environnementaux d'une espèce pour subsister sans émigration (Grinnell, 1917) et est généralement représenté comme l'espace défini par l'hypervolume de variables environnementales qui décrit les limites environnementales d'une espèce. Un postulat important pour l'utilisation des modèles de niche écologique est que l'espèce doit être à l'équilibre avec son environnement (Guisan et Zimmermann, 2000; Guisan et Thuiller, 2005). En effet, les SDMs ne sont pas construits de manière à séparer la réponse transitoire de la réponse à l'équilibre d'une espèce avec son environnement. Cependant, la distribution actuelle d'un grand nombre d'espèces est correctement pré-

dite par des modèles bioclimatiques validant ainsi *a posteriori* la capacité prédictive des modèles et suggérant que les espèces considérées sont à l'équilibre avec leur environnement (Peterson *et al.*, 2002; Pearson *et al.*, 2002).

Il existe deux concepts de la niche écologique d'une espèce, le premier étant la niche fondamentale ou théorique et le second la niche réalisée. La niche fondamentale est une fonction des performances physiologiques associées aux contraintes de l'environnement. La niche réalisée ajoute à la niche fondamentale les relations biotiques qui affectent l'espèce étudiée (compétition, prédation, parasitisme). Les modèles de niche fondamentale sont construits pour comprendre les processus responsables de la distribution d'une espèce et sont généralement des modèles mécanistes. Inversement, les modèles s'intéressant à la niche écologique réalisée sont généralement des modèles statistiques associant variables environnementales et relations biotiques (Kearney et Porter, 2009). De plus, les modèles de niche fondamentale simulent une distribution selon des contraintes théoriques physiologiques, contrairement aux modèles de niche réalisée qui s'appuie sur des données récoltées sur le terrain pour simuler la répartition d'une espèce (Guisan et Zimmermann, 2000). Cependant, il est parfois difficile d'inclure ces relations biotiques (compétition, prédation) dans la modélisation de la distribution d'une espèce. La niche réalisée est alors approchée par des modèles d'habitats qui sont des modèles statistiques de corrélation associant variables environnementales abiotiques (p. ex. température, altitude, *etc.*) et parfois biotiques (p. ex. couverture végétale) (Hirzel et Le Lay, 2008). Lorsque ces modèles sont utilisés pour prédire la distribution spatiale d'une espèce, ils peuvent également être appelés «predictive habitat distribution models» (Guisan et Zimmermann, 2000) et «spatially explicit habitat suitability models» (Rotenberry *et al.*, 2006). Les fonctions de sélection des ressources, proches du concept de modélisation de l'habitat (Manly *et al.*, 1992; Boyce *et al.*, 2002), sont des fonctions qui sont proportionnelles à l'utilisation d'un habitat par un organisme ou une espèce. Pour construire une telle fonction, la zone d'étude est généralement divisée en sous-unités de surface. Ces unités sont conçues comme des ressources utilisées ou non par l'espèce associées à des variables prédictives (p. ex. altitude, couverture végétale). Il s'agit ensuite de proposer une fonction (p. ex. un modèle statistique) permettant de décrire l'utilisation

des ressources disponibles (sous-unités géographique) par l'espèce étudiée. Si la fonction proposée est pertinente, il est alors possible de prédire l'utilisation par l'espèce des ressources disponibles d'une autre zone d'étude ou d'explorer les changements d'utilisation face à des modifications des ressources sur la zone d'étude considérée (p. ex. changements climatiques) (Boyce *et al.*, 2002; Franklin, 2009).

1.1 LA MODÉLISATION STATISTIQUE : UNE SIMPLIFICATION UTILE

Les modèles et plus particulièrement les modèles statistiques sont utilisés pour extraire du savoir ou approfondir la connaissance du phénomène étudié, mais sont également utilisés en écologie quantitative comme outils de prédiction et/ou de décision (Clark, 2007; Parent et Rivot, 2012). En effet, la modélisation statistique de processus écologiques est généralement guidée par les objectifs de l'étude et les hypothèses que l'on souhaite tester. Pour répondre à ces questions, il est nécessaire de construire des modèles, structures mathématiques simplifiées, qui approchent la réalité. Les modèles statistiques simples, souvent éloignés des réalités écologiques, expliquent une faible partie de la variance observée rendant l'interprétation des résultats difficile (Clark, 2007). Il en résulte une volonté de construire des modèles écologiques de plus en plus réalistes et donc une complexification des modèles statistiques proposés. Cette propension à proposer des modèles complexes est directement due à la nature des données écologiques. En effet, les données écologiques sont par nature complexes, car non indépendantes en temps et espace (Scheffer et Carpenter, 2003), mais également influencées par des processus à petites et grandes échelles. Ces deux aspects des données écologiques reflètent les difficultés rencontrées lors de la construction de modèles statistiques adaptés à ces données.

Les modèles complexes tiennent compte de nombreux phénomènes comme la structure ou la dynamique spatiale, des interactions multiples entre individus, populations ou espèces, mais également des phénomènes et dynamiques temporelles. Tous ces éléments sont généralement associés à des variables environnementales caractérisant l'environnement abiotique de l'étude. Cependant, les modèles trop complexes sont parfois difficiles à interpréter et l'inférence de leurs paramètres peut

poser problème. De plus, la complexité et la spécificité de certains modèles ne permettent pas de généraliser à d'autres populations, espèces ou écosystèmes l'approche statistique développée. De ce fait, les conclusions de la modélisation statistique sont restreintes au phénomène étudié. Lors de la construction d'un modèle, il est important d'équilibrer entre réalisme écologique de la modélisation et parcimonie afin de produire des outils statistiques pertinents et généralisables à d'autres études.

Parfois, certaines hypothèses des modèles utilisés ne sont plus respectées (p. ex. indépendance des observations en temps et espace) à la suite d'une simplification excessive du processus étudié. De même, certaines sources d'incertitudes sont volontairement ignorées (p. ex. biais d'échantillonnage). D'un côté, les propriétés mathématiques du modèle sont ignorées pour étudier les données naturelles, de l'autre, certains aspects des données sont ignorés pour satisfaire les hypothèses nécessaires à l'application de la méthode statistique considérée (Clark, 2007). Il est donc primordial de construire des modèles parcimonieux prenant en compte de la meilleure manière possible les différentes sources d'incertitudes des données et garantissant l'identifiabilité des paramètres du modèle. En effet, il faut garder à l'esprit que le modèle doit être adapté aux données et non le contraire.

Pendant ces travaux de thèse, j'ai gardé à l'esprit cet équilibre difficile entre données écologiques à grandes dimensions, complexité du modèle, identifiabilité des paramètres, interprétations et applications écologiques. Je me suis donc attaché à proposer des modèles parcimonieux permettant de répondre aux objectifs et questionnements rencontrés pendant ces travaux de thèse.

1.2 Les données écologiques

Pour étudier les processus écologiques responsables de la distribution des espèces, les écologues ont recours à des échantillonnages d'espèces cibles sur la zone d'étude. Les données ainsi récoltées peuvent être de plusieurs natures :

1. l'espèce est notée comme présente ou absente du site échantillonné, les observations sont donc binaires (0 si l'espèce est absente, 1 si elle est présente).

2. Les individus de l'espèce cible sont comptés en chaque site échantillonné, les observations sont alors discrètes, on parle alors d'abondance de l'espèce.

3. Enfin la biomasse de l'espèce est mesurée, les observations sont donc continues. Ce type de données est relativement rare en écologie terrestre, mais plus fréquent en écologie marine. En effet, la procédure usuelle d'échantillonnage lors d'étude halieutique est effectuée par des relevés au chalut de fond pour lesquels la biomasse de chaque espèce est mesurée. De plus, les données de captures des pêches commerciales sont fréquemment utilisées pour étudier l'évolution de la biomasse des espèces de poissons commercialisés (Maunder et Punt, 2004).

Des variables environnementales sont généralement associées à ces données d'observations. Elles sont mesurées à l'échelle de l'observation, mais peuvent aussi être extraites d'études précédentes. Il faut alors s'assurer de la conformité de ces variables avec les observations effectuées sur le terrain. En effet, plusieurs auteurs ont alerté sur le fait d'effectuer des études écologiques à l'échelle de l'individu (environ 1 m) et d'utiliser des variables environnementales issues d'études à grande échelle ($10 - 100$ km) créant ainsi un décalage et des difficultés d'ajustement entre ces deux échelles (Schneider, 1994; Bell *et al.*, 1997; Edgar et Barrett, 2002).

De plus, contrairement à des disciplines comme la médecine ou l'agronomie, où tous les facteurs de l'étude sont contrôlés et dont les variables potentiellement explicatives sont construites volontairement contrastées de manière à quantifier les effets d'une substance sur un groupe d'individus, les études écologiques réalisées sur le terrain ne permettent pas un tel contrôle. L'écologue échantillonnant une zone d'étude est dépendant de l'environnement, il est alors parfois difficile d'extraire les effets de tel ou tel facteurs sur la présence de l'espèce. Il est notamment compliqué de séparer les effets de variables physiques dans les environnements marins comme le type de sédiment, la profondeur, la salinité, la turbidité ou la température (Dethier et Schoch, 2005), on parle alors de confusion des d'effets. Par exemple, dans les milieux marins côtiers le type de sédiment est fortement lié à l'action des vagues de même que la salinité varie avec la température (Clarke et Green, 1988). Cet effet, appelé le paradoxe de Simpson, est illustré en section 2.2.5.1.

Pour éviter ce genre de problèmes, des stratégies d'échantillonnage ont été développées minimisant ainsi cette source de biais. L'échantillonnage aléatoire stratifié permet notamment de prendre en compte des sources de variabilité connue à priori (Ferrier *et al.*, 2002; Latimer *et al.*, 2006). De plus, il permet une optimisation de la précision tout en réduisant l'effort d'échantillonnage. Pour ce faire, il faut subdiviser la zone d'étude en strates homogènes, mutuellement exclusives et collectivement exhaustives. Chacune de ces strates est ensuite échantillonnée avec la même intensité. Cependant, cette méthode d'échantillonnage nécessite de connaitre préalablement la zone d'étude, ce qui n'est pas toujours le cas.

1.2.1 *Données de relevés au chalut de fond*

Le Ministère des Pêches et Océan canadien (MPO) est chargé de la conservation et de la gestion durable des ressources halieutiques canadiennes. Son rôle est d'établir des programmes et stratégies de gestion répondant aux intérêts économiques et environnementaux du Canada. Un enjeu majeur dans la gestion des ressources halieutiques est d'obtenir des estimations fiables de l'abondance ou de la biomasse des espèces étudiées. L'abondance réelle est difficilement estimée du fait notamment de l'étendue des zones d'études et des mouvements des individus. Pour étudier l'évolution de l'abondance ou de la biomasse, des indices relatifs sont communément utilisés (Walsh, 1997; Candy, 2004; Maunder et Punt, 2004). Ces indices sont construits à partir des captures de pêches commerciales ou récréatives (en kg) et sont communément appelés Captures Par Unité d'Effort (CPUE). Les CPUE représentent généralement les captures par longueur de trait standard ou par unité de surface. L'utilisation de tels indices suppose que les captures sont proportionnelles au produit de la densité et de l'effort de pêche pour de petites échelles :

$$C = qEN \tag{1.1}$$

où C sont les captures, E est l'effort de pêche, N la densité et q le taux de capture pour une unité d'effort standard. On peut alors réécrire l'équation 1.1 :

$$\frac{C}{E} = qN \tag{1.2}$$

L'équation 1.2 est utilisée pour créer des indices relatifs d'abondance ou de biomasse (Walsh, 1997; Maunder et Punt, 2004). En effet, les données de captures associées à leur effort de pêche sont modélisées, généralement avec des modèles linéaires généralisés, afin de produire des estimations de séries de CPUE standardisées avec leurs intervalles de confiance. Ces CPUE standardisées sont ensuite utilisées afin d'émettre des avis sur la gestion des stocks considérés (Candy, 2004; Maunder et Punt, 2004).
Avant les années 1970, ces indices étaient estimés avec des données fournies par les pêches commerciales exploitant les ressources halieutiques. Cependant, de nombreux effets, comme le type de bateaux ou d'engin de pêche, mais également l'expérience de l'équipage étaient ignorés produisant des estimations d'indices peu fiables. De plus, l'échantillonnage était souvent considéré comme aléatoire alors que les bateaux de pêche ont une tendance à se concentrer sur des zones de forte densité de poissons surestimant alors les indices d'abondance. Ainsi, des parties entières de la zone d'étude étaient dépourvues de données rendant une image incomplète de l'abondance. Pour résoudre ces différents problèmes, des pêches scientifiques au chalut de fond utilisant un protocole d'échantillonnage standard et constant au court du temps ont été adopté par un grand nombre d'organisations (Doubleday et Rivard, 1981), malgré un cout élevé et la difficulté de récolter les données (Maunder et Punt, 2004). Ces relevés, généralement effectués une fois par an, permettent de produire des indices et des estimations de l'abondance plus fiables, mais très variables du fait du faible nombre de données récoltées par rapport aux captures commerciales. Néanmoins, de nouvelles méthodes sont développées afin de tenir compte de ce biais d'échantillonnage dans les données de pêches commerciales (Diggle *et al.*, 2010) , mais également pour associer les deux types de données (cf. chapitre 6).
Lors de mes travaux de thèse, je me suis intéressé à deux relevés scientifiques particuliers, exécutés par le MPO dont les caractéristiques sont présentées ci-après (sections 1.2.2 et 1.2.3).

1.2.2 *Invertébrés du golfe du Saint-Laurent*

La région du sud du Golfe du Saint-Laurent (sGSL), une des six régions administrées par le MPO, est gérée par le Centre des Pêches du Golfe (Figure 1.1). Chaque mois de septembre

depuis 1971, le MPO organise une campagne de pêche scientifique dans cette région (Chadwick *et al.*, 2007; Benoît *et al.*, 2009). Depuis sa création, l'objectif principal de ce relevé annuel est de quantifier l'abondance et la distribution d'espèces de poissons marines ainsi que d'espèces d'invertébrés à forte valeur économique. Cependant, depuis 1988, des données de biomasses d'invertébrés épibenthiques (p. ex. oursin, étoile de mer, anémone de mer) sont également collectées pour chaque site échantillonné. Ces espèces n'ont pas de valeur économique directe et ne sont pas exploitées par les pêcheries canadiennes. Elles jouent cependant plusieurs rôles nécessaires au bon fonctionnement des écosystèmes marins côtiers, comme le recyclage des nutriments, la dispersion et la fixation des sédiments ou encore participent à la production secondaire (Snelgrove, 1999). De plus, ces espèces de macro-invertébrés marins sont considérées comme de bons indicateurs écologiques du fait de leur faible capacité de migration et donc de leur difficulté de changer d'habitat, mais également des différences entre ces espèces de leur tolérance au stress (Dauer, 1993).

Le sGSL a été divisé en 27 zones qui diffèrent notamment par leur profondeur et leur substrat sédimentaire. Le plan d'échantillonnage pour cette étude est donc un plan aléatoire stratifié selon ces 27 secteurs restés identiques depuis 1971. Chaque année, entre 140 et 200 sites sont échantillonnés, en nombre globalement proportionnel à la surface de chaque zone. Un site échantillonné correspond à un trait de chalut d'une durée ciblée à 30 minutes à 3,5 nœuds. Tous les invertébrés capturés sont identifiés, au niveau de l'espèce lorsque c'est possible, puis pesés. La température de fond et la profondeur sont également mesurées à chaque site. Le type de sédiment est interpolé à chaque site à partir d'une carte géologique du GSL établie en 1973 (Loring et Nota, 1973). Cette interpolation est possible du fait que le type de sédiment change très peu au cours du temps à l'échelle du golfe.

1.2.3 *Poissons de fond en Colombie-Britannique*

Côté pacifique, le MPO a également mis en œuvre des relevés de pêches par chalutage pour étudier les populations de poissons de fond vivant dans le bassin de la Reine-Charlotte, une étendue maritime le long du littoral de la Colombie-Britannique

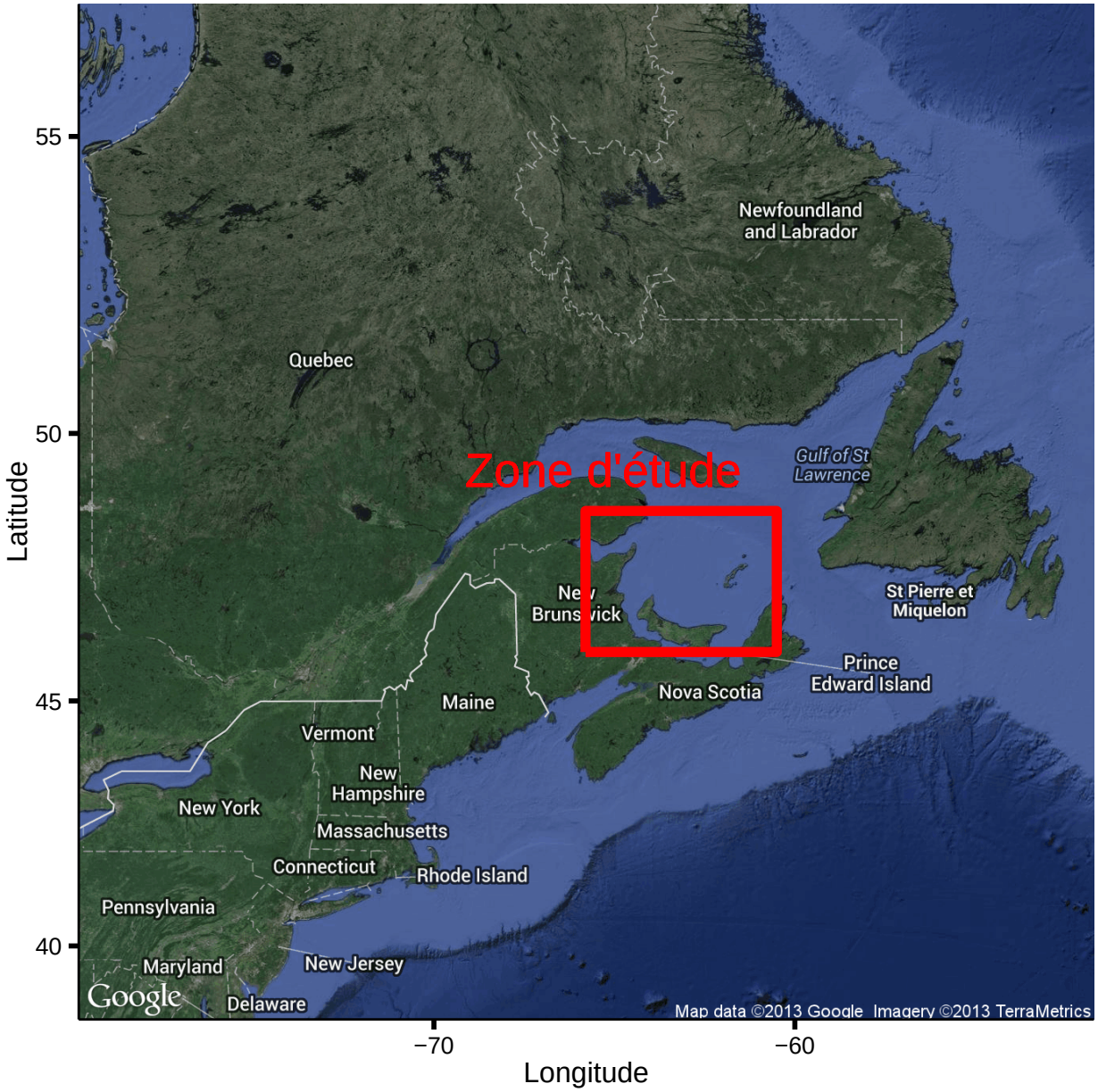

FIGURE 1.1 : Zone d'étude dans le Golfe du Saint-Laurent, Nouveau-Brunswick, Canada.

au Canada (Figure 1.2). Contrairement aux invertébrés épibenthiques du sGSL, ces espèces de poissons sont exploitées par l'Homme et présentent donc un intérêt commercial et un impact économique dans la région. Les quatre espèces de poissons considérées sont deux espèces de soles (*Microstomus pacificus* et *Errex zachirus*), une espèce de flet (*Atheresthes stomias*) et une espèce de sébaste (*Sebastes alutus*).

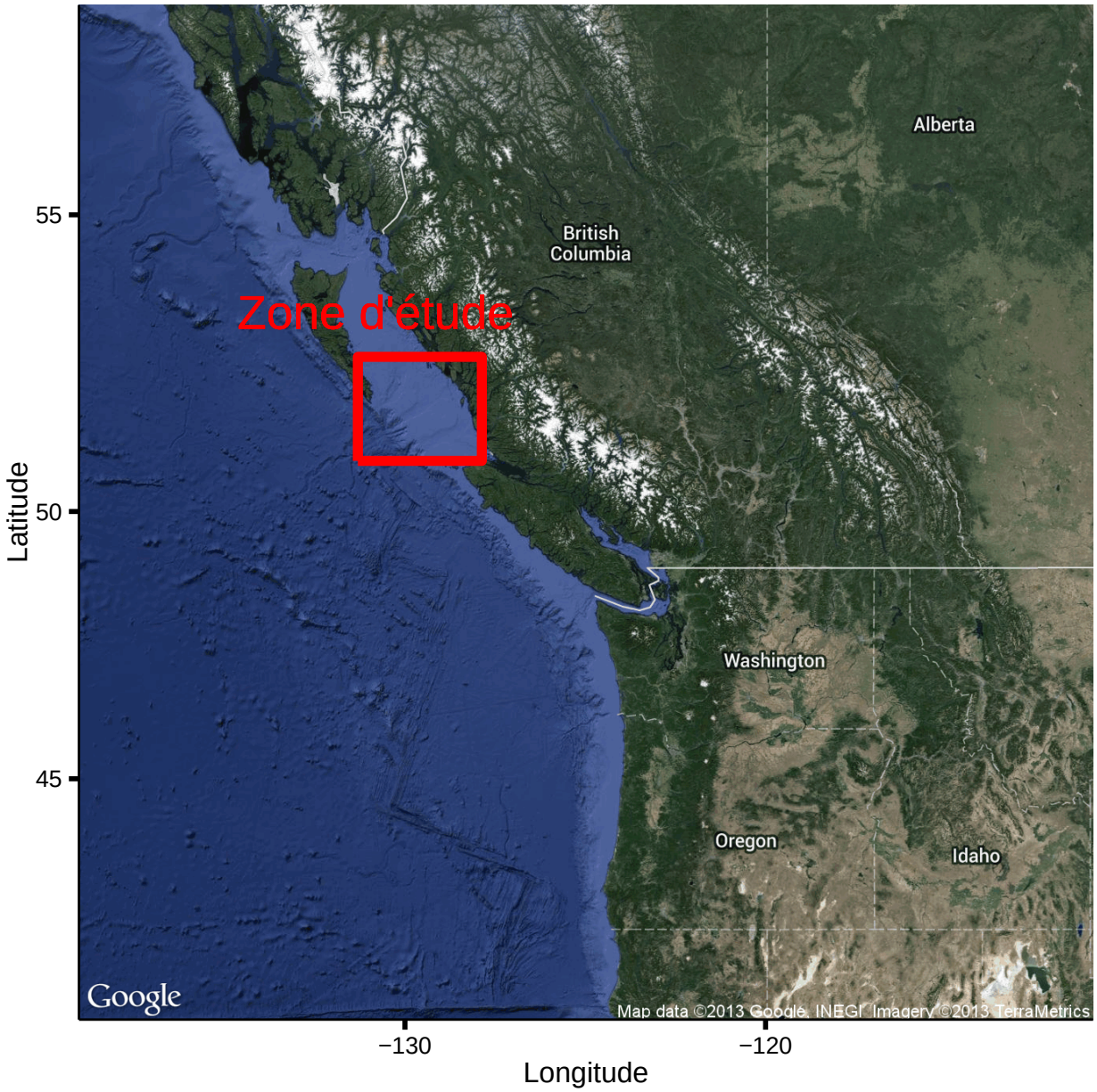

FIGURE 1.2 : Zone d'étude dans le bassin de la Reine-Charlotte, Colombie-Britannique, Canada.

Les relevés scientifiques au chalut débutent en 2003 suite à la recommandation du comité «Pacific Scientific Advice Review Committee» pour le développement d'indices d'abondance issus de pêches scientifiques et non de données de pêcheries commerciales pour les espèces de poissons de fonds (Sinclair *et al.*, 2003). Les indices d'abondance ont pour but de suivre l'évolution des stocks de ces différentes espèces de poissons de fond et pourront servir d'aide à la décision dans la gestion de ces populations exploitées.

Quatre relevés scientifiques au chalut ont été effectués depuis 2003 et couvrent collectivement l'ensemble du bassin de la Reine-Charlotte (Olsen *et al.*, 2007). De 2003 à 2005, un relevé a été effectué chaque année, mais depuis 2005 un relevé est effectué tous les deux ans. Le jeu de données disponible regroupe donc les années 2003, 2004, 2005, 2007 et 2009. Un plan aléatoire stratifié est également utilisé pour ce relevé. La zone d'étude est partagée en deux principales zones (nord et sud) puis divisée en quatre blocs de profondeurs différentes, délimitant ainsi des strates homogènes, pour une surface totale de 6 920 km^2. Afin de réduire l'erreur d'observation, 240 traits de chalut sont effectués à chaque relevé (Stanley *et al.*, 2005). Un site échantillonné correspond à une durée de chalutage de minimum 15 minutes et d'une durée de 20 minutes maximum. Le nombre de sites échantillonnés est proportionnel à la taille de chaque strate. De plus, les relevés de pêches commerciales sont à notre disposition permettant ainsi de confronter les deux types de données (scientifiques et commerciales), mais surtout de les associer afin d'améliorer les estimations des indices d'abondance.

1.3 OBJECTIFS, ENJEUX ET DIFFICULTÉS

Lors de ces travaux de thèse, je me suis intéressé à améliorer l'estimation des indices permettant de fournir des outils d'aide à la décision pour les décisionnaires et gestionnaires. En effet, les estimations d'indices comme les estimations de CPUE standardisées peuvent être utilisées directement comme conseil de gestion, mais sont fréquemment utilisées pour calibrer un modèle d'évaluation des stocks (Maunder et Punt, 2004). Il est donc primordial de proposer en amont des estimations de CPUE les plus fiables et les plus précises possible.

Dans un premier temps, j'ai souhaité comparer deux approches statistiques permettant d'estimer des indices de CPUE. Ces deux

approches sont adaptées aux données présentant une forte proportion de valeurs nulles : le modèle delta gamma (Stefansson, 1996) et le modèle poisson composé gamma (Jorgensen, 1987; Ancelet *et al.*, 2009; Foster et Bravington, 2012; Lecomte *et al.*, 2013b). Ces deux modèles permettent de prendre en compte des données à forte proportion de zéros, ce qui est fréquemment le cas pour les données de pêches par chalutage et plus particulièrement pour les deux jeux de données présentés précédemment (Figure 1.3).

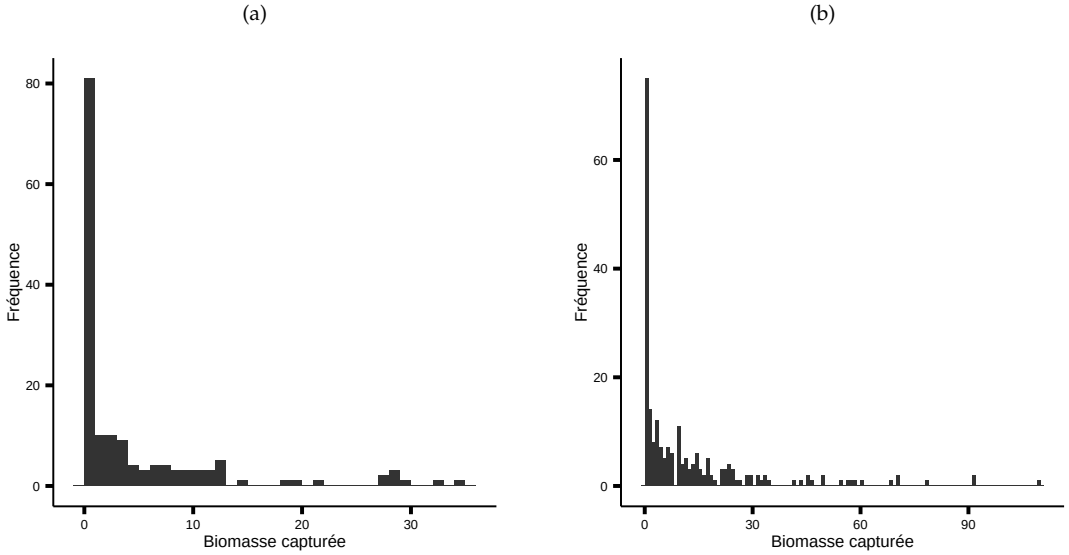

FIGURE 1.3 : Histogrammes des captures (en *kg*) effectuées par le MPO en 2005. (a) Captures d'oursins effectuées dans le sGSL. (b) Captures de sole (*Microstomus pacificus*) effectuées dans le bassin de la Reine-Charlotte.

Dans un second temps, j'ai développé des modèles capables de décrire la distribution spatiale d'une espèce, et ce en gardant comme objectif l'amélioration d'indices ou d'outils d'aide à la décision. En effet, il est nécessaire de comprendre et donc d'étudier la répartition spatiale d'une population afin d'en tirer les informations utiles pour leur gestion (Engler *et al.*, 2004). De ce fait de nombreuses approches statistiques ont été développées pour étudier la distribution des espèces (Latimer *et al.*, 2006). Les données utilisées sont souvent associées à des variables environnementales afin de prendre en compte les connaissances écologiques associées aux espèces étudiées (Austin, 2002). Cependant, modéliser la distribution d'une population pose deux

difficultés majeures : (i) les statisticiens et écologues ont souvent à leur disposition des jeux de données avec une forte quantité de zéros, comme présentés précédemment. Cet excès de zéros implique l'utilisation d'outils statistiques spécifiques, (ii) les valeurs de variables prélevées en des sites voisins sont spatialement corrélées. Ignorer cette autocorrélation amène à de fausses conclusions (Legendre et Legendre, 1998). Pour répondre à ces deux difficultés, le modèle Poisson composé gamma est utilisé au sein d'un modèle hiérarchique bayésien, qui prend en compte plusieurs variables environnementales ainsi que la corrélation spatiale.

Enfin, une approche de modélisation spatio-temporelle est proposée. Cette approche permet d'étudier la distribution spatiale de l'espèce cible comme évoquée précédemment, mais a pour avantage d'étudier la dynamique temporelle de l'espèce étudiée sur la zone d'étude. Cette méthode appliquée aux invertébrés épibenthiques du sGSL est généralisable à d'autres espèces et constitue donc un outil statistique intéressant pour la production d'indices de CPUE standardisées, qui doivent être fournis aux modèles d'évaluation des stocks.

2

DES OUTILS STATISTIQUES POUR REPRÉSENTER DES DONNÉES ZÉRO-INFLATÉES ET DES STRUCTURES DE DÉPENDANCE COMPLEXES POUR L'ÉCOLOGIE

Dans ce chapitre, je présente les outils statistiques mis en oeuvre lors de mes travaux de thèse.

2.1 LES MODÈLES HIÉRARCHIQUES BAYÉSIENS

2.1.1 *Origine et principe de base*

L'intérêt pour les modèles hiérarchiques bayésiens en écologie est grandissant, et ce depuis les années 90 (McCarthy, 2007). En effet, l'essor de ces modèles est aujourd'hui possible par deux faits conjugués que sont le développement et la démocratisation de la puissance de calcul associé à plusieurs avancées théoriques notamment celle de Gelfand et Smith (1990) qui présente des méthodes d'échantillonnage de distributions de probabilités marginales. Ces deux avancées, technologique et théorique, ont permis le renouvèlement de la statistique bayésienne par l'intermédiaire des techniques d'échantillonnage Monte-Carlo par Chaines de Markov (MCMC). Ainsi l'approche bayésienne est basée sur l'idée que l'expérimentateur possède une idée *a priori* sur le système qu'il étudie et que ses connaissances sont mises à jour par les données qu'il observe sur ce même système. La statistique bayésienne est fondée sur travaux du Révérend Thomas Bayes, qui a vécu au $18^{\text{ème}}$ siècle en Angleterre, et dont le théorème éponyme établit les bases de la statistique bayésienne :

$$[\theta|y] = \frac{[\theta][y|\theta]}{[y]} \tag{2.1}$$

où y représente un jeu de données et θ le vecteur de paramètres du modèle. $[y]$ est la probabilité marginale et s'écrit $[y] = \int [y|\theta][\theta]d\theta$. Ensuite, $[\theta]$ est le *prior* du vecteur de paramètres θ, qui est également appelé distribution *a priori*. Enfin, $[\theta|y]$ la distribution *a posteriori* du vecteur de paramètres θ également appelé *posterior*. Il est important de noter que contrairement à l'approche statistique classique (ou fréquentiste), l'estimation des paramètres du modèle donne lieu à une distribution *a posteriori* et non une estimation ponctuelle.

Une analyse bayésienne se décompose en trois grandes étapes. La première consiste à choisir un modèle approprié qui décrit la relation entre les données y et les paramètres du modèle θ. Si l'on considère l'équation 2.6, cette étape revient à spécifier le terme $[y|\theta]$. Ensuite, il faut définir les distributions *a priori* des paramètres du modèle $[\theta]$. Enfin, à l'aide du théorème de Bayes la distribution *a posteriori* $[\theta|y]$ est obtenue. Cette démarche peut être résumée dans le digramme suivant :

$$prior + \text{data} \overset{\text{model}}{\longrightarrow} posterior \tag{2.2}$$

Les modèles hiérarchiques bayésiens s'inscrivent donc dans ce cadre général qu'est la statistique bayésienne. Dans les sections suivantes, je vais présenter les intérêts majeurs de cette modélisation hiérarchique pour les problèmes rencontrés en écologie.

2.1.2 *Décomposer la complexité*

Les modèles hiérarchiques bayésiens permettent d'appréhender la complexité des processus écologiques en les décomposant en sous-problèmes, également appelés niveaux ou couches. Cette formulation a été proposée par Berliner (1996) et a rapidement été adoptée en écologie (Wikle *et al.*, 1998; Clark, 2004). Les modèles hiérarchiques sont donc tous composés de trois couches ou niveaux :

1. Le niveau des observations définit la distribution de probabilité associée aux données sachant les paramètres du modèle et le processus écologique latent.

2. Le niveau du processus latent décrit le processus écologique étudié (p. ex. croissance d'une population, recrutement, *etc*.).

3. Le niveau des paramètres contrôle le niveau du processus latent et le niveau des observations par l'intermédiaire de paramètres fixés permettant de reproduire les données observées.

Les modèles hiérarchiques bayésiens utilisent donc une décomposition conditionnelle de problèmes complexes pour traiter une série de problèmes simples reliés entre eux de façon probabiliste. Cette décomposition est illustrée par l'équation 2.3.

$$\begin{aligned} p(\text{paramètres}|\text{modèle}, \text{données}) \propto\ & p(\text{données}|\text{processus}, \text{données}, \text{paramètres}) \\ & \times p(\text{processus}|\text{processus}, \text{paramètres}) \\ & \times p(\text{paramètres}) \end{aligned} \tag{2.3}$$

En d'autres termes, la modélisation hiérarchique bayésienne consiste à définir successivement la loi *a priori* $[\theta]$, le modèle du processus latent $[Z|\theta]$ puis le modèle des observations $[Y|Z, \theta]$. Ceci permet de définir la distribution jointe des observations, processus latent et des paramètres :

$$[Y, Z, \theta] = [\theta][Z|\theta][Y|Z, \theta] \tag{2.4}$$

En changeant de conditionnement, on obtient :

$$[Y, Z, \theta] = [\theta, Z|Y][Y] \tag{2.5}$$

Il est possible d'introduire dans la formule de Bayes (équation 2.6) les variables latentes Z :

$$[\theta, z|y] = \frac{[\theta][z|\theta][y|z, \theta]}{[y]} \tag{2.6}$$

La couche latente ou processus latent est une idée centrale des modèles hiérarchiques bayésiens. Il existe deux types de processus latents et donc deux types de modèles hiérarchiques (Royle et Dorazio, 2008). Le premier spécifie de manière explicite le processus latent qui est alors construit sur la base de connaissances biologiques ou écologiques (p. ex. abondance, présence). Le deuxième type de modèles hiérarchiques utilise un processus latent implicite et est alors basé sur un ensemble d'effets aléatoires généralement spatialisés ou indexés sur le temps. Un processus latent implicite est généralement et de-

vrait être uniquement utilisé lorsque l'information sur le processus écologique étudié est inconnue ou difficile à appréhender.

2.1.3 *Intégration de l'incertitude et des différentes sources de variabilité*

La complexité des processus écologiques étudiés et des modèles étudiant ces processus nécessitent de considérer les différentes sources de variabilité et d'incertitudes qui leur sont associées. Ces deux aspects (variabilité et incertitude) doivent être pris en compte lors de la construction du modèle afin que ce modèle approche le plus possible les connaissances du phénomène étudié (variabilité), mais également pour quantifier de la meilleure manière possible la qualité des prédictions faites par le modèle étudiant ce phénomène (incertitude).

En écologie, trois sources majeures sont responsables de l'incertitude :

1. **Les erreurs de modèle**, également appelées erreurs de structures, résultent de l'imperfection du modèle et des simplifications nécessaires effectuées pour étudier le processus écologique choisi. Ces erreurs peuvent poser problème lorsque le modèle est utilisé à des fins prédictives, et plus particulièrement si le modèle est utilisé en dehors de la zone d'étude (Parent et Rivot, 2012).

2. **Les erreurs de processus** permettent au modèle de prendre en compte le caractère stochastique du processus écologique étudié. En effet, le processus latent d'un modèle hiérarchique bayésien doit prendre en compte cette variabilité imprévisible et ceci est d'autant plus important si le processus latent considéré est issu d'un modèle déterministe (Parent et Rivot, 2012).

3. **Les erreurs d'observations ou de mesures** apparaissent du fait de la difficulté d'observer la nature dans toute sa complexité. Ces erreurs peuvent avoir comme origine une procédure d'échantillonnage inadaptée aux questions posées par l'étude, mais également des erreurs de mesures ou de détection (p. ex. espèce présente en un site, mais non détectée). La présence de données manquantes

est également fréquente dans les jeux de données écologiques. Il est alors nécessaire de prendre en compte dans l'approche de modélisation les difficultés intrinsèques au jeu de données utilisé.

2.1.4 *Des modèles très utiles en prédictions*

L'un des intérêts de la modélisation bayésienne est la facilité d'inférence prédictive lorsque la distribution *a posteriori* a été obtenue. Supposons que nous disposions de $Y = (y_1, ..., y_N)$ les données observées, nous souhaitons prédire une nouvelle observation y_{new}. À l'aide des données, nous avons obtenu la distribution *a posteriori* $[\theta|Y]$. Nous souhaitons désormais prédire une nouvelle observation y_{new}, et donc calculer sa distribution prédictive *a posteriori* notée $[y_{new}|Y]$. Cette distribution prédictive *a posteriori* est obtenue avec :

$$[y_{new}|Y] = \int_{\theta} [y_{new}|\theta] \times [\theta|Y] d\theta \tag{2.7}$$

Lorsque l'on considère un modèle hiérarchique bayésien, il est préférable d'introduire les variables latentes Z pour obtenir la distribution prédictive *a posteriori* :

$$[y_{new}|Y] = \int_{\theta} \int_{z} [y_{new}|\theta, z] \times [\theta, z|Y] d\theta dz \tag{2.8}$$

Avec l'équation 2.8, il est relativement facile d'obtenir une prédiction y_{new}. Pour ce faire, la distribution *a posteriori* $[\theta, z|Y]$ doit avoir été préalablement obtenue. La prédiction d'une nouvelle observation y_{new} se réalise en deux étapes :

1. Génération aléatoire de $\theta_{new}, Z_{new} \sim [\theta, z|Y]$,
2. avec θ_{new}, Z_{new} tiré aléatoirement $y_{new} \sim [y|\theta_{new}, z_{new}]$

2.2 Les données de biomasse à forte proportion de zéros

Les jeux de données mesurant la présence-absence, le nombre d'individus ou la biomasse d'une espèce présentent souvent une forte quantité de zéros (Martin *et al.*, 2005). Ces données sont appelées zéro-inflatées lorsque les distributions statistiques standards (p. ex. distribution normale, Poisson, binomiale) ne permettent pas de les modéliser correctement (Heilbron, 1994;

Tu, 2002; Martin *et al.*, 2005). Dans ces travaux de thèse, les données étudiées sont continues et présentent une masse en zéro, ainsi des modèles spécifiques ont alors été mis en oeuvre pour analyser ces données.

2.2.1 *Origine et difficultés de traitement statistique*

Il est possible de distinguer deux types d'absences. La première est due à la nature même du processus écologique étudié et est appelée *vraies absences* ou *vrais zéros* et résulte généralement d'une faible fréquence de présence de l'espèce étudiée. En effet, des phénomènes de compétition, de processus démographiques ou d'effet de l'habitat peuvent induire ces *vraies absences* (Martin *et al.*, 2005). Ces zéros témoignent directement du processus écologique à l'étude, mais un zéro peut également être un résultat aléatoire (stochasticité) du fait que l'espèce ne soit pas à l'équilibre avec son environnement ou d'effet démographique (extinction locale). Le second type d'absences est appelé *fausses absences* ou *faux zéros* et peut également être séparé en deux catégories. L'espèce est présente, mais elle n'est pas observée. Cette situation peut se produire lorsque le temps d'échantillonnage est faible ou lorsque l'échantillonnage n'est pas adapté à l'écologie de l'espèce étudiée (Tyre *et al.*, 2003). Le second type de *fausses absences* se produit lorsque l'espèce est présente, mais n'est pas détectée par l'observateur.

Il est important de noter que les objectifs de l'étude définissent si un zéro est *vrai* ou *faux*. Si l'objectif de l'étude est de cartographier la présence d'une espèce à la période de l'échantillonnage, une absence temporaire de l'espèce à un site donné sera un *vrai zéro*. Inversement pour une étude à long terme pour laquelle l'objectif est de cartographier l'utilisation de la zone d'étude par l'espèce, ce même zéro sera considéré comme une *fausse absence* (Martin *et al.*, 2005).

Les deux types d'absences posent des problèmes statistiques, car les hypothèses de base (p. ex. normalité des données) ne sont pas respectées. Dans ces travaux de thèse, seules les *vraies absences* ont été étudiées (voir la revue de Martin *et al.* (2005) pour plus de détail sur les méthodes adaptées aux *fausses absences*).

2.2.2 *Le modèle Log-Normal*

Une idée simple et naturelle pour traiter les données à forte proportion de zéros est de supprimer ces zéros et de s'intéresser uniquement aux valeurs strictement positives. Cette procédure a pour conséquence de biaiser positivement les résultats de l'étude. De plus, si l'étude est pratiquée sur plusieurs années le biais peut varier entre ces années (Maunder et Punt, 2004). Une seconde approche facile à mettre en oeuvre est d'ajouter une constante δ à toutes les données Y. Une fois cette opération effectuée, il est possible d'utiliser une distribution log-normale pour représenter ces données transformées :

$$log(Y + \delta) \sim Normale(\mu, \sigma^2) \tag{2.9}$$

Cette méthode n'est cependant pas la meilleure pour modéliser ce type de données. En effet, le choix de la constante est arbitraire et les résultats peuvent varier en fonction de celle-ci (Berry, 1987; Porch et Scott, 1994; Stefansson, 1996; Maunder et Punt, 2004).

2.2.3 *Les approches Delta*

Les modèles Delta, également appelés modèles en deux parties (*two parts model*, ou encore *hurdle mode*l), permettent de modéliser des données continues à forte proportion de zéros sans transformer les données. Comme l'indique leur nom, ces modèles sont constitués de deux parties ou deux sous modèles. Le premier décrit la présence-absence de l'espèce : X est une variable binaire égale à 1 si l'espèce est présente et 0 si elle est absente :

$$X \sim Bernoulli(\pi) \tag{2.10}$$

où π est la probabilité de présence de l'espèce. La seconde partie du modèle delta est conditionnée à la précédente :

$$\begin{aligned} Y|X = 1 &\sim Gamma(\alpha, \beta) \\ Y|X = 0 &\sim \delta_0 \end{aligned} \tag{2.11}$$

(α, β) étant les paramètres de forme et d'intensité et δ_0 une distribution Dirac en zéro. Les données strictement positives

sont donc distribuées selon une loi gamma de paramètre α et β. L'espérance et la variance de ces quantités de biomasses non nulles sont respectivement obtenues par $\mathbb{E}(Y|X=1) = \frac{\alpha}{\beta}$ et $\mathbb{V}ar(Y|X=1) = \frac{\alpha}{\beta^2}$. Le modèle delta gamma est composé de trois paramètres $DG(\pi, \alpha, \beta)$, avec lesquels nous pouvons déduire l'espérance des quantités de biomasses $\mathbb{E}(Y) = \pi\frac{\alpha}{\beta}$, ainsi que leur variance $\mathbb{V}ar(Y) = \pi\frac{\alpha}{\beta^2}(1+\alpha(1-\pi))$.

La distribution gamma peut être remplacée par une autre distribution strictement positive comme une distribution log-normale construisant ainsi un modèle delta log-normal (DLogN) :

$$\begin{aligned} Log(Y|X) = 1 &\sim Normale(\mu, \sigma^2) \\ Y|X = 0 &\sim \delta_0 \end{aligned} \tag{2.12}$$

Dans ce cas, le nombre de paramètres est inchangé avec π la probabilité de présence,μ la moyenne de la loi normale et σ^2 sa variance. L'introduction de variables explicatives est possible par l'utilisation d'un modèle linéaire généralisé (GLM, (Zuur *et al.*, 2009)). Pour ce faire, les deux parties d'un modèle delta peuvent être modélisées séparément. Pour commencer, la probabilité de présence π est généralement modélisée par une régression logistique :

$$log(\frac{\pi}{1-\pi}) = V\tau \tag{2.13}$$

avec V le vecteur des variables explicatives et τ le vecteur de paramètres associé. Dans cet exemple, la fonction de lien choisie est une fonction *logit*, mais il est également possible d'utiliser une fonction *probit*. L'introduction de variables explicatives dans la partie du modèle delta s'intéressant aux données non nulles se fait généralement avec une fonction de lien logarithmique :

$$log(\mathbb{E}(Y|X=1)) = W\eta \tag{2.14}$$

avec W le vecteur des variables explicatives et η le vecteur de paramètres associé. Les variables explicatives utilisées dans chaque partie du modèle delta peuvent être identiques ($V \equiv W$)

ou volontairement différentes. L'inférence d'un modèle delta peut s'effectuer en deux temps. Dans un premier temps, les données doivent être converties en présence-absence (0 ou 1) afin d'estimer la probabilité de présence π. Dans un deuxième temps, les données strictement positives sont utilisées pour inférer les paramètres α et β. Cette séparation facilite l'inférence d'un modèle delta, mais pose le problème du traitement séparé des valeurs nulles et non nulles. En effet, ce modèle ne tient pas compte d'une possible cohérence entre les absences et les quantités non nulles de biomasses de l'espèce étudiée. Si l'on considère un gradient de biomasse composé de faibles densités de l'espèce étudiée, on s'attend alors à ce que la probabilité de présence de l'espèce soit faible. En d'autres termes, l'information disponible dans les données non nulles n'est pas prise en compte lors de l'inférence de la probabilité de présence.

Une seconde limite du modèle delta est qu'il n'est pas stable par addition (Stefansson, 1996). Cette propriété d'additivité est très utile en écologie. Il est par exemple désirable que deux échantillons d'une durée respective d'une heure suivent la même loi de probabilité qu'un échantillon d'une durée de deux heures. Cette propriété permet de traiter au sein d'un même modèle des échantillons qui n'ont pas été obtenus avec la même durée d'échantillonnage. Le modèle delta ne respectant pas cette propriété, il est nécessaire de proposer des méthodes permettant de prendre en compte ces différences d'échantillonnage. La méthode usuelle consiste à standardiser en amont les données afin d'obtenir le même effort d'échantillonnage pour toutes les observations. Une seconde méthode propose d'inclure l'effort d'échantillonnage comme variable explicative dans la partie strictement positive du modèle (Stefansson, 1996). L'inconvénient de ces deux approches est que l'effort d'échantillonnage des valeurs nulles n'est pas pris en compte. Dans le premier cas, les valeurs nulles sont standardisées comme les valeurs non nulles, mais le résultat de la standardisation d'une valeur nulle est toujours une valeur nulle. Dans le deuxième cas, l'effort d'échantillonnage est ajouté comme variable explicative aux seules valeurs non nulles. Or, il est naturel de penser que lorsque la durée d'échantillonnage augmente la probabilité d'obtenir une valeur non nulle augmente elle aussi. C'est pourquoi je propose d'inclure l'effort d'échantillonnage comme variable explicative dans les deux parties du modèle delta (cf chapitre 3). De cette manière, les différentes durées d'échantillonnage

sont prises en compte de la meilleure façon possible, alors que la structure même du modèle ne le permet pas simplement.

2.2.4 *Les modèles Tweedie*

Le troisième type de modèle utilisé pour traiter les données à forte proposition de zéro est regroupé dans une famille appelée Tweedie (Jorgensen, 1997). Cette classe de modèle a pour avantage de considérer au sein d'un même modèle les données strictement positives et les données nulles. Cette distribution Tweedie est très utilisée en science halieutique (Candy, 2004; Shono, 2008; Tascheri *et al.*, 2010; Ancelet *et al.*, 2009; Peel *et al.*, 2012; Foster et Bravington, 2012; Lecomte *et al.*, 2013a). Le modèle Tweedie le plus couramment utilisé est le modèle linéaire généralisé-Tweedie (Smyth, 1996), mais plusieurs alternatives ont été proposées (Ancelet *et al.*, 2009; Foster et Bravington, 2012; Lecomte *et al.*, 2013a).

2.2.4.1 *Le modèle linéaire généralisé-Tweedie*

Le modèle linéaire généralisé Tweedie (TGLM) proposé par (Smyth, 1996) est un modèle linéaire généralisé définit par une relation entre moyenne et variance telle que :

$$\begin{aligned} \mathbb{E}(y) &= \mu \\ \mathbb{V}ar(y) &= \phi\mu^p \end{aligned} \tag{2.15}$$

ϕ est un paramètre de dispersion et p un paramètre de puissance. La distribution Tweedie inclut plusieurs distributions qui sont produites avec différentes valeurs du paramètre de puissance (Tableau 5.6).

En écologie ou halieutique, le paramètre de puissance est généralement compris entre 1 et 2 (Kendal, 2004). Lorsque $1 < p < 2$ une variable aléatoire suivant une distribution Tweedie peut être décomposée en une somme de variables aléatoires distribuées selon un Poisson composé :

$$Y = \sum_{i=1}^{N} M_i \tag{2.16}$$

Tableau 2.1

Distributions	p
normale	$p = 0$
n'existe pas	$1 < p < 2$
Poisson	$p = 1$
Poisson-gamma	$1 < p < 2$
Gamma	$p = 2$
positive et stable	$p > 2$
extrême et stable	$p < 0$

avec $N \sim Poisson(\lambda)$ et $M_i \sim Gamma(a, b)$. Il est alors possible de paramétrer la distribution Tweedie en fonction de λ, a et b (Smyth, 1996) :

$$\begin{aligned} \lambda &= \frac{1}{\phi}\frac{\mu^{2-p}}{2-p} \\ a &= \frac{2-p}{p-1} \\ b &= \phi(p-1)\mu^{p-1} \end{aligned} \tag{2.17}$$

Ainsi avec cette paramétrisation, l'espérance de la biomasse observée est $\mathbb{E}(y) = \frac{\lambda a}{b}$ et sa variance est $\mathbb{V}ar(y) = \frac{a\lambda}{b}\frac{a+1}{b}$. La formulation proposée en équation 2.16 représente de manière intuitive l'échantillonnage d'organismes au chalut de fond : les organismes sont regroupés en patchs, chacun de ces patchs possède une biomasse propre. Lors d'un trait de chalut, un nombre de patchs est collecté et la biomasse échantillonnée lors de ce trait de chalut est la somme de la biomasse contenue dans chacun de ces patchs. Cette approche est également intéressante pour les écologues quantitatifs puisque la relation moyenne-variance respecte la loi empirique de Taylor (Taylor, 1961) où la variance est une fonction de l'espérance :

$$\mathbb{V}ar(y) = c\mathbb{E}(y)^d \tag{2.18}$$

où c et d sont des constantes.

Le modèle TGLM est un modèle linéaire généralisé définit par sa relation moyenne-variance, les procédures d'estimation standard des GLM peuvent donc être utilisées (Smyth, 1996).

L'introduction de variables explicatives (p. ex. type de substrat, profondeur) se fait également de façon naturelle :

$$g(\mathbb{E}(y)) = g(\mu) = X\tau \tag{2.19}$$

où X représente les variables explicatives et τ le vecteur de paramètres qui leur sont associés. Dans le contexte de l'écologie ou de l'halieutique, la fonction $g(.)$ est généralement une fonction logarithme, car elle permet de construire un modèle multiplicatif (Foster et Bravington, 2012).

2.2.4.2 *Le modèle Poisson composé gamma*

Le modèle Poisson composé gamma (CPG) est basé sur la formulation du modèle Tweedie présenté à l'équation 2.16. Il est donc composé de trois paramètres :

$$Y \sim CPG(\lambda, a, b) \tag{2.20}$$

Sans variables explicatives, le modèle CPG est équivalent à un modèle Tweedie. Les différences entre les deux modèles apparaissent lorsque des variables explicatives sont ajoutées aux modèles. En effet, contrairement au modèle TGLM qui modélise directement l'espérance des observations ($\mathbb{E}(Y)$), le modèle CPG permet de travailler avec l'espérance de la loi de Poisson ($\mathbb{E}(N)$) et l'espérance de la loi gamma ($\mathbb{E}(M_i)$. De cette façon, les variables explicatives affectant le nombre de patchs peuvent être différentes de celles affectant la masse de ses patchs :

$$\begin{aligned} log(\mathbb{E}(N)) &= X\tau \\ log(\mathbb{E}(M_i)) &= Z\beta \\ log(\mathbb{E}(Y)) &= log(\mathbb{E}(N)) + log(\mathbb{E}(M_i)) \end{aligned} \tag{2.21}$$

avec X et Z des vecteurs de variables explicatives et τ et β des vecteurs de paramètres associés. Si les mêmes variables explicatives sont utilisées pour expliquer le nombre de patchs et la biomasse de ses patchs alors le modèle CPG est équivalent au modèle TGLM :

$$
\begin{aligned}
log(\mathbb{E}(Y)) &= X\tau + Z\beta \text{ avec } X \equiv Z \\
log(\mathbb{E}(Y)) &= X\beta^* \text{ avec } \beta^* = \tau + \beta \\
\mathbb{V}ar(Y) &= \frac{a+1}{b}\mathbb{E}(M_i)\mathbb{E}(Y) = \phi\mu^p
\end{aligned} \tag{2.22}
$$

Le modèle CPG peut être pensé comme un modèle hiérarchique, le nombre de patchs et leur masse constituent la couche latente du modèle :

$$
\begin{aligned}
N_s &\sim \text{Poisson}(\lambda_s) \qquad \forall\, s \in \{1, ..., S\} \\
M_{s,i} &\sim \text{Gamma}(a, b)
\end{aligned} \tag{2.23}
$$

La somme des patchs d'organismes conduisant à la biomasse totale observée Y :

$$
Y = \begin{cases} \sum_{p=1}^{N} M_p & \text{if } N > 0 \\ \\ 0 & \text{if } N = 0 \end{cases} \tag{2.24}
$$

Contrairement au modèle delta, le modèle CPG est stable par addition (Jorgensen, 1997). Cette propriété permet donc de prendre en compte des efforts d'échantillonnage variables V de façon linéaire avec le paramètre d'intensité de la loi de Poisson :

$$
Y \sim CPG(\lambda V, a, b) \tag{2.25}
$$

Ainsi, il n'est pas nécessaire de standardiser les données ou d'inclure à l'aide d'un GLM l'effort d'échantillonnage comme variable explicative. On peut également noter que le modèle CPG permet de prendre en compte conjointement la probabilité de présence et les valeurs non nulles. Il est donc possible avec cette approche de modéliser un gradient décroissant de la biomasse dans la distribution de l'espèce ciblée en raison de la faible densité d'organismes ou d'une faible détectabilité.

Dans le chapitre 3, une comparaison entre le modèle DG et le modèle CPG est proposée lorsque l'effort d'échantillonnage est variable au sein d'un même jeu de données.

2.2.5 *Variables explicatives*

L'introduction de variables explicatives dans les modèles bayésiens hiérarchiques est très courante en écologie. Ces variables explicatives décrivent généralement l'habitat dans lequel l'espèce est échantillonnée (p. ex. substrat, température, pH, *etc.*), mais peuvent également mesurer des effets biotiques comme la présence de prédateurs ou de parasites. Il est cependant nécessaire de rester prudent quant à l'interprétation des effets environnementaux estimés. Comme évoqué dans le chapitre d'introduction, il est parfois difficile de séparer les effets de variables physiques dans les environnements marins (Dethier et Schoch, 2005). On parle alors de confusion des effets dont une illustration est proposée ci-dessous.

2.2.5.1 *Le paradoxe de Simpson*

Imaginons que nous notions Y la variable binaire de présence d'une espèce d'intérêt, soumise à un prédateur. On notera P l'évènement correspondant à une forte présence du prédateur et $\bar{P}$ l'évènement complémentaire correspondant à niveau de prédation plus faible. Imaginons, pour simplifier que l'espèce d'intérêt vive dans deux groupes de milieux possibles, l'habitat le plus favorable à son développement F et un milieu moins riche $\bar{F}$. Supposons enfin qu'un membre de l'espèce ait une probabilité de présence (évènement $Y = 1$) fonction du milieu dans lequel il vit (F ou $\bar{F}$) et de l'intensité de prédation (P ou $\bar{P}$) selon le tableau suivant de probabilités conditionnelles :

$$\begin{array}{ll} [Y = 1|\bar{P}, F] = 0.8 & [Y = 1|\bar{P}, \bar{F}] = 0.7 \\ [Y = 1|P, F] = 0.4 & [Y = 1|P, \bar{F}] = 0.3 \end{array}$$

Les valeurs relatives des probabilités de tableau sont très logiquement conformes à nos connaissances scientifiques : une forte intensité de prédation limite plus le développement de l'espèce d'intérêt qu'une faible intensité, quelle que soit la qualité du milieu ; dans chacun des deux groupes d'habitat, on a moins de chances de trouver l'espèce si de nombreux prédateurs sont présents.

Imaginons maintenant que les prédateurs ont eux aussi tendance à préférer le milieu le plus favorable de telle sorte que :

$$[P|F] = 0.75$$
$$[P|\bar{F}] = 0.35$$

Regardons maintenant la conséquence de ces choix sur la probabilité de présence de l'espèce dans le milieu le plus favorable :

$$\begin{aligned}[Y = 1|F] &= [Y = 1 \text{ et } \bar{P}|F] + [Y = 1 \text{ et } P|F] \\ &= [Y = 1|\bar{P}, F] \times [\bar{P}|F] + [Y = 1|P, F] \times [P|F] \\ &= 0.8 * 0.25 + 0.4 * 0.75 = 0.5\end{aligned}$$

Et par un raisonnement analogue pour le milieu le moins favorable :

$$\begin{aligned}[Y = 1|\bar{F}] &= [Y = 1|\bar{P}, \bar{F}] \times [\bar{P}|\bar{F}] + [Y = 1|P, \bar{F}] \times [P|\bar{F}] \\ &= 0.7 * 0.65 + 0.3 * 0.35 = 0.56\end{aligned}$$

L'espèce a donc une probabilité de présence plus forte dans le milieu le moins favorable. Cet exemple simpliste illustre le paradoxe de Simpson qui devrait tempérer l'engouement pour toute méthode automatique d'analyse statistique descriptive des grosses bases de données. Dans cet exemple, le paradoxe s'explique bien sûr par le fait que les prédateurs sont également présents dans le milieu favorable. En d'autres termes, il a confusion entre l'effet du milieu et celui du prédateur.

Deux remarques s'imposent. En premier lieu, on ne peut échapper à la cohérence du calcul probabiliste, les calculs précédents sont corrects. En second lieu, les deux probabilités marginalisées $[Y = 1|F]$ et $[Y = 1|\bar{F}]$ sont celles dont disposerait le scientifique s'il ne mesure pas la grandeur P. Le paradoxe disparait si le prédateur est présent avec la même probabilité 50% dans chacun des deux milieux : $[P|F] = 0.50 = [P|\bar{F}]$. Or cette hypothèse est fausse dans notre exemple, et il est illusoire lorsque l'on travaille en écologie avec des données observationnelles, sans réelles possibilités de contrôler l'expérience, de faire varier suffisamment voire d'obtenir la gamme complète des variables potentiellement explicatives.

Ce paradoxe est une source d'ennuis expérimentaux bien reconnue dans de nombreux domaines d'application de la statistique. En écologie, l'article (Clark *et al.*, 2011) détaille le pa-

radoxe de Simpson et en explique les conséquences avec des représentations et des schémas explicatifs convaincants. Les valeurs numériques de l'exemple utilisé ci-dessus pour illustrer notre propos ont été adaptées de Kadane (2011, p. 41-42), qui les présentent dans un autre contexte d'application que l'écologie.

2.2.6 *Sélection de modèles*

Les modèles hiérarchiques bayésiens permettent de tester un grand nombre de modèles dont les différences (p. ex. de structure, de variables explicatives, de processus écologique) permettent de tester des hypothèses écologiques. Plusieurs méthodes ont été utilisées lors de ces travaux de thèse. La première catégorie de méthodes consiste à trouver un équilibre entre adéquation du modèle aux données et parcimonie de ce modèle. Dans ce cas, deux critères sont disponibles pour comparer un groupe de modèles. Le *Bayesian Information criterion* (BIC) proposé par Schwarz (1978) permet de mesurer l'adéquation du modèle aux données tout en pénalisant les modèles avec un grand nombre de paramètres :

$$BIC = -2 \times \log(L(y|\theta)) + k \times \log(n) \tag{2.26}$$

En pratique, lorsque plusieurs modèles sont comparés avec le BIC, le modèle le plus parcimonieux et le plus en adéquation avec les observations est le modèle possédant le plus petit score de BIC. Il est généralement considéré qu'une différence entre 0 et 2 ne permet pas de sélectionner un modèle par rapport à l'autre. Cependant, une différence supérieure à 2 permet de sélectionner le modèle avec le score de BIC le plus faible.

Le second critère utilisé lors de ces travaux est le *Deviance Information Criterion* DIC (Spiegelhalter *et al.*, 2002) :

$$DIC = p_D + \bar{D} \tag{2.27}$$

ou de manière équivalente :

$$DIC = D(\bar{\theta}) + 2p_D \tag{2.28}$$

Ce critère mesure l'adéquation aux données par l'intermédiaire de la déviance $D(\theta)$:

$$D(\theta) = -2 \times \log(L(y|\theta)) \tag{2.29}$$

La complexité du modèle est mesurée par le «nombre effectif de paramètres» noté p_D :

$$p_D = \bar{D} - D(\bar{\theta}) \qquad \text{avec } \bar{D} = \mathbb{E}^{\theta}[D(\theta)] \tag{2.30}$$

La seconde approche consiste à mesurer la qualité des prédictions effectuées par le modèle. Le *Posterior predictive checking* est une méthode pour simuler des données dites répliquées qui sont ensuite comparées aux données observées (Gelman *et al.*, 1996). La procédure peut se décomposer en trois grandes étapes :

1. prendre N tirages $(\theta^{(1)}, \theta^{(2)}, ..., \theta^{(N)})$ dans la distribution *a posteriori* $[\theta|y]$.

2. Pour chaque tirage $\theta^{(i)}$, simuler les données répliquées selon le modèle testé.

3. Enfin, Gelman *et al.* (1996) propose d'utiliser des statistiques capables de mettre en évidence les différences entre les données observées et les données répliquées. Le modèle dont les données répliquées sont le plus proches des données observées sera donc sélectionné.

Une méthode alternative consiste à séparer le jeu de données observées en deux parties, la première est utilisée pour estimer les paramètres du modèle et le second groupe de données sert à mesurer la capacité de prédiction du modèle. Le critère ainsi utilisé est l'erreur quadratique moyenne de prédictions $MSPE$:

$$\begin{aligned} \text{MSPE}_s &= \sum_{s=1}^{S} \mathbb{E}((\hat{Y}_s - Y_s)^2) \\ &= \sum_{s=1}^{S} \{(\mathbb{E}(\hat{Y}_s) - Y_s)^2 + \mathbb{V}ar(\hat{Y}_s)\} \end{aligned} \tag{2.31}$$

Pour chaque observation Y_s, le critère $MSPE$ est calculé, le modèle possédant les plus petites $MSPE$ est considéré comme le modèle ayant les meilleures capacités prédictives. Ce critère a pour avantage de prendre en compte la variance de prédiction $\mathbb{V}ar(\hat{Y})$, mais également un terme de biais $(\mathbb{E}(\hat{Y}) - Y)^2$.

3

MODÉLISATION DE DONNÉES DE BIOMASSE À FORTE PROPORTION DE ZÉROS ET EFFORT D'ÉCHANTILLONNAGE VARIABLE

Dans ce chapitre, une étude comparative est proposée afin de mesurer les performances de deux approches statistiques adaptées aux données présentant une forte proportion de zéros. Cette étude s'intéresse plus particulièrement à l'impact de l'effort d'échantillonnage (p. ex. durée, volume échantillonné) lorsque celui-ci est variable au sein d'une même étude. Dans un premier temps, une étude par simulation est réalisée pour comparer les capacités d'inférence des modèles delta gamma (DG) et Poisson composé gamma (CPG). Plusieurs critères ont été utilisés afin d'explorer les différences entre les deux modèles. Enfin des données de pêches commerciales de poissons de fond au large de Vancouver ont été utilisées pour tester les deux modèles. Cette étude a été publiée dans la revue à comité de lecture *Methods in Ecology and Evolution*. On montre par l'étude de simulation que le modèle DG n'estime pas correctement les quantités d'intérêt (p. ex. biomasse moyenne, probabilité de présence) lorsque l'effort d'échantillonnage est très variable. A l'inverse, le modèle CPG est très robuste à la variabilité de l'effort d'échantillonnage et estime correctement les quantités d'intérêt. Ces résultats sont confirmés par l'étude de cas des données de pêches commerciales de poissons de fond au large de Vancouver, qui présentent une forte variabilité d'effort d'échantillonnage.

Methods in Ecology and Evolution

British Ecological Society

Methods in Ecology and Evolution 2013 doi: 10.1111/2041-210X.12122

Compound Poisson-gamma vs. delta-gamma to handle zero-inflated continuous data under a variable sampling volume

Jean-Baptiste Lecomte[1,2]*, Hugues P. Benoît[3], Sophie Ancelet[4], Marie-Pierre Etienne[1,2], Liliane Bel[1,2] and Eric Parent[1,2]

[1]*INRA, UMR 518 Math. Info. Appli., Paris F-75005 France;* [2]*AgroParisTech, UMR 518 Math. Info. Appli., Paris F-75005, France;* [3]*Gulf Fisheries Centre, Fisheries and Oceans Canada, Moncton, NB E1C 9B6,Canada; and* [4]*Institut de Radioprotection et de Sûreté Nucléare, Laboratoire d'épidémiologie, Fontenay-aux-Roses, France*

Summary

1. Ecological data such as biomasses often present a high proportion of zeros with possible skewed positive values. The Delta-Gamma (DG) approach, which models separately the presence–absence and the positive biomass, is commonly used in ecology. A less commonly known alternative is the compound Poisson-gamma (CPG) approach, which essentially mimics the process of capturing clusters of biomass during a sampling event.

2. Regardless of the approach, the effort involved in obtaining a sample (henceforth called the sampling volume, but could also include swept areas, sampling durations, etc.), which can potentially be quite variable between samples, needs to be taken into account when modelling the resulting sample biomass. This is achieved empirically for the DG approach (using a generalized linear model with sampling volume as a covariate), and theoretically for the CPG approach (by scaling a parameter of the model). In this study, the consequences of this disparity between approaches were explored first using theoretical arguments, then using simulations and finally by applying the approaches to catch data from a commercial groundfish trawl fishery.

3. The simulation study results point out that the DG approach can lead to poor estimates when far from standard idealized sampling assumptions. On the contrary, the CPG approach is much more robust to variable sampling conditions, confirming theoretical predictions. These results were confirmed by the case study for which model performances were weaker for the DG.

4. Given the results, care must be taken when choosing an approach for dealing with zero-inflated continuous data. The DG approach, which is easily implemented using standard statistical softwares, works well when the sampling volume variability is small. However, better results were obtained with the CPG model when dealing with variable sampling volumes.

Key-words: commercial fishery catches, compound Poisson, estimation of biomass, sampling variability, two-part model

Introduction

Ecological data for species population densities are often characterized by a large proportion of zero values accompanied by a skewed distribution of remaining values, including occasional extremes (Pennington 1996; Martin *et al.* 2005). Ignoring these features could lead to incorrect estimates of quantities of interest (e.g. mean biomass, probability of presence) and their associated uncertainty, and possibly to incorrect conclusions (Martin *et al.* 2005). Zero values in species population densities can originate from two general sources, with consequences for the appropriate analytical approach used to make inferences [see review in Martin *et al.* (2005)]. *True zeros* can occur as a direct result of the effect under study (e.g. suitability of a given 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 or observer effects. Our interest here lies in *true zeros*.

Standard continuous probability distributions such as the normal, gamma or log-normal are often inappropriate for the analysis of zero-inflated biomass data, even with ad hoc assumptions such as the addition of constants to create a mass at zero. A better approach is to use so-called two parts, hurdle or Delta models, which assume that zero and nonzero data arise, respectively, from separate processes (Stefansson 1996; Punt *et al.* 2000; Ortiz & Arocha 2004; Maunder & Punt 2004). This method does not require the addition of a constant, which can introduce a bias in the data. This model is also very flexible as covariates can be added in the zero and nonzero parts of the model using conventional generalized linear modelling techniques. However, the break between zero and

*Correspondence author. E-mail: jean-baptiste.lecomte@agroparistech.fr

nonzero values presents a particularly unnatural discontinuity in density data, where many zeros are actually stochastic clues of a strong gradient of decreasing biomass quantities. A second approach is the use of a positive distribution that simultaneously incorporates zeros and positive quantities. Jorgensen (1987) proposed the exponential dispersion model, with a power variance function. This model, also known as the Tweedie distribution, handles zero-inflated data without treating the zero and nonzero values separately. The Tweedie model and its variants have been applied to fisheries data (Candy 2004; Shono 2008; Foster & Bravington 2012; Lecomte *et al.* 2013). In this article, we rely on a gamma marked compound Poisson, named compound Poisson-gamma model (CPG), a member of the Tweedie family. Foster & Bravington (2012) extended it to be more flexible when covariates can affect parameters. They showed that the CPG mean–variance relationship is not necessarily constant, conversely to the Tweedie distribution (Foster & Bravington 2012). Parsimonious variant of this distribution, using exponential rather than gamma variables, has also been used [e.g. Ancelet *et al.* (2010)].

In many studies, the effort involved in obtaining a sample (henceforth called the sampling volume, but could also include swept areas, sampling durations, etc) can vary among sampling events. These differences in the sampling volume have to be accounted for in the analysis. Variable sampling volume is accounted for directly in the modelling for the CPG approach by scaling a parameter, whereas recourse to a generalized linear model to take into account the sampling volume as a covariate or an offset is required for the delta-gamma (DG) approach (Maunder & Punt 2004). Such different approaches to dealing with variable sampling volumes are likely to affect estimation reliability for quantities of interest (e.g. mean quantity, probability of presence).

This study evaluates the relative robustness of the DG and CPG approaches for estimating biomasses and presence probabilities under variable sampling volumes conditions in three ways. Firstly, the form and analytical properties of the two models are presented and contrasted from a theoretical perspective. Secondly, simulations were used to evaluate the robustness of the proposed models and compare their fitting abilities with variable volumes with different variances. Thirdly, the two approaches are applied to catch data from a commercial groundfish trawl fishery. Theory and analyses of simulated and observed data have all indicated that the CPG approach outperforms the DG approach under variable sampling volumes.

Materials and methods

THE DELTA-GAMMA MODEL

The delta modelling approach is based on the specification of two submodels to represent the biomass (Stefansson 1996). Let X be a binary variable that equals to 1 if the species of interest is present and 0 otherwise.

$$X \sim Bernoulli(\pi) \qquad \text{eqn 1}$$

π being the probability of the presence of the species. Conditionally, let Y be a positive sampled quantity of interest (e.g. species density or biomass) after a sampling event:

$$\begin{aligned} Y|X=1 &\sim \Gamma(\alpha,\beta) \\ Y|X=0 &\sim \delta_0 \end{aligned} \qquad \text{eqn 2}$$

with shape and rate parameters, (α,β) and δ_0 the Dirac distribution at zero. This yields the DG model, DG(π,α,β), with other distributional assumptions for strictly positive quantities yielding other models in the delta family, such as the delta log-normal. The expected value for the biomass under the DG model is $\mathbb{E}(Y)=\frac{\alpha}{\beta}\pi$, and the other main derived quantities (e.g. variance of the biomass, probability of presence) are summarized in Table 1.

Table 1. Quantities of interest (probability of presence, expected positive biomass, expected biomass and variance of biomass) for the DG and the CPG model under a standard sampling volume

	DG	CPG
Probability of presence	π	$1-e^{-\lambda}$
Expected positive biomass	$\frac{\alpha}{\beta}$	$\frac{a\lambda}{b(1-e^{-\lambda})}$
Expected biomass	$\frac{\alpha}{\beta}\pi$	$\lambda\frac{a}{b}$
Variance of the biomass	$\pi\frac{\alpha}{\beta^2}(1+\alpha(1-\pi))$	$\frac{a\lambda}{b}\frac{a+1}{b}$

A useful model property in statistical ecology is additivity with regards to the sampling volumes, in which the sum of two independent sampling events follows the same distribution type as each sampling event. For example, sampling during two hours follows the same distribution as two samplings process of one hour. It allows gathering data with different sampling volumes in the same model, as their sum is obtained according to a distribution in the same family. Unfortunately, the DG model is not additively coherent as pointed out by Stefansson (1996). As a consequence, it is not clear how the DG parameters vary in time or space when sampling volumes vary among sampling events. In practice, a simple way to deal with a non-constant sampling volume is to perform a pre-standardization of the data. The biomass collected is divided by the sampling volume. The problem with this method is that it only standardizes the positive data and leaves the presence–absence part unscaled, ignoring the fact that, as sampling volume increases, the probability of observing zero biomass should decrease when the species is present. A more relevant solution uses generalized linear modelling (e.g. Zuur *et al.* 2009) with the sampling volume as a covariate in each part of the DG model. The probability of presence is usually modelled with a logistic regression:

$$\log\left(\frac{\pi}{1-\pi}\right)=\zeta+\phi V \qquad \text{eqn 3}$$

where V is the sampling volume. A log-linear function is used to add the sampling volume in the expected positive biomass given the species is present :

$$\mathbb{E}(Y|X=1)=\frac{\alpha}{\beta}=\exp(\gamma+\delta V) \qquad \text{eqn 4}$$

A criticism of the delta approach is the separate modelling of the presence–absence and the strictly positive quantities. Consequently, gradients of biomass including low or null densities of the species are modelled disjointly in practice, which is a rather unnatural representation of the phenomenon being modelled.

THE COMPOUND POISSON-GAMMA MODEL

Conceptually, the CPG mimics the process involved when sampling most living organisms in nature for which the observed variable of interest is a continuous variable, such as the total

biomass captured during a sampling event (Foster & Bravington 2012; Lecomte *et al.* 2013). Simply put, the model assumes that a Poisson distributed number N of aggregations (i.e. patches or lumps) of organisms are collected, each patch containing a mass M_p modelled using a gamma distribution. It should be noted that an aggregation could contain only one organism (Foster & Bravington 2012). The sum of the individual masses of captured aggregations yields the total observed biomass Y:

$$Y = \begin{cases} \sum_{p=1}^{N} M_p & \text{if} \quad N > 0 \\ 0 & \text{if} \quad N = 0 \end{cases} \qquad \text{eqn 5}$$

The CPG is characterized by three parameters: λ the Poisson intensity, a and b the shape and rate gamma parameters:

$$Y \sim \mathrm{CPG}(\lambda, a, b) \qquad \text{eqn 6}$$

The main derived quantities for the CPG model are summarized in Table 1.

Due to additivity properties (Jorgensen 1987), the sampling volume V may be straightforwardly incorporated in a CPG model by scaling the Poisson intensity parameter:

$$Y \sim \mathrm{CPG}(\lambda V, a, b) \qquad \text{eqn 7}$$

The CPG approach jointly models the probability of presence and the nonzero sampled quantity. This capacity allows one to model a gradient of decreasing biomass in the distribution of the targeted species due to low density of organisms or low detectability.

There is no disjoint treatment of null and positive values as in the DG model. Foster & Bravington (2012) note that when no covariates are included in either the Poisson or gamma latent components, the CPG model belongs to the Tweedie family, and, in addition, a reviewer has noted that this is still the case when the set of covariates is identical in each of the Poisson and gamma components.

A SIMULATION STUDY TO COMPARE THE IMPACT OF VARIABLE SAMPLING VOLUME

The abilities of the DG and CPG models to reliably estimate quantities of interest when sampling volume is variable were compared using simulations. The trawls are divided into small fractions, or microvolumes, that could conceptually be although of as the sweeping of one unit of area by the trawl. Each microvolume contains a small amount of biomass produced according to a DG process. The observed sampled biomass is the sum of the biomass collected over those small microvolumes. Because the DG does not possess additivity, a biomass amount summed over all microvolumes constituting a complete trawl haul does not conform with either the DG or CPG model. The simulation proceeded as follows:

1. Biomass values were generated with a DG model of parameters α_{micro}, β_{micro} and π_{micro} corresponding to a sampled fraction V_{micro}. These biomasses are denoted as 'microbiomasses'.

2. The total collected biomass of a sample is the sum of N_{V} microbiomasses captured across all sampled microvolumes for that sample to result in a total volume V:

$$N_{\mathrm{V}} = \frac{V}{V_{\mathrm{micro}}}$$

3. The total volumes V are simulated according to a log-normal distribution:

$$V \sim \log N(0, \sigma^2)$$

with one of several variances σ^2 which varied between simulations: 0·1, 0·2, 0·3, 0·5, 0·7, 0·9, 1.00, corresponding, respectively, to a coefficient of variation of the sampling volumes C_{V}: 0·10, 0·20, 0·31, 0·53, 0·80, 1·12, 1·31 and a constant median of 1. For each σ^2 value, $n = 100$ data sets composed of 150 full samples were generated of the population of microvolumes.

Three quantities of interest can be expressed analytically as a function of the microvolume parameters. The expected biomass collected over N_{V} microvolumes, each producing a microbiomass from a DG distribution with parameters (α_{micro}, $\beta_{\mathrm{micro}}\pi_{\mathrm{micro}}$), equals:

$$Q = \mathbb{E}(Y) = \sum_{i=1}^{N_{\mathrm{V}}} \mathbb{E}(Y_{\mathrm{micro}_i}) = N_{\mathrm{V}}\pi_{\mathrm{micro}}\frac{\alpha_{\mathrm{micro}}}{\beta_{\mathrm{micro}}}.$$

The probability π of presence is obtained by noticing that:

$$\pi = P(Y > 0) = 1 - P(Y = 0) = 1 - \prod_{i=1}^{N_{\mathrm{V}}} P(Y_{\mathrm{micro}_i} = 0) = 1 - (1 - \pi_{micro})^{N_{\mathrm{V}}}. \qquad \text{eqn 8}$$

Finally, the strictly positive expected biomass is given by:

$$\mathrm{QP} = \mathbb{E}(Y|Y > 0) = \frac{\mathbb{E}(Y)}{P(Y > 0)} = N_V\pi_{\mathrm{micro}}\frac{\alpha_{micro}}{\beta_{\mathrm{micro}}(1 - (1 - \pi_{\mathrm{micro}})^{N_{\mathrm{V}}})}.$$

To simulate zero-inflated biomass data with a variation in the sampling volume, a very small microvolume $V_{\mathrm{micro}} = 0{\cdot}001$ and large numbers N_{V} were considered. According to equation 8, π_{micro} has to be chosen very small to simulate a realistic probability of presence. Three contrasting sets of the parameters (α_{micro}, β_{micro}, π_{micro}) were considered as follows: (200,2,0·001), (200,2,0·0005), (20,2,0·001). In those cases, if $N_{\mathrm{V}} = 1000$, the resulting sampled volume is $V = 1$. Thus, when $N_{\mathrm{V}} = 1000$, the mean biomass of a data set generated with parameters (200,2,0·001) was $Q = 100$, and the probability of presence was $\pi = 0{\cdot}63$. This data set presented a reasonable proportion of zeros associated with large positive biomasses that is often encountered in ecological surveys. Data sets generated with parameters (200,2,0·0005) were intended to investigate higher proportion of zeros ($Q = 50$ and $\pi = 0{\cdot}39$), whereas data sets simulated with parameters (20,2,0·001) were representative of situation with lower quantities of biomass ($Q = 10$ and $\pi = 0{\cdot}63$). Summing over a large number of microvolumes N_{V} allowed to simulate realistic continuous zero-inflated data with a variation in the sample volume. However, one may object that the previous large sum of microvolumes could unduly favour the additively consistent CPG model. That is why, a fourth set of parameters with a larger π_{micro} was chosen to test the robustness of the CPG model in situation far from the addition of a large number of microvolumes. In this case, a larger microvolume was chosen $V_{\mathrm{micro}} = 0{\cdot}3$, and a small number of microvolumes N_{V} were summed to ensure a realistic overall probability of presence. When $N_{\mathrm{V}} = 3$, data sets generated with this set of parameters (200,2,0·15) presented a mean biomass $Q = 45$ and a probability of presence $\pi = 0{\cdot}39$.

BAYESIAN INFERENCE

We choose to use Bayesian inference and computation, using Markov chain Monte Carlo methods. For both models, the Bayesian model specification requires prior distributions. Commonly, vague normal distributions, with mean zero and standard deviation 100, were chosen for

all regression parameters. For the positive parameters, weakly informative gamma prior distributions, *Gamma*(1,0.001), were chosen. The inference was carried out using OpenBUGS, the open version of WinBUGS (Ntzoufras 2011). For each model, three chains were run for 60,000 iterations, with a burn-in period of 30 000 iterations. A thinning of 100 iterations was performed to avoid autocorrelations in each chain. Convergence was assessed using the Gelman–Rubin convergence test. Maximum likelihood estimation of both models can be found in Foster & Bravington (2012).

MODEL EVALUATION

The effect of variable sampling volume was examined for three quantities of interest θ, namely mean biomass, mean strictly positive biomass and probability of absence. The results obtained for the two models were explored using four performances indices for each quantity. The first was the root mean squared error, which accounts for the common trade-off between variance and bias of the posterior mean of the quantity for the *ith* data set, $\widehat{\theta}_i$. It is defined as:

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\theta_{\mathrm{true}} - \widehat{\theta}_i)^2} \qquad \text{eqn 9}$$

where θ_{true} is the 'true' value of θ used in the simulations. The second was the estimated average coefficient of variation computed for each unknown quantity of interest, which highlights the relative estimated dispersion, and is defined as:

$$\hat{C}_{\mathrm{V}} = \frac{1}{n}\sum_{i=1}^{n}\frac{\widehat{\sigma}_{\theta_i}}{\widehat{\theta}_i} \qquad \text{eqn 10}$$

where $\widehat{\sigma}_{\theta_i}$is the posterior standard deviation of θ_i related to data set *i*. The third is the recovery ratio, $R_{90\%}$, (sometimes called the confidence coefficient), which is obtained by counting over the 100 data sets, how many times the true value falls within the 90% credible interval. It highlights the fitting capacity of the model. Finally, the average posterior median, $\bar{\widehat{\theta}}$, over the n=100 replicated data sets was computed as an estimator of the three quantities of interest.

CASE STUDY: COMMERCIAL FISHERY GROUNDFISH DATA

The consequences for how the CPG and DG approaches deal with variable sampling volumes were explored by applying the methods to commercial fishery catches, which are known to present variable sampling volumes between sampled sites and a high proportion of zeros. This case study is particularly pertinent because the synthesis of commercial fishery catches is routinely used to assess relative stock abundance in fisheries worldwide (e.g. Maunder & Punt 2004).

The data consisted of bottom trawl catches for two years, 2006 and 2009, from a commercial fishery that covered the continental shelf off the west coast of Canada. The two years of data were chosen because they presented a contrast in annual dispersions of the sampling duration. The mean duration of a sampling event for both years was 120 minutes, and all sampling volumes were scaled accordingly so that one unit of sampling effort corresponds to two hours of towing. Histograms of the sampling duration after rescaling by the mean are provided in Fig.1. The variation observed in these fisheries is commensurate with variation observed in other fisheries elsewhere (e.g. Fig.2). Such scaling by the mean led to the following contrasted variance between the selected years:

- 2006, with empirical variance $\hat{\sigma}^2_{V_{2006}} = 0{\cdot}31$ and empirical coefficient of variation $C_{\mathrm{V}} = 0{\cdot}56$,
- year 2009: with empirical variance $\hat{\sigma}^2_{V_{2009}} = 0{\cdot}14$ and empirical coefficient of variation $C_{\mathrm{V}} = 0{\cdot}37$.

The data for two species exhibited differences in mean sampled density between the dover sole (*Microstomus pacificus*, $Q_{\mathrm{sole}_{2009}} = 31$ in kg per tow) and the Pacific Ocean perch (*Sebastes alutus*, $Q_{\mathrm{perch}_{2009}} = 267$ in kg per tow). Both models were applied to data from each species and year separately. Depth (in metres) was added to both models as a covariate to account for its well-known effect on catch rates. Depth, which ranged from 50 to 500 *m*, was split into three classes to account for a possible nonlinear response with bin cut points at 125 *m* and 200 *m*. The most prevalent class (50, 125) was defined as the baseline effect. The resulting model for the delta approach was as follows:

$$\begin{aligned} Y &\sim \mathrm{DG}(\pi, \alpha, \beta) \\ \log\left(\frac{\pi}{1-\pi}\right) &= \zeta + \phi V + \kappa_{\mathrm{Depth}} \\ \frac{\alpha}{\beta} &= \exp(\gamma + \delta V + \eta_{\mathrm{Depth}}) \end{aligned} \quad . \qquad \text{eqn 11}$$

where κ_{Depth} and η_{Depth} account for the depth effect. The depth was incorporated via the Poisson intensity parameter in the CPG (consistent with Lecomte *et al.* 2013) although the effect of covariates can be added to either or both of the CPG parameters (Foster & Bravington 2012). The resulting model was as follows:

$$\begin{aligned} Y &\sim CPG(\lambda V, a, b) \\ \log(\lambda) &= \mu + \tau_{\mathrm{Depth}}. \end{aligned} \qquad \text{eqn 12}$$

Where μ is the intercept and τ_{Depth} denotes the depth effect. The same priors and estimation procedure , as the ones used in the simulation study, were considered for the Bayesian inference of the case study. The fitting ability of the two approaches was compared using the deviance information criterion (DIC) (Spiegelhalter *et al.* 2002). The posterior coefficients of variation $\hat{C}_{\mathrm{V}}$, 90% credible intervals CI , and the posterior medians $\hat{\theta}$ were computed for both approaches.

Results

SIMULATION STUDY

The simulation results of the data set generated with parameters set (200,2,0,001) are presented in this section. The three other data sets generated with the sets of parameters (200,2,0,0005), (20,2,0,001) and (200,2;0,15) are provided in Tables S1–S3 as their results are very similar to those of the first data set.

When sampling volume variability was small ($C_{\mathrm{V}} < 0{\cdot}8$), the estimates for the three quantities of interest were good and quite similar for both models, with well calibrated $R_{90\%}$, small RMSE and $\hat{C}_{\mathrm{V}}$ (Table 2). It is worth noting that for a log-normally distributed sampling volume with a unit median and variance σ^2_V, the mean is an increasing function of the variance σ^2_V. Consequently, for the data sets with a small variance, the results did not differ much from those obtained for a constant sampling volume equal to 1. As the C_{V} increases, DG estimates of the probability of absence and positive biomass are overestimated, the recovery ratios decrease, and relative uncertainties surrounding estimated parameters based on the $\hat{C}_{\mathrm{V}}$ increase. In contrast, the CPG approach was able to estimate correctly

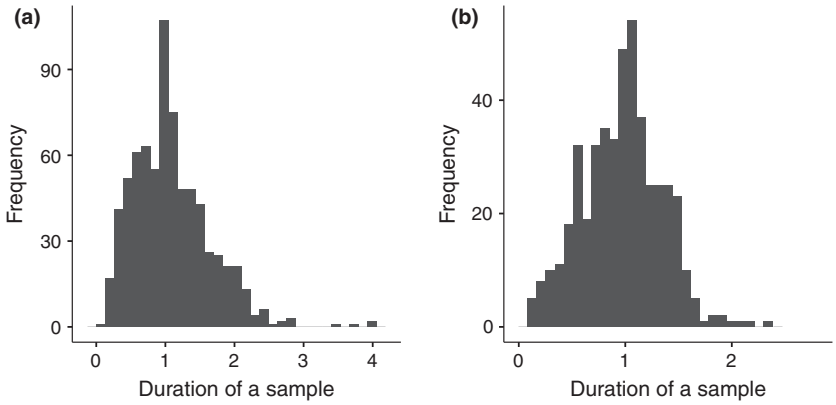

Fig. 1. Histograms of the duration of sampling events of the groundfish commercial catches after rescaling by its mean for the years (a) 2006 and (b) 2009.

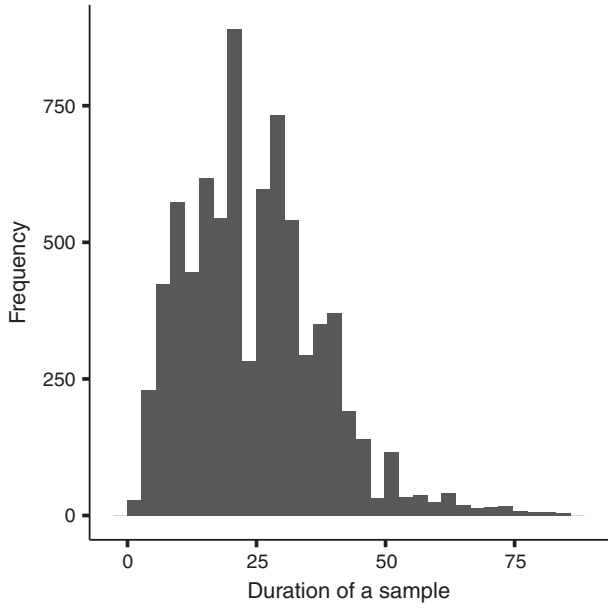

Fig. 2. Histogram of the fishing effort (hours) in the bottom-trawl fisheries of the southern Gulf of St Lawrence (Canada) for Atlantic cod and American plaice, in 1992, the year prior to a moratorium on cod fishing.

the simulated values, with recovery ratios that remained generally correct and constant. Overall, *RMSE* values were lower for the CPG compared to the DG model even when C_V was small. These general patterns remained for different choices of simulated parameters (see Tables S1–S2 in Supporting information).

CASE STUDY: COMMERCIAL FISHERY GROUNDFISH DATA

The probability of absence of dover sole estimated by the two models was high and similar for both years (Table 3). Estimated depth parameters were in accordance between models with depth classes (125, 200) and, (200, 500) having a positive effect on the probability of presence relative to shallow depths, as it was observed with the CPG parameters (Tables 4 and 5). No depth effects were detected with the DG approach for the modelling of the positive biomass (Table 5). In contrast to absence probability, estimates of the overall mean and of the mean positive biomass differed between models for the 2006 data although not the 2009 data (Table 3). Recall that sampling volumes were more variable in 2006 than in 2009.

Results for the ocean perch were similar to those for dover sole. Depth parameters estimates were in accordance between models (Tables 6 and 7). Depth classes (125, 200) and (200, 500) had a positive effect on the presence of the Pacific Ocean perch regarding shallow depths for both years.

Both models similarly estimated a high probability of absence (Table 8), but estimates of the overall mean and of the mean positive biomasses differed dramatically between models for the 2006 data, and to a much less extent for the 2009 data. DIC scores were lower for the CPG than the DG model for both years (Table 9), indicating that the fitting capacity of the CPG model was better than that of the DG model. The CPG model remains a model of choice even in situations where the observed biomass is the sum of a small number of DG-distributed microvolumes biomasses as shown in Table S1, Supporting information.

Discussion

The simulations used in this study allowed for a comparison of two statistical approaches for continuous zero-inflated data by relying on simulated data that mimics the catches of organisms in a uniform habitat with zero-inflation and continuous values of abundance. Based on the simulations, variable sampling volumes were found to produce inference challenges for the DG but not for the CPG distribution. This is consistent with the theoretical arguments we presented concerning the additivity property.

The case study and simulations confirmed that under a variable sampling duration, as it is often encountered in fisheries and other ecological data, the CPG model outperforms the DG overall, providing better fits to data and correct inferences on estimated quantities. The DG model in such situations

Table 2. Estimation of mean biomass Q, mean positive biomass QP and probability of absence $1-\pi$ with a variable sampling volume, for the simulated parameter set ($\alpha_{micro} = 200, \beta_{micro} = 2, \pi_{micro} = 0{\cdot}001$)

		$\overline{\hat{\theta}}$		R		RMSE		$\hat{C_V}$	
Volume	θ_{true}	CPG	DG	CPG	DG	CPG	DG	CPG	DG
$C_V = 0{\cdot}1$									
Q	100	**95·83**	95·6	**78**	73	8·05	8·15	0·08	0·08
QP	158·15	**155·42**	154·2	77	**81**	**5·41**	7·62	0·03	0·05
$1-\pi$	0·37	0·38	0·38	77	**88**	**0·03**	0·04	0·08	0·11
$C_V = 0{\cdot}2$									
Q	100	**95·82**	95·56	**85**	84	**7·06**	7·28	0·08	0·08
QP	158·15	**155·41**	155·06	**86**	78	**4·66**	7·69	0·03	0·05
$1-\pi$	0·37	0·38	0·38	85	**87**	0·03	0·03	0·08	0·11
$C_V = 0{\cdot}31$									
Q	100	**96·33**	95·67	87	86	**6·71**	7·19	0·08	0·09
QP	158·15	155·75	**156·04**	88	91	**4·45**	6·49	0·03	0·05
$1-\pi$	0·37	0·38	0·39	87	85	**0·03**	0·04	0·08	0·11
$C_V = 0{\cdot}53$									
Q	100	**96·3**	95·86	88	84	**6·45**	7·58	0·08	0·09
QP	158·15	155·59	**158·8**	87	86	**4·32**	6·43	0·03	0·06
$1-\pi$	0·37	0·38	0·4	87	85	**0·02**	0·04	0·08	0·11
$C_V = 0{\cdot}8$									
Q	100	96·27	**97·28**	84	**86**	**6·37**	8·64	0·07	0·1
QP	158·15	**155·62**	164·61	83	83	**4·23**	10·67	0·03	0·06
$1-\pi$	0·37	**0·38**	0·41	**84**	71	**0·02**	0·05	0·07	0·11
$C_V = 1{\cdot}12$									
Q	100	94·04	**99·77**	76	90	**7·13**	8·34	0·07	0·11
QP	158·15	**154·12**	173·66	**76**	53	**4·74**	17·45	0·03	0·06
$1-\pi$	0·37	**0·39**	0·42	**75**	69	**0·03**	0·06	0·06	0·12
$C_V = 1{\cdot}31$									
Q	100	**95·99**	106·95	**88**	82	**5·88**	12·27	0·07	0·11
QP	158·15	**155·36**	186·94	88	23	**3·97**	31·81	0·03	0·06
$1-\pi$	0·37	**0·38**	0·43	**86**	64	**0·02**	0·06	0·06	0·12

C_V is the coefficient of variation of the simulated sampling volume. θ_{true} is the value used to produce the simulations, $\overline{\hat{\theta}}$ is the posterior median, $R_{90\%}$ is the recovery ratio and should be 90%, RMSE is the root mean squared error and $\hat{C_V}$ is the average estimated coefficient of variation and have to be the lowest. Values in bold denote the best fit.

Table 3. Estimation of mean biomass Q, mean positive biomass QP and probability of absence $1-\pi$ for the dover sole sampled in 2006 and 2009

	$\hat{\theta}$		CI		$\hat{C_V}$	
Year	PG	DG	PG	DG	PG	DG
2006						
Q	4·69	7·86	2·95–7·03	4·6-1-2·09	0·27	0·28
QP	100·22	154·09	88·36–113·21	129·38–183·36	0·07	0·1
$1-\pi$	0·95	0·95	0·93–0·97	0·92–0·97	0·01	0·01
2009						
Q	40·9	38·06	31·04–53·54	27·21–52·54	0·17	0·2
QP	154·81	157·76	131·47–184·35	122·24–199·08	0·11	0·14
$1-\pi$	0·74	0·76	0·68–0·79	0·7–0·8	0·05	0·04

$\hat{\theta}$ is the posterior median, CI is the credible interval at 95% and $\hat{C_V}$ is the coefficient of variation.

Table 4. Parameter estimates for the CPG model fitted to the dover sole biomass data sampled in 2006 and 2009

		2006		2009	
Parameter	Term	Mean	SD	Mean	SD
Intercept		−3·047	0·255	−1·187	0·157
Depth	(125,200)	3·13	0·273	−3·401	0·811
Depth	(200,500)	2·985	0·269	0·206	0·226
a		0·528	0·047	0·93	0·138
b		0·005	0·001	0·007	0·001

tends to overestimate mean biomass values, potentially leading to incorrect conclusions, which in the case of fisheries may mean incorrect stock management recommendations. These differences in fitting capacity could be explained by the structure of the CPG model, which can handle variable sampling volumes easily because of the additivity property, whereas the DG approach takes variable sampling volume empirically with the help of a generalized linear model. However, when the sampling volume variability is small, the models performed comparably in the simulation. Fortunately, small sampling volume variability is more the rule than the exception in standardized surveys, and the DG approach therefore remains a valid standard practice in those cases. It is in cases where data do not come from planned surveys, or when data are from two or more surveys with different sampling durations and for which a joint analysis is desired that model choice becomes very

important. This choice can have important ecological and economic consequences. For example, commercial fishery catch-rate data, such as those analysed here for groundfish or the ones exemplified in Fig. 2 for cod, provide the data required to estimate relative abundance indices. These indices form the basis for a large number of stock assessments world-wide, including tuna and cod fisheries that are both highly lucrative and that pose important conservation concerns (e.g. Ahrens (2010); Carruthers *et al.* (2011)). Incorrect inferences drawn from the data are liable to lead to incorrect stock assessment advice and a potential that conservation or economic objectives for a fishery will not be achieved.

In the case study, sampling volume and depth were modelled to affect only the number of patches for the CPG models as, for example, increasing the duration of a sampling event results in an increased number of captured patches. Patch size should vary randomly with respect to changes in sampling volume if a sample is taken in a generally homogeneous habitat. Of course, if increasing sampling volume causes a sample to span more than one area of homogeneous habitat, then both patch number and size can vary in complex ways, and the underlying assumptions of both the CPG and DG could be violated.

Ancelet *et al.* (2010) pointed out a high correlation between the two quantities (number of patches, biomass in one patch) in a special case of the CPG approach. This result suggests that when the CPG distribution is used to model the effect of covariates on the property of interest, such as in generalized linear models (Stefansson 1996; Shono 2008; Zuur *et al.* 2009; Foster & Bravington 2012) or additive models (Zuur *et al.* 2009), it is appropriate to link only one of these two hidden quantities to the explanatory covariates. We suggest that it is most appropriate to model the effect of covariates on the number of patches only, because it tunes both the presence–absence and the quantity of biomass sampled. The parameters are heuristically defined as the number and biomass of patches

Table 5. Parameter estimates for the DG model fitted to the dover sole biomass data sampled in 2006 and 2009

			2006		2009	
Part	Parameter	Term	Mean	SD	Mean	SD
Bernoulli	Intercept		−3·431	0·346	−5·711	0·882
	Volume		0·485	0·176	1·128	0·345
	Depth	(125, 200)	3·459	0·304	3·439	0·799
	Depth	(200, 500)	3·339	0·297	3·753	0·807
Gamma	Intercept		−0·72	0·131	−0·272	0·245
	Volume		0·221	0·091	0·21	0·189
	Depth	(125, 200)	−0·071	0·127	−0·044	0·195
	Depth	(200, 500)	−0·02	0·12	−0·087	0·225

Table 6. Parameter estimates for the CPG model fitted to the pacific ocean perch biomass data sampled in 2006 and 2009

		2006		2009	
Parameter	Term	Mean	SD	Mean	SD
Intercept		−3·434	0·31	−8·452	2·932
Depth	(125, 200)	4·009	0·317	7·194	2·939
Depth	(200, 500)	4·759	0·319	7·849	2·933
a		0·405	0·039	1·568	0·23
b		0·001	0	0·002	0

Table 7. Parameter estimates for the DG model fitted to the pacific ocean perch biomass data sampled in 2006 and 2009

			2006		2009	
Part	Parameter	Term	Mean	SD	Mean	SD
Bernoulli	Intercept		−3·035	0·495	−10·817	3·634
	Volume		−0·077	0·317	1·293	0·344
	Depth	(125, 200)	4·393	0·359	8·356	3·591
	Depth	(200, 500)	7·409	0·647	9·053	3·612
Gamma	Intercept		−0·498	0·101	0·042	0·212
	Volume		0·17	0·075	0·531	0·147
	Depth	(125, 200)	0·014	0·102	−0·306	0·173
	Depth	(200, 500)	0·032	0·098	0·18	0·173

Table 9. Deviance information criterion (DIC) scores related to the DG and CPG models fitted to the data sets of the two species collected in 2006 and 2009 by commercial fisheries

	Perch		Sole	
	2006	2009	2006	2009
DG	7946	2066	4319	1539
CPG	7092	1597	3490	1100

Table 8. Estimation of mean biomass Q, mean positive biomass QP and probability of absence $1-\pi$ for the pacific ocean perch sampled in 2006 and 2009. $\hat{\theta}$ is the posterior median, CI is the credible interval at 95% and $\hat{C_V}$ is the coefficient of variation

	$\hat{\theta}$		CI		$\hat{C_V}$	
Year	PG	DG	PG	DG	PG	DG
2006						
Q	17·35	69·32	10·11–27·2	41·06–107·74	0·29	0·3
QP	538·73	1604·82	473·58–610·62	1417·96–1835·28	0·08	0·08
$1-\pi$	0·97	0·96	0·95–0·98	0·94–0·98	0·01	0·01
2009						
Q	281·14	222·35	207·97–358·88	160·41–296·36	0·16	0·18
QP	1136·87	943·78	980·41–1309·96	755·52–1149·25	0·09	0·13
$1-\pi$	0·75	0·76	0·7–0·81	0·71–0·81	0·04	0·04

although these ecological properties are not actually being estimated. Foster & Bravington (2012) used a data set which was composed of biomass and abundance data to explore the relationship between patch size and the size of one typical fish coming from this patch. They showed that the size and the number of patches could have a different relationship to those for the size and number of individual fish. However, such an hypothesis about the size and numbers of patches collected during a sampling event need to be checked. Even if the conjunction of the parameters yields a distribution of biomass values possessing the properties of interest, that is, zero-inflation as well as continuous values with occasional extremes and additivity with respect to variable sampling volume, one must not over interpret an ecological meaning for the individual parameters.

We conclude with practical recommendations arising from this work. When facing zero-inflated data with a constant sampling volume or a sampling volume with a low variability, the DG approach is likely to be understandably preferred by many because of its ease of implementation. However, when working with variable sampling volumes, the analyst should be wary of the DG model. We suggest the CPG structure as a better alternative, even at the cost of some increased complexity of implementation. If not, the simulation study developed in this study shows that, conversely to the CPG, the DG estimates may provide fallacious conclusions, unduly overestimating the biomass quantities.

Acknowledgements

We are indebted to the insightful comments of two anonymous reviewers, which greatly improve the manuscript. We also want to thank an anonymous reviewer for proposing the use of the sampling volumes as a covariate in both parts of the DG approach, which allows for a fair comparison on a more balanced basis.

References

Ahrens, R. (2010) *Global Analysis of Apparent Trends in Abundance and Recruitment of Large Tunas and Billfishes Inferred from Japanese Longline Catch and Effort Data*. PhD thesis. The University of British Columbia, Vancouver, BC.

Ancelet, S., Etienne, M.-P., Benoît, H. P. & Parent, E. (2010) Modelling spatial zeroinated continuous data with an exponentially compound Poisson process. *Environmental and Ecological Statistics*, **17**, 347–376.

Candy, S. (2004) Modelling catch and effort data using generalised linear models, the Tweedie distribution, random vessel effects and random stratum-by-year effects. *CCAMLR Science*, **11**, 59–80.

Carruthers, T.R., Ahrens, R.N., McAllister, M.K. & Walters, C.J. (2011) Integrating imputation and standardization of catch rate data in the calculation of relative abundance indices. *Fisheries Research*, **109**, 157–167.

Foster, S.D. & Bravington, M.V. (2012) A PoissonGamma model for analysis of ecological non-negative continuous data. *Environmental and Ecological Statistics*, **19**, 1–20.

Jorgensen, B. (1987) Exponential dispersion models. *Journal of the Royal Statistical Society. Series B (Methodological)*, **49**, 127–162.

Lecomte, J.B., Benoît, H.P., Etienne, M.P., Bel, L. & Parent, E. (2013) Modeling the habitat associations and spatial distribution of benthic macroinvertebrates, A hierarchical Bayesian model for zeroinated biomass data. *Ecological Modelling*, **265**, 74–84.

Martin, T., Wintle, B., Rhodes, J., Kuhnert, P., Field, S., Low-Choy, S., Tyre, A. & Possingham, H. (2005) Zero tolerance ecology, improving ecological inference by modelling the source of zero observations. *Ecology Letters*, **8**, 1235–1246.

Maunder, M.N. & Punt, A.E. (2004) Standardizing catch and effort data: a review of recent approaches. *Fisheries Research*, **70**, 141–159.

Ntzoufras, I. (2011) *Bayesian Modeling using WinBUGS, volume 698*. Wiley, Hoboken, NJ.

Ortiz, M. & Arocha, F. (2004) Alternative error distribution models for standardization of catch rates of non-target species from a pelagic longline fishery: billfish species in the Venezuelan tuna longline fishery. *Fisheries Research*, **70**, 275–297.

Pennington, M. (1996) Estimating the mean and variance from highly skewed marine data. *Fishery Bulletin*, **94**, 498–505.

Punt, A.E., Walker, T.I., Taylor, B.L., & Pribac, F. (2000) Standardization of catch and effort data in a spatially-structured shark fishery. *Fisheries Research*, **45**, 129–145.

Shono, H. (2008) Application of the Tweedie distribution to zero-catch data in CPUE analysis. *Fisheries Research*, **93**, 154–162.

Spiegelhalter, D.J., Best, N.G., Carlin, B.P. & der Linde, A. (2002) Bayesian measures of model complexity and fit. *Journal of the Royal stsistical Society: Series B (Statistical Methodology)*.**64**, 583–639.

Stefansson, G. (1996) Analysis of groundfish survey abundance data: combining the GLM and delta approaches. *ICES Journal of Marine Science*, **53**, 577–588.

Zuur, A.F., Ieno, E.N., Walker, N., Saveliev, A.A., & Smith, G.M. (2009) *Mixed Effects Models and Extensions in Ecology with R. Statistics for Biology and Health*. Springer, New York.

Received 27 August 2013; accepted 27 September 2013
Handling Editor: Robert B. O'Hara

Supporting Information

Additional Supporting Information may be found in the online version of this article.

Table S1. Estimation of mean biomass Q, mean positive biomass QP and probability of absence $1-\pi$ with a variable sampling volume, for the simulated parameter set ($\alpha_{\text{micro}} = 200, \beta_{\text{micro}} = 2, \pi_{\text{micro}} = 0{\cdot}0005$).

Table S2. Estimation of mean biomass Q, mean positive biomass QP and probability of absence $1-\pi$ with a variable sampling volume, for the simulated parameter set ($\alpha_{\text{micro}} = 20, \beta_{\text{micro}} = 2, \pi_{\text{micro}} = 0{\cdot}001$).

Table S3. Estimation of mean biomass Q, mean positive biomass QP and probability of absence $1-\pi$ with a variable sampling volume, for the simulated parameter set ($\alpha_{\text{micro}} = 200, \beta_{\text{micro}} = 2, \pi_{\text{micro}} = 0{\cdot}15$).

4

MODÉLISATION DE DONNÉES DE BIOMASSE GÉORÉFÉRENCÉES À FORTE PROPORTION DE ZÉROS

On s'intéresse dans cette partie à la modélisation de la distribution spatiale d'espèces d'invertébrés dans le sud du golfe du Saint-Laurent (sGSL).

Dans un premier travail, Ancelet *et al.* (2009) ont étudié la distribution spatiale d'espèces d'invertébrés dans le sGSL avec un modèle hiérarchique bayésien. Le modèle ainsi proposé est un proche du modèle Poisson composé gamma. En effet, Ancelet *et al.* (2009) ont utilisé un Poisson composé à marques exponentielles comme modèle d'observations. Le modèle latent est donc composé des deux quantités N et M :

$$\begin{aligned} N &\sim Poisson(\lambda) \\ M_p &\sim Exp(\rho), \qquad p = 1, ..., N \end{aligned} \tag{4.1}$$

où N est le nombre de patchs et M la biomasse contenue dans un patch p. Afin de prendre en compte la dimension spatiale de la distribution des espèces étudiées, Ancelet *et al.* (2009) ont proposé de partitionner le sGSL en I sous-unités géographiques d'habitat homogène et d'introduire des dépendances entre sous-unités proches géographiquement. Pour ce faire, un modèle BYM (Besag *et al.*, 1991), développé originellement pour les études épidémiologiques, est mis en œuvre. Ce modèle permet d'introduire des effets aléatoires non gaussiens et spatialement structurés dans la couche latente :

$$\begin{aligned} \lambda_i &= exp(\mu + \Phi_i + \epsilon_i) \\ \epsilon_i &\overset{i.i.d.}{\sim} Normal(0, \sigma_\epsilon^2) \end{aligned} \tag{4.2}$$

où μ est la moyenne commune à toutes les sous-unités géographiques, ϵ_i sont les effets aléatoires normaux *i.i.d*. La structure spatiale est définie par le vecteur Φ selon un modèle gaussien conditionnel autorégressif intrinsèque défini par les lois conditionnelles :

$$[\Phi_i|\Phi_j, j \sim i] \propto \exp\left\{ -\frac{(\Phi_i - \frac{1}{n_i}\sum_{j\sim i}\Phi_j)^2}{2\frac{\sigma^2_{IAR}}{n_i}} \right\} \tag{4.3}$$

où $j \sim i$ signifie que les deux unités j et i sont voisines. Ces deux unités seront voisines si elles partagent une frontière en commun. n_i est le nombre de voisins de l'unité i. Enfin, le paramètre de variance locale est σ^2_{IAR}.

La principale limite de ce type de modélisation est le partitionnement arbitraire de la zone d'étude. En effet, le modèle CPG permet une cohérence distributionnelle à grande échelle, or ce découpage implique une cohérence distributionnelle au sein d'une même sous-unité, mais pas entre sous-unité. La cohérence spatiale à grande échelle est donc perdue. Afin de conserver cette cohérence spatiale, il est nécessaire de travailler directement au niveau des observations et non plus entre sous unités géographiques.

Dans ces travaux de thèse, une approche alternative spatialement explicite a été mise en œuvre. Elle repose sur le modèle CPG, mais la structure spatiale est modélisée avec les outils de la géostatistique et plus particulièrement le variogramme exponentiel. Ces outils permettent d'introduire une cohérence spatiale à grande échelle et ne requièrent pas le découpage de la zone d'étude en unités homogènes. L'approche ainsi proposée permet d'introduire plusieurs variables environnementales (types de sédiments, profondeur, température) associées à des erreurs spatialement structurées dans la partie latente du modèle hiérarchique. Les répartitions spatiales de trois espèces d'invertébrés épibenthiques ont été étudiées pour les quantités de biomasses récoltées sur l'ensemble du sGSL en 1997 par Pêche et Océan Canada. Cette approche de modélisation permet de créer des cartes de quantités d'intérêt (p. ex. biomasse moyenne, probabilité de présence) sur l'ensemble du sGSL. Ce travail a été publié dans la revue scientifique Ecological Modelling en juillet 2013.

Ecological Modelling 265 (2013) 74–84

Contents lists available at SciVerse ScienceDirect

Ecological Modelling

journal homepage: www.elsevier.com/locate/ecolmodel

Review

Modeling the habitat associations and spatial distribution of benthic macroinvertebrates: A hierarchical Bayesian model for zero-inflated biomass data

J.B. Lecomte[a,b,*], H.P. Benoît[c], M.P. Etienne[a,b], L. Bel[a,b], E. Parent[a,b]

[a] INRA, UMR 518 Math. Info. Appli., F-75005 Paris, France
[b] AgroParisTech, UMR 518 Math. Info. Appli., F-75005 Paris, France
[c] Gulf Fisheries Centre, Fisheries and Oceans Canada, Moncton, NB E1C 9B6, Canada

ARTICLE INFO

Article history:
Received 10 April 2013
Received in revised form 7 June 2013
Accepted 8 June 2013
Available online 5 July 2013

Keywords:
Zero-inflated data
Bayesian hierarchical modeling
Habitat associations
Spatial dependencies
Macro-invertebrates

ABSTRACT

Biomass samples from marine scientific surveys are commonly used to investigate spatial and temporal variations in stock abundances. Biomass records are often characterized by a high proportion of zeros on the one hand, and occasional large catches on the other. These features induce a modeling challenge when trying to understand the state of populations and their ecological associations with one another and with habitat. We develop a hierarchical Bayesian model to represent the spatial structure of biomass and analyze the spatial distribution and habitat associations of three species of macro-invertebrates sampled in the southern Gulf of St. Lawrence (Canada). A zero-inflated distribution based on a compound Poisson with Gamma marks is used for the observation layer, and a linear model with spatial correlated errors accounts for the role of habitat variables (temperature, depth and sediment type) in the process layer. Maps of quantities of interest (e.g. probability of presence, quantity of biomass) are produced, taking into account the uncertainty of the estimated parameters and observation errors. This hierarchical Bayesian modeling approach provides a useful tool for spatial management of human activities that may affect living resources that may affect living resources, such as marine protected areas.

Contents

* Corresponding author at: INRA, UMR 518, 16 rue Claude Bernard, 75005 Paris, France. Tel.: +33 144087292.
E-mail address: jean-baptiste.lecomte@agroparistech.fr (J.B. Lecomte).

0304-3800/$ – see front matter.
http://dx.doi.org/10.1016/j.ecolmodel.2013.06.017

1. Introduction

Understanding species spatial distribution and habitat associations are key challenges when managing harvested, endangered or invasive species (Welsh et al., 1996; Engler et al., 2004; Cook et al., 2007). Marine spatial management measures, in which the spatial and temporal distribution of human activities is restricted to achieve ecological, social and economic objectives (e.g., marine protected areas), have been the focus of many studies in the recent decades (Shea, 1998; Hilborn et al., 2004; Hobday and Hartmann, 2006; Hartog et al., 2011). In many applications, these management approaches require knowledge of habitat use by the targeted species to be effective (Perry and Smith, 1994; Williams and Bax, 2001).

Linear or additive models are often developed to infer distributions and habitat use and preferences using survey or other ecological data (as reviewed by Guisan and Thuiller, 2005). Efficient models must be able to address two common characteristics of ecological data: observations can be dominated by a large number of null values combined with skewed positive values, and abundance can be strongly spatially correlated. Failure to address both of these characteristics is well known to impact model parameter estimates and their uncertainty, leading to incorrect statistical inference and therefore, in turn, potentially inappropriate management actions (Zuur et al., 2009; Sileshi et al., 2009). Ideally, the models should also be able to address possible spatial misalignment between the available data for abundances and for habitat characteristics.

High proportions of zeros in survey data stem from three general causes. An observed zero value can be a true zero if the species is not present in the studied area, while a false zero, also called pseudo-absence, results from a low probability of detection even though the species is present. A third class of zeros results from an observer effect, whereby a species normally found in the study area is frightened away by some inappropriate data collection procedure. Numerous approaches exist for such zero-inflated data when dealing with counts, as reviewed in Martin et al. (2005). The two main approaches, Zero-inflated Poisson (ZIP) and Zero-inflated binomial (ZIB), are mixture models and the presence–absence is modeled separately from the number of counts (i.e. individuals). The development of zero-inflated models for continuous abundance data (i.e. densities or biomasses) has also received attention (Stefansson, 1996; Maunder and Punt, 2004; Fletcher et al., 2005; Shono, 2008; Ancelet et al., 2010). The simplest approach consists in adding a positive constant to all the observations, typically followed by a logarithmic transformation, as is often performed in generalized linear modeling (GLM). This approach requires choosing an arbitrary constant that could severely bias model estimates (Maunder and Punt, 2004; Shono, 2008). An alternative is to remove the zero catches from data prior to the analysis. However removing zero values often affects the results and can also bias the analysis (Martin et al., 2005), though this is not necessarily the case (Maunder and Punt, 2004). A common and slightly more complex approach for continuous data, named the delta approach (Stefansson, 1996; Shono, 2008), models separately the presence–absence using a binomial distribution and positive values using a standard probability distribution function such as the log-normal (leading to a delta-lognormal model) or the gamma (delta-gamma). The approach reduces bias since the expected biomass is the product of the probability of presence and the average positive biomass. This family of models treats all absences as true zeros. Furthermore, sampling effort, which can vary between sites for a number of logistical and operational reasons, is mostly addressed by a prior standardization of the data (Stefansson, 1996). However, performing such a standardization may obscure the relationship that exists between expected values (for a given sampling effort) and their associated variance for count probability density functions.

In this paper, we develop a hierarchical Bayesian spatial model for biomass data that overcomes these shortcomings. We apply this approach to describe the distribution and habitat associations of epibenthic invertebrates in the southern Gulf of St. Lawrence (sGSL), Canada. The biomass records come from an annual bottom trawl survey in which invertebrates and fish are collected at randomly chosen locations by sweeping the ocean floor over targeted distances which can vary between sites. We use a model based on two substructures that are linked probabilistically using a hierarchical approach. The first substructure, the observation layer, consists of a compound Poisson model with Gamma marks, which heuristically models the process of observing a Poissonian number of patches of a species, each containing a random biomass given by the Gamma mark. This approach constitutes a generalization of the one proposed by Bernier and Fandeux (1970) and applied in ecology by Ancelet et al. (2010) which used exponential marks. It also allows for explicit accounting for the duration or volume of sampling for individual sampling events. The second model substructure explicitly models habitat associations using a linear model that accounts for spatial autocorrelation using a geostatistical approach. Jointly, these model substructures result in a modeling approach that is very flexible, likely making it a useful tool for spatial analysis and planning.

2. Methods

2.1. Data description

Fisheries and Oceans Canada has conducted an annual bottom-trawl survey in the sGSL each September since 1971 (Chadwick et al., 2007; Benoît et al., 2009). Since its inception, the main objective of this survey has been to quantify the abundance and the distribution of marine fishes and certain commercially important invertebrates. Since 1988, data for epibenthic invertebrates such as urchins, starfish, whelks and anemones have been collected. The domain for the sGSL survey is split into 27 strata defined so as to be homogeneous in terms of depth and geographic location. Every year, since the mid 1980s, 140–200 sites have been chosen according to a stratified random design. The number of sites per stratum is generally proportional to stratum size, making the selection of sites at the survey level approximately randomly balanced. Sites are sampled using a straight-line tow for a target duration of 30 min. at 3.5 knots. All captured organisms are identified to the lowest taxonomic level possible and weighed in kilograms per tow. Habitat information, such as bottom temperature (°C) and depth (m), is also collected at each bottom-trawl site. Moreover the type of sediments is interpolated at each sampling site from an existing map of surficial geology for the Gulf of St. Lawrence (Loring and Nota, 1973). This study focuses on three epibenthic macroinvertebrates sampled during the 1997 survey to illustrate the modeling approach: green sea urchin (*Strongylocentrotus droebachiensis*), starfish (*Asterias sp*), and sea cucumber (*Cucumaria frondosa*). These three taxa were chosen for their differences in density distribution and habitat preferences so as to demonstrate the model's ability to confront different data situations (Figs. 1 and 2). In fact, the majority of epibenthic macroinvertebrates in the sGSL are distributed in patches of localized variable abundance, interspersed by numerous and relatively large areas where the species is absent. Consequently, the dataset contains a very large proportion of sites where the species are not observed.

2.2. The statistical model for zero-inflated continuous positive data

The model description is split into two parts, as is classically done in hierarchical Bayesian modeling. The first section describes

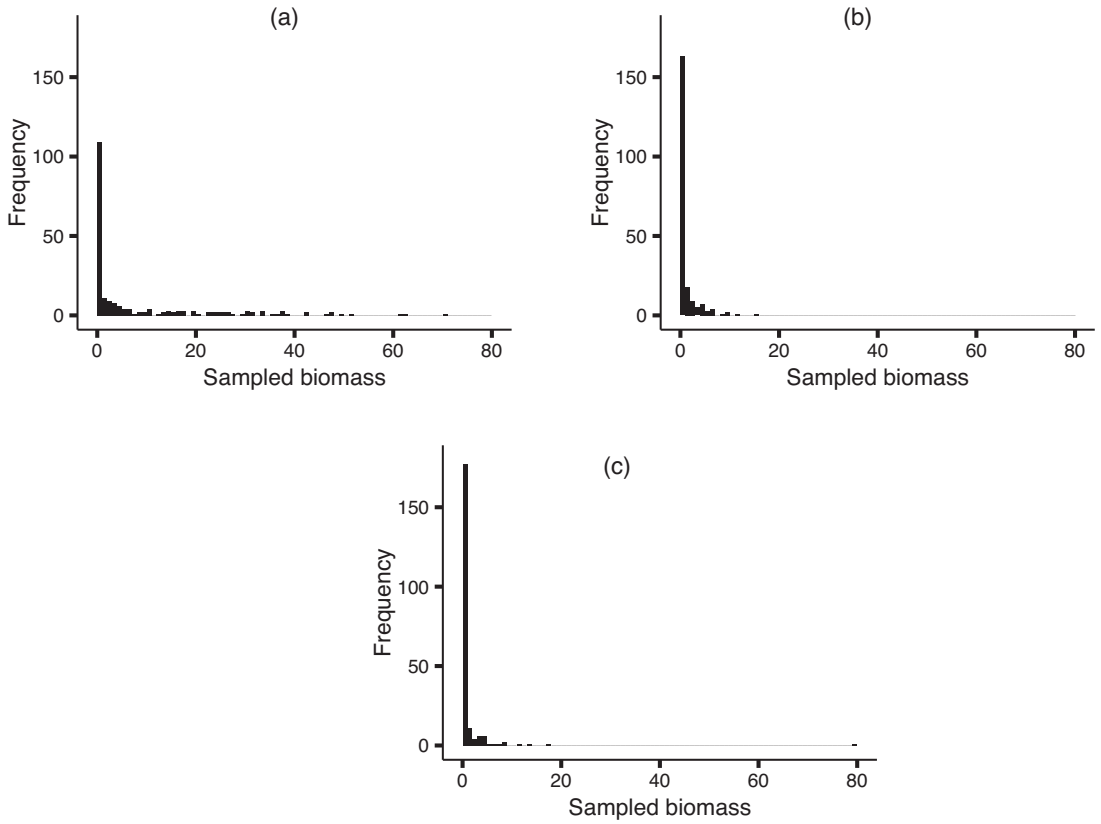

Fig. 1. Histogram of the sampled biomass from individual tows in the 1997 sGSL bottom-trawl survey year (in kg/tow) for the three selected species: (a) urchin, (b) starfish, and (c) cucumber.

how the observation is linked to the actual biomass in a given site. The second section models the spatial distribution of the latent biomass field at the scale of the survey.

2.2.1. Observation layer

The observation layer consists of a compound Poisson process. Let us define N_s, the unknown number of patches of organisms sampled at a site s, which results from a possibly non-homogeneous Poisson process:

$$N_s \sim \text{Poisson}(E_s \mu_s) \qquad \forall \ \ s \in \{1, \ldots, S\} \tag{1}$$

where μ_s is the locally expected number of patches and E_s is the sampling effort at site s. Every sampled patch i is defined as containing an unknown random quantity of biomass $M_{s,i}$. We assume that all the marks $M_{s,i}$ are independent and identically Gamma distributed with scale and rate parameters a and b:

$$M_{s,i} \sim \text{Gamma}(a, b). \tag{2}$$

The average biomass of a patch is a/b. Note that when $a = 1$, $M_{s,i}$ is exponentially distributed, which corresponds to the observation model used by Ancelet et al. (2010).

The quantities N_s and $M_{s,i}$ can be interpreted heuristically in terms familiar to ecologists and allow modeling of observed biomass at a location s, denoted Y_s:

$$Y_s = \begin{cases} \sum_{i=1}^{N_s} M_{s,i} & \text{if } N_s > 0 \\ 0 & \text{if } N_s = 0 \end{cases} \tag{3}$$

If there is at least one patch at the sampling site, the observed biomass Y_s is the random sum of the existing biomass in each patch. Conversely, if there are no patches at the sampled site, then nothing is caught. The main quantities of interest of the model are summarized in Table 1.

2.2.2. Spatial distribution layer

The spatial coherence of biomass distribution relies on the spatial distribution of the covariates and some unexplained latent spatial structure. In the following, we first describe how covariates are introduced to the model, and then how additional spatial structure is included.

2.2.2.1. Covariate effects. Bottom temperature, depth and sediment type are selected as available covariates related to habitat that potentially explain the distribution of the invertebrates. Their effects on μ_s, the average number of patches at site s, are included in the model through a logarithm link function. The full specification of μ_s is then given by:

$$\log(\mu_s) = \alpha_0 + \beta_{Sed_s} + \gamma_{Depth_s} + \zeta_{Temp_s} + \epsilon_s \tag{4}$$

where β_{Sed_s}, γ_{Depth_s} and ζ_{Temp_s} are the site specific effects of sediment type, depth and temperature, respectively, and ϵ_s is a Gaussian noise that accounts for potentially spatially structured process error. Four classes of sediments, based on the granulometry, are distinguished in the model: pelite, fine sand, coarse sand, gravel with occasional sand patches. Depth (in meters) is split into

Table 1
Quantities of interest for the proposed model.

Probability of presence	$1 - \exp^{(-\mu_s)}$
Expected positive biomass	$\left(\frac{\mu_s a}{b}\right)\left(\frac{1}{1-\exp^{(-\mu_s)}}\right)$
Expected biomass	$\frac{(\mu_s a)}{b}$
Variance of the biomass	$\left(\frac{\mu_s a}{b}\right)\left(\frac{a+1}{b}\right)$

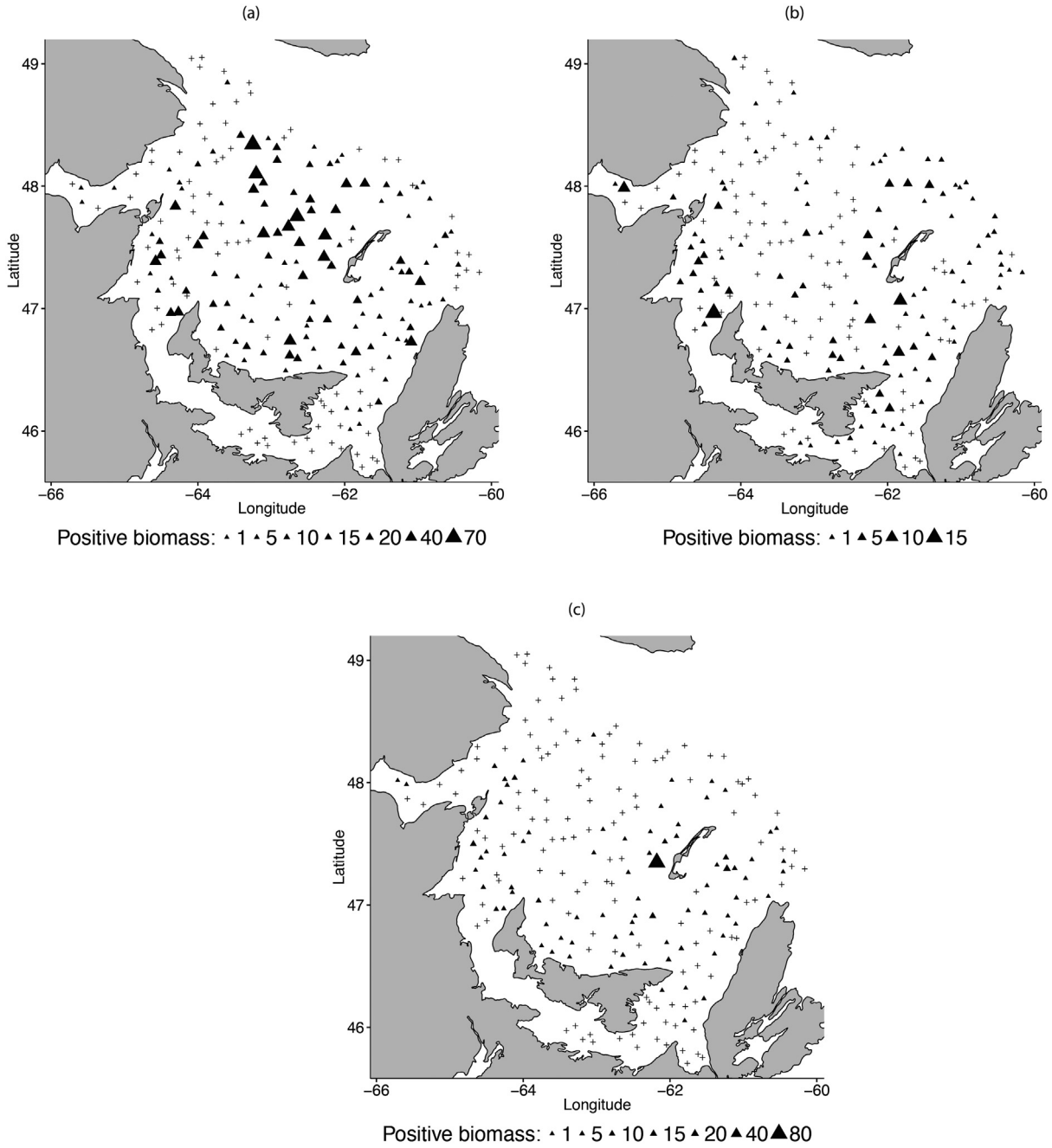

Fig. 2. Spatial distribution of macro-invertebrate biomass from individual tows in the 1997 sGSL bottom-trawl survey. Triangles represent sites where positive biomass was collected. Their widths are proportional to the biomass sampled (in kg/tow). Crosses represents sites where no biomass was caught: (a) urchin, (b) starfish, and (c) sea cucumber.

four classes to account for a possible non-linear response: [0, 50[; [50, 100[; and, [100, 400[. Likewise, bottom temperature (in °C) is split into three classes: [− 1, 1[; [1, 5[; and, [5, 15[. For the purpose of identifiability, some covariate classes have to be chosen as references. The most prevalent class for each covariate is defined as the baseline effect: fine sand for sediment type, [50, 100[for depth and [− 1, 1[for temperature.

2.2.2.2. Spatial effects. Covariates are often able to capture a part of the spatial dimension of a species' distribution, but they are unlikely to fully explain the spatial distribution. Even after accounting for the deterministic effect of environmental covariates, nearby locations are much more likely to resemble each other compared to more distant stations. In order to account for this residual spatial structure, spatially correlated errors are included. This approach has the benefit of improving inferences on the covariates by accounting for correlation and is very useful for creating interpolated maps of biomass in the study area (Lichstein et al., 2002). In practice, the random spatial process is included as a spatial Gaussian noise, $(\epsilon_s)_{s=1,\ldots,S}$, already defined in Eq. (4). The simple exponential covariance function, known to be robust in environmental applications, is used:

$$Cov(\epsilon_s, \epsilon_{s'}) = \sigma^2 \left(\exp\left(-\frac{h}{\Phi} \right) \right) \qquad \text{with} \quad h = d(s, s'). \tag{5}$$

It describes a decreasing exponential neighborhood covariation with increasing distance h between two sites s and s'. Φ is the

parametric range, which controls the rate of correlation decline as a function of distance, and σ^2 is the variance parameter.

Modeling the effect of the covariates and the spatial correlation could also have been done on the expected biomass in a patch (a/b). The inference of expected number of patch μ and the expected biomass in one patch a/b produces highly correlated estimates (Ancelet et al., 2010). Therefore, it is preferable to include covariates in only one of these quantities. Because the expected number of patches controls species presence as well as abundance, we choose to add both covariates and spatial effects to this latent layer.

2.3. Bayesian inference

Hierarchical models such as the one proposed in this paper have been developed under the Bayesian paradigm (Gelman et al., 2004). Bayesian analysis requires setting prior distributions for all the parameters (*i.e.* $a, b, \alpha_0, \zeta, \beta, \gamma, \sigma^2$, and Φ). We choose standard flat priors for the regression parameters ($\alpha_0, \zeta, \beta, \gamma$) because the number of patches and the effect of the three covariates are not known *a priori*. We choose a uniform prior distribution for the standard deviation σ as recommended by Gelman (2006).

The range parameter Φ is known to be difficult to estimate in hierarchical models (Cressie, 1993; Stein, 1999; Zhang, 2004; Zhang and Wang, 2009). We therefore devise a strategy based on data from other years for direct standard estimation. It is assumed that the latent spatial structure of the organisms changes little between neighboring years because of their limited dispersal abilities over large spatial scales and their slow biological turnover rates. For that matter, preliminary analyses involving data from 1996, 1997 and 1998 confirmed that the constant latent spatial structure assumption is reasonable (unpublished results). Therefore inference is conducted on the data sampled in 1996 by first fitting a model including all the covariates but no latent spatial structure and then applying a classical spatial analysis (kriging) to the resulting residuals. The range estimate is then plugged into the hierarchical analysis for 1997.

We choose a sufficiently vague prior for the expected biomass in a patch ($\mathbb{E}(M) = a/b$), with 5%, 50% and 95% quantiles that are respectively: 0.005, 1, and 135 kg. This prior distribution allows the expected biomass of one patch to take realistic values, which helps remove some of the possible confounding between a situation with a high number of small patches each containing a small amount of biomass and a low number of high biomass patches.

Eqs. (4) and 5 describe the full process model, but variants of this full model, with subsets of the explanatory covariates, are also considered as candidate alternatives (described below). Each model run is implemented in OpenBUGS, the open source version of WinBUGS (Ntzoufras, 2011) and its add-on GeoBUGS. The code for the model is presented in Appendix A. Previous tries with a reduced set of simulated data (to make sure that the inference algorithm worked) showed a strong autocorrelation of the MCMC iterations but the Gelman–Rubin convergence test became acceptable after 30,000 iterations of each of the three MCMC chains. Due to computational costs (about 3 days for each model), only two chains are launched for 60,000 iterations with a burn-in period of 30,000 iterations and MCMC convergence checked by visual inspection. A thinning of 100 iterations is performed in order to get rid of within-chain autocorrelation. Prediction and validation steps are computed in R 2.15.0 with the geoR package for the spatial correlation (Diggle and Ribeiro, 2001).

2.4. Validation and model selection

2.4.1. Posterior predictive checking

Posterior predictive checking is used to evaluate the model's ability to fit the observed data as recommended by Gelman et al. (1996). N draws $(\theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(N)})$ are sampled from the posterior distribution for vector $\theta = (a, b, \alpha_0, \beta, \gamma, \zeta, \sigma^2, \Phi)$. For each draw $\theta^{(i)}$, artificial replicated data $(\hat{Y}_s^{(i)})_{s=1,\ldots,S}$ are generated using the model $Y_s|\theta^{(i)}$. When the model fits well the data, the replicated data $\hat{Y}_s$ are expected to be close to the observed data Y_s. The discrepancy between $\hat{Y}_s$ and Y_s can be calculated using the omnibus statistics denoted $T(Y, \theta)$ as suggested in Gelman et al. (1996). In that case, the chosen T statistic is a Bayesian residual sum of squares, which is asymptotically distributed as a χ^2 distribution Gelman et al. (1996). Model fit is assessed by comparing the observed $T(Y, \theta^{(i)})$ with the $T(Y, \theta^{(i)})$ distributions. A Bayesian p-value is then computed to account for the cases in which the $T(\hat{Y}^{(i)}, \theta^{(i)})$ statistic exceeds the $T(Y, \theta^{(i)})$. Bayesian p-value close to 0.5 indicate a well fitted model.

$$T(Y, \theta^{(i)}) = \sum_{s=1}^{S} \left\{ \frac{(Y_s - \mathbb{E}(Y_s|\theta^{(i)}))^2}{\mathbb{V}(Y_s|\theta^{(i)})} \right\} = \sum_{s=1}^{S} \left\{ \frac{(Y_s - \frac{a^{(i)}\mu_s^{(i)}}{b^{(i)}})^2}{\frac{\mu_s^{(i)}a^{(i)}}{b^{(i)}}\frac{a^{(i)}+1}{b^{(i)}}} \right\}$$
$$T(\hat{Y}^{(i)}, \theta^{(i)}) = \sum_{s=1}^{S} \left\{ \frac{(\hat{Y}_s^{(i)} - \mathbb{E}(\hat{Y}_s^{(i)}|\theta^{(i)}))^2}{\mathbb{V}(\hat{Y}_s^{(i)}|\theta^{(i)})} \right\} = \sum_{s=1}^{S} \left\{ \frac{(\hat{Y}_s^{(i)} - \frac{a^{(i)}\mu_s^{(i)}}{b^{(i)}})^2}{\frac{\mu_s^{(i)}a^{(i)}}{b^{(i)}}\frac{a^{(i)}+1}{b^{(i)}}} \right\} \quad (6)$$

2.4.2. Model comparison

Many competing submodels of Eq. (4) can be defined based on combinations of the covariates and the spatially correlated errors. The full model (M_{s1}) includes all covariates (depth, temperature and sediment type) and the spatial correlation. For each species, this full model is compared with several submodels summarized in Table 2. Two criteria are used to compare the models. The first is the Bayesian Information Criterion (*BIC*) proposed by Schwarz (1978):

$$BIC = -2 \times \log(L) + k \times \log(n) \quad (7)$$

where n is the number of observations, k is the number of parameters to be estimated in the submodel and L is the maximized value of the likelihood function for the estimated submodel. The *BIC* measures how well the model fits the data, with respect to the number of model parameters.

The second comparison criterion, the mean squared prediction error (*MSPE*) is used to assess the accuracy of the model predictions:

$$MSPE_S = \sum_{s=1}^{S} \mathbb{E}((\hat{Y}_s - Y_s)^2) = \sum_{s=1}^{S} \{(\mathbb{E}(\hat{Y}_s) - Y_s)^2 + \mathbb{V}(\hat{Y}_s)\}. \quad (8)$$

This criterion takes into account the bias of the prediction relative to the true value, and includes a term for the predictive variance. Here, we consider the spatial average of this criterion over the prediction dataset.

2.5. Predictions

The main advantage of models such as the ones proposed in this paper is the capability of making predictions, Y_{new}, conditional on the observations, while preserving the inferred spatial structure. The predictive distribution of the biomass quantity is given by:

$$[Y_{new}|Y_{obs}] = \iint [Y_{new}, \theta, \mu|Y_{obs}] d\mu \, d\theta. \quad (9)$$

In practice, for each iteration i of the MCMC chains, we perform a conditional simulation $[\epsilon_{new}^{(i)}|\sigma^{2^{(i)}}, \Phi, \epsilon^{(i)}]$ to obtain by kriging a realization of the latent Gaussian field $\epsilon_{new}^{(i)}$ at the sites where predictions are wanted. Prediction for a new site s_0 requires knowledge of the values of the covariates at this site. These are obtained using a linear interpolation of values at neighboring sites for the

Table 2
Alternate submodels for the spatial hierarchical model and results of the model fit for the three taxa based on the model selection criteria BIC and MSPE, and the Bayesian *p*-value model checking criterion. *s* denotes a model with spatial correlation and a lack of *s* means no spatial correlation. M_{s1} is the more complex model, which includes all covariates: sediment type (*Sed*), depth (*Dep*) and bottom temperature (*Temp*) and a spatial correlation. Model M_1 include all covariates without spatial structure.

	Covariates			*BIC*	B. *p*-value	*MSPE*
Model	*Sed*	*Dep*	*Temp*			
Strongylocentrotus droebachiensis						
M_{s1}	×	×	×	844.41	0.51	38.41
M_{s2}	×	×		888.76	0.46	53.13
M_{s3}	×		×	853.28	0.43	38.69
M_{s4}	×			874.16	0.41	42.45
M_{s5}		×	×	927.55	0.47	74.187
M_{s6}		×		961.34	0.49	91.10
M_{s7}			×	999.78	0.54	122.67
M_{s8}				1078.47	0.53	225.21
M_1	×	×	×	1022.67	0.54	160.91
Asterias sp						
M_1	×	×	×	317.76	0.48	0.69
M_2	×	×		343.33	0.45	0.71
M_3	×		×	338.61	0.45	0.73
M_4	×			365.63	0.41	0.82
M_5		×	×	392.24	0.49	1.54
M_6		×		399.77	0.55	1.36
M_7			×	461.06	0.41	2.76
M_8				412.38	0.53	2.56
Cucumaria frondosa						
M_{s1}	×	×	×	215.38	0.54	0.11
M_{s2}	×	×		270.12	0.55	0.12
M_{s3}	×		×	244.05	0.46	0.15
M_{s4}	×			263.09	0.44	0.17
M_{s5}		×	×	244.80	0.52	0.14
M_{s6}		×		270.12	0.42	0.18
M_{s7}			×	282.09	0.41	0.22
M_{s8}				305.34	0.56	0.26
M_1	×	×	×	482.19	0.57	5.65

temperature and depths, while sediment type is obtained from an interpolated map from the study of Loring and Nota (1973). Then, the latent layer $\mu_{new}^{(i)}$ is generated to account for the effects of the included covariates. The biomass of the studied species in unsampled locations, $Y_{new}^{(i)}$, is then merely drawn from the observation model sub-component. These posterior predictive joint distributions can be summarized by maps showing various statistics of the species distribution (e.g. mean, or median biomass, or the proportion of zeros). Given the assumption of little movement between successive years, we rely on data from the 1998 survey (not used for model fitting) to evaluate the predictive ability of the competing models.

3. Results

The results are presented by species, following a common presentation format.

1. Results of analyses to establish whether there is spatial structure in the residuals, and in the affirmative case, estimating the range parameter using data from the 1996 survey.
2. The model is fitted to the 1997 data under the Bayesian paradigm and its predictive ability is checked using the 1998 data.
3. Results of submodel comparisons and the implied effect of covariates are presented.

3.1. Green sea urchin

Spatial structure is apparent in the residuals of the inference performed on the green sea urchin biomass sampled in 1996 (Fig. 3a). The estimated value for the parametric range is $\hat{\Phi} = 23$ km. This range is considerably smaller than the average inter-station distance in the annual survey of the sGSL, 156 km. *BIC* and *MSPE* scores are considerably smaller for the full spatialized model, M_{s1}, compared to a similar model without spatial correlation, M_1 (Table 2), indicating that adding a spatial structure improves model fit. The evaluation of competing models with different subsets of covariates is therefore limited to models that include spatial correlation. Detailed results of submodel comparisons for the green sea urchin are provided in Table 2.

The validation of the models by posterior predictive checking gave acceptable results for all models, with the *p*-values around 0.5. Based on *BIC*, the best fitting model is the one that includes all three covariates. This model also has the best *MSPE*. The different effects of the covariates included in this model are presented by their posterior distributions in Fig. 4. Sediment type has an important effect on the biomass of the green sea urchin, with pelite having a negative effect and both coarse sand

Table 3
Proportion of predictions of the best model for the three species (urchin: M_{s1}, starfish: M_1, cucumber: M_{s1}) for the biomass sampled in 1998.

	Predictions			
	True zero	False positive	True positive	False zero
Strongylocentrotus droebachiensis	0.71	0.29	0.76	0.24
Asterias sp	0.55	0.45	0.51	0.49
Cucumaria frondosa	0.79	0.21	0.52	0.48

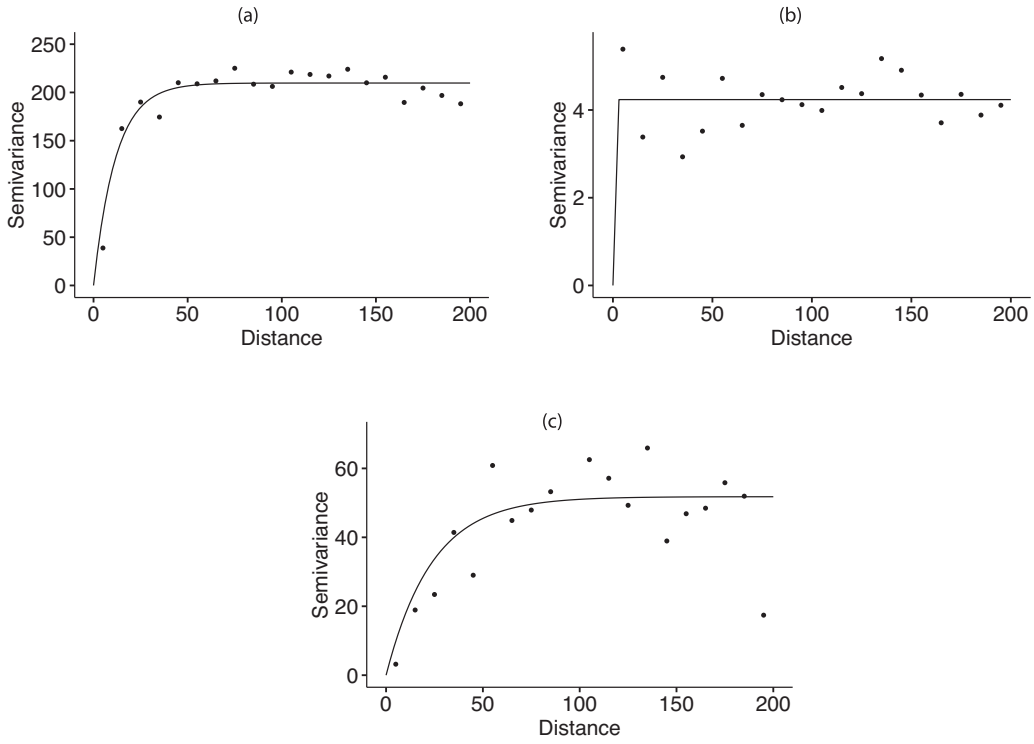

Fig. 3. Variogram for the biomass sampled in 1996 obtained from the residuals of the model including all covariates without spatial correlation for the three species: (a) green sea urchin, (b) starfish, and (c) sea cucumber. Distance in kilometers.

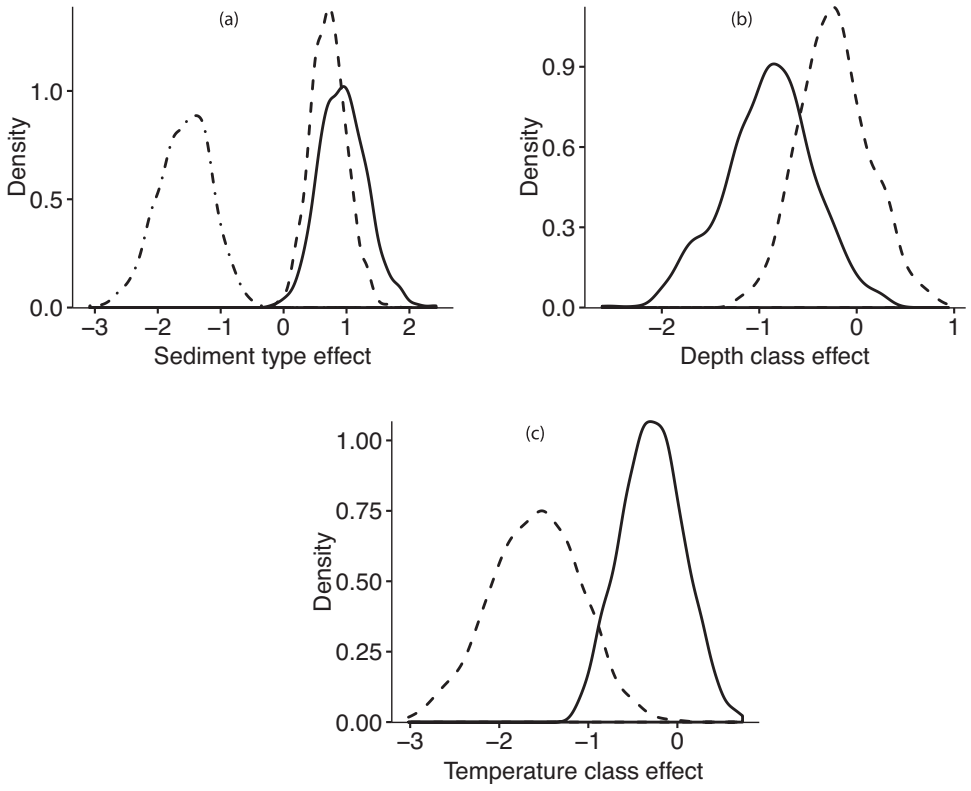

Fig. 4. Posterior distributions of the parameters included in the model M_{s1} for green sea urchin (a) sediment type, gravel (solid line), coarse sand (dashed line) and pelite (dot-dashed line). (b) depth class, [0, 50[(dotted line), [100, 400[(solid line). (c) temperature, [1, 5[(solid line), [5, 15[(dotted line).

and gravel having a positive effect relative to fine sand (Fig. 4a). Depth also appears to be important with large depths having a negative effect on the urchin biomass (Fig. 4b). The effect of bottom temperature is negative for both classes, [1, 5[and [5, 15[, relative to the effect of the lower temperature class. The negative effect is strongest for the warmest temperature class (Fig. 4c).

The majority of the 1998 biomass records are predicted well, though there are some misclassifications with the positive biomass (Table 3). Qualitatively, the interpolated map of the median biomass quantities predicted with Model M_{s1} matches well the survey observations (Fig. 5). This interpolation allows a good identification of areas with a high quantity of biomass as well as areas without biomass. Note that urchins are widely distributed in the sGSL.

3.2. Starfish

No spatial structure is detected in the residuals of the model fitted to the starfish data in 1996 (Fig. 3b). The favored model, based on the two model selection criteria includes the effects of the sediment type, depth and temperature classes (Table 2; Fig. 6).

Pelite type sediment has a negative effect and the gravel type has a positive effect on starfish biomass, relative to fine sand (Fig. 6a). The effect of depth is manifested by a small positive effect of the shallow depths relative to intermediate depths (Fig. 6b). Temperature classes [1, 5[and [5, 15[have a positive effect on the starfish biomass relative to temperatures between −1 and 1 °C (Fig. 6c). The model is not able to provide good predictions of starfish biomass sampled in 1998 (Table 3).

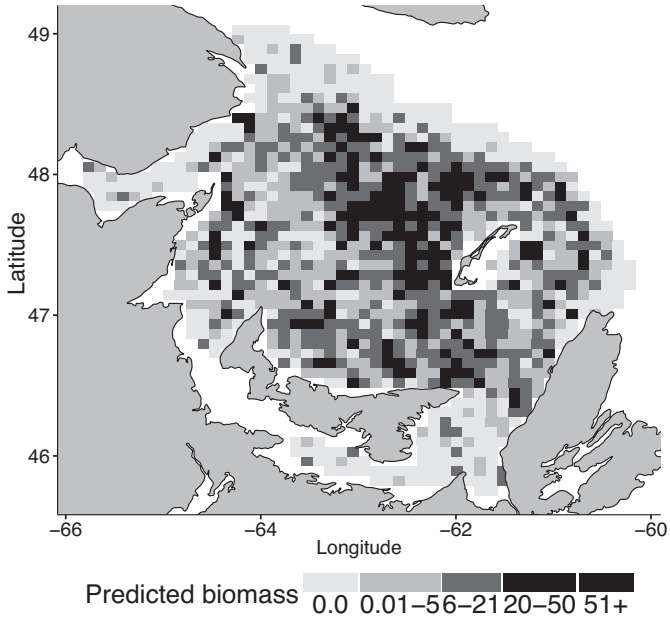

Fig. 5. Prediction of the median quantity of green sea urchin biomass (in kg per standard tow) on a grid in the sGSL.

3.3. Sea cucumber

Some spatial structure is detected in the 1996 distribution of sea cucumber biomass (Fig. 3c). As for the urchins, the estimated parametric range $\hat{\Phi} = 22$ km, is considerably smaller than the average interstation range in the survey. *BIC* and *MSPE* scores confirm that

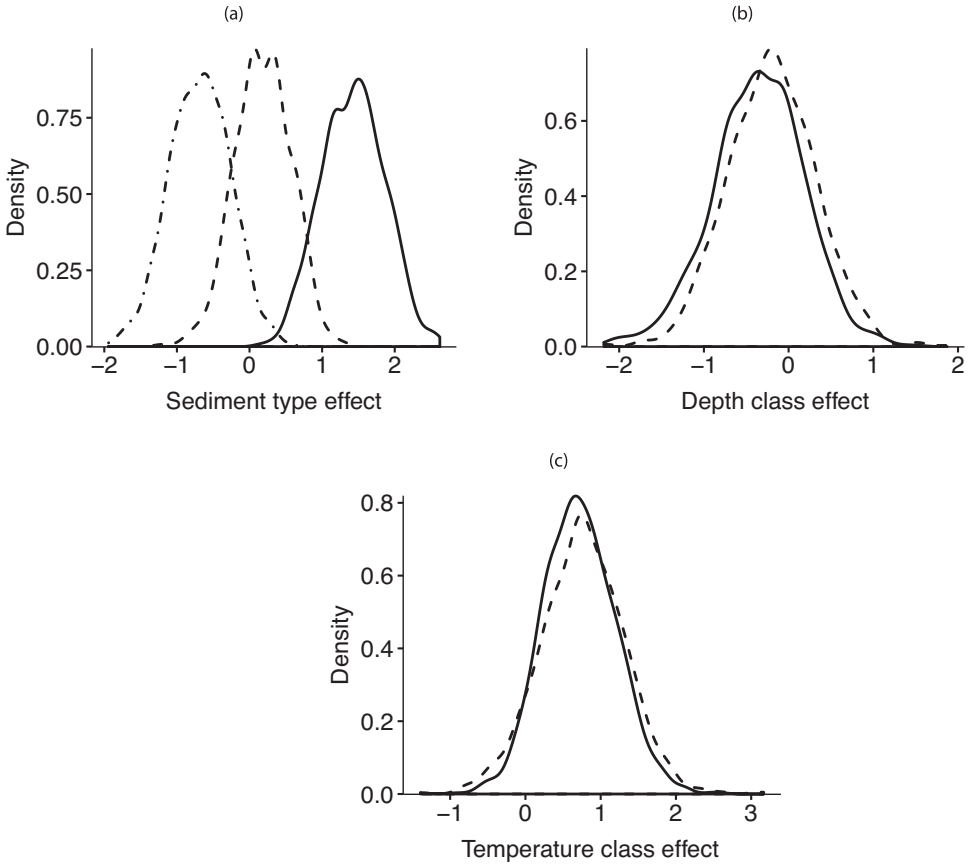

Fig. 6. Posterior distributions of the parameters included in the model M_1 for starfish (a) sediment type, gravel (solid line), coarse sand (dashed line) and pelite (dot-dashed line). (b) depth class, [0, 50[(dotted line), [100, 400[(solid line). (c) temperature, [1, 5[(solid line), [5, 15[(dotted line).

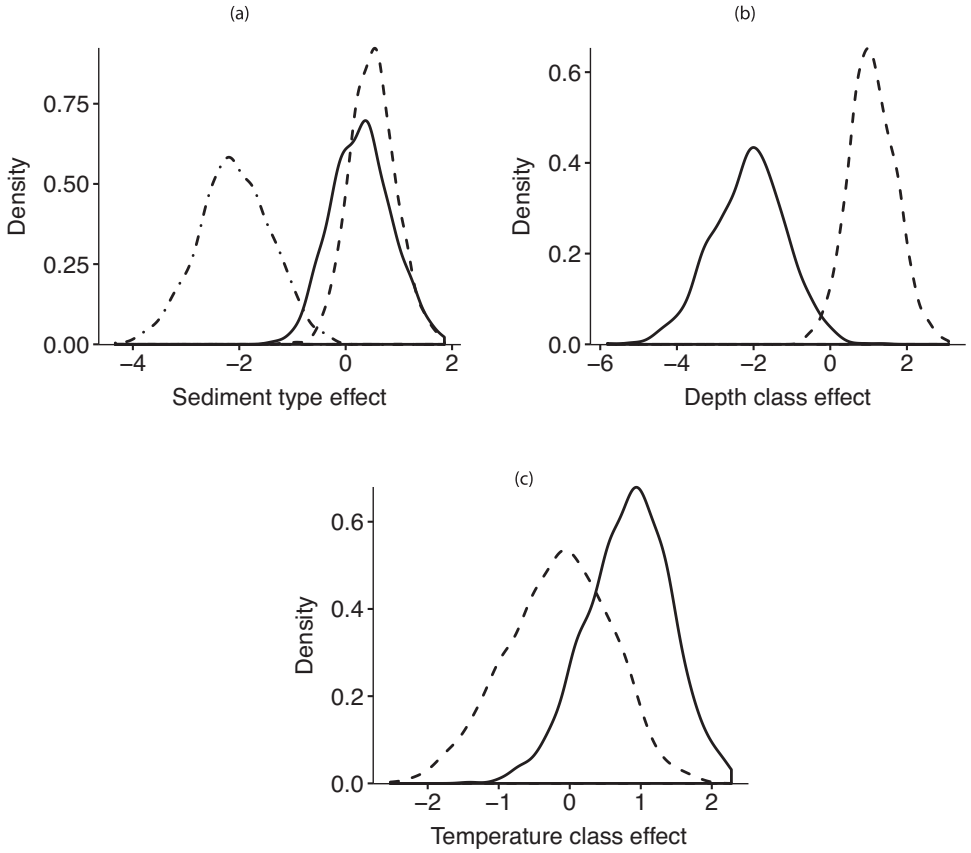

Fig. 7. Posterior distributions of the parameters included in the model M_{s1} for sea cucumber (a) sediment type, gravel (solid line), coarse sand (dashed line) and pelite (dot-dashed line). (b) depth class, [0, 50[(dotted line), [100, 400[(solid line). (c) temperature, [1, 5[(solid line), [5, 15[(dotted line).

a model including spatially correlated error is preferred (Table 3). The analysis of competing models reveals that a model including all three covariates is favored (Table 3). Relative to fine sand, pelite sediment has a negative effect on sea cucumber biomass, while coarse sand and gravel has a positive effect (Fig. 7a). Sea cucumber biomass is predicted to be greater in shallow depths, and lower in deeper areas, relative to intermediate depths (Fig. 7b). The effect of temperatures is positive for the [1, 5[class and null for the [5, 15[class, compared to the coldest temperature class. (Fig. 7c).

The sites without sea cucumber in 1998 are well predicted by the model (Table 3). However, the model predicts a high number of false positive values. Fig. 8 represents the median quantity of cucumber biomass predicted by Model M_{s1}. This map allows the identification of a small area of high biomass in the center of the area near the Magdalen Islands archipelago and points out large areas without sea cucumbers in the northern portion of the survey area where depth declines rapidly into a large channel, the Laurentian channel.

4. Discussion

In this work, we propose a Bayesian hierarchical model to handle spatialized continuous zero-inflated data. We expand the model proposed by Ancelet et al. (2010) by considering Gamma marks instead of Exponential ones, which allows more flexibility in the modeling approach. We also add locally recorded covariates to the model in order to describe the habitat associations as well as complex spatial dependence structures.

The distributions of the three species appear to be affected by the sediment type and depth, which is expected given that both are key variables affecting the distribution of epibenthic invertebrates in other marine ecosystems (Freeman and Rogers, 2003; Hily et al., 2008). The species are also affected by temperature, a variable known to rule the distribution of marine biota. Noticeable residual spatial variation remains for urchins and sea cucumbers, but not for starfish. Model performance for starfish was weaker than for the other two species. A probable explanation is that contrary to urchins and sea cucumbers, the starfish data represent catches for an assemblage of species, each with potentially unique habitat

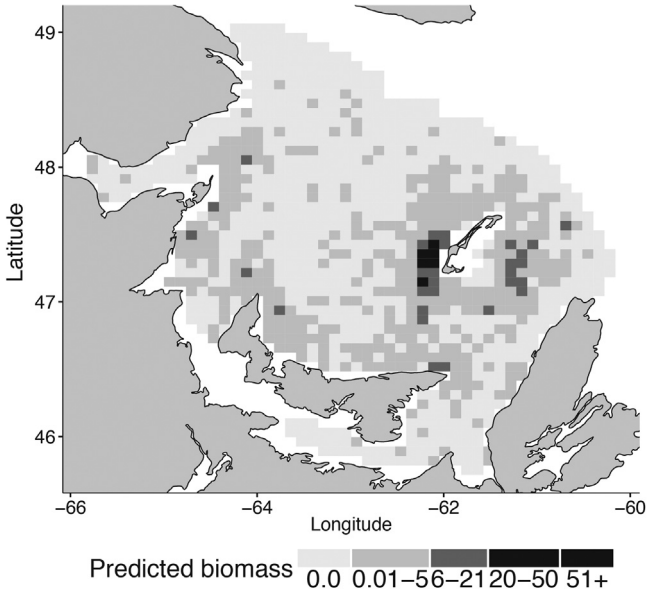

Fig. 8. Prediction of the median quantity of sea cucumber biomass (in *kg* per standard tow) on a grid in the sGSL.

associations. Modeling the assemblage as a unit effectively means modeling a weighted average habitat association, which may not be strong and which will change as the relative abundance of the component species changes. This is likely to lead to a weakened ability to predict distribution patterns over time based on habitat covariates. Furthermore, the performance of this model will be influenced by sampling stochasticity given that the data from any one year of the survey represent a single sampling realization that will only on average reflect the true patterns in distribution and habitat associations even if the model is correct. Such an effect may explain some of the deficiencies in model performance observed for all three invertebrate taxa.

The small estimated values for the range parameters are consistent with the limited dispersal abilities of the organisms considered here. For example, the maximum displacement of green sea urchin is estimated at 20 cm per day *i.e.* approximately 17 km per year (Lauzon-Guay and Scheibling, 2007). Possible explanations for this small scale effect are small scale habitat features, other habitat features not included in the model, or a result of demographic or behavioral characteristics of the organisms. This small correlation distance helps to capture a spatial structure at small scales, which is not otherwise related to the deterministic effect of covariates that vary over much broader scales. Taking into account this spatial correlation greatly improves the fitting and the predictive capacity of the model. However, the model still misclassifies some predictions of the biomass sampled in 1998. These errors could be the result of changes in the hydrodynamic and physico-chemical properties of the water column, which impacts macroinvertebrates over short temporal scales relative to depth or sediment types (Warwick and Uncles, 1980; Ysebaert and Herman, 2002; Freeman and Rogers, 2003; Bolam et al., 2008). Another possible explanation is that the range parameter ϕ is not well estimated because it is based on a single realization of the survey and because the survey sampling density is considerably coarser than the apparent range (Zimmerman, 2006; Irvine et al., 2007). Additional surveys with finer grain sampling might help in better defining the range parameter.

An attractive feature of the approach is its flexibility. For example, instead of a Poisson distribution for the number of patches, it is possible to use the Negative Binomial distribution, a Gamma mixture of Poisson. This alternative extension would allow for over-dispersion of the number of patches in the observation submodel, resulting in an over-dispersed quantity of biomass. A GLM with a logarithm link function is used here to model the intensity of the Poisson distribution but it would also be possible to use a generalized additive model in more complex cases (Guisan et al., 2002; Zuur et al., 2009).

The two latent quantities (number of patches and biomass quantity in each patch) must not be over-interpreted. These two conceptual quantities reflect a heuristic construction of a zero-inflated model but their direct ecological interpretation cannot be confirmed by observation. In addition, the two hidden quantities μ and a/b are highly correlated (Ancelet et al., 2010). While this would hamper the interpretation of their individual values, it is not an issue in the application presented here because we are interested in their joint effect, *i.e.* the predicted biomass. By linking μ instead of a/b to the covariates, covariates play a role in both the probability of presence and the quantity of biomass, also partially controlled by the number of patches.

Interpolation relying on covariates and spatial structure exploits a large amount of information from the data collected during marine surveys. Maps predicting ecological properties such as the mean biomass, areas of high density, or the presence–absence of organisms can be easily produced, together with their uncertainties. Such maps are commonly used for ecological analyses and data-based approaches (Shea, 1998; Hilborn et al., 2004; Hobday and Hartmann, 2006; Hartog et al., 2011; Dutertre et al., 2012). Furthermore, the approach provides preliminary answers with probabilistic predictions for the quantitative effect of long-term changes in important habitat variables, for example as might be expected under global warming.

Appendix A. Model code

```
model{
### Latent Layers
for (s in 1:nsample){
  log(mu[s])       <       - alpha0 + beta2[Tp[s]] +
beta1[Sed[s]] + beta3[Pr[s]] + w[s]
  w[s] <- v[s] - mean(v[1:nsample])
  mupres[s] <- mu[s]*(dtow[s]/dstandard)
}
v[1:nsample]      ~      spatial.exp(mu_v[],x[],y[],
tau_v,phi_v,kappa_v)
### Observation Model
### Strictly positive biomass data
for(k in 1:npres){
  ### Number of patches
  ngis[k] ~ dpois(mupres[pres[k]])C(1,)
  ### Observed positive biomass
  Y_a[k] <-a*ngis[k]
  Y[pres[k]] ~ dgamma(Y_a[k],b)
}
### Evaluation of probabilities of zeros
for(j in 1:nabs){
  ### Probability of presence at site j
  proba[j] <- 1 - exp(-mupres[abs[j]])
  Y[abs[j]] ~ dbern(proba[j])
}
### Priors
alpha0 ~ dnorm(0,0.01)
beta1[2] <- 0
beta1[1] ~ dnorm(0,0.01)
beta1[3] ~ dnorm(0,0.01)
beta1[4] ~ dnorm(0,0.01)
beta3[2] <- 0
beta3[1] ~ dnorm(0,0.01)
beta3[3] ~ dnorm(0,0.01)
beta2[1] <- 0
beta2[2] ~ dnorm(0,0.01)
beta2[3] ~ dnorm(0,0.01)
a ~ dgamma(2,5)
b ~ dgamma(2,5)
sigma ~ dunif(0,100)
sigma2 <-sigma*sigma
tau_v <-1/(sigma2)
   }
```

References

Ancelet, S., Etienne, M.-P., Benoît, H.P., Parent, E., 2010. Modelling spatial zero-inflated continuous data with an exponentially compound Poisson process? Environmental and Ecological Statistics 17 (3), 347–376.

Benoît, H.P., Swain, D.P., Chouinard, G.A., 2009. Using the long-term bottom-trawl survey of the southern Gulf of St. Lawrence to understand marine fish populations and community change. AZMP Bulletin 8, 19L 27.

Bernier, J., Fandeux, D., 1970. Théorie du renouvellement application à l'étude statistique des précipitations mensuelles? Revue de Statistiques Appliquées 18 (2), 75–87.

Bolam, S., Eggleton, J., Smith, R., Mason, C., Vanstaen, K., Rees, H., 2008. Spatial distribution of macrofaunal assemblages along the English Channel. Journal of the Marine Biological Association of the UK 88 (04), 675–687.

Chadwick, E., Brodie, W., Colbourne, E., Clark, D., Gascon, D., Hurlbut, T., 2007. History of annual multi-species trawl surveys on the Atlantic coast of Canada. Atlantic Zonal Monitoring Program Bulletin 6, 25L 42.

Cook, A., Marion, G., Butler, A., Gibson, G., 2007. Bayesian inference for the spatio-temporal invasion of alien species? Bulletin of Mathematical Biology 69 (6), 2005–2025.
Cressie, N., 1993. Statistics for Spatial Data, revised edition. New York, Wiley.
Diggle, P., Ribeiro, P., 2001. geoR: a package for geostatistical analysis? R news 1 (2), 14–18.
Dutertre, M., Hamon, D., Chevalier, C., Ehrhold, A., 2012. The use of the relationships between environmental factors and benthic macrofaunal distribution in the establishment of a baseline for coastal management? ICES Journal of Marine Science 70 (2), 294–308.
Engler, R., Guisan, A., Rechsteiner, L., 2004. An improved approach for predicting the distribution of rare and endangered species from occurrence and pseudo-absence data? Journal of Applied Ecology 41 (2), 263–274.
Fletcher, D., MacKenzie, D., Villouta, E., 2005. Modelling skewed data with many zeros: a simple approach combining ordinary and logistic regression? Environmental and Ecological Statistics 12 (1), 45–54.
Freeman, S., Rogers, S., 2003. A new analytical approach to the characterisation of macro-epibenthic habitats: linking species to the environment. Estuarine. Coastal and Shelf Science 56 (3-4), 749–764.
Gelman, A., 2006. Prior distributions for variance parameters in hierarchical models (Comment on Article by Browne and Draper). Bayesian Analysis 1, 515L 534.
Gelman, A., Carlin, J., Stern, H., Rubin, D., 2004. Bayesian Data Analysis. CRC press.
Gelman, A., Meng, X.-l., Stern, H., 1996. Posterior predictive assessment of model fitness via realized discrepancies. Statistica Sinica 6, 733L 807.
Guisan, A., Edwards, T.C., Hastie, T., 2002. Generalized linear and generalized additive models in studies of species distributions: setting the scene? Ecological Modelling 157 (2-3), 89–100.
Guisan, A., Thuiller, W., 2005. Predicting species distribution: offering more than simple habitat models? Ecology Letters 8 (9), 993–1009.
Hartog, J.R., Hobday, A.J., Matear, R., Feng, M., 2011. Habitat overlap between southern bluefin tuna and yellowfin tuna in the east coast longline fishery-implications for present and future spatial management? Deep Sea Research Part II: Topical Studies in Oceanography 58 (5), 746–752.
Hilborn, R., Stokes, K., Maguire, J.-J., Smith, T., Botsford, L.W., Mangel, M., Orensanz, J., Parma, A., Rice, J., Bell, J., Cochrane, K.L., Garcia, S., Hall, S.J., Kirkwood, G., Sainsbury, K., Stefansson, G., Walters, C., 2004. When can marine reserves improve fisheries management? Ocean & Coastal Management 47 (3–4), 197–205.
Hily, C., Le Loc'h, F., Grall, J., Glémarec, M., 2008. Soft bottom macrobenthic communities of North Biscay revisited: long-term evolution under fisheries-climate forcing. Estuarine, Coastal and Shelf Science 78 (2), 413–425.
Hobday, A.J., Hartmann, K., 2006. Near real-time spatial management based on habitat predictions for a longline bycatch species? Fisheries Management and Ecology 13 (6), 365–380.
Irvine, K.M., Gitelman, A.I., Hoeting, J.A., 2007. Spatial designs and properties of spatial correlation: effects on covariance estimation. Journal of Agricultural, Biological, and Environmental Statistics 12 (4), 450–469.
Lauzon-Guay, J.-S., Scheibling, R.E., 2007. Seasonal variation in movement, aggregation and destructive grazing of the green sea urchin (*Strongylocentrotus droebachiensis*) in relation to wave action and sea temperature. Marine Biology 151 (6), 2109–2118.
Lichstein, J.W., Simons, T.R., Shriner, S.A., Franzreb, K.E., 2002. Spatial autocorrelation and autoregressive models in ecology? Ecological Monographs 72 (3), 445–463.
Loring, D., Nota, D., 1973. Morphology and sediments of the Gulf of St. Lawrence. Fisheries and Marine Service, Ottawa.
Martin, T., Wintle, B., Rhodes, J., Kuhnert, P., Field, S., Low-Choy, S., Tyre, A., Possingham, H., 2005. Zero tolerance ecology: improving ecological inference by modelling the source of zero observations? Ecology Letters 8 (11), 1235–1246.
Maunder, M.N., Punt, A.E., 2004. Standardizing catch and effort data: a review of recent approaches? Fisheries Research 70 (2–3), 141–159.
Ntzoufras, I., 2011. Bayesian Modeling using WinBUGS, vol. 698. Wiley, Hoboken, NJ.
Perry, R.I., Smith, S.J., 1994. Identifying habitat associations of marine fishes using survey data: an application to the Northwest Atlantic? Canadian Journal of Fisheries and Aquatic Sciences 51 (3), 589–602.
Schwarz, G., 1978. Estimating the dimension of a model? The Annals of Statistics 6 (2), 461–464.
Shea, K., 1998. Management of populations in conservation, harvesting and control. Trends in Ecology & Evolution 13 (9), 371–375.
Shono, H., 2008. Application of the Tweedie distribution to zero-catch data in CPUE analysis? Fisheries Research 93 (1-2), 154–162.
Sileshi, G., Hailu, G., Nyadzi, G.I., 2009. Traditional occupancy-abundance models are inadequate for zero-inflated ecological count data? Ecological Modelling 220 (15), 1764–1775.
Stefansson, G., 1996. Analysis of groundfish survey abundance data: combining the GLM and delta approaches? ICES Journal of Marine Science 53 (3), 577–588.
Stein, M., 1999. Interpolation of Spatial Data: Some Theory for Kriging. Springer-Verlag, New York.
Warwick, R., Uncles, R., 1980. Distribution of benthic macrofauna associations in the Bristol channel in relation to tidal stress. Marine Ecology Progress Series 3, 97L 103.
Welsh, A., Cunningham, R., Donnelly, C., Lindenmayer, D., 1996. Modelling the abundance of rare species: statistical models for counts with extra zeros? Ecological Modelling 88 (1-3), 297–308.
Williams, A., Bax, N., 2001. Delineating fish-habitat associations for spatially based management: an example from the south-eastern Australian continental shelf. Marine and Freshwater Research 52 (4), 513–536.
Ysebaert, T., Herman, P., 2002. Spatial and temporal variation in benthic macrofauna and relationships with environmental variables in an estuarine, intertidal soft-sediment environment. Marine Ecology Progress Series 244, 105L 124.
Zhang, H., 2004. Inconsistent estimation and asymptotically equal interpolations in model-based geostatistics? Journal of the American Statistical Association 99 (465), 250–261.
Zhang, H., Wang, Y., 2009. Kriging and cross-validation for massive spatial data? Environmetrics 21 (3-4), 290–304.
Zimmerman, D.L., 2006. Optimal network design for spatial prediction, covariance parameter estimation, and empirical prediction. Environmetrics 17 (6), 635–652.
Zuur, A.F., Ieno, E.N., Walker, N., Saveliev, A.A., Smith, G.M., 2009. Mixed Effects Models and Extensions in Ecology with R. Statistics for Biology and Health. Springer, New York.

5

MODÉLISATION SPATIO-TEMPORELLE DE DONNÉES DE BIOMASSE GÉORÉFÉRENCÉES À FORTE PROPORTION DE ZÉROS PAR CONVOLUTION

Les études de distributions spatio-temporelles de populations ou d'espèces se sont développées ces dernières décennies. En effet, le besoin de prédire les changements de la répartition spatiale de populations face notamment aux changements climatiques a encouragé l'extension des modèles à structure spatiale vers des modèles spatio-temporels. De plus, la nécessité de prédire la dispersion d'espèces envahissantes a également encouragé le développement de méthodes capables d'appréhender ces problèmes (Hooten et Wikle, 2007; Cook *et al.*, 2007).

Les méthodes développées pour étudier la dispersion spatio-temporelle d'une espèce envahissante sur un territoire donné utilisent différents types de données. Par exemple, Cook *et al.* (2007) se sont intéressés à la colonisation de l'Angleterre par une espèce végétale envahissante à l'aide de données de présence-absence. De plus, comme proposé par Hooten et Wikle (2007), il est également possible d'utiliser des données de comptage pour l'étude de l'invasion d'une espèce de tourterelle aux États-Unis. Les données de comptage, courantes en écologie, ne sont pas restreintes à l'étude d'espèces envahissantes, par exemple Ver Hoef et Jansen (2007) ont étudié la distribution spatio-temporelle d'une espèce de phoque en Alaska.

Les trois études évoquées précédemment (Hooten et Wikle, 2007; Cook *et al.*, 2007; Ver Hoef et Jansen, 2007) utilisent des modèles hiérarchiques bayésiens (HBM) pour analyser et prédire la distribution spatio-temporelle des espèces considérées. Cependant, les choix de modélisation diffèrent entre ces trois études et illustrent la grande disponibilité de méthodes actuellement à disposition. L'approche proposée par Cook *et al.* (2007) nécessite le découpage de l'Angleterre en petites zones carrées

($10 \times 10\ km$) caractérisées par un habitat homogène. La colonisation d'un site est quant à elle supposée résulter d'un processus stochastique qui varie entre sites et dans le temps. Une fonction prenant en compte les variables environnementales et la distance au premier site occupé par l'espèce envahissante permet de construire un taux de colonisation pour chaque site. Le modèle utilise donc un noyau de dispersion associé à des variables explicatives caractérisant l'habitat pour prédire la dispersion de la plante invasive. L'approche développée par Hooten et Wikle (2007) est relativement différente, elle consiste à utiliser des équations de réaction-diffusions au sein de la couche latente d'un HBM pour prédire l'invasion des États-Unis par une espèce de tourterelle. Le modèle des observations est alors une loi de Poisson et permet de prendre en compte des variables environnementales comme la densité humaine, facteur déterminant dans la colonisation de nouveaux territoires par cette espèce. L'étude de Ver Hoef et Jansen (2007) ne s'intéresse pas à une espèce envahissante, mais repose également sur des comptages avec la particularité de présenter une forte proportion de zéros. Le modèle des observations du modèle bayésien hiérarchique est donc adapté à ce type de données. La partie temporelle du modèle est modélisée par un modèle autorégressif d'ordre 1 et constitue avec la partie spatiale la couche latente du HBM. L'aspect spatial de l'approche est modélisé à l'aide d'un modèle autorégressif conditionnel (CAR), également appelé modèle à champ de Markov Gaussien. Ces modèles sont couramment utilisés pour la modélisation de processus spatiaux (Besag et Kooperberg, 1995; Rue et Held, 2005). La dépendance spatiale est alors modélisée grâce aux caractéristiques markoviennes du champ : un site est corrélé à ses plus proches voisins selon un graphe non orienté. Afin de modéliser un phénomène spatial avec un modèle à champ de Markov Gaussien, il est nécessaire de définir une structure de voisinage et donc de diviser l'espace étudié en sous-unités géographiques. Cette approche peut donc conduire à des simplifications de la structure spatiale étudiée et nécessite un travail préalable d'identification de zones homogènes vis-à-vis de la quantité d'intérêt. Pour éviter les problèmes lié à cette simplification, il est possible d'utiliser les outils géostatistiques développés dans le chapitre 4 pour modéliser la distribution spatio-temporelle d'une espèce. Axis-Arroyo et Mateu (2004) ont proposé un HBM pour étudier la distribution spatio-temporelle du nombre d'espèces d'invertébrés épi-

benthiques au sud du golfe du Mexique en utilisant les outils de la géostatistique usuelle. Pour ce faire deux années sont modélisées (1996, 1999) mais chaque année est utilisée séparément. L'analyse temporelle de la modélisation est simplement effectuée en comparant les résultats de chaque modèle. Cette difficulté à utiliser les modèles géostatistiques au sein d'un HBM spatio-temporel est principalement due aux jeux de données utilisés pour cette modélisation. En effet, ces jeux de données présentent généralement un grand nombre d'observations (p. ex. 150 sites échantillonnés pendant 5 ans : $n = 150 * 5 = 750$ sites) et posent des problèmes computationnels lorsque les modèles géostatistiques sont utilisés. Ces problèmes d'inférences sont essentiellement liés à l'inversion d'une matrice de variance covariance, qui lorsque le nombre d'observations est important, rend l'inférence difficile (Zhu et Wu, 2010).

Higdon (1998, 2002) a proposé une approche alternative consistant à créer un processus gaussien dans $\mathbb{R}^2$ en générant plusieurs variables aléatoires gaussiennes *i.i.d.* sur une grille, elle aussi dans $\mathbb{R}^2$, puis en convoluant ces variables par un noyau de lissage préalablement choisi. Ainsi, les prédictions spatiales obtenues par cette méthode par convolution sont similaires à celles obtenues par les variogrammes (Higdon, 2002). De plus, l'approche par convolution permet une bien plus grande flexibilité que les modèles géostatistiques. Elle peut, par exemple, prendre en compte des effets de non-stationnarité de la dépendance spatiale ou des effets de bords. D'ailleurs, cette méthode n'est pas restreinte aux champs gaussiens ni à une grille préalablement localisée : par exemple Wolpert (1998) utilise un processus de Poisson marqué pour obtenir un processus latent sur un «réseau». Enfin, cette méthode est également adaptée aux jeux de données à grandes dimensions comme la modélisation spatio-temporelle ; cette dernière capacité est détaillée dans la description de l'approche.

Ces outils statistiques de modélisation spatio-temporelle permettent de répondre aux demandes d'analyse et de prédictions des distributions d'espèces. La modélisation spatio-temporelle des espèces d'invertébrés dans le golfe du Saint-Laurent (sGSL) pose cependant plusieurs challenges. Les données disponibles couvrent une longue période de temps (1989 à 2005) et sont continues avec une forte proportion de zéros. L'approche de modélisation proposée par Ver Hoef et Jansen (2007) est adaptable aux données du sGSL en changeant le modèle des ob-

servations, qui est une distribution discrète adaptée aux fortes proportions de zéros, en distribution positive continue. Cependant, la partie spatiale du modèle proposé dans cette étude (comme celle de (Cook *et al.*, 2007)) nécessite de découper la région d'étude en zones d'habitat homogène. Or, nous souhaitons ici développer un modèle spatio-temporel qui ne requiert pas le découpage préalable du sGSL en strates homogènes. L'approche de Hooten et Wikle (2007) semble également inadaptée aux données du sGSL car le but de l'étude de leur distribution n'était pas de caractériser leur dispersion sur l'ensemble du sGSL. L'alternative aux modèles géostatistiques développée par Higdon (1998, 2002) présente quant à elle plusieurs avantages pour la modélisation spatio-temporelle de la biomasse des invertébrés épibenthiques développée dans ce chapitre : dans un premier temps, cette méthode de modélisation parcimonieuse nous sert à construire un champ gaussien univarié. Ensuite, nous proposons d'adapter cette approche aux données géoréférencées d'invertébrés épibenthiques du sGSL. Enfin, nous ajoutons une composante temporelle autorégressive pour obtenir un premier modèle hiérarchique bayésien spatio-temporel pour modéliser l'évolution de la biomasse de ces espèces d'invertébrés.

5.1.1 *Construction par moyenne mobile*

Un processus gaussien stationnaire $z(s)$ défini sur un domaine spatial $S \subset \mathbb{R}^2$ peut être construit par convolution d'un champ gaussien $x(s)$, $s \in S$, avec un noyau de lissage κ :

$$z(s) = \int_S \kappa(u-s)x(u)du, \quad \text{pour } s \in S. \tag{5.1}$$

En supposant le champ gaussien de moyenne nulle et de fonction de variance covariance en masse de Dirac («bruit blanc continu»), la fonction de covariance du processus gaussien z dépend uniquement du noyau κ et est obtenue par :

$$\begin{aligned} cov(z(s), z(s')) &= \int_S \kappa(u-s)\kappa(u-s')du \\ &= \int_S \kappa((u-s+s')-s')\kappa(u-s')du \\ &= \int_S \kappa(u-(s-s'))\kappa(u)du \\ &= c(s-s') \end{aligned} \tag{5.2}$$

La relation entre le noyau κ et la fonction de covariance $c(s-s')$ est basée sur le théorème de convolution des transformées de Fourier (Pinkus, 1997; Higdon, 2002) :

$$\begin{aligned} &\kappa(\cdot) \overset{\text{TF}}{\rightarrow} K(u) \overset{.^2}{\rightarrow} K^2(u) \overset{\text{TFI}}{\rightarrow} c(\cdot) \\ &\kappa(\cdot) \overset{\text{TFI}}{\leftarrow} C^{\frac{1}{2}}(u) \overset{\sqrt{.}}{\leftarrow} C(u) \overset{\text{TF}}{\leftarrow} c(\cdot) \end{aligned} \tag{5.3}$$

TF et TFI représentent la transformée de Fourier et son inverse, et les fonctions $.^2$ et $\sqrt{.}$ correspondent au carré et à sa racine. Cette relation permet d'exprimer le spectre, $C(u)$, de covariance, $c(s)$, en fonction du noyau κ. Ce noyau est la transformée de Fourier de la racine carrée de la transformée inverse de $c(s-s')$. Dans le cas isotrope, chaque noyau κ est associé à une unique fonction de covariance $c(s-s')$. Cependant, si le noyau κ est anisotrope alors il existe plusieurs noyaux de lissage qui sont associés à la fonction de covariance $c(s-s')$. Barry et Ver Hoef (1996) et Kern (2000) ont décrit et approfondi les relations entre le noyau de lissage κ et la fonction de covariance $c(s-s')$.

5.1.2 *Construction par convolution discrète*

Il est pratique de construire un champ spatial gaussien z par «convolution discrète». Dans ce cas, le processus latent x n'est plus un champ de type bruit blanc continu, mais il est défini sur une grille latente recouvrant le domaine S. Le champ z est alors construit par convolution du bruit blanc discret x par un noyau de lissage κ. Le champ spatial z s'obtient par :

$$\begin{aligned} z(s|x) &= \mu + \sum_{g=1}^{G} \kappa(\omega_g - s)x(\omega_g) \\ x &\sim Normal(0, \sigma_x^2) \end{aligned} \tag{5.4}$$

où μ est la moyenne du champ z, κ le noyau de lissage et ω_g les points de grille composée de G points. Un bruit blanc $x(\omega_g)$ est associé à chaque point de grille. Le champ spatial z est donc déterminé par G variables latentes, $x(\omega_g)$ avec $g = 1, 2, \ldots, G$. Seules ces G variables latentes contrôlent le champ spatial z réduisant la taille du problème lorsque l'on construit et estime les structures de corrélation spatiale, et ceci, indépendamment du nombre d'observations du champ spatial z ; le prix à payé étant bien sûr l'inférence de ces G variables latentes. Cette méthode, permettant une réduction remarquable de dimension du problème, a notamment été appliquée avec succès pour la modélisation spatio-temporelle de températures moyennes océaniques (Higdon, 1998), de concentration de nitrogène (Kern, 2000), de concentration d'ozone (Higdon, 2002) et pour la modélisation spatiale d'un système proie-prédateur (Allen *et al.*, 2001).

Une partie délicate de ce type de modélisation est le choix de la grille sur laquelle le processus est construit. Il est possible de régler la distance entre deux points de grille en fonction de la portée effective d'un variogramme exponentiel appliqué aux données observées. En effet, Higdon (2002) et Kern (2000) ont montré que si l'on considère un variogramme exponentiel dans le plan, la portée effective de ce variogramme ϕ est simplement égale à la fenêtre du noyau exponentiel associé. Dans la suite de ce chapitre, seuls les noyaux exponentiels sont considérés afin d'être consistants avec les résultats du chapitre 4 et une certaine pratique de la géostatistique qui recommande, en première approche empirique, d'éviter les variogrammes gaussiens qui ont généralement tendance à lisser fortement le champ spatial. Il est cependant possible, de construire une grille régulière à partir

de la portée effective d'un variogramme gaussien (Kern, 2000; Calder, 2003). Lorsque les observations sont directement issues de réalisation du champ spatial en des sites de mesures avec une variance de bruit normal σ_ϵ^2 :

$$Y(s) \sim Normal(z(s), \sigma_\epsilon^2) \tag{5.5}$$

On se trouve en présence d'un modèle hiérarchique normal-normal. L'inférence de tels modèles normaux à convolution discrète peut être réalisée par les méthodes classiques d'inférence fréquentistes : il suffit de considérer le modèle à convolution discrète comme un modèle mixte qui s'écrit :

$$\begin{aligned} Y_s &= \mu + \sum_{g=1}^{G} K_{sg} x_g + \epsilon_s \\ x &\sim Normal(0, \sigma_x^2) \\ \epsilon &\sim Normal(0, \sigma_\epsilon^2) \end{aligned} \tag{5.6}$$

avec μ l'effet fixe, x les effets aléatoires et ϵ un terme d'erreur. Higdon (2002) utilise la méthode du maximum de vraisemblance restreint (REML) pour ce type de modèles. Il est également naturel d'utiliser le cadre bayésien pour estimer les paramètres de ce modèle (cf Annexe B), en particulier quand la couche d'observation n'est plus normale, ni dans la famille exponentielle. En effet, le modèle normal à convolution discrète peut être considéré comme un modèle hiérarchique bayésien, les processus x et z étant des processus latents qui permettent d'observer les quantités Y. La formalisation du modèle à convolution discrète en modèle hiérarchique bayésien permet alors une flexibilité de l'approche, ainsi qu'une adaptabilité à différentes situations comme présentées dans les prochaines sections. Cette forme hiérarchique bayésienne a notamment été utilisée avec succès en environnement Kern (2000) et Calder (2003, 2004).

5.1.3 *Construction par convolution discrète à résolutions multiples*

Il est possible de construire un modèle spatial à résolutions multiples en considérant le processus spatial $z(s)$ comme une somme de processus spatiaux (Higdon, 2002) :

$$z(s) = \sum_{l=1}^{p} z_l(s) \tag{5.7}$$

où p est le nombre de processus spatiaux $z_l(s)$ qu'il est nécessaire de sommer pour obtenir $z(s)$. Chaque $z_l(s)$ est construit par convolution avec le noyau $\kappa_l(s)$ associé aux points de grille $x_l(s)$. La position spatiale des points de grille de chaque processus $x_l(s)$ est définie par $(\omega_{l1}, ..., \omega_{lm_l})$ où m_l est le nombre de points de grille de la composante l. La position des points de grille est généralement différente pour chaque processus $x_l(s)$. Le modèle ainsi construit pour la composante l s'écrit :

$$z_l(s) = \sum_{j=1}^{m_l} \kappa_l(\omega_{lj} - s) x(\omega_{lj}) \tag{5.8}$$

En pratique, deux résolutions sont généralement suffisantes, $L = 2$ (Higdon, 2002) : une résolution fine permettant de décrire les changements locaux de la quantité considérée, et une résolution plus large permettant d'appréhender les tendances à larges échelles. Les quantités d'intérêts Y sont alors modélisées par :

$$Y(s) = \mu + \sum_{l=1}^{L} \sum_{j=1}^{m_l} \kappa_l(\omega_{lj} - s) x(\omega_{lj}) + \epsilon(s) \tag{5.9}$$

5.1.4 *Construction par convolution discrète multidimensionnelle*

Une seconde extension du modèle par convolution discrète est de considérer une modélisation spatiale multivariée (Ancelet, 2008, Chapitre 7). Considérons Y_1 et Y_2 deux quantités d'intérêt échantillonnées sur la même zone d'étude. Un modèle spatial bivarié peut être construit pour modéliser de façon dépendante les deux quantités Y_1 et Y_2 :

$$Y_1(s) = \mu_1 + \sum_{g=1}^{G} \kappa_1(\omega_g - s)x(\omega_g) + \epsilon_1(s)$$
$$Y_2(s) = \mu_2 + \sum_{g=1}^{G} \kappa_2(\omega_g - s)x(\omega_g) + \epsilon_2(s) \qquad (5.10)$$

Dans cet exemple, les deux quantités Y_1 et Y_2 partagent la même grille latente ainsi que le processus x. Cependant, deux noyaux de convolution différents κ_1 et κ_2 sont utilisés. Cette approche permet ainsi d'utiliser deux noyaux de convolution différents et donc deux structures spatiales spécifiques pour chaque quantité, néanmoins corrélées, car elles partagent une même influence des nœuds de grille.

5.1.5 *Construction par convolution discrète avec variables explicatives*

L'ajout de variables explicatives est souhaitable pour ce type de modèle et se fait de manière naturelle. Si l'on considère le modèle spatial par convolution discrète présenté à l'équation 5.6, l'ajout de variables explicatives se traduit simplement par exemple par :

$$Y(s) = \mu + \alpha_1 F^1(s) + \alpha_2 F^2(s) + \sum_{g=1}^{G} \kappa(\omega_g - s)x(\omega_g) \qquad (5.11)$$

où D et F représentent deux variables explicatives associées aux paramètres α_1 et α_2.

5.2 APPROCHE PAR CONVOLUTION POUR DES DONNÉES ZÉRO-INFLATÉES

Le modèle à convolution discrète proposé par Higdon (2002) et mis en œuvre par Calder (2003) est inadapté aux données continues strictement positives présentant une grande proportion de zéros. Cependant, la construction du processus latent spatial par convolution discrète est généralisable à ce type de données. Nous proposons ici d'adapter cette approche à des données zéro-inflatées. Pour ce faire, nous adoptons la forme

hiérarchique bayésienne du modèle à convolution discrète proposé par Calder (2003). Le modèle des observations Y ne peut être une loi normale, comme présenté précédemment (équation 5.5), pour modéliser des données strictement positives avec occurrence de valeurs nulles, telles celles des invertébrés du sGSL. Le modèle Poisson-Composé Gamma (CPG) est donc adopté (Foster et Bravington, 2012; Lecomte *et al.*, 2013b). Le processus de convolution discrète modélise le log du paramètre d'intensité de la loi de Poisson, λ, à interpréter comme le nombre de «patchs» moyen récoltés au site s :

$$\begin{aligned} Y(s) &\sim CPG(\lambda(s), a, b) \\ \lambda(s) &= exp(z(s)) \\ z(s) &= \mu + \sum_{g}^{G} \kappa(s, \omega_g) x(\omega_g), \end{aligned} \tag{5.12}$$

Ce modèle est plus complexe que celui proposé par Higdon (2002) ou Calder (2003). En effet, une couche latente supplémentaire a été ajoutée au modèle proposé dans l'équation 5.6 si on considère que le modèle CPG est lui même un modèle hiérarchique (cf. chapitre 2.1). Les paramètres de ce modèle comprennent donc μ la moyenne du champ spatial latent z, ϕ le paramètre du noyau de convolution contrôlant la distance de corrélation, a et b les paramètres de la taille de patchs, et σ_x l'écart type du bruit blanc gaussien x sur grille. Les variables latentes sont les valeurs des nœuds de la grille, ainsi que le nombre de patchs récoltés en chaque site lorsque la forme hiérarchique du modèle CPG est considérée.

5.2.1 *Simulation de données à forte proportion de zéros avec structure spatiale*

Pour simuler des données zéro-inflatées et structurées spatialement, le modèle présenté ci-dessus (équations 5.12) est mis en œuvre. Après avoir choisi $\theta = (\mu, \phi, a, b, \sigma_x)$, la simulation se déroule en quatre étapes :

1. dans un premier temps, une grille de G points est construite sur la zone d'étude. Une valeur dans une loi normale est tirée de façon *i.i.d.* pour chaque point de la grille x_g :

$$x_g \sim Normal(0, \sigma_x^2), \tag{5.13}$$

2. le noyau exponentiel κ est utilisé pour structurer spatialement la couche latente :

$$\kappa(s-\omega_g) = exp(-\frac{d(s-\omega_g)}{\phi}), \tag{5.14}$$

où d est la distance séparant un point de grille et une observation et ϕ le paramètre qui contrôle la vitesse de décroissance de la corrélation. La couche latente z structurée spatialement est simplement calculée en effectuant la pondération en chaque point s :

$$z(s) = \mu + \sum_{g=1}^{G} \kappa(s-\omega_g)x(\omega_g). \tag{5.15}$$

Matriciellement :

$$\underset{(n,1)}{Z} = \underset{(n,1)}{\mu} + \underset{(n,g)}{K}\underset{(g,1)}{X} \tag{5.16}$$

3. Puis le nombre de patchs est tiré dans une loi de Poisson, la masse de chaque patch est, quant à elle, tirée dans une loi Gamma :

$$\begin{aligned} N_(s) &\sim \text{Poisson}(\lambda(s)) \qquad \text{avec } \lambda(s) = exp(z(s)) \\ M(s,i) &\sim \text{Gamma}(a,b) \end{aligned} \tag{5.17}$$

4. Les observations sont finalement créées selon les marques du modèle CPG :

$$Y(s) = \begin{cases} \sum_{i=1}^{N(s)} M(s,i) & \text{si } N(s) > 0 \\ 0 & \text{si } N(s) = 0 \end{cases} \tag{5.18}$$

5.2.2 *Inférence*

L'inférence d'un modèle hiérarchique bayésien par convolution peut être menée par les méthodes classiques MCMC (Calder, 2003). Il est alors nécessaire de définir des distributions *a priori* pour les paramètres (a, b, μ, σ_x^2). Les deux paramètres contrôlant la taille des patchs (a et b) ont comme prior une distri-

bution gamma $G(2, 0.01)$. Une distribution normale, $N(0, 100)$, est utilisée pour la moyenne du champ latent z. L'écart type, σ_x du bruit blanc x a pour distribution *a priori* une loi uniforme $U(0, 10)$. La portée efficace ϕ est supposée connue. Le logiciel OpenBUGS est utilisé pour effectuer l'inférence du modèle (Lunn *et al.*, 2009; Ntzoufras, 2011). Les distributions *a posteriori* de ces paramètres ont été obtenues à partir de trois chaines MCMC, dont les 30 000 premières itérations ont été écartées afin de pouvoir considérer la convergence des chaines comme atteinte. Cette convergence est également contrôlée par le test de Gelman-Rubin (Gelman et Rubin, 1992). Ensuite, 10 000 itérations ont été gardées, mais afin d'éliminer l'autocorrélation au sein d'une chaine, une itération sur cent a été conservée, soit un total de 3×100 itérations pour obtenir une distribution *a posteriori*.

5.2.3 *Étude par simulations*

5.2.3.1 *Données simulées*

La procédure de simulation présentée précédemment a permis de générer des données « observées » Y avec $S = 250$ sites et $G = 25$ points de grille. La Figure 5.1 représente les données «observées» Y, le champ latent z et les valeurs du bruit blanc sur grille x. Les paramètres utilisés pour générer ces données sont : $a = 1, b = 0.2, \mu = 0.1, \sigma_x^2 = 2, \phi = 0.5$.

5.2.3.2 *Résultats*

Les données simulées représentées en Figure 5.1 ont été utilisées pour estimer les paramètres du modèle (μ, a, b, σ_x^2). La méthode d'inférence utilisée est celle proposée en section 5.2.2. Le Tableau 5.1 montre les moyennes *a posteriori* de chaque paramètre, ainsi que l'intervalle *a posteriori* à 90% et la racine de l'erreur quadratique moyenne.

Les paramètres μ la moyenne du champ spatial et la variance des points de grille σ_x ont ici tendance à être surestimés. Néanmoins, les faibles valeurs de RMSE indiquent que les distributions *a posteriori* sont proches des valeurs simulées pour les paramètres a et b. La Figure 5.2 représente le champ latent λ simulé et le champ latent prédit sur l'ensemble de la zone d'étude.

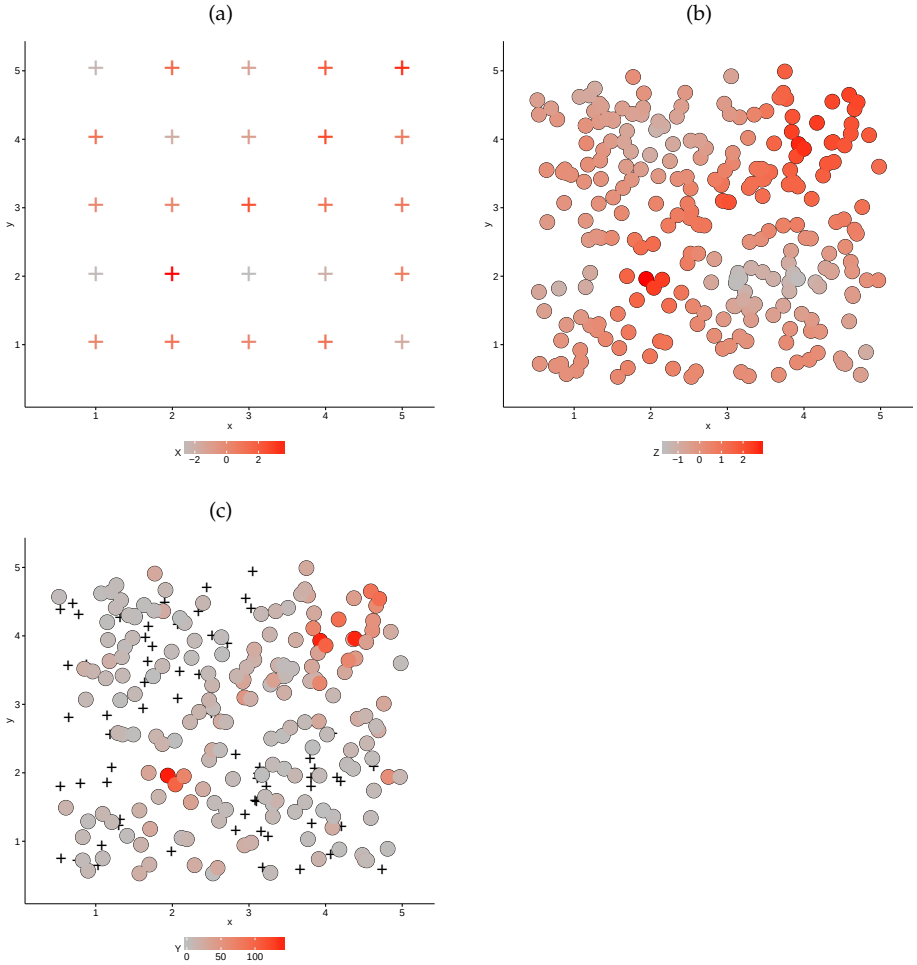

FIGURE 5.1 : Données simulées ($S = 250$ et $G = 25$) avec les paramètres ($a = 1, b = 0.2, \mu = 0.1, \sigma_x^2 = 2, \phi = 0.5$). (a) Première étape simuler sur une grille , (b) deuxième étape simuler le champ latent , (c) enfin simuler les observations (les croix représentent des valeurs nulles).

Tableau 5.1 : Estimation des paramètres du modèle (nombre d'années 1, nombre de sites 250). θ est la valeur simulée de chaque paramètre. Prior représente l'espérance et la variance des distributions *a priori* de chaque paramètre. $\hat{\theta}$ est la valeur moyenne *a posteriori*, les bornes de l'intervalle à 95% sont $\theta_{5\%}$ $\theta_{95\%}$. RMSE est l'erreur quadratique moyenne.

	θ	Prior	$\hat{\theta}$	$\theta_{5\%}$	$\theta_{95\%}$	RMSE
μ	0.10	0 ;100	0.34	-0.08	0.84	0.50
a	1.00	5 ;10	1.28	0.96	1.63	0.35
b	0.20	5 ;10	0.27	0.20	0.35	0.08
σ_x	2.00	5 ;8.3	2.21	1.70	2.95	0.74

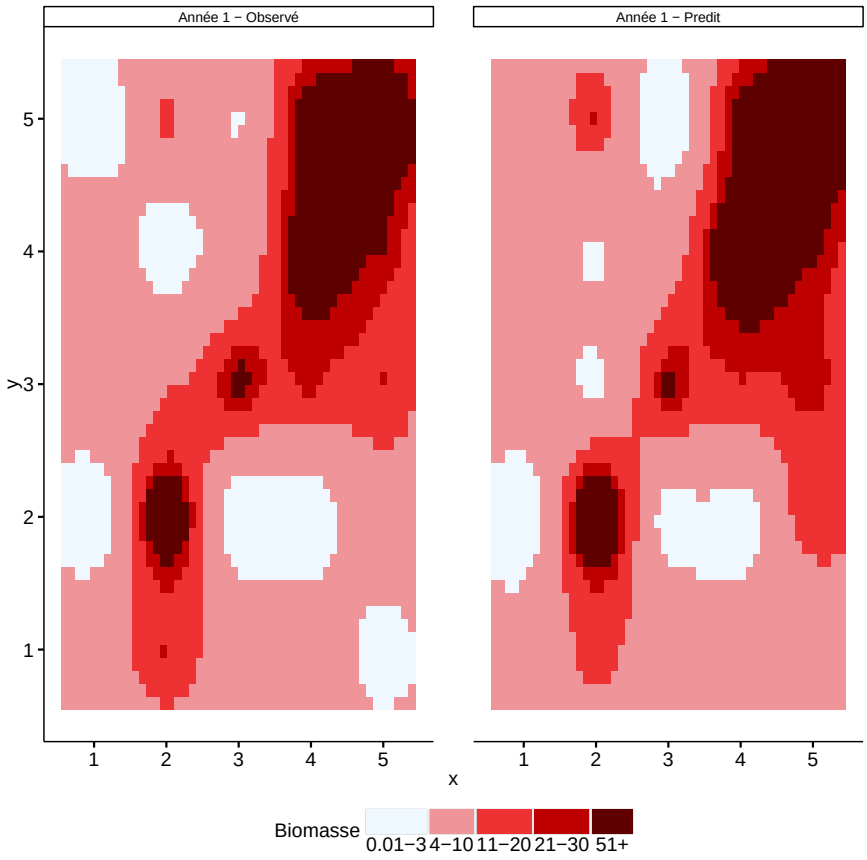

FIGURE 5.2 : Champ latent simulé et champ latent estimé sur l'ensemble de la zone d'étude.

Le champ latent prédit ignore quelques zones de faible densité (Figure 5.2). Ce phénomène peut être expliqué par la surestimation des paramètres présentés dans le Tableau et notamment la surestimation de la moyenne du champ latent z (Tableau 5.1 et équation 5.12). Les zones de fortes densités sont quant à elles correctement estimées.

5.3 UN MODÈLE SPATIO-TEMPOREL PAR CONVOLUTION DISCRÈTE

La modélisation spatiale proposée par Kern (2000), Higdon (2002) et Calder (2003) peut être appliquée à des phénomènes spatio-temporels. Il existe deux principaux types de données spatio-temporelles :

1. les données sont observées à des stations fixes à chaque pas de temps. Néanmoins, certaines données peuvent être manquantes au cours de la période de temps étudiée (p. ex. appareils de mesures défectueux, stations fermées, *etc*.). Pour traiter ce type de données, une nouvelle structure latente est généralement utilisée et les données manquantes sont modélisées comme des données aléatoirement manquantes pendant la durée de l'étude (Gelman *et al.*, 2004, Chapitre 21).

2. Le second type de données spatio-temporelles correspond à des données dont la localisation spatiale change à chaque pas de temps. Ce type de données, encore appelé *misalignment* dans la littérature anglo-saxonne, est fréquemment utilisé pour la modélisation de la distribution spatio-temporelle d'espèces. Ici, nous nous intéressons particulièrement à ce type de données.

La modélisation spatio-temporelle par convolution est adaptée à ces deux types de données (Calder, 2003). Deux approches ont été proposées pour modéliser des données spatio-temporelles par convolution. La première, proposée par Higdon (1998) pour étudier des températures océaniques, consiste à utiliser un noyau de convolution à trois dimensions pour convoluer le processus x défini sur une grille. Les variables gaussiennes $x(t, g)$ sont supposées indépendantes en temps et espace et le noyau de lissage κ est fonction du temps et de l'espace. Le champ spatio-

temporel z est donc modélisé par convolution discrète du processus x par le noyau κ :

$$z(t,s) = \sum_{g=1}^{G} \kappa(t - \tau_g, s - \omega_g,) x_g \tag{5.19}$$

Les coordonnées spatio-temporelles des points de grilles sont données par $(\tau_1, \omega_1), ..., (\tau_g, \omega_g)$. Higdon (1998) a choisi de séparer la partie spatiale de la partie temporelle du noyau κ :

$$\kappa(\Delta_t, \Delta_s) = R(\Delta_t)C(\Delta_s) \tag{5.20}$$

Dans cette étude, le noyau de convolution de la partie temporelle $C(\Delta_s)$ est un noyau gaussien choisi pour des raisons de commodité mathématique ainsi que pour l'interprétation physique du phénomène qu'il apporte. La dépendance temporelle est déterminée à l'aide de variogrammes empiriques, le paramètre de portée ainsi estimé est ensuite utilisé pour construire un noyau gaussien adapté à la corrélation temporelle présente dans les données.

La seconde approche proposée par Calder (2003) et présentée comme la méthode la plus prometteuse au chapitre 6 de Cressie et Wikle (2011), est une modélisation dynamique par convolution. Cette approche modélise la partie temporelle du phénomène par l'intermédiaire du processus latent défini sur grille et non à l'aide de composantes temporelles du noyau de convolution. La forme du noyau et la grille sont constantes au cours du temps, mais la valeur des points de grille évolue au cours de celui-ci. Cette approche, quoique permettant de donner naissance à un processus non stationnaire, présente plusieurs avantages dont une grande flexibilité dans la modélisation. En effet, avec cette approche il est possible, en premier lieu, d'introduire des variables explicatives dans la modélisation de l'évolution temporelle de la quantité observée. Il est également possible d'utiliser des modèles dynamiques conceptuels décrivant l'évolution de cette quantité. En effet, la couche latente d'un modèle bayésien hiérarchique peut être définie par des modèles aux dérivées partielles de la physique ou de la biologie (Cressie et Wikle, 2011, Chapitre 7), ceci permet d'introduire des connaissances supplémentaires sur le phénomène étudié (cf chapitre 6). Cette flexibilité de modélisation est limitée, voire inexistante, quand un noyau tridimensionnel est utilisé. De plus, l'inférence des paramètres du modèle présenté à l'équation 5.19 nécessite

l'inversion d'une matrice de dimensions $TM \times TM$. Cette inversion de matrice peut augmenter les besoins computationnels lorsque le jeu de données est grand. Ce n'est pas le cas pour la modélisation dynamique du fait de l'indépendance, supposée par hypothèse, du processus temporel par rapport au processus spatial.

5.3.1 *Modélisation dynamique par convolution*

La modélisation dynamique par convolution permet de modéliser des phénomènes spatio-temporels à pas de temps discret. La quantité d'intérêt $Y(t,s)$ observée au pas de temps t et au site s est modélisée par :

$$\begin{aligned} Y(t,s|x) &= \mu + \sum_{g=1}^{G} \kappa(s-\omega_g)x(t,\omega_g) + \epsilon(t,s) \\ x(t,\omega_g) &= \delta x(t-1,\omega_g) + \nu(t,g) \end{aligned} \tag{5.21}$$

avec x le processus latent temporel défini sur une grille composée d'un nombre G de points, κ le noyau de convolution, μ la moyenne du champ spatial. Les deux termes d'erreur $\epsilon(t,s)$ et $\nu(t,s)$ sont définis par :

$$\begin{aligned} \text{Erreur de mesure :} \quad & \epsilon(t,s) \overset{i.i.d.}{\sim} Normal(0,\sigma_\epsilon^2) \\ \text{Erreur du processus :} \quad & \nu(t,g) \overset{i.i.d.}{\sim} Normal(0,\sigma_\nu^2) \end{aligned} \tag{5.22}$$

Ici, l'évolution temporelle est modélisée par un processus autorégressif (AR1). Cependant, une modélisation plus complexe de cette évolution est envisageable comme par exemple un membre plus élaboré de la famille des ARIMA (Box *et al.*, 1967).

5.3.2 *Modélisation dynamique par convolution pour des données zéro-inflatées*

5.3.2.1 *Poisson Composé Gamma*

Comme mentionné dans la section 5.2, le modèle hiérarchique bayésien CPG est utilisé comme modèle des observations (Foster et Bravington, 2012; Lecomte *et al.*, 2013b) :

$$Y(t,s) \sim CPG(\lambda(t,s), a, b) \tag{5.23}$$

Les observations, Y, suivent un Poisson Composé Gamma qui reçoit une interprétation phénoménologique interpelant l'écologue si l'on fait apparaitre les deux variables latentes N, le nombre de patchs collectés, et M, la masse de ces patchs. Nous avons choisi d'utiliser l'approche par convolution discrète uniquement dans la modélisation du nombre de patchs : il est spatialisé et varie au cours du temps, contrôlant à la fois la probabilité de zéros et l'intensité de la biomasse récoltée :

$$\begin{aligned}
\lambda(t,s) &= exp(z(t,s)) \\
z(t,s|x) &= \mu + \sum_{g=1}^{G} \kappa(s-\omega_g) x(t,\omega_g) \\
x(t,\omega_g) &= \delta x(t-1,\omega_g) + \nu(t,g) \\
\nu(t,g) &\overset{i.i.d.}{\sim} Normal(0,\sigma_\nu^2)
\end{aligned} \tag{5.24}$$

Le paramètre λ, l'intensité de la loi de Poisson, varie donc avec le temps et l'espace de manière similaire à ce que faisaient Higdon (2002) ou Calder (2003) pour leurs paramètres d'intérêt respectifs. La partie spatiale est modélisée par convolution discrète. Les variations temporelles de la quantité de biomasses Y sont quant à elles modélisées par un processus autorégressif d'ordre 1 qui mime cette évolution temporelle par l'intermédiaire de la grille latente. Il serait cependant possible de complexifier le processus dynamique afin de prendre en compte les connaissances écologiques, comme nous l'évoquons en discussion (cf. chapitre 7).

5.3.3 *Simulation de données à fortes proportions de zéros avec structure spatiale et pas de temps discret*

Pour le premier pas de temps, la simulation d'un jeu de données temporel structuré spatialement est identique à la simulation proposée en section 5.2.1. Ce point obtenu, il est alors possible de générer des quantités de biomasses Y aux pas de temps suivants :

1. au pas de temps t, on tire une valeur dans une loi normale pour chaque point de la grille $x_{t,g}$ avec une loi normale centrée sur la valeur du point de grille au pas de temps précédent $t-1$:

$$\begin{aligned} x(1,\omega_g) &\sim Normal(0,\sigma_x^2) \\ x(t,\omega_g) &= \delta x(t-1,\omega_g) + \nu(t,g) \\ \nu(t,g) &\overset{i.i.d.}{\sim} Normal(0,\sigma_\nu^2). \end{aligned} \tag{5.25}$$

2. Un noyau exponentiel κ est utilisé pour structurer spatialement la couche latente :

$$\kappa(s-\omega_g) = exp(-\frac{d(s-\omega_g)}{\phi}), \tag{5.26}$$

où d est la distance séparant les points de grilles des points échantillonnés et ϕ le paramètre qui contrôle la vitesse de décroissance de la corrélation. Nous considérons que la structure spatiale conditionnellement aux nœuds est inchangée au cours du temps, le paramètre ϕ est donc constant. La couche latente z structurée spatialement par convolution est simulée avec :

$$z(t,s) = \mu + \sum_{g=1}^{G} \kappa(s-\omega_g)x(t,\omega_g). \tag{5.27}$$

La moyenne du champ spatiale est elle aussi considérée comme constante au cours du temps. Ensuite, on tire le nombre de patchs dans une loi de Poisson et la taille des patchs dans une loi Gamma (constante au cours du temps) :

$$
\begin{aligned}
N(t,s) &\sim \text{Poisson}(\lambda(t,s)) \quad \text{avec } \lambda(t,s) = exp(z((t,s))) \\
M(s,i) &\sim \text{Gamma}(a,b)
\end{aligned} \tag{5.28}
$$

3. Les observations sont ensuite générées avec le modèle *CPG* :

$$
Y_{t,s} = \begin{cases} \sum_{i=1}^{N(t,s)} M(s,i) & \text{si } N(t,s) > 0 \\ \\ 0 & \text{si } N(s) = 0 \end{cases} \tag{5.29}
$$

Il est important de noter que la localisation des observations varie au cours du temps. Les poids accordés à chaque site par les nœuds de grille du noyau de convolution doivent ainsi être calculés à chaque pas de temps, même si conditionnellement aux nœuds la structure spatiale est constante ($\phi = constante$).

5.3.4 *Inférence*

L'inférence d'un modèle spatio-temporel hiérarchique bayésien par convolution est identique à celle proposée pour le modèle spatial. Les paramètres (a, b, μ, σ_x^2) ont les mêmes distributions *a priori* (cf section 5.2.2). Les distributions *a priori* des paramètres de dépendance temporelle σ_v et δ sont deux lois uniformes $U(0, 10)$. Le nombre de chaines MCMC utilisées est au nombre de trois. Pour chaque chaine, 1000 itérations sont conservées sur 10 000 afin d'éviter l'autocorrélation et ceci après 30 000 itérations de période de chauffe.

5.3.5 *Étude par simulations*

5.3.5.1 *Données simulées*

Les données utilisées lors de cette étude ont été générées par la procédure de simulation proposée en section 5.3.3. La variance du processus x pour la première année est égale à $\sigma_x^2 = 2$, la variance de la marche aléatoire est égale $\sigma_v^2 = 0.1$. Les paramètres a et b sont égaux respectivement à 1 et 0.2, la moyenne du champ latent $\mu = 0.1$, le paramètre de dépendance temporelle est égal à $\delta = 0.9$ et la valeur du paramètre de portée

$\phi = 0.5$ est supposée connue. L'évolution du champ latent λ est représentée par la Figure 5.3 où les 5 années sont représentées.

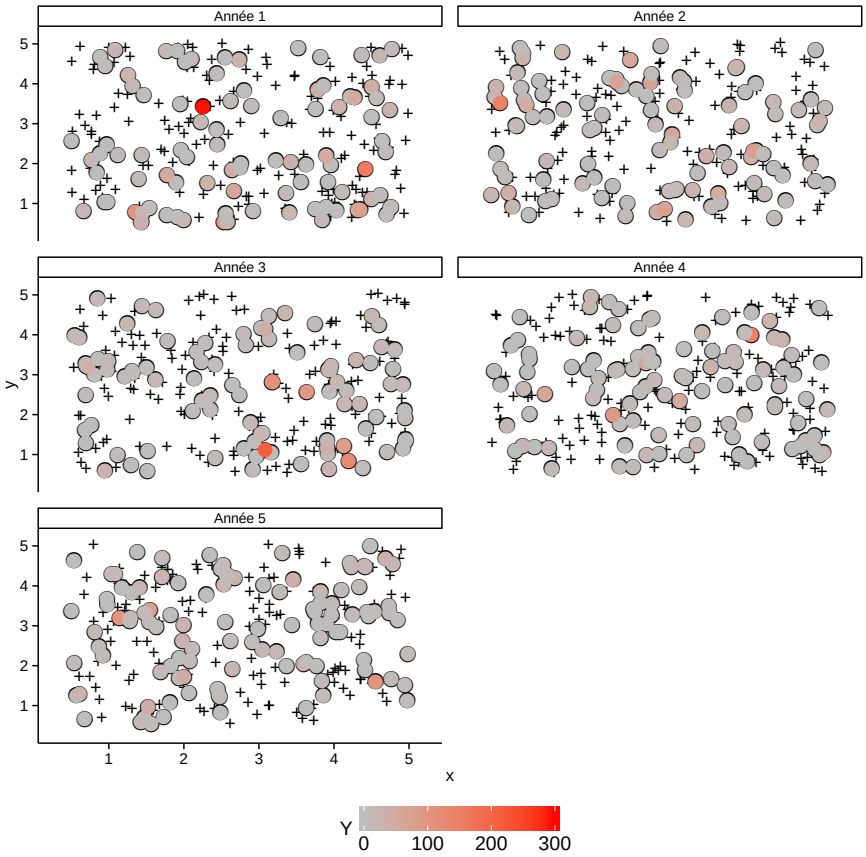

FIGURE 5.3 : Données simulées pendant 5 années ($a = 1$, $b = 0.2$, $\mu = 0.1$, $\sigma_x = 2$, $\phi = 0.5$, $\sigma_v = 0.1$, et $\delta = 0.9$).

5.3.5.2 *Résultats*

Le Tableau 5.2 donne les statistiques obtenues *a posteriori* des paramètres du modèle $(\mu, a, b, \sigma_x, \sigma_v)$.

La moyenne du champ latent μ est sous-estimée, les paramètres de variances σ_x et σ_v sont quant à eux surestimés. Les autres paramètres sont estimés correctement (Tableau 5.2). La Figure 5.4 représente les champs latents simulés et prédits pour les 5 années.

Comme pour le modèle strictement spatial les prédictions du champ latent permettent de retrouver correctement les zones à fortes densités. Cependant, les zones de densités moyennes sont quelque peu rognées. Le fait que la moyenne du champ

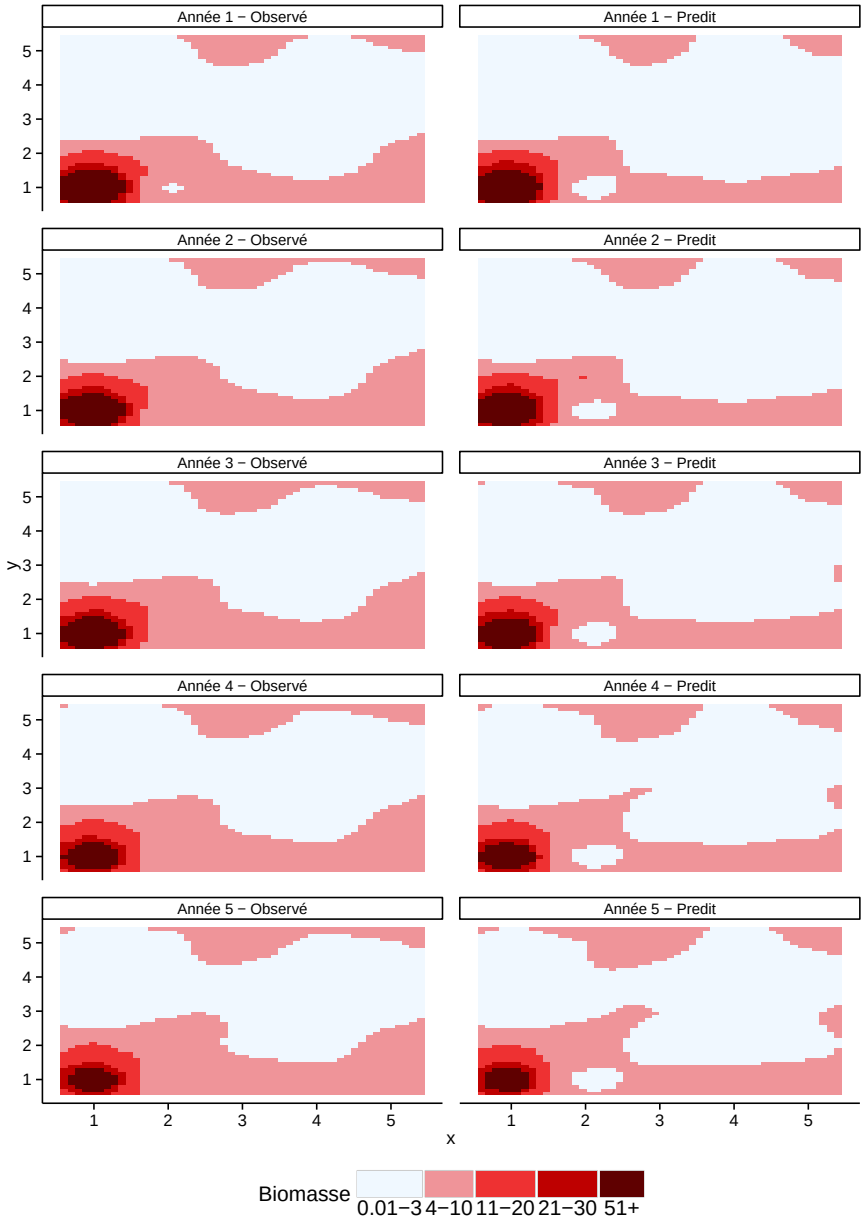

FIGURE 5.4 : Champ latent simulé et champ latent estimé sur l'ensemble de la zone d'étude pour les 5 années ($a = 1$, $b = 0.2$, $\mu = 0.1$, $\sigma_x = 2$, $\phi = 0.5$, $\sigma_v = 0.1$, et $\delta = 0.9$).

Tableau 5.2 : Estimation des paramètres du modèle (nombre d'années 5, nombre de sites par année 250). θ est la valeur simulée de chaque paramètre, Prior représente l'espérance et l'écart type des distributions *a priori* de chaque paramètre. $\hat{\theta}$ est la valeur moyenne *a posteriori*, les bornes de l'intervalle à 90% sont $\theta_{5\%}$ $\theta_{95\%}$. RMSE est la racine de l'erreur quadratique moyenne.

	θ	Prior	$\hat{\theta}$	$\theta_{5\%}$	$\theta_{95\%}$	RMSE
μ	0.10	0 ;100	-0.78	-0.87	-0.70	0.08
a	1.00	5 ;10	0.97	0.87	1.09	0.12
b	0.20	5 ;10	0.20	0.18	0.23	0.03
σ_X	2.00	5 ;8.3	2.28	1.74	3.05	0.77
σ_ν	0.10	5 ;8.3	0.22	0.06	0.46	0.24
δ	0.90	5 ;8.3	0.94	0.90	0.97	0.03

latent est sous-estimée est certainement compensé par la surestimation du paramètre de variance des nœuds de grille. Le paramètre de dépendance temporelle δ étant proche de 1, l'évolution temporelle des quantités de biomasses simulées est conservative. Cette stabilité temporelle peut expliquer la «tache blanche» de faible densité proche de la partie à fortes quantités de biomasses. Le paramètre de bruit σ_ν est surestimé et permet de légèrement compenser cette stabilité temporelle.

5.4 Application aux données d'invertébrés épibenthiques du golfe du Saint-Laurent

5.4.1 *Construction d'une grille régulière*

Le choix des points de grilles est déterminant dans la modélisation spatio-temporelle de biomasses des espèces étudiées. Pour ce faire, la distance de corrélation estimée par Lecomte *et al.* (2013b) peut être utilisée. Cette distance maximale de corrélation permet la construction d'une grille régulière sur l'ensemble de notre zone d'étude, le sud du golfe du Saint-Laurent. Plusieurs grilles régulières, représentées sur la Figure 5.5, ont ainsi été proposées, elles diffèrent par le nombre de points qui les compose.

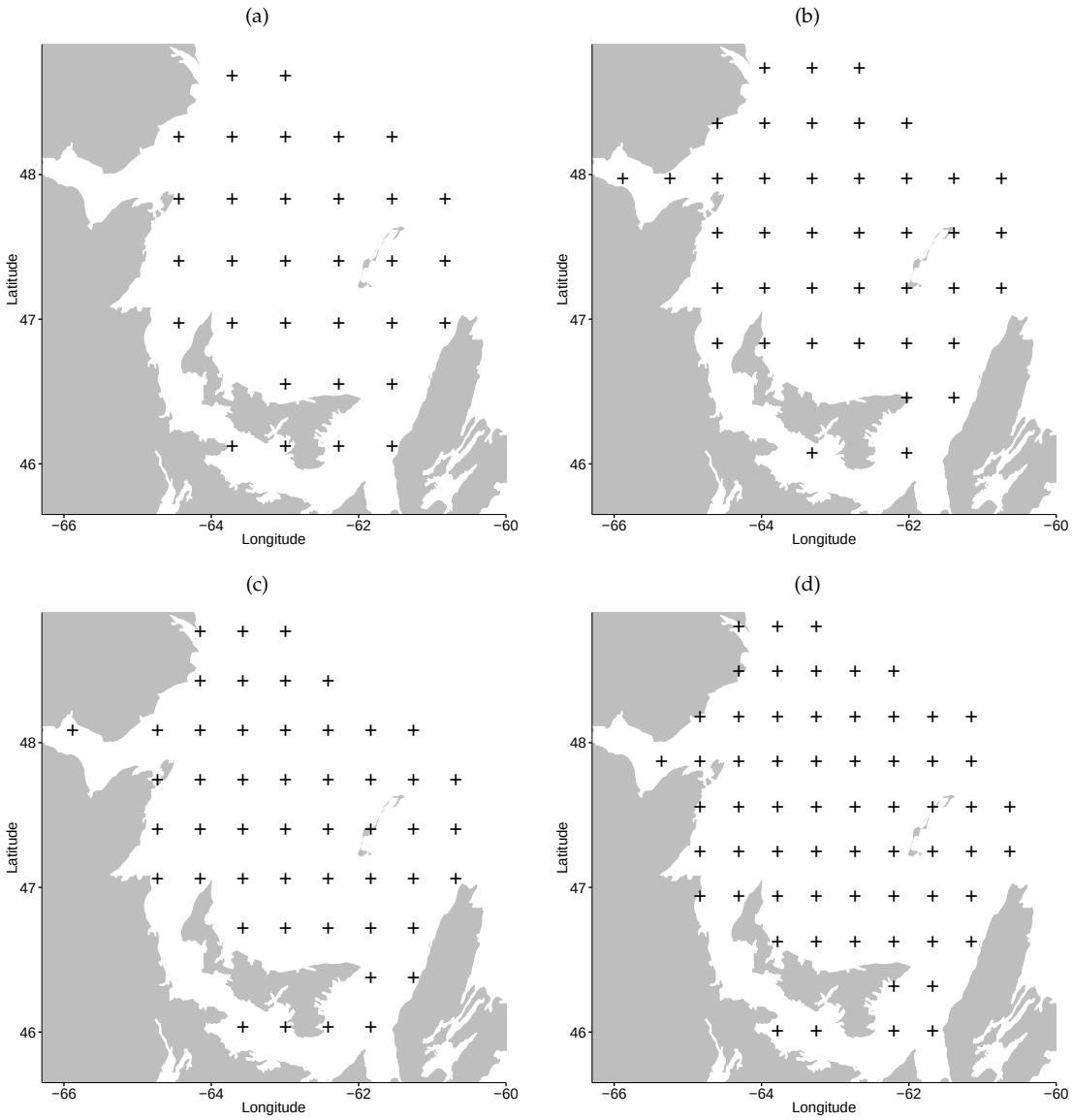

FIGURE 5.5 : Grilles régulières construites sur l'ensemble du sud du golfe du Saint-Laurent. (a) Grille régulière de 32 points séparés de 47 *km*. (b) Grille régulière de 41 points séparés de 42 *km*. (c) Grille régulière de 50 points séparés de 38 *km*. (d) Grille régulière de 63 points séparés de 35 *km*.

5.4.2 *Construction d'une grille irrégulière*

L'utilisation d'une grille non régulière est également envisageable. Afin de pouvoir comparer les différences apportées par une grille irrégulière à une grille régulière, les nombres de points de grille sont identiques à ceux proposés en section 5.4.1. La Figure 5.6 présente quatre grilles irrégulières construites avec la fonction *coverdesign* du package R *fields* (Furrer *et al.*, 2012). Cette fonction permet de construire une grille irrégulière dont les nœuds sont répartis de manière homogène sur l'ensemble de la zone d'échantillonnage.

5.4.3 *Modèle spatial par convolution*

Dans un premier temps, un modèle spatial par convolution est construit à partir de l'équation 5.12. À ce modèle sont ajoutées les variables explicatives utilisées dans le chapitre 4 (types de sédiments, profondeur et température). Ces variables environnementales sont ajoutées à la couche latente :

$$\begin{aligned} \lambda(s) &= exp(z(s)) \\ z(s) &= \mu + \sum_{g}^{G} \kappa(s, \omega_g) x(\omega_g) + \beta_{Sed_s} + \gamma_{Prof_s} + \zeta_{Temp_s} \end{aligned} \quad (5.30)$$

où β_{Sed_s}, γ_{Prof_s} et ζ_{Temp_s} sont les paramètres associés respectivement aux types de sédiments, à la profondeur et à la température. Comme proposée dans le chapitre 4, la profondeur (en m) est divisée en trois classes ($[0, 50[$, $[50, 100[$, et $[100, 400[$), la température (en °C) est également séparée en trois classes ($[-1, 1[$, $[1, 5[$, et $[5, 15[$). Enfin, quatre types de sédiments sont considérés en fonction de leur granulométrie (pélite, sable fin, sable grossier et gravier avec des patchs occasionnels de sable).

Les données de biomasse récoltées en 1997 sont considérées pour effectuer l'inférence de ce modèle. La procédure d'inférence proposée en section 5.2.2 est utilisée. La portée ϕ est constante et est fixée à 18 km (Lecomte *et al.*, 2013b). Plusieurs grilles régulières qui diffèrent par leur nombre de points sont comparées. Les grilles irrégulières sont ensuite étudiées et également comparées entre elles. Enfin, les deux types de grille sont comparés.

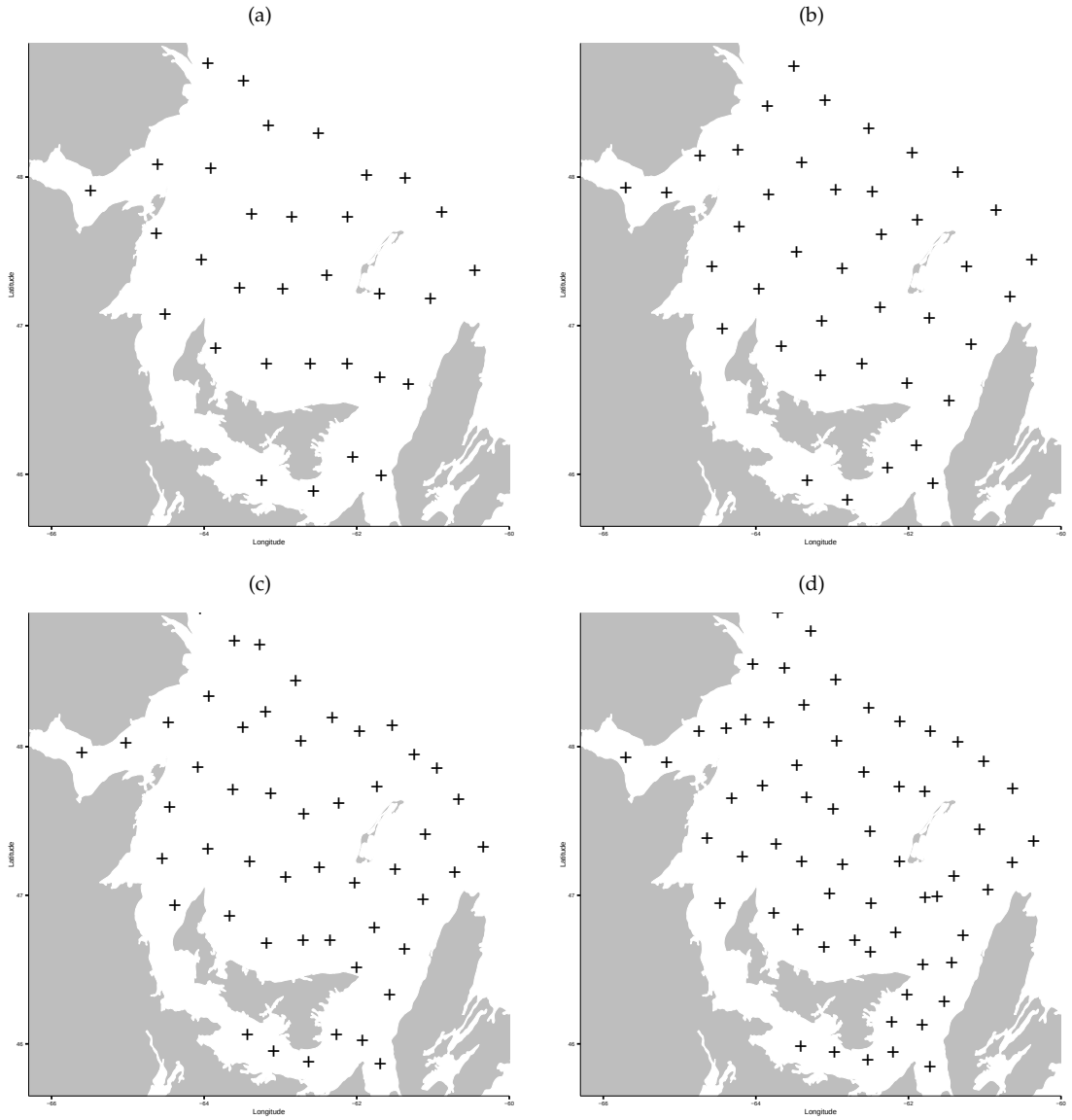

FIGURE 5.6 : Grilles irrégulières construites sur l'ensemble du sud du golfe du Saint-Laurent. (a) Grille de 32 points séparés en moyenne de 43 *km*. (b) Grille régulière de 41 points séparés en moyenne de 41 *km*. (c) Grille régulière de 50 points séparés en moyenne de 34 *km*. (d) Grille régulière de 63 points séparés en moyenne de 29 *km*.

5.4.3.1 *Résultat avec une grille régulière*

Quatre modèles sont proposés, ils se distinguent par la grille latente utilisée support de la dépendance spatiale et temporelle. Les grilles diffèrent par leur nombre de points (32, 41, 50 et 63, Figure 5.5). Les prédictions effectuées sur l'ensemble du sGSL sont présentées en Figure 5.7.

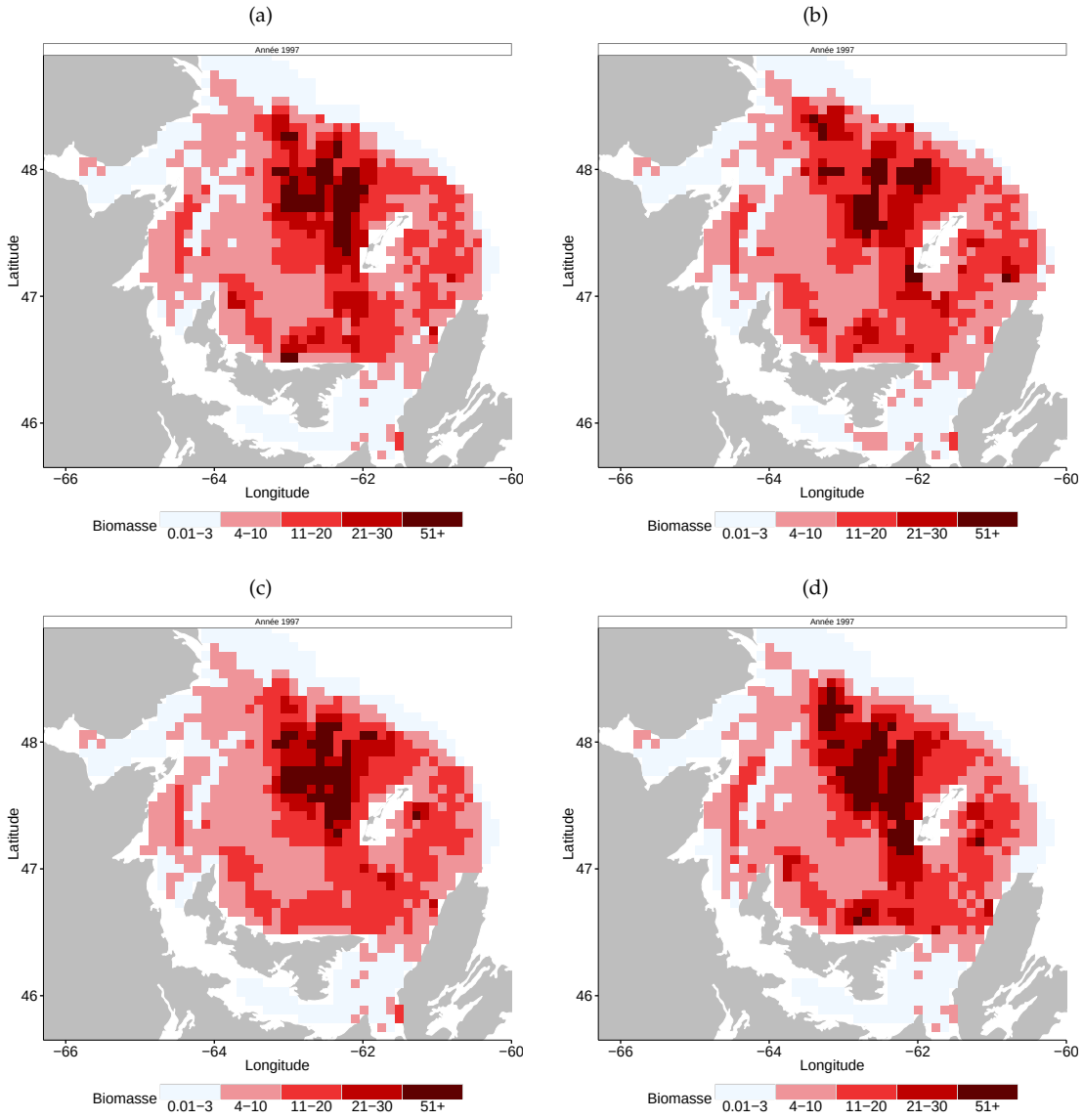

FIGURE 5.7 : Biomasse d'oursins attendue sur l'ensemble du sGSL pour l'année 1997. (a) Grille régulière de 32 points. (b) Grille régulière de 41. (c) Grille régulière de 50 points. (d) Grille régulière de 63 points.

Les prédictions de la répartition spatiale de la biomasse d'oursins sur l'ensemble du sGSL sont similaires entre les quatre grilles utilisées (Figure 5.7). Cette Figure montre une zone de fortes quantités de biomasses dans la partie centrale-nord du sGSL, ainsi que plusieurs zones d'absences d'oursins comme à l'embouchure du Saint-Laurent, dans la partie extrême nord ou encore au sud des iles du Prince Edward.

Afin de sélectionner un modèle et donc une grille sur les quatre proposées, le critère DIC (Spiegelhalter *et al.*, 2002) a été calculé (Tableau 5.3). Le modèle sélectionné par ce critère est le modèle dont la grille latente est composée de 50 points.

Tableau 5.3 : Scores DIC obtenus avec le modèle spatial par convolution pour les biomasses d'oursins récoltés en 1997 en fonction du nombre de points de grille régulière.

Nb points de grille	32	41	50	63
DIC	1155.7	1146.2	**1137.8**	1175.7

Les estimations des paramètres du modèle à 50 points de grille sont résumées dans le Tableau 5.4. Les effets des variables environnementales sont cohérents avec ceux estimés avec le modèle à variogramme (cf. chapitre 4). Le type de sédiment pélite a un effet négatif sur la biomasse d'oursins, contrairement aux types sable grossier et gravier qui ont un effet positif sur cette biomasse comparativement à l'effet de base qui est le sable fin. La même tendance est également retrouvée pour l'effet de la profondeur, avec un effet négatif des profondeurs les plus grandes. Les températures les plus fortes ont quant à elles un effet négatif sur la biomasse d'oursins.

Les capacités prédictives de ce modèle sont étudiées pour l'année suivante, c'est-à-dire l'année 1998. La proportion de biomasses positives correctement prédites ainsi que le nombre de zéros correctement prédits pour cette année 1998 sont résumés dans le Tableau 5.5. La RMSE est également calculée pour cette année sur l'ensemble des sites échantillonnés en 1998. La majorité des sites sont correctement prédits, cependant des améliorations sont possibles notamment pour les faux zéros.

5.4.3.2 *Résultat avec une grille irrégulière*

Quatre grilles irrégulières ont également été testées (Figure 5.6). Les modèles, dont les grilles sont composées de 41 points, 50

Tableau 5.4 : Estimations des paramètres du modèle spatial des quantités de biomasses d'oursins (41 nœuds de grille régulière). $\hat{\theta}$ est la valeur moyenne *a posteriori*, les bornes de l'intervalle à 90% sont $\theta_{5\%}$ $\theta_{95\%}$.

Paramètres	Termes	$\hat{\theta}$	$\theta_{5\%}$	$\theta_{95\%}$
μ		0.45	0.16	0.74
a		0.46	0.36	0.56
b		0.09	0.07	0.12
σ_x		1.2	0.83	1.63
Sédiment	pélite	-1.02	-1.6	-0.5
	sable fin	0	0	0
	sable grossier	0.56	0.2	0.91
	gravier	0.64	0.17	1.12
Température	$[-1, 1[$	0	0	0
	$[1, 5[$	-0.29	-0.79	0.23
	$[5, 15[$	-1.24	-1.88	-0.61
Profondeur	$[0, 50[$	0.06	-0.51	0.62
	$[50, 100[$	0	0	0
	$[100, 400[$	-0.71	-1.46	-0.04

Tableau 5.5 : Proportion des prédictions de la biomasse d'oursins pour les années 1998 et RMSE moyenne sur l'ensemble des sites échantillonnés.

Année	Vrai positif	Faux zéro	Vrai zéro	Faux positif	RMSE
1998	0.69	0.31	0.62	0.38	8.93

points et 63 points, capturent la même répartition générale de la biomasse d'oursins que les modèles à grilles régulières (Figures 5.8 b, c et d). Cependant, le modèle à grille irrégulière de 32 ne reproduit pas le patron de répartition observé précédemment (Figures 5.8 a). Les mêmes zones d'absences de l'espèce oursin sont identifiées par ce modèle à grille irrégulière (embouchure du Saint-Laurent, dans la partie extrême nord ou encore au sud des iles du Prince Edward), ainsi que la zone de forte densité située au centre-nord du golfe (Figures 5.8 b, c et d).

Le critère DIC confirme les moins bonnes performances du modèle à 32 points de grille. Le modèle sélectionné est le modèle dont la grille irrégulière est composée de 41 points. Les estimations des paramètres pour ce modèle sont présentées dans le Tableau 5.7. Les effets des variables explicatives sont similaires à ceux estimés avec une grille régulière. Les prédictions

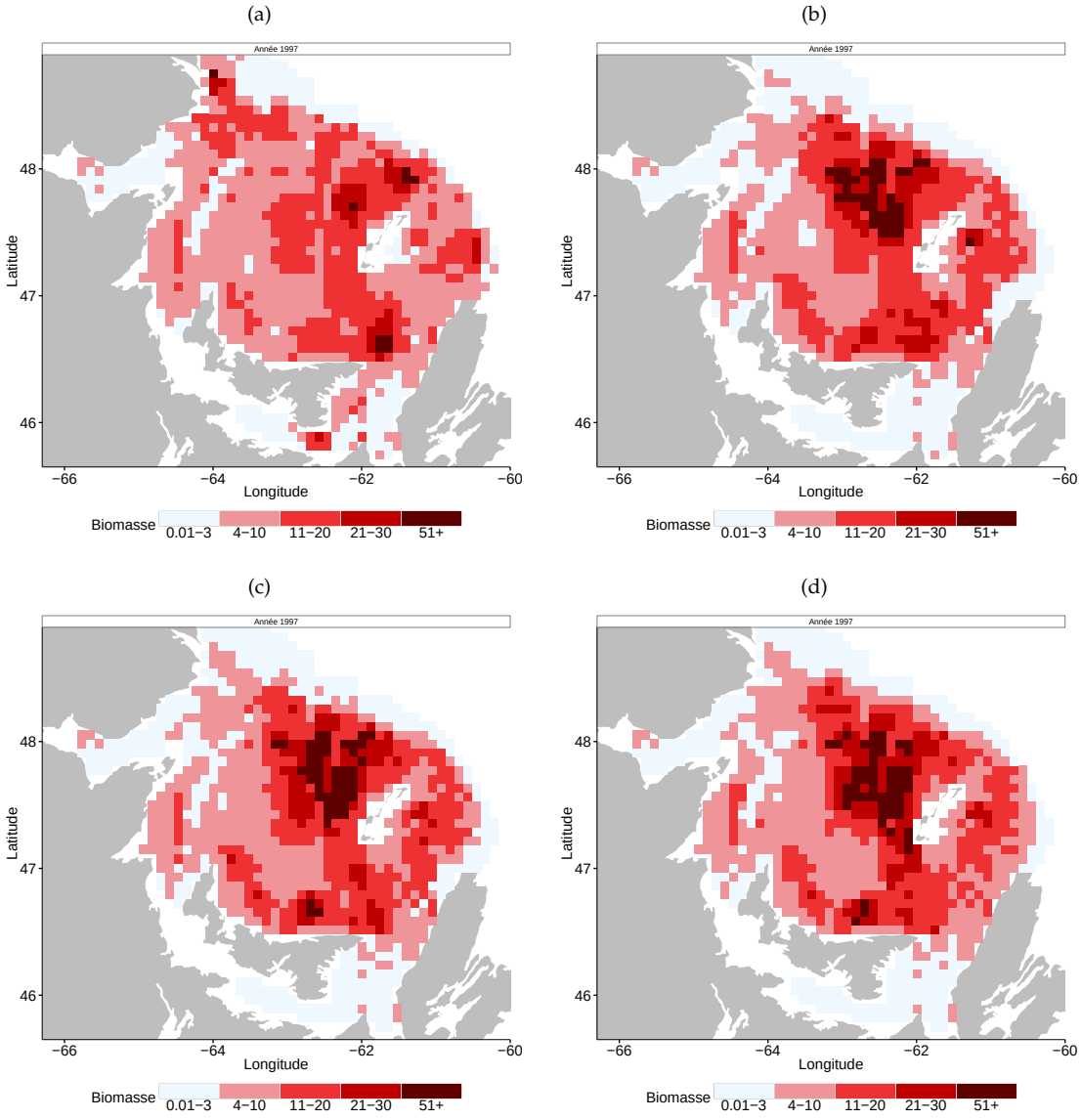

FIGURE 5.8 : Biomasse d'oursins attendue sur l'ensemble du sGSL pour l'année 1997. (a) Grille irrégulière de 32 points. (b) Grille irrégulière de 4. (c) Grille irrégulière de 50 points. (d) Grille irrégulière de 63 points.

de la biomasse récoltées en 1998 sont également proches de celles obtenues avec une grille régulière (Tableau 5.8).

Tableau 5.6 : Scores de BIC et DIC obtenus avec le modèle spatial par convolution pour les biomasses d'oursins récoltés en 1997 en fonction du nombre de points de grille irrégulière.

Nb points de grille	32	41	50	63
DIC	1159.7	**1136.0**	1146.0	1153.5

Tableau 5.7 : Estimation des paramètres du modèle spatial des quantités de biomasses d'oursins (41 nœuds de grille irrégulière). $\hat{\theta}$ est la valeur moyenne *a posteriori*, les bornes de l'intervalle à 90% sont $\theta_{5\%}$ $\theta_{95\%}$.

Paramètres	Termes	$\hat{\theta}$	$\theta_{5\%}$	$\theta_{95\%}$
μ		0.44	0.11	0.75
a		0.46	0.35	0.57
b		0.1	0.07	0.12
σ_x		1.4	0.92	1.98
Sédiment	pélite	-0.92	-1.53	-0.35
	sable fin	0	0	0
	sable grossier	0.49	0.16	0.83
	gravier	0.71	0.28	1.14
Température	$[-1, 1[$	0	0	0
	$[1, 5[$	-0.27	-0.71	0.17
	$[5, 15[$	-1.26	-1.85	-0.68
Profondeur	$[0, 50[$	0.02	-0.52	0.56
	$[50, 100[$	0	0	0
	$[100, 400[$	-0.74	-1.45	-0.18

Tableau 5.8 : Proportion des prédictions de la biomasse d'oursins pour les années 1998 et RMSE moyenne sur l'ensemble des sites échantillonnés.

Année	Vrai positif	Faux zéro	Vrai zéro	Faux positif	RMSE
1998	0.68	0.32	0.61	0.39	9.22

5.4.3.3 *Comparaison grille régulière et irrégulière*

Le modèle spatial par convolution peut être construit avec une grille régulière comme une grille irrégulière. Comme présenté précédemment, les résultats obtenus pour chacun des deux types de grilles sont similaires en terme d'effets estimés des variables environnementales, mais également en terme de qualité de prédiction. Le paramètre de variance des nœuds de grille σ_x est légèrement plus grand lorsque l'on considère une grille irrégulière (1.2 pour la grille régulière et 1.4 pour la grille irrégulière), cette différence peut s'expliquer par le nombre de points de grille différent entre les deux grilles. Les points de la grille irrégulière étant moins nombreux, la variance de ceux ci tend à augmenter pour capturer les différences de quantités de biomasses. Les deux modèles estiment cependant une moyenne du champ μ très similaire, comme les paramètres a et b contrôlant la biomasse présente dans un patch. La principale différence entre les deux meilleurs modèles à grille régulière et irrégulière est le nombre de points composant leur grille respective. Le modèle à grille régulière est composé d'un nombre plus élevé de points (50) que le modèle à grille irrégulière (41). Cette différence peut s'expliquer par le fait que les points de la grille irrégulière sont disposés astucieusement sur l'ensemble du sGSL. Cette répartition astucieuse permet une économie de points de grille qui n'est pas possible avec une grille régulière composée du même nombre de points. Il est cependant difficile de sélectionner un modèle par rapport à l'autre (différence de DIC égale à 1), néanmoins la parcimonie avantage le modèle a grille irrégulière.

5.4.4 *Modèle spatio-temporel par convolution pour étudier la biomasse d'oursins*

Le modèle spatio-temporel par convolution proposé dans les équations 5.23 et 6.18 est utilisé pour modéliser la répartition de la biomasse d'oursins sur l'ensemble du sGSL pour les années 1989 à 2003 (Figures 5.9 et 5.10). Les variables environnementales utilisées précédemment sont également ajoutées à la couche latente (équation 5.30). La dépendance temporelle est donc modélisée par un processus autorégressif d'ordre 1. Deux paramètres (δ et σ_v) ont donc été ajoutés par rapport au modèle strictement spatial présenté précédemment (section 5.4.3). Comme précédemment, plusieurs modèles sont compa-

rés. Dans un premier temps les modèles dont la grille est régulière, mais dont le nombre de points qui la composent diffère, sont comparés. Ensuite, des modèles à grilles latentes irrégulières sont étudiés. Enfin, les deux types de grille sont comparés.

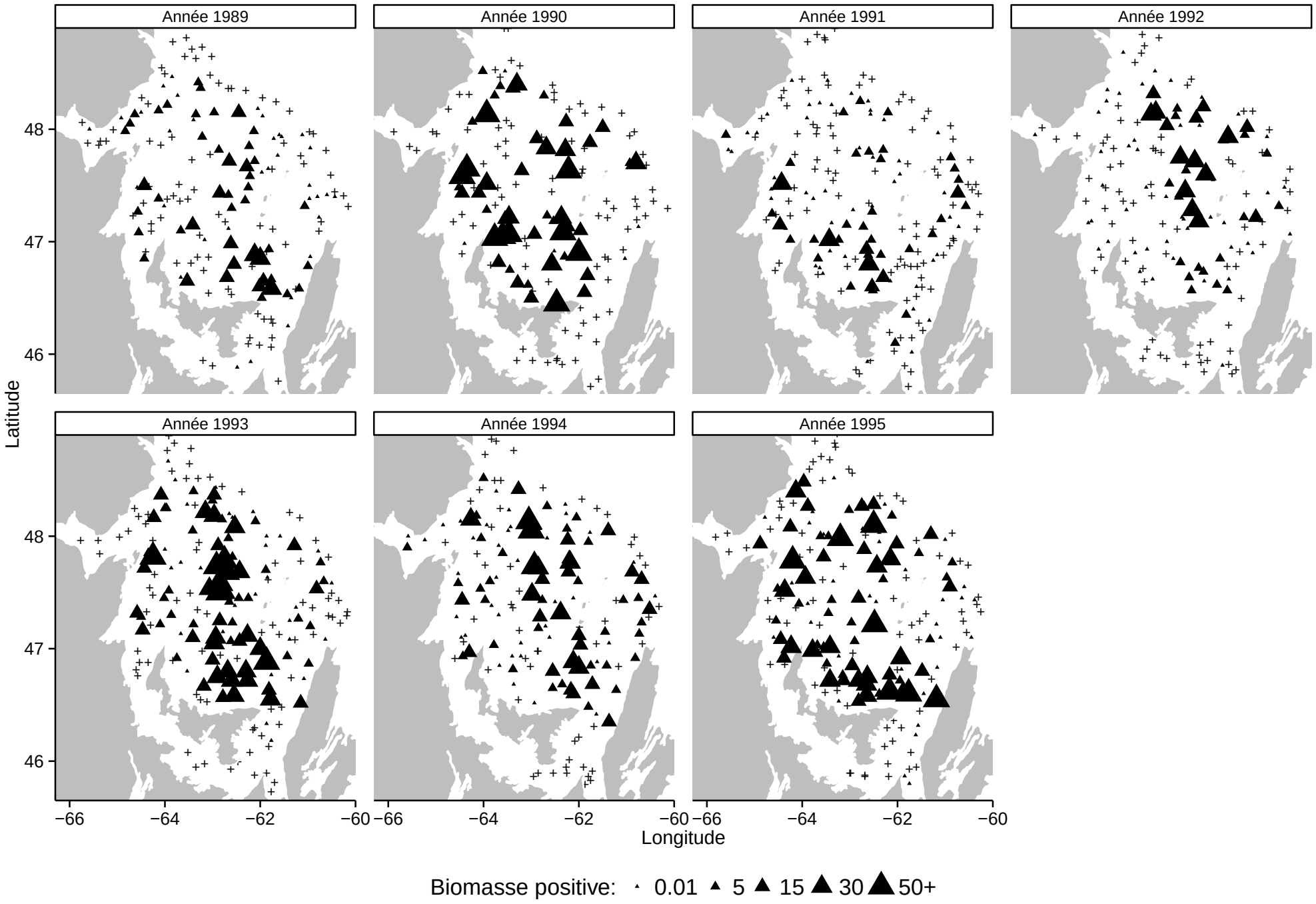

FIGURE 5.9 : Quantité de biomasses d'oursins récoltées sur l'ensemble su sGSL pour les années 1989 à 1995. Les triangles représentent les quantités de biomasses non nulles, les croix représentent l'absence de l'espèce.

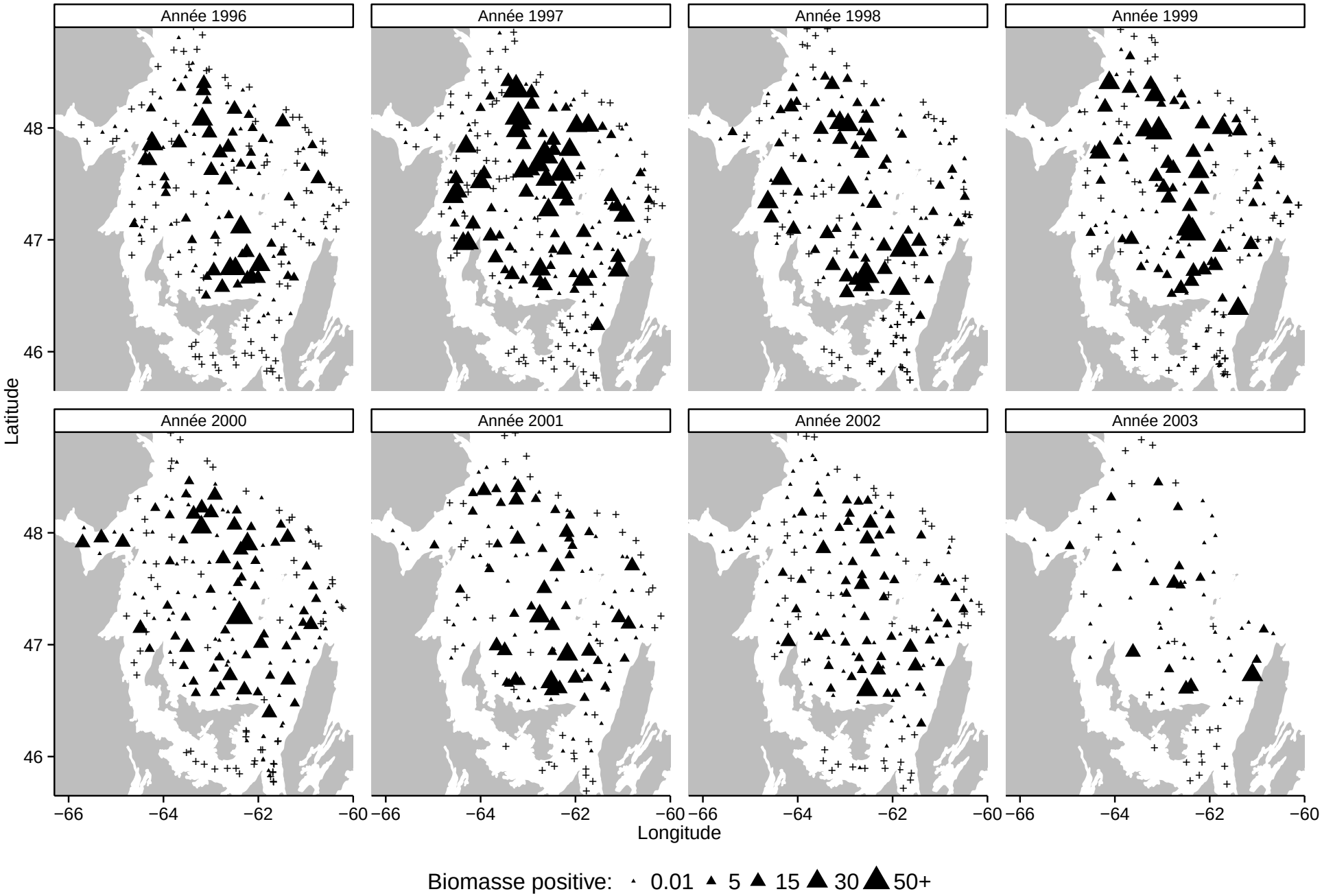

FIGURE 5.10 : Quantité de biomasses d'oursins récoltées sur l'ensemble su sGSL pour les années 1996 à 2003. Les triangles représentent les quantités de biomasses non nulles, les croix représentent l'absence de l'espèce.

5.4.4.1 *Grille régulière*

Les grilles régulières considérées sont composées de 32, 41, 50 et 63 points (Figure 5.5). Le critère DIC est utilisé pour sélectionner le meilleur modèle (Tableau 5.9), le modèle ainsi sélectionné est basé sur une grille composée de 41 points.

Tableau 5.9 : Scores de DIC obtenus avec le modèle spatio-temporel par convolution pour les biomasses d'oursins récoltés entre 1989 et 2003 en fonction du nombre de points de grille régulière.

Nb points de grille	32	41	50	63
DIC	12 692.1	**12 586.2**	12 738.7	13 073.4

Ce modèle permet de construire l'évolution de la biomasse d'oursins récoltée depuis l'année 1989 jusqu'à l'année 2003 (Figure 5.11). La distribution des quantités de biomasses d'oursins en 1989 était concentrée au nord-est des iles du Prince Edward. De plus, de nombreuses zones présentaient une absence de cette espèce. À partir des années 1992-1993, la distribution des quantités de biomasses d'oursins recouvre l'ensemble du golfe excepté la zone de l'embouchure du Saint-Laurent, la partie extrême nord ou encore le sud des iles du Prince Edward. L'interpolation des quantités de biomasses d'oursins effectuée pour les années 2002 et 2003 montre un éclatement des zones à forte densité sur l'ensemble du golfe.

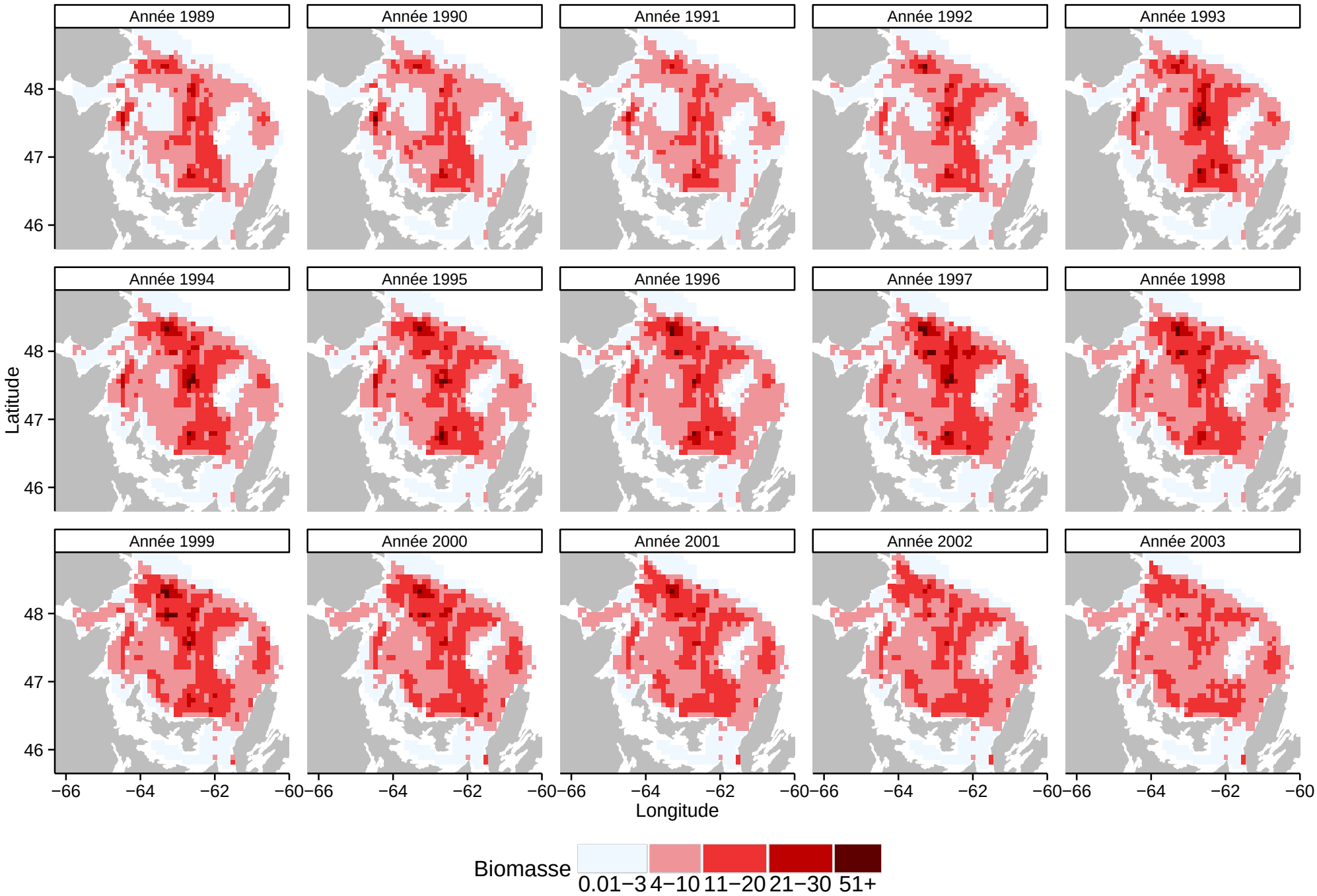

FIGURE 5.11 : Biomasse d'oursins prédite sur l'ensemble su sGSL pour les années 1989 à 2003 (grille régulière, $\phi = 18$ et $G = 41$).

L'estimation des paramètres du modèle est présentée dans le Tableau 5.10. L'effet du type de sédiment reste inchangé par rapport au modèle strictement spatial avec un effet négatif du type de sédiment pépite et un effet positif pour les deux types de sédiments sable grossier et gravier. On constate également la même relation avec la profondeur où les plus grandes profondeurs ont un effet négatif sur la biomasse d'oursins. Les effets de la température restent inchangés par rapport au modèle strictement spatial. Les paramètres de dépendance temporelle traduisent une grande corrélation entre les années (δ proche de 1) et une faible variabilité entre les années ($\sigma_\nu = 0.47$).

Tableau 5.10 : Estimations des paramètres du modèle spatio-temporel des quantités de biomasses d'oursins (41 nœuds de grille régulière).

Paramètres	Termes	$\hat{\theta}$	$\theta_{5\%}$	$\theta_{95\%}$
μ		0.25	0.03	0.46
a		0.49	0.46	0.53
b		0.11	0.1	0.12
σ_x		2.35	1.59	3.25
σ_ν		0.47	0.36	0.61
δ		0.89	0.84	0.94
Sédiment	pelite	-0.65	-0.82	-0.49
	sable fin	0	0	0
	sable grossier	0.55	0.42	0.65
	gravier	0.78	0.62	0.94
Température	$[-1, 1[$	0	0	0
	$[1, 5[$	-0.14	-0.28	-0.03
	$[5, 15[$	-1.33	-1.51	-1.16
Profondeur	$[0, 50[$	0.12	-0.03	0.28
	$[50, 100[$	0	0	0
	$[100, 400[$	-0.93	-1.12	-0.71

Les capacités prédictives du modèle spatio-temporel dont la grille régulière est composée de 41 points sont étudiées pour les années 2004 et 2005, années consécutives aux années ayant permis l'estimation du modèle (Tableau 5.11). La proportion de vrai positif est comparable à celle du modèle spatial par convolution (Tableau 5.8). Néanmoins, le gain de capacité prédictive d'un modèle spatio-temporel par rapport à un modèle strictement spatial est observé par la plus faible valeur de RMSE obtenue. Il est également important de noter que les capacités

prédictives diminuent entre l'année 2004 et l'année 2005. Une illustration de ces prédictions est proposée dans la figure 5.12.

Tableau 5.11 : Proportion des prédictions de la biomasse d'oursins pour les années 2004 et 2005 et RMSE moyenne sur l'ensemble des sites échantillonnés.

Année	Vrai positif	Faux zéro	Vrai zéro	Faux positif	RMSE
2004	0.62	0.38	0.79	0.21	3.80
2005	0.58	0.42	0.79	0.21	4.72

5.4.4.2 *Grille irrégulière*

Comme pour le modèle spatial, quatre modèles dont les grilles irrégulières diffèrent par leur nombre de points sont considérés pour étudier la biomasse d'oursins sur l'ensemble du sGSL pour les années 1989 à 2003. Le modèle sélectionné par le critère DIC est le modèle dont la grille est composée de 41 points (Tableau 5.12).

Tableau 5.12 : Scores de DIC obtenus avec le modèle spatio-temporel par convolution pour les biomasses d'oursins récoltés entre 1989 et 2003 en fonction du nombre de points de grille irrégulière.

Nb points de grille	32	41	50	63
DIC	12 653.6	**12 630.5**	12 684.3	12 685.0

Les estimations des paramètres de ce modèle à grille irrégulière de 41 points sont résumées dans le Tableau 5.13. Les effets des variables environnementales sont identiques aux estimations précédentes. L'estimation des paramètres contrôlant la dépendance temporelle est comparable au modèle à grille régulière, c'est-à-dire une estimation du paramètre δ proche de 1, traduisant une forte dépendance temporelle entre les années, associée à une faible variabilité entre ces mêmes années ($\sigma_v = 0.48$). Cette évolution temporelle est illustrée par la Figure 5.13 qui propose une succession de cartes représentant la biomasse d'oursins sur l'ensemble du sGSL.

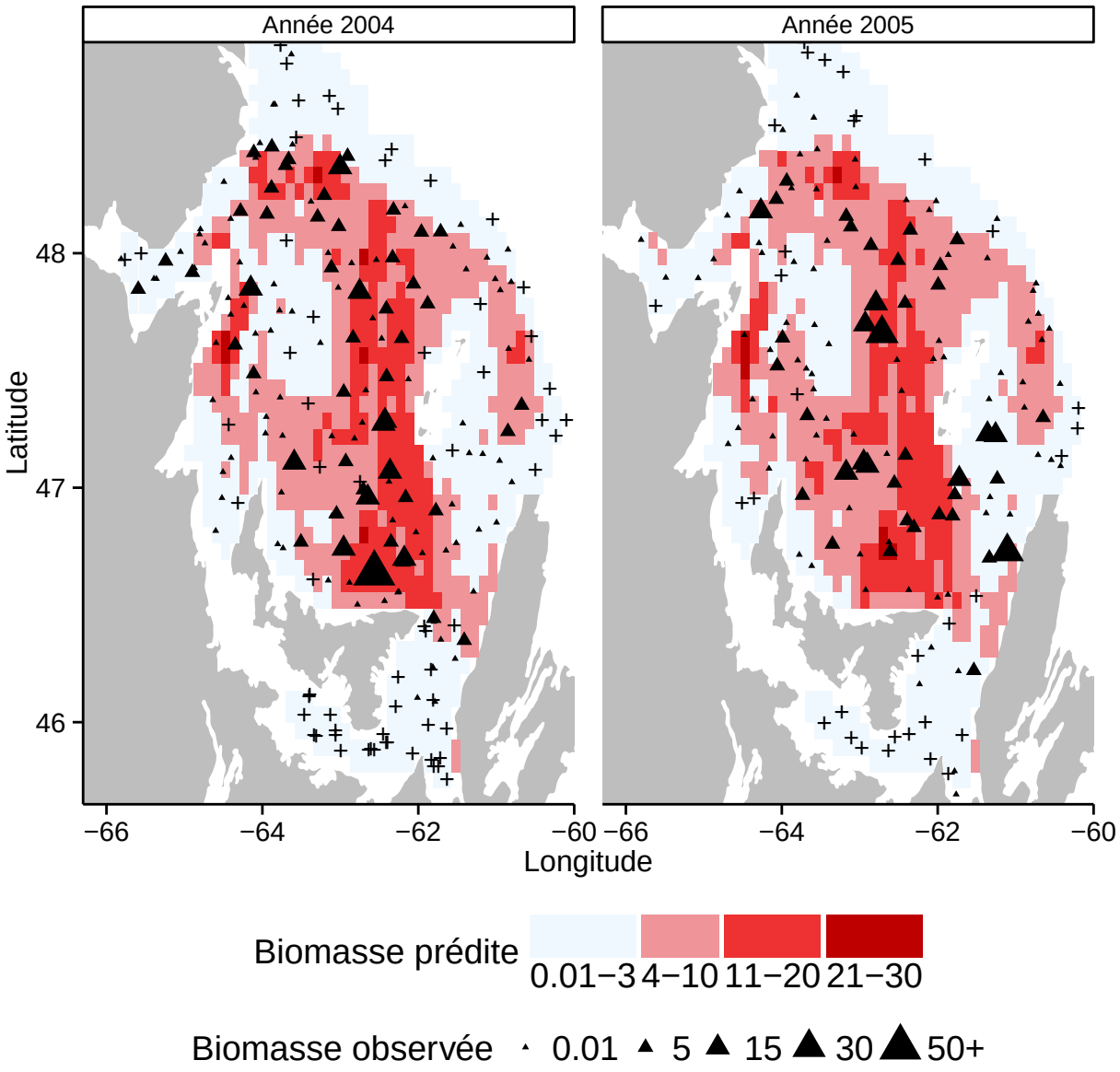

FIGURE 5.12 : Prédictions et observations des quantités de biomasses d'oursins sur l'ensemble du sGSL pour les années 2004 et 2005 (les croix représentent les quantités de biomasses observées nulles).

Tableau 5.13 : Estimations des paramètres du modèle spatio-temporel des quantités de biomasses d'oursins (41 nœuds de grille irrégulière).

Paramètres	Termes	$\hat{\theta}$	$\theta_{5\%}$	$\theta_{95\%}$
μ		0.17	-0.03	0.37
a		0.49	0.46	0.53
b		0.11	0.1	0.12
σ_x		1.55	1.08	2.1
σ_v		0.48	0.36	0.6
δ		0.89	0.84	0.94
Sédiment	pelite	-0.62	-0.79	-0.45
	sable fin	0	0	0
	sable grossier	0.45	0.34	0.55
	gravier	0.7	0.56	0.84
Température	$[-1, 1[$	0	0	0
	$[1, 5[$	-0.15	-0.29	-0.04
	$[5, 15[$	-1.24	-1.42	-1.03
Profondeur	$[0, 50[$	0.09	-0.08	0.23
	$[50, 100[$	0	0	0
	$[100, 400[$	-0.94	-1.17	-0.72

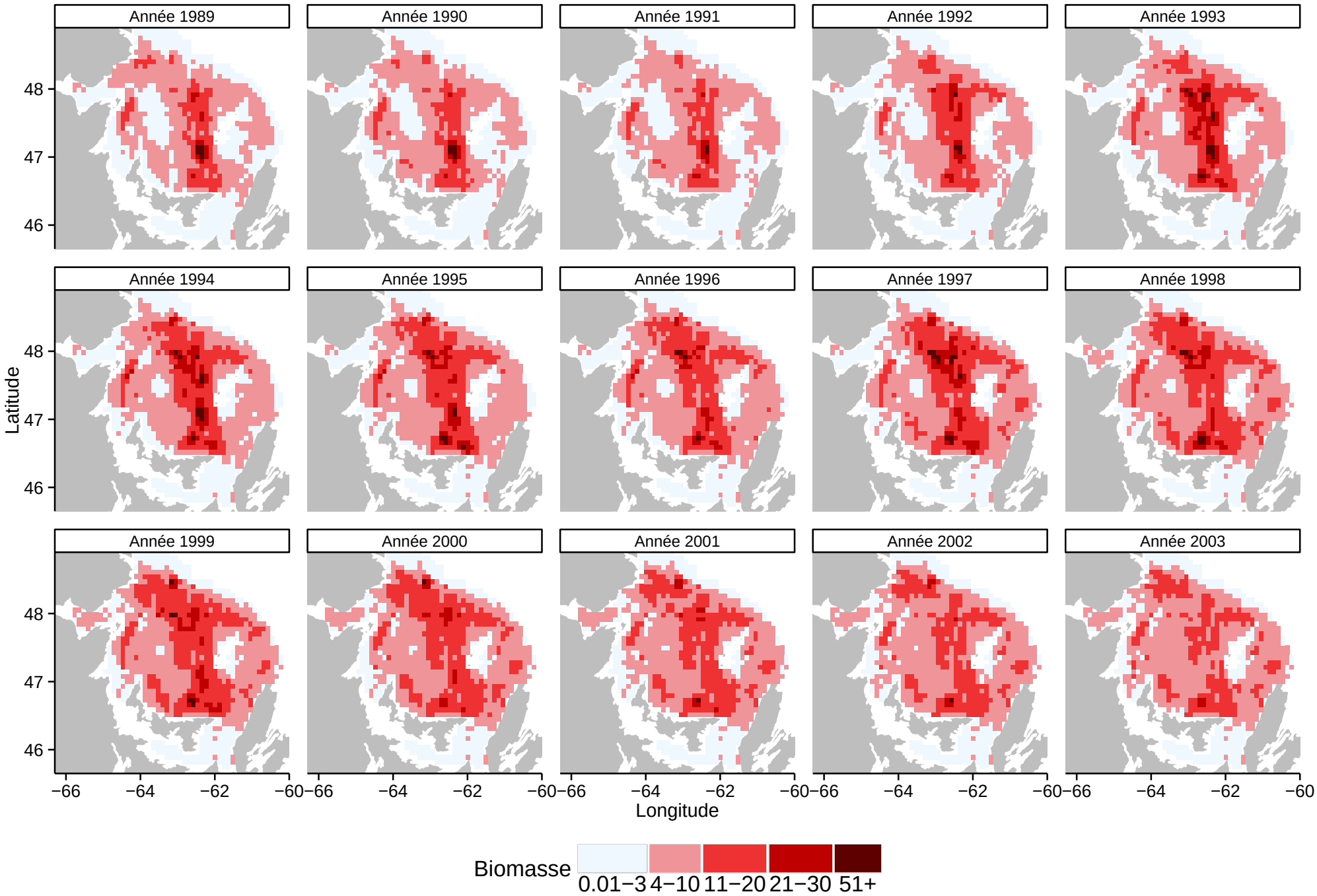

FIGURE 5.13 : Biomasse d'oursins attendue sur l'ensemble su sGSL pour les années 1989 à 2003 (grille irrégulière, $\phi = 18$ et $G = 41$).

Les prédictions de la biomasse d'oursins effectuée pour les années 2004 et 2005 sont peu satisfaisantes (Tableau 5.14). En effet, le nombre de faux zéros est relativement élevé et pose problème. Néanmoins, les vrais zéros sont bien prédits. Les RMSE obtenus pour chaque année restent cependant acceptables et sont plus faibles comparativement au modèle strictement spatial.

Tableau 5.14 : Proportion des prédictions de la biomasse d'oursins pour les années 2004 à 2005.

Année	Vrai positif	Faux zéro	Vrai zéro	Faux positif	RMSE
2004	0.63	0.37	0.79	0.21	3.90
2005	0.58	0.42	0.79	0.21	4.67

5.4.4.3 *Comparaison grille régulière et irrégulière*

La comparaison entre les modèles spatio-temporels à grille régulière ou à grille irrégulière est plus aisée que dans le cas des modèles strictement spatiaux. En effet, le critère DIC sélectionne nettement le modèle à grille régulière, malgré des performances prédictives comparables pour les deux modèles.

5.4.5 *Modèle spatio-temporel par convolution pour étudier la biomasse de concombre de mer*

La modélisation spatio-temporelle par convolution discrète (équations 5.23 et 6.18) est appliquée ici aux quantités de biomasses de concombre de mer récoltées entre 1989 à 2003 dans le sGSL (Figures 5.14 et 5.15). Le type de sédiments, la température et la profondeur sont ajoutés à la couche latente de ce modèle 5.30). Comme pour les biomasses d'oursins, un modèle autorégressif d'ordre 1 définit la dépendance temporelle. La grille latente utilisée est composée de 63 points répartis régulièrement sur l'ensemble du sGSL (Figure 5.5.d). L'inférence des paramètres est réalisée avec le logiciel OpenBUGS et est décrite en section 5.3.4.

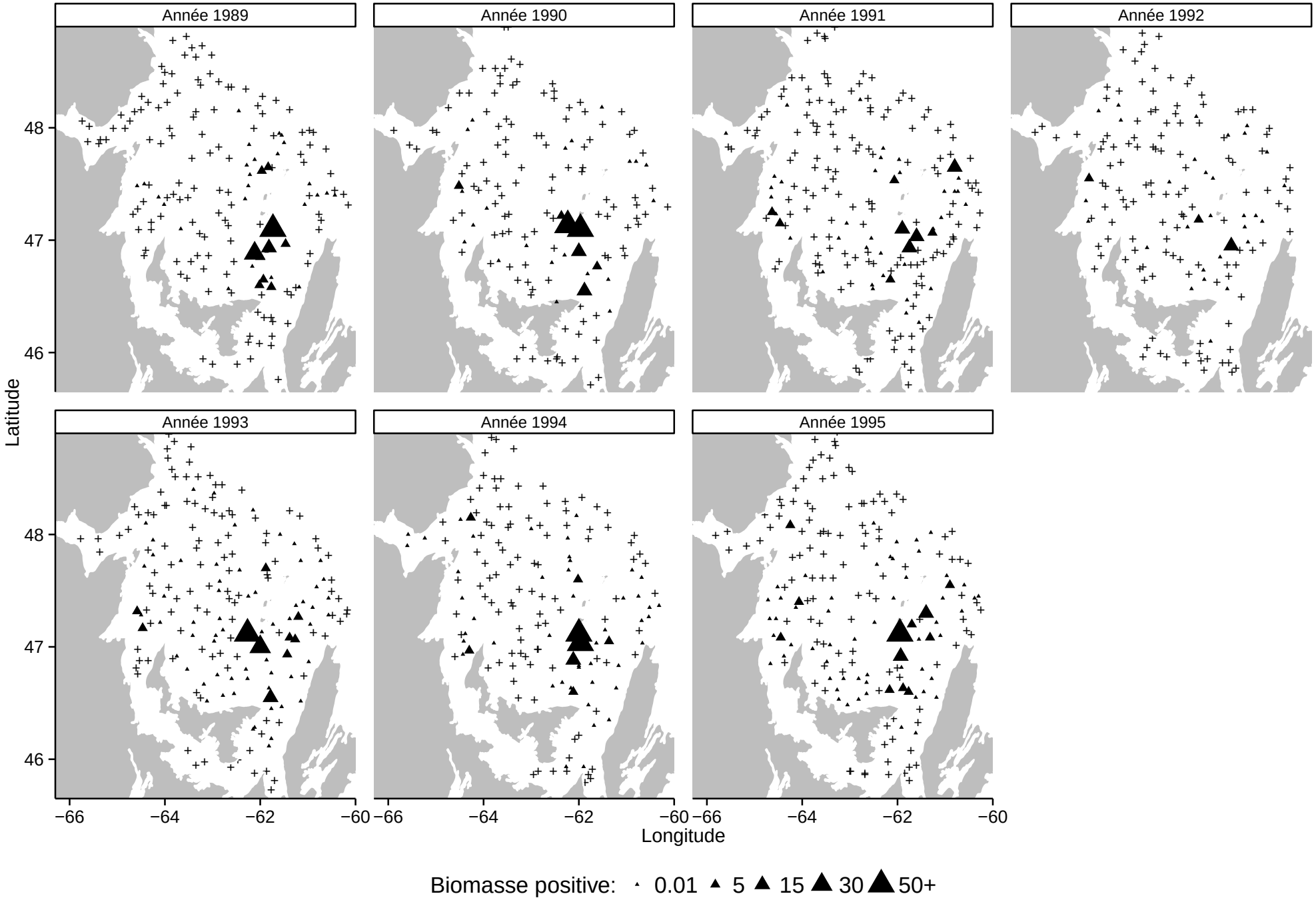

FIGURE 5.14 : Quantité de biomasses de concombre de mer récoltées sur l'ensemble su sGSL pour les années 1989 à 1995. Les triangles représentent les quantités de biomasses non nulles, les croix représentent l'absence de l'espèce.

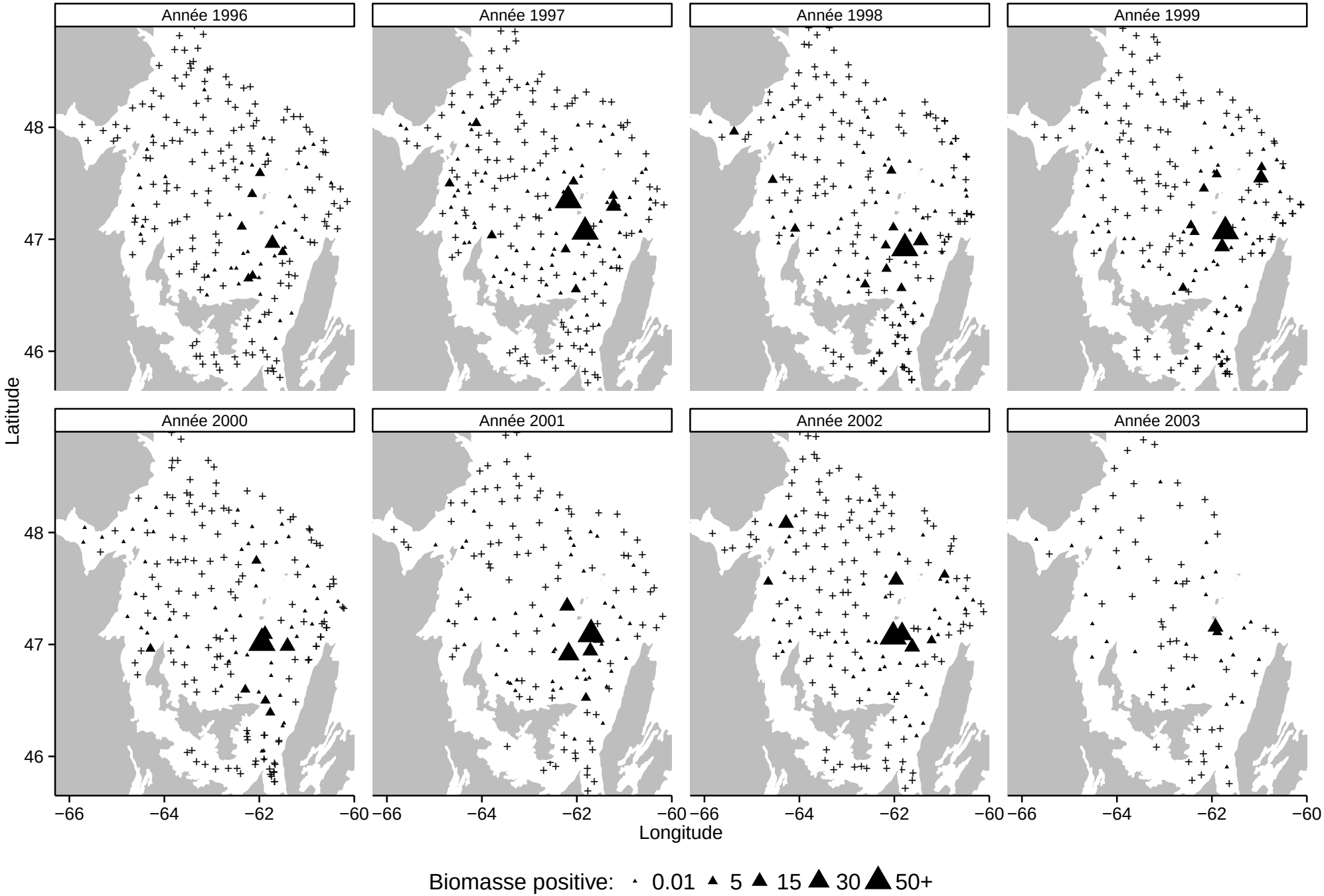

FIGURE 5.15 : Quantité de biomasses de concombre de mer récoltées sur l'ensemble su sGSL pour les années 1996 à 2003. Les triangles représentent les quantités de biomasses non nulles, les croix représentent l'absence de l'espèce.

5.4.5.1 *Résultats*

Les estimations des paramètres du modèle spatio-temporel analysant la distribution des quantités de biomasses de concombre de mer sont résumées dans le Tableau 5.15. Les effets des variables environnementales sont similaires à ceux estimer dans l'étude du chapitre 4 : on observe un effet négatif du type de sédiment pélite, contrairement aux sédiments de types sable grossier et gravier qui ont un effet positif sur les quantités de biomasses de concombre de mer. Les faibles profondeurs ont également un effet positif à l'inverse des profondeurs les plus fortes. On remarque également que les températures les plus élevées ont un effet négatif sur cette espèce.

Tableau 5.15 : Estimations des paramètres du modèle spatio-temporel des quantités de biomasses de concombre de mer (grille régulière de 63 nœuds).

Paramètres	Termes	$\hat{\theta}$	$\theta_{5\%}$	$\theta_{95\%}$
μ		-1.97	-2.35	-1.61
a		0.35	0.32	0.39
b		0.14	0.12	0.16
σ_x		2.28	1.72	2.99
σ_ν		0.56	0.4	0.74
δ		0.95	0.91	0.98
Sédiment	pélite	-0.36	-0.68	-0.02
	sable fin	0	0	0
	sable grossier	0.45	0.24	0.66
	gravier	0.8	0.58	1.02
Température	$[-1, 1[$	0	0	0
	$[1, 5[$	0.27	0.09	0.47
	$[5, 15[$	-0.38	-0.61	-0.11
Profondeur	$[0, 50[$	1.01	0.79	1.21
	$[50, 100[$	0	0	0
	$[100, 400[$	-1.61	-2.05	-1.19

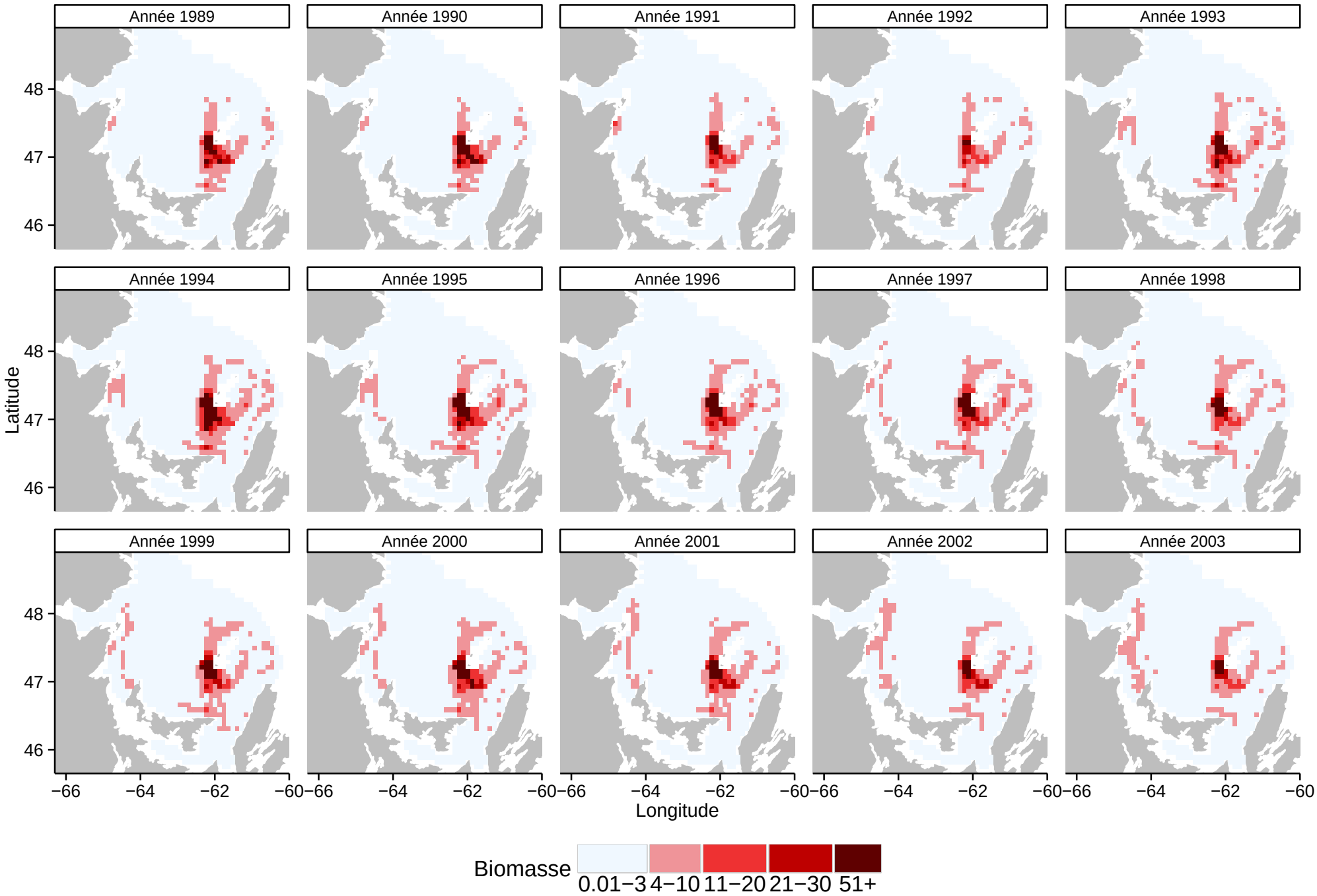

FIGURE 5.16 : Biomasse de concombre de mer prédite sur l'ensemble su sGSL pour les années 1989 à 2003 (grille régulière, $\phi = 18$ et $G = 63$).

Le paramètre δ, dont la moyenne *a posteriori* est proche de 1, atteste d'une certaine stabilité temporelle entre deux années consécutives. Cette stabilité temporelle est également illustrée par la faible valeur de σ_v, l'écart type du bruit blanc. L'interpolation des quantités de biomasses de concombre est représentée en Figure 5.16. Ces cartes illustrent la persistance des concombres de mer au nord-est des iles du Prince Edward. De faibles densités sont également prédites à l'ouest du sGSL, elles ont tendance à occuper une plus grande surface au cours du temps. Les prédictions pour les années 2004 et 2005 sont illustrées par la figure et le tableau 5.16. La prédictions des quantités de biomasses positives est compliquée. Néanmoins, les zones à fortes biomasses sont convenablement prédites comme les zones d'absences. Cependant, quelques quantités positives ne sont pas correctement prédites comme celles présentes à l'embouchure du Saint-Laurent.

Tableau 5.16 : Proportion des prédictions de la biomasse de concombre de mer pour les années 2004 à 2005.

Année	Vrai positif	Faux zéro	Vrai zéro	Faux positif	RMSE
2004	0.42	0.58	0.85	0.15	1.00
2005	0.47	0.53	0.85	0.15	2.12

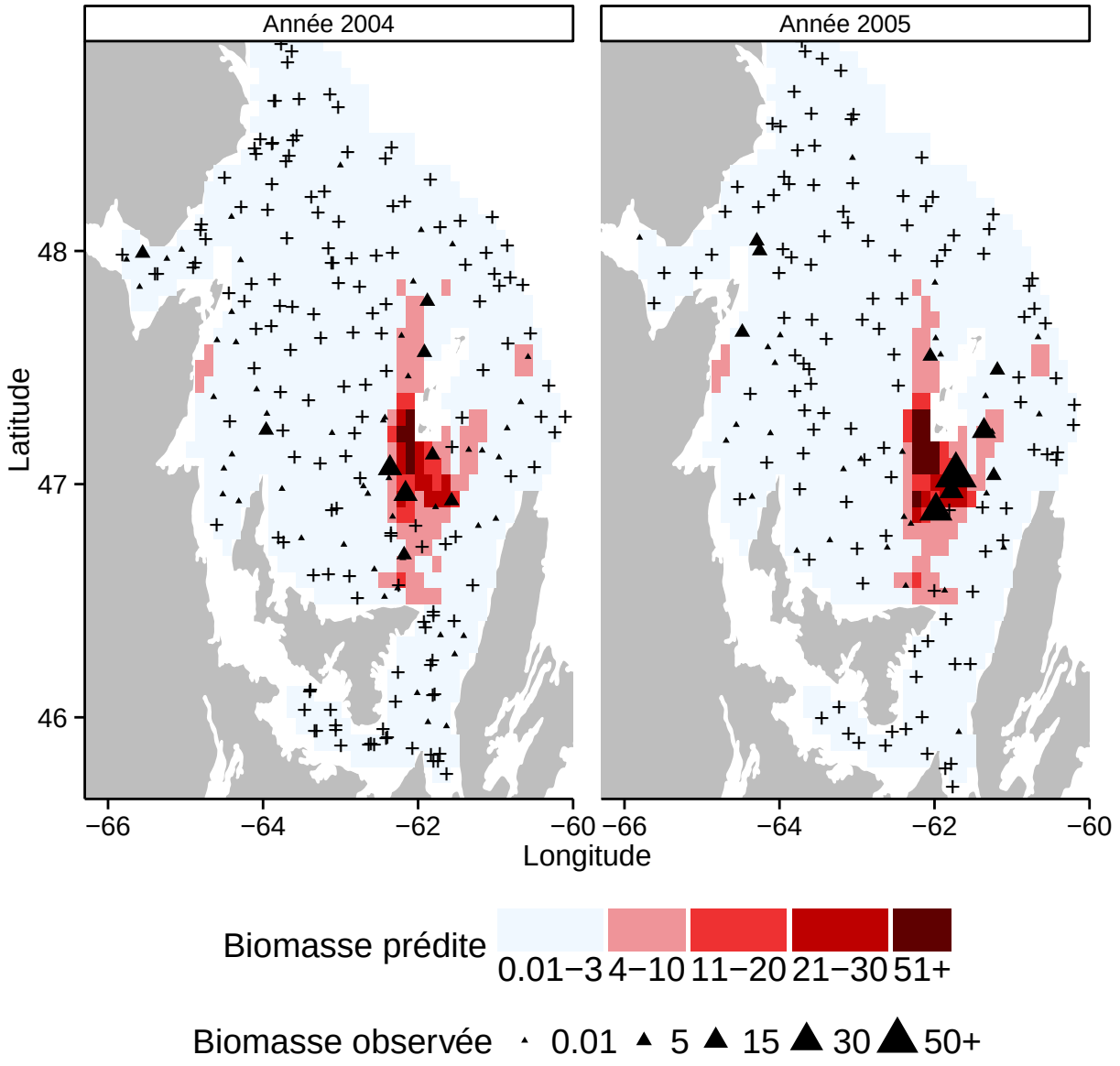

FIGURE 5.17 : Prédictions et observations des quantités de biomasses de concombre de mer sur l'ensemble du sGSL pour les années 2004 et 2005 (les croix représentent les quantités de biomasses observées nulles).

5.5 CONCLUSION

L'approche par convolution discrète est un outil très prometteur pour la modélisation spatiale, mais surtout pour la modélisation spatio-temporelle. Les premières applications dans le domaine de l'océanographie (Higdon, 1998) ou des études sur la qualité de l'air (Calder, 2007) n'ont que peu été développées pour écologie (Allen *et al.*, 2001). Néanmoins, les adaptations proposée dans ce chapitre permettent de prendre en compte les spécificités des données écologiques (p. ex. forte proportion de zéros, dépendance spatiale et temporelle, effets de variables environnementales) au sein d'une approche de modélisation flexible et performante. En effet, la réduction de dimension effectuée par l'utilisation d'une grille latente, support de la dépendance spatiale et temporelle, permet de travailler avec un grand jeu de données (2628 observations entre l'année 1989 et l'année 2003), ce qui d'un point de vue computationnel est difficile avec les outils usuels de la géostatistique. Les perspectives et améliorations possibles à court terme, mais également à long terme sont discutées dans le chapitre 6.

6

PERSPECTIVES

La modélisation sur grille d'un processus spatio-temporel latent proposée dans le chapitre précédent (chapitre 5) ouvre beaucoup de perspectives en exploitant les possibilités d'extensions modulatoires de la structure hiérarchique. Ainsi dans ce chapitre, je propose plusieurs voies de recherches permettant d'approfondir et de complexifier la construction de modèles spatio-temporels utiles pour l'écologie.

6.1 INFLUENCE DE DIFFÉRENTES ÉCHELLES SPATIALES

L'échelle spatiale à laquelle la distribution d'une espèce est modélisée est d'une très grande importance (Pearson *et al.*, 2004; Guisan et Thuiller, 2005). En effet, les relations entre l'espèce et son environnement induisent des structures spatiales de distribution qui peuvent être observées à différentes échelles. Ces structures spatiales observées à larges échelles sont généralement attribuées à des variables climatiques. La répartition spatiale à petite échelle, qui peut prendre la forme de patch, est quant à elle généralement due à des modifications physico-chimiques, biotiques très locales du milieu ou de la fragmentation d'habitat (Scott, 2002; Guisan et Thuiller, 2005).

Les relevés effectués au chalut de fond par Pêche et Océan Canada recouvrent l'ensemble du sud du golfe du Saint-Laurent et permettent de modéliser la distribution spatio-temporelle d'espèces d'invertébrés épibenthiques à l'échelle du golfe du Saint-Laurent (chapitre 5). Cependant, les études de distribution d'espèces benthiques marines sont généralement effectuées sur de faibles surfaces et à l'échelle de l'organisme ($< 1\ m$). Les processus océanographiques utilisés dans ces études couvrent généralement une large échelle spatiale ($10 - 100\ km$), créant ainsi un déséquilibre entre les deux échelles (Dethier et Schoch, 2005). Afin de prendre en compte les effets environnementaux à

échelles régionale et locale, des adaptations de protocole d'échantillonnage ont été proposées (Edgar et Barrett, 2002; Ysebaert et Herman, 2002). Le protocole d'échantillonnage ainsi proposé consiste à visiter à plusieurs reprises un même site tout en échantillonnant la plus grande zone possible. Ceci permet de mesurer conjointement la présence ou l'abondance de l'espèce avec plusieurs variables environnementales locales supposées explicatives à une petite échelle. Cependant, ce type de protocole peut être difficilement réalisable (p. ex. coût, temps). C'est pourquoi il est possible d'utiliser l'approche de modélisation par convolution discrète afin de prendre en compte les différentes échelles spatiales influençant la distribution d'une espèce.

Les modèles spatio-temporels par convolution discrète à résolutions multiples sont utiles pour considérer les variations du processus étudié à différentes échelles spatiales (Higdon, 2002). Cette approche permet d'améliorer la représentation du phénomène spatio-temporel analysé par la prise en compte d'effets à petite et grande échelles. Dans cet exemple, le modèle des observations sélectionné est le modèle CPG avec :

$$\begin{aligned} Y(t,s) &\sim CPG(\lambda(t,s),a,b) \\ \lambda(t,s) &= exp(z(t,s)) \end{aligned} \tag{6.1}$$

La couche latente du modèle spatio-temporel composée de deux résolutions spatiales différentes s'écrit par exemple :

$$z(t,s) = \mu + \sum_{l=1}^{G_L} \kappa^L(s-\omega_l)x_L(t,\omega_l) + \sum_{r=1}^{G_R} \kappa^R(s-\omega_r)x_R(t,\omega_r)$$

$$\begin{aligned} x_L(t,\omega_l) &= x_L(t-1,\omega_l) + \nu(t,l), \quad \forall\, l \\ \nu_{l,g} &\overset{i.i.d.}{\sim} Normal(0,\sigma_\nu^2) \\ x_R(t,\omega_r) &= x_R(t-1,\omega_r) + \tau(t,r), \quad \forall\, r \\ \tau_{l,g} &\overset{i.i.d.}{\sim} Normal(0,\sigma_\tau^2) \end{aligned} \tag{6.2}$$

où G^L est le nombre de points de grille à résolution locale associé au noyau de convolution κ^L et G^R est le nombre de points de grille à résolution régionale associé au noyau de convolution κ^R. Dans l'équation 6.2, les deux résolutions (locale et régionale) sont dépendantes du temps. Or, les processus étudiés montrent

généralement peu de dépendance temporelle à résolution locale (Calder, 2003). Cependant, cette résolution locale peut améliorer les capacités prédictives du modèle en prenant en compte des dépendances spatiales à petite échelle (Wikle, 1999). Dans l'exemple précédent, deux résolutions spatiales sont utilisées, mais il est parfaitement possible d'utiliser plus de deux résolutions. Néanmoins, pour que la réduction de dimension effectuée par l'utilisation d'une ou de plusieurs grilles soit profitable, le nombre de points de grilles et donc le nombre de grilles doient rester limité.

La construction des deux grilles latentes à résolutions locale et régionale peut se faire par grille régulière recouvrant l'ensemble de la zone d'étude. Cette méthode permet de contrôler facilement la position d'une grille par rapport à l'autre et ainsi de proposer un ensemble de points de grille réparti uniformément sur l'ensemble de la zone d'étude.

Comme détaillé précédemment (cf chapitre 5), l'utilisation de deux grilles irrégulières est également envisageable et permet de placer astucieusement les nœuds qui les composent. Pour ce faire, nous proposons une méthode scindée en deux étapes. Dans un premier temps, la grille à résolution locale est construite à l'aide d'un algorithme des *k-means*. L'ensemble des positions des observations (longitudes-latitudes) est utilisé pour définir la position des points de la grille à résolution fine. À cette grille est associée un noyau de convolution κ^R dont la portée effective est volontairement choisie petite. Ensuite, la seconde grille à résolution régionale est construite afin de recouvrir le plus entièrement possible la zone d'étude. Un algorithme de type recuit simulé avec marche aléatoire pour la construction de cette grille régionale pourrait être le suivant :

1. à l'itération i, G^R localisations $(\omega_1, \omega_r, ..., \omega_R)$ sont proposées.

2. Le produit suivant qui évalue la valeur de la distribution cible à l'itération i est calculé :

$$Q = \prod_{r=1}^{G^R} \prod_{l=1}^{G^L} \prod_{r'=1}^{G^R} (\omega_r - \omega_l)^2 (\omega_r - \omega_{r'})^2$$

3. Une nouvelle localisation est ensuite proposée au hasard pour le site r et le produit Q est à nouveau calculé.

4. Le rapport $h = \frac{Q_{i-1}}{Q_i}$ est calculé, la localisation proposée est alors conservée avec une probabilité $min(1, h)$; sinon la localisation précédente est conservée.

5. Cette procédure est répétée pour chacun des G^R points de la grille.

6. Après avoir proposé une nouvelle localisation pour chacun des points, la procédure est répétée un grand nombre de fois. Les localisations des G^R points sélectionné est le design qui minimise le produit Q.

Les deux grilles ainsi construites recouvrent astucieusement la zone d'étude. Cependant, le nombre de points de chacune des grilles est déterminant si l'on souhaite que la réduction de dimension soit efficace. En considérant l'application de cette méthode aux invertébrés épibenthiques du sGSL le nombre total de nœuds ne devrait pas excéder 85 étant donné que le nombre moyen d'observations par année est de 176 sur la période 1989-2003. Un partage intéressant serait par exemple de composer une grille locale avec 60 points, puis de répartir 25 points sur l'ensemble du golfe qui représenterait alors la grille régionale. Cependant, plusieurs grilles composées d'un nombre de points différents peuvent être testées et évaluées par des critères d'adéquation aux données observées comme le DIC, ou encore par des critères mesurant les capacités prédictives des modèles considérés.

6.2 LA DISPERSION

Le modèle spatio-temporel présenté dans le chapitre 5 ne prend pas en compte de manière explicite la dispersion éventuelle de l'espèce sur la zone d'étude. Or, la plupart des espèces marines libèrent leurs gamètes directement dans l'océan, permettant alors une grande diffusion (Kinlan et Gaines, 2003). De ce fait, prendre en considération la dispersion, qui est donc un facteur important dans la détermination de l'aire de répartition d'une espèce, permet d'obtenir des modèles plus précis (Robinson *et al.*, 2011). C'est pourquoi nous travaillons avec une approche de modélisation spatio-temporelle par convolution discrète, qui permet de prendre en compte cette dispersion. Le modèle des observations CPG est encore une fois choisi comme exemple : la dispersion est alors ajoutée dans la couche latente

et plus particulièrement au niveau des points de grille. Cette approche proposée par Cressie et Wikle (2011) est proposée sous forme matricielle :

$$\begin{aligned} Z_t &= \mu + KX_t \\ X_t &= MX_{t-1} + \eta_t \end{aligned} \tag{6.3}$$

où η_t est un terme d'erreur, X_t sont les valeurs des points de grille qui permettent de construire le champ spatio-temporel latent z_t par convolution discrète avec le noyau K. La dispersion est modélisée par l'intermédiaire de la matrice M, qui est appelée matrice de propagation (Cressie et Wikle, 2011). Cette matrice peut être définie par une équation de réaction-diffusion comme proposée par Wikle (2003) ou Hooten et Wikle (2007). Dans cet exemple les points de grille X sont remplacés par u pour une meilleure lisibilité (x et y étant ici les coordonnées des points de grilles) :

$$\frac{\partial u}{\partial t} = \frac{\partial}{\partial x}\left(\delta(x,y)\frac{\partial u}{\partial x}\right) + \frac{\partial}{\partial y}\left(\delta(x,y)\frac{\partial u}{\partial y}\right) \tag{6.4}$$

$\delta(x,y)$ représente le coefficient de diffusion variant spatialement. En considérant le processus latent u_t sur une grille d'espacement Δ_x pour la longitude et Δ_y pour la latitude avec un pas de temps Δ_t, on peut discrétiser l'équation 6.4 pour obtenir u_t sous la forme :

$$\begin{aligned} u_t(x,y) = {} & u_{t-\Delta_t}(x,y)\left[1 - 2\delta(x,y)\left(\frac{\Delta_t}{\Delta_{x^2}} + \frac{\Delta_t}{\Delta_{y^2}}\right)\right] \\ & + u_{t-\Delta_t}(x-\Delta_x,y)\left[\frac{\Delta_t}{\Delta_{x^2}}\left\{\delta(x,y) - \frac{(\delta(x+\Delta_x,y) - \delta(x-\Delta_x,y))}{4}\right\}\right] \\ & + u_{t-\Delta_t}(x+\Delta_x,y)\left[\frac{\Delta_t}{\Delta_{x^2}}\left\{\delta(x,y) + \frac{(\delta(x+\Delta_x,y) - \delta(x-\Delta_x,y))}{4}\right\}\right] \\ & + u_{t-\Delta_t}(x,y+\Delta_y)\left[\frac{\Delta_t}{\Delta_{y^2}}\left\{\delta(x,y) + \frac{(\delta(x,y+\Delta_y) - \delta(x,y-\Delta_y))}{4}\right\}\right] \\ & + u_{t-\Delta_t}(x,y-\Delta_y)\left[\frac{\Delta_t}{\Delta_{y^2}}\left\{\delta(x,y) - \frac{(\delta(x,y+\Delta_y) - \delta(x,y-\Delta_y))}{4}\right\}\right] \\ & + \eta_t(x,y) \end{aligned} \tag{6.5}$$

η_t est le terme d'erreur précédemment évoqué qui prend notamment en compte les erreurs dues à la discrétisation. Le processus u obtenu s'écrit sous la forme matricielle par :

$$u_t = M(\delta, \Delta_t, \Delta_x, \Delta_y)u_{t-\Delta_t} + M_B(\delta, \Delta_t, \Delta_x, \Delta_y)u^B_{t-\Delta_t} + \eta_t \quad (6.6)$$

u_t correspond toujours au processus latent, M représente la matrice de propagation dépendante du coefficient de diffusion δ, du pas de temps Δ_t ainsi que les pas de discrétisation de la grille Δ_x et Δ_y. M est une matrice clairsemée de taille $n \times n$ avec cinq diagonales non nulles correspondant aux coefficients entre crochets de l'équation 6.5. Afin de prendre en compte les effets de bords, le produit $M_B(\delta, \Delta_t, \Delta_x, \Delta_y)u^B_{t-\Delta_t}$ est ajouté dans la définition du processus u_t. Ce produit est composé d'une matrice M_B de dimension $n \times n_B$ ainsi que du processus u_t aux limites de la grille de discrétisation notée u^B_t, un vecteur de taille $n \times n_B$. Cette approche de modélisation est utile pour représenter la dispersion d'une espèce sur un territoire et est particulièrement adaptée aux espèces invasives (Wikle, 2003; Hooten et Wikle, 2007).

6.3 COMMENT PRENDRE EN COMPTE LES INTERACTIONS ENTRE ESPÈCES ?

La majorité des modèles permettant de prédire la distribution d'une espèce n'inclut pas de manière explicite les interactions entre celles-ci (Guisan *et al.*, 2006). Ignorer ces interactions peut affaiblir les prédictions et projections effectuées par ce type de modèle (Austin, 2002). En effet, un grand nombre d'interactions (p. ex. compétition, mutualisme, parasitisme, prédation, *ect*) peuvent être responsables de la distribution spatio-temporelle d'une espèce. Plus particulièrement, il a été montré que la compétition joue un rôle important lors de la construction de modèles de distribution d'espèces terrestres (Guisan et Thuiller, 2005; Araújo et New, 2007; Heikkinen *et al.*, 2007). Les relations trophiques (p. ex. prédation) ont quant à elles plutôt été étudiées pour des modèles s'intéressant aux espèces marines (Redfern *et al.*, 2006; Torres *et al.*, 2008).

L'ensemble des espèces d'invertébrés échantillonnées par Pêche et Océan Canada présentent plusieurs prédateurs généralistes (p. ex. étoiles de mer, buccins) dont la répartition spatiale peut

être modélisée conjointement avec des espèces de proies (p. ex. oursins, ophiures). En effet, les étoiles de mer se nourrissent généralement de bivalves, mais sont également des prédateurs d'oursins juvéniles et d'ophiures (Himmelman et Dutil, 1991; Nishizaki et Ackerman, 2006). La figure 6.1 présente les quantités de biomasses récoltées au cours de l'année 1997 pour l'espèce étoile de mer et ces deux proies potentielles : oursins juvéniles et ophiures.

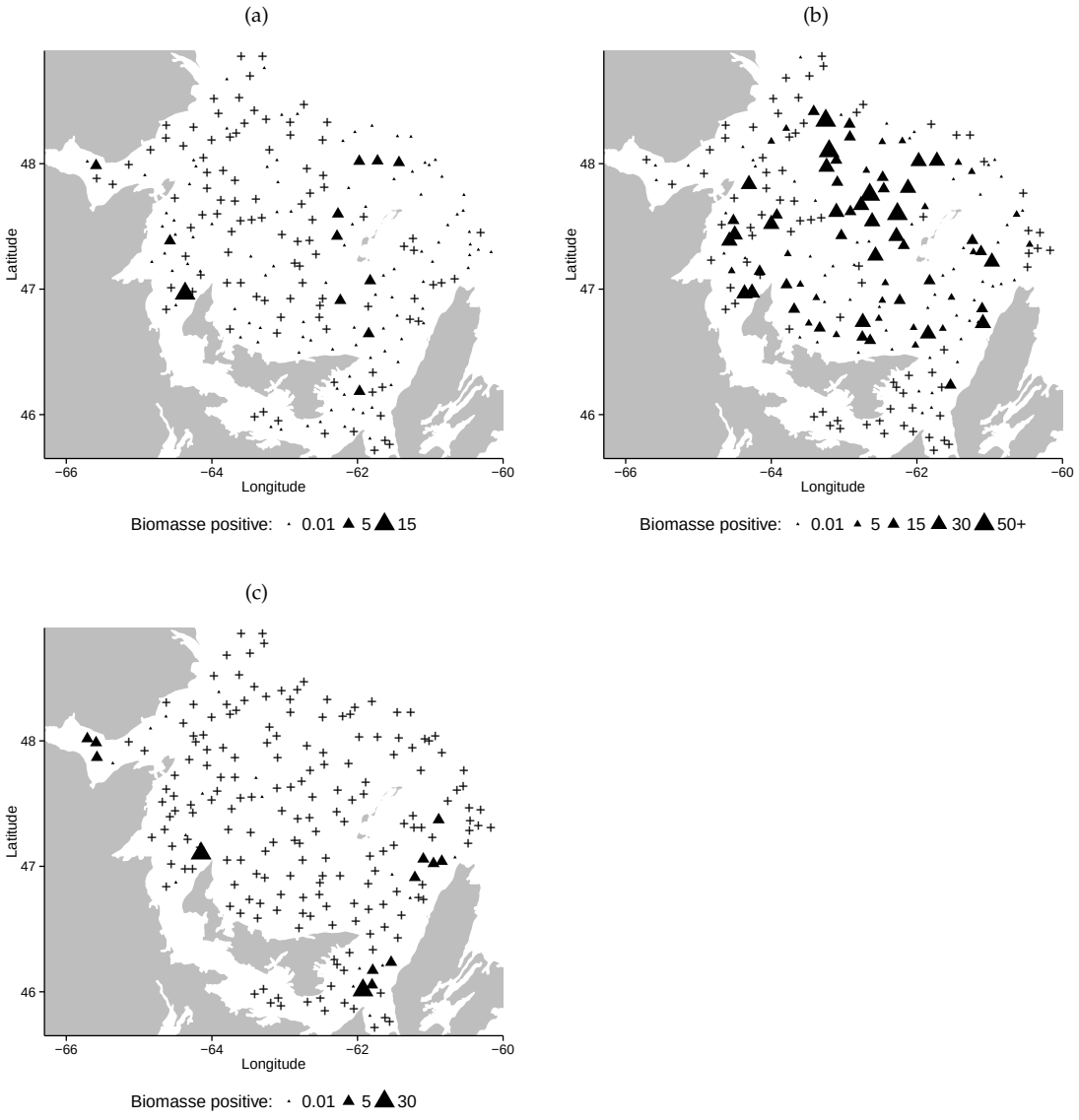

FIGURE 6.1 : Quantités de biomasses récoltées en 1997 dans le golfe du Saint-Laurent : (a) étoile de mer , (b) oursin, (c) ophiure.

Dans cette section, un modèle spatial et un modèle spatio-temporel basés sur l'approche par convolution discrète sont proposés pour tenir compte des interactions entre espèces et notamment du phénomène de prédation.

6.3.1 *Modèle spatial multi-espèces*

L'approche par convolution offre également une piste intéressante pour la construction d'un modèle spatial multi-espèces. Considérons deux espèces d'invertébrés dont les biomasses Y_1 et Y_2 sont récoltées dans le sGSL. Le modèle des observations pour chacune des quantités Y_1 et Y_2 est un modèle CPG avec :

$$\begin{aligned} Y_1(s) &\sim CPG(\lambda_1(s), a_1, b_1) \\ Y_2(s) &\sim CPG(\lambda_2(s), a_2, b_2) \end{aligned} \tag{6.7}$$

ici $\lambda_1(s)$ désigne le nombre de patchs moyen récolté au site s pour l'espèce 1, a_1 et b_1 contrôlent la biomasse contenue dans chacun de ces patchs. $\lambda_2(s)$ désigne quant à lui le nombre de patchs moyen récolté au site s pour l'espèce 2 associée aux paramètres a_2 et b_2. L'introduction de la dépendance spatiale s'effectue au niveau du nombre de patchs moyen pour chacune des espèces :

$$\begin{aligned} log(\lambda_1(s)) &= \mu_1 + \sum_{g=1}^{G} \kappa(s-\omega)x(\omega) \\ log(\lambda_2(s)) &= \mu_2 + \sum_{g=1}^{G} \kappa(s-\omega)x(\omega) \end{aligned} \tag{6.8}$$

où μ_1 et μ_2 représentent les effets fixes de chacune des espèces sur l'ensemble du sGSL, x sont les effets aléatoires communs à chacune des espèces. Dans cet exemple, les deux espèces partagent la même grille latente et le même noyau κ, et sont dépendantes spatialement par l'intermédiaire des effets aléatoires x. Cependant, il semble peu vraisemblable que les deux espèces partagent les mêmes effets aléatoires. Ancelet (2008) a proposé l'utilisation d'un champ gaussien bivarié permettant de modéliser de façon plus réaliste la relation entre les deux quantités Y_1 et Y_2 :

$$
\begin{aligned}
log(\lambda_1(s)) &= \mu_1 + \sum_{g=1}^{G} \kappa(s-\omega)x_1(\omega) \\
log(\lambda_2(s)) &= \mu_2 + \frac{\rho}{\sqrt{\rho^2 + (1-|\rho|)^2}} \sum_{g=1}^{G} \kappa(s-\omega)x_1(\omega) \\
&\quad + \frac{(1-|\rho|)}{\sqrt{\rho^2 + (1-|\rho|)^2}} \sum_{g=1}^{G} \kappa(s-\omega)x_2(\omega) \qquad (6.9)
\end{aligned}
$$

où ρ contrôle la relation entre les deux champs spatiaux dont la valeur est comprise entre -1 et 1. Lorsque ce paramètre est égal à 1, les deux espèces sont corrélées positivement, inversement si $\rho = -1$ alors les deux espèces sont corrélées négativement. Enfin, si $\rho = 0$ la variabilité spatiale de l'espèce 2 est indépendante de celle de l'espèce 1.

Dans un souci de parcimonie, la biomasse contenue dans un patch (M) n'est pas considérée comme spatialement structurée pour chacune des espèces :

$$
\begin{aligned}
M_1 &\sim Gamma(a_1, b_1) \\
M_2 &\sim Gamma(a_2, b_2) \qquad (6.10)
\end{aligned}
$$

Pour une modélisation plus fine entre chaque espèce, il est possible de spécifier deux noyaux de convolution différents pour chacune des espèces étudiées. En effet, la structure spatiale des champs latents peut varier selon l'espèce considérée. Il est par exemple envisageable de proposer deux noyaux exponentiels dont les portées respectives diffèrent.

6.3.2 *Modèle spatio-temporel multi-espèces*

Le modèle spatial multi-espèces présenté précédemment peut être généralisé à un modèle spatio-temporel. Les quantités de biomasses de chacune des espèces sont désormais indicées par le temps $Y_1(t,s)$ et $Y_2(t,s)$. La dépendance temporelle est ajoutée au nombre de patchs moyen :

$$\lambda_1(t,s) = \mu_1 + \sum_{g=1}^{G} \kappa_1(s-\omega)x_1(t,\omega)$$
$$\lambda_2(t,s) = \mu_2 + \sum_{g=1}^{G} \kappa_2(s-\omega)x_2(t,\omega) \qquad (6.11)$$

Dans cet exemple, les distributions spatiales des deux espèces sont indépendantes si x_1 et x_2 le sont. Cependant, la dépendance temporelle est modélisée conjointement pour les deux espèces. Par exemple, Allen *et al.* (2001) décrivent un modèle hôte-parasitoïde où la densité d'une espèce à l'instant t joue un rôle direct sur la densité de l'autre espèce et inversement. Dans l'exemple des invertébrés épibenthiques, il est envisageable de considérer un modèle proie-prédateur avec l'espèce oursin (proie) et l'espèce étoile de mer (prédateur). Un modèle temporel simple peut s'inspirer des équations de Lotka-Volterra pour décrire cette relation non linéaire sous forme stochastique :

$$x_1(t,\omega) = x_1(t-1,\omega(A - Bx_2(t-1,\omega) + \nu_1(t)$$
$$x_2(t,\omega) = -x_2(t-1,\omega)(C - Dx_2(t-1,\omega)) + \nu_2(t) \qquad (6.12)$$

où A est le taux de croissance de la proie, B le taux de prédation exercé par le prédateur sur cette proie, C le taux de mortalité du prédateur et D le taux de croissance de ce même prédateur. $\nu_1(t)$ et $\nu_2(t)$ sont des perturbations aléatoires indépendantes qui représentent les facteurs inconnus non pris en compte, mais qui interviennent dans la dynamique de ces deux espèces.

6.4 COMMENT PRENDRE EN COMPTE PLUSIEURS SOURCES DE DONNÉES ?

Pour la gestion des stocks de poissons de fond au large de Vancouver, la station biologique du Pacifique de Pêches et Océans Canada organise des relevés scientifiques au chalut. La localisation des zones chalutées est tirée au hasard au début de chaque relevé pour assurer un échantillonnage aléatoire simple. Ces relevés au chalut sont ensuite utilisés pour construire des in-

dices d'abondance qui alimentent des modèles de dynamique de population. Par ailleurs, les pêcheries commerciales fournissent une importante quantité de données supplémentaires sur l'abondance de ces espèces de poissons de fonds, mais les zones de captures ne sont, bien sûr, pas choisies au hasard. La figure 6.2 représente les quantités de biomasses récoltées en 2009 par pêches scientifiques et pêches commerciales pour une espèce de sole (*Microstomus pacificus*). Ces cartes illustrent la différence des zones visitées par les deux types de pêches, mais également la différence entre les quantités récoltées par chacune.

Pour améliorer les indices d'abondances construits avec les pêches scientifiques grâce aux données des pêches scientifiques, il est souhaitable de proposer une approche de modélisation capable d'utiliser toute l'information disponible sur les espèces étudiées (pêches scientifiques et commerciales). Dans ce cas, les modèles hiérarchiques bayésiens et plus particulièrement les modèles hiérarchiques spatiaux par convolution discrète sont des outils performants qui permettent de prendre en compte plusieurs sources de données.

6.4.1 *Variables latentes*

Les variables latentes communes, représentant l'abondance potentielle de la zone, assurent la cohérence spatiale entre les deux sources de données. Un champ latent spatial z est construit par convolution discrète de variables normales indépendantes sur grille x avec un noyau de convolution κ (section 5.1.2) :

$$z(s) = \mu + \sum_{g=1}^{G} \kappa(s - \omega_g) x(\omega_g). \tag{6.13}$$

avec μ la moyenne du champ spatial z. L'utilisation d'une fonction de lien logarithmique, $\lambda_s = exp(z_s)$, permet de s'intéresser à l'intensité de la loi de Poisson utilisée dans le modèle CPG (section 5.2).

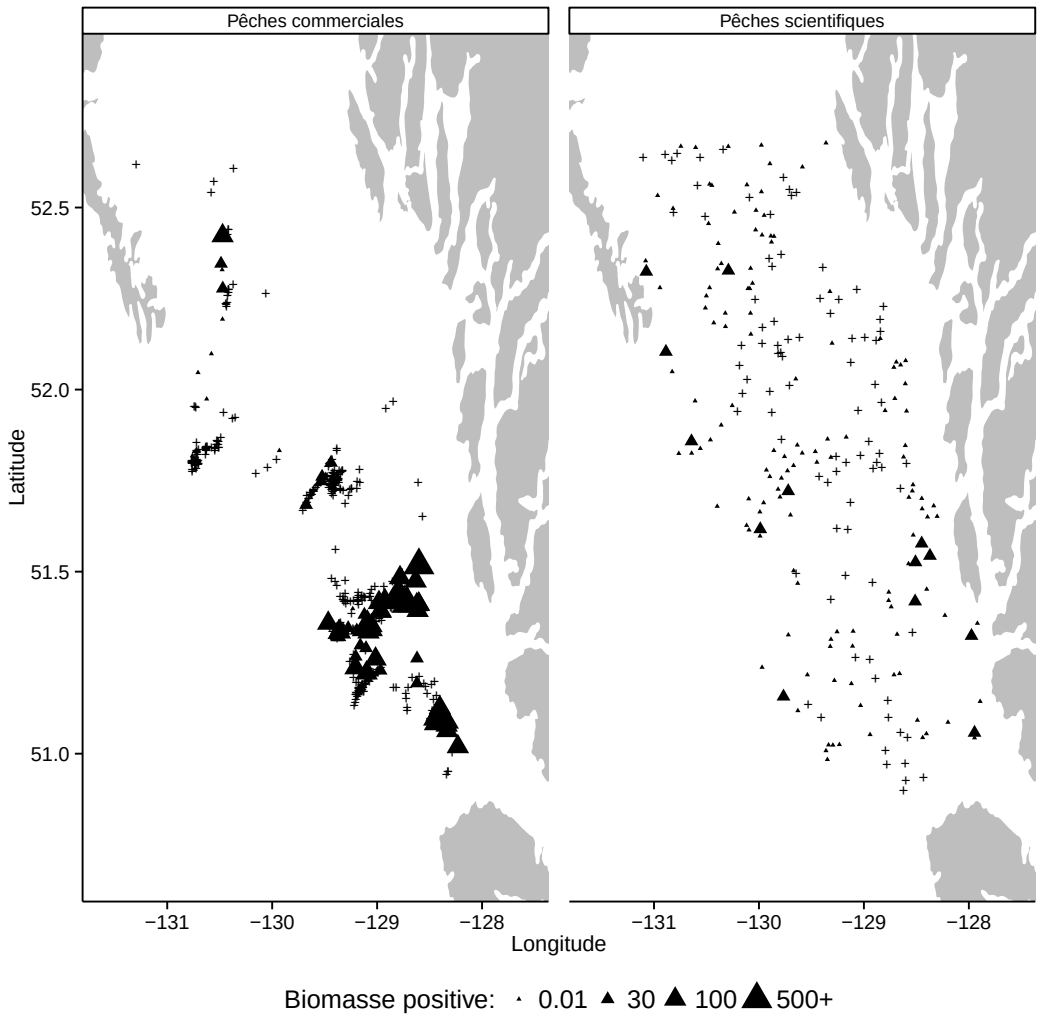

FIGURE 6.2 : Quantités de biomasses de sole (*Microstomus pacificus*) récoltées par pêches commerciales et pêches scientifiques en 2009 dans le bassin de la Reine Charlotte et la partie méridionale du détroit de Hecate (Colombie Britannique, Canada).

6.4.2 *Modèles des observations*

6.4.2.1 *Données de pêches scientifiques*

Les quantités de biomasses récoltées par les pêches scientifiques sont notées Y_1 et l'effort d'échantillonnage (p. ex. durée, distance) est noté E_1. Le nombre de patchs récoltés lors d'un trait de chalut est donc :

$$N_1(s) \sim Poisson(E_1(s) \times \lambda(s)) \tag{6.14}$$

Chaque patch de biomasse est distribué selon une loi gamma de paramètres a_1 et b_1. La biomasse totale récoltée au site s est donc distribuée selon un modèle CPG :

$$\begin{aligned} &Y_1(s) \sim Gamma(a_1(s) \times N_1(s), b_1(s)) \text{ si } N_1(s) > 0 \\ &Y_1(s) = 0 \text{ si } N_1(s) = 0 \end{aligned} \tag{6.15}$$

6.4.2.2 *Echantillonnage préférentiel des pêches commerciales*

On note $s \in S_2$ les sites de pêches commerciales (S_1 pour les sites échantillonnés par pêches scientifiques, et S pour l'ensemble de la zone). En supposant que les sites de pêches commerciales ne sont pas visités au hasard, une manière de prendre en compte cet échantillonnage préférentiel est de modéliser le processus ponctuel formé par les sites de pêches S_2 par un processus de Poisson inhomogène d'intensité $\lambda(.)$ et donc réglé par la biomasse :

$$P_\theta(s) = \frac{\lambda(s)}{\lambda(S)}, \quad \text{où} \quad \lambda(S) = \int_{s \in S} \lambda(s) ds$$

de sorte que la composante de la vraisemblance de l'échantillonnage préférentiel est le produit

$$[S_1|\mu] = \left(\prod_{s_1 \in S_1} P(s_1) \right)$$

Remarque : on travaille à nombre de points total fixés, et donc la vraisemblance du semis de points est exprimé conditionnellement on nombre de sites de pêches présents dans les données.

6.4.2.3 *Données de pêches commerciales*

Les données de pêches commerciales notées Y_2 présentent généralement des efforts d'échantillonnage (E_2) très variables. Cet effort d'échantillonnage permet de prendre en compte la différence de pratique des pêcheries commerciales et scientifiques par l'intermédiaire du nombre de patchs récoltés :

$$N_2(s) \sim Poisson(E_2(s) \times \lambda(s)) \tag{6.16}$$

Ensuite, les quantités de biomasses récoltées par les pêches commerciales sont également distribuées selon un modèle CPG :

$$\begin{aligned} &Y_2(s) \sim Gamma(a_2(s) \times N_2(s), b_2(s)) \text{ si } N_2(s) > 0 \\ &Y_2(s) = 0 \text{ si } N_2(s) = 0 \end{aligned} \tag{6.17}$$

Afin de prendre en compte les capacités supérieures de captures des pêcheries commerciales comparativement aux pêches scientifiques, la biomasse contenue dans un patch récolté lors d'un trait de chalut commercial est distribuée selon une loi gamma de paramètres a_2 et b_2. La biomasse contenue dans un patch est donc différente selon le type de pêche considéré.

Ce modèle spatial peut être étendu à un modèle spatio-temporel en définissant par exemple un modèle de transition temporelle autorégressif AR1 :

$$z(t,s) = \mu + \sum_{g=1}^{G} \kappa(s - \omega_g) x(t, \omega_g)$$

$$\begin{aligned} x(t,\omega_g) &= \delta x(t-1,\omega_g) + \nu(t,g) \\ \nu(t,g) &\overset{i.i.d.}{\sim} Normal(0, \sigma_\nu^2) \end{aligned} \tag{6.18}$$

Il est bien entendu possible (et souhaitable) d'introduire des variables environnementales dans le champ latent z comme présenté dans le chapitre 5.

6.5 VARIABLES ENVIRONNEMENTALES

Dans les chapitres précédents (cf chapitres 4 et 5) plusieurs variables environnementales ont été incluses dans la modélisation du nombre de patchs N récoltés lors d'un trait de chalut. Foster et Bravington (2012) ont proposé d'introduire également des variables explicatives dans la seconde partie de la couche latente du modèle CPG, c'est-à-dire dans la biomasse contenue dans chaque patch M_i. Les choix adoptés lors de ces travaux de thèse ont été de modéliser uniquement le nombre de patchs. Néanmoins, il est par exemple envisageable d'introduire la température dans la modélisation de la biomasse contenue dans chaque patch étant donné le faible effet de cette dernière sur le nombre de patchs (cf chapitres 4 et 5). Il est alors possible d'utiliser un modèle linéaire généralisé (Zuur *et al.*, 2009) :

$$\mathbb{E}(M(s)) = \frac{a}{b}(s) = exp(\gamma + \eta Temp(s)) \tag{6.19}$$

avec γ l'effet moyen et η l'effet associé à la température. Il est également possible d'introduire toutes les variables explicatives disponibles dans chaque partie du modèle CPG. Si tel est le cas, le modèle proposé s'apparente alors à un modèle Tweedie (cf chapitres 2.1 et 3).

6.6 DELTA LOG-NORMAL

Des alternatives au modèle des observations CPG peuvent également être utilisées dans le cadre de la modélisation spatio-temporelle par convolution pour des données à forte proportion de zéros. Le modèle delta log-normal qui fait partie de la famille des «two parts model» est une de ces alternatives. Comme présentée dans le chapitre 2.1, la modélisation de la biomasse est séparée en deux parties. La première partie consiste à modéliser la présence de l'espèce, *Pres*, qui est une variable binaire prenant la valeur 1 si l'espèce est présente et 0 si elle est absente :

$$Pres(t,s) \sim Bernoulli(\pi(t,s)) \tag{6.20}$$

$\pi_{t,s}$ est la probabilité de présence de l'espèce à l'année t et au site s. Cette probabilité de présence peut être modélisée

par les méthodes de convolution discrète présentée au chapitre précédent. Une fonction de lien probit est alors utilisée pour construire le champ spatial par convolution discrète dont α_{pres} est la moyenne :

$$\begin{aligned} probit(\pi(t,s)) &= \alpha_{pres} + \sum_{g=1}^{G} \kappa(s-\omega_g) x_{pres}(t,\omega_g) \\ x_{pres}(t,\omega_g) &= \delta\ x_{pres}(t-1,\omega_g) + \nu(t,g) \\ \nu(t,g) &\overset{i.i.d.}{\sim} Normal(0,\sigma_\nu^2) \end{aligned} \tag{6.21}$$

De cette manière, la probabilité de présence de l'espèce est modélisée spatialement et temporellement. La seconde partie du modèle s'intéresse uniquement aux quantités de biomasses strictement positives Y^{pos} qui sont distribuées selon une loi log-normale :

$$log(Y^{pos}(t,s)) \sim Normal(\mu(t,s),\sigma^2) \tag{6.22}$$

où μ et σ^2 sont respectivement la moyenne et la variance de la loi normale. Ces quantités strictement positives sont également modélisées grâce à un modèle spatio-temporel par convolution discrète :

$$\begin{aligned} log(\mu(t,s)) &= \alpha_{pos} + \sum_{g=1}^{M} \kappa(\omega_g - s) x_{pos}(\omega_g,t) \\ x_{pos}(t,\omega_g) &= x_{pos}(t-1,\omega_g) + \eta(t,g) \\ \eta(t,g) &\overset{i.i.d.}{\sim} Normal(0,\sigma_\eta^2) \end{aligned} \tag{6.23}$$

La moyenne du champ spatial est α_{pos} et le modèle de dépendance temporel est un autorégressif d'ordre 1. Dans cet exemple, les variations temporelles sont modélisées dans chaque partie du modèle delta, néanmoins cette dépendance temporelle peut être incluse dans une seule partie de ce modèle. Afin d'être consistant avec la modélisation spatio-temporelle proposée pour le modèle des observations CPG, dont les variations temporelles sont prises en compte dans la partie présence-absence, mais aussi dans la partie strictement positive par l'in-

termédiaire du paramètre λ, les variations temporelles sont incluses dans chaque partie du modèle delta.

Cette approche présente plusieurs avantages : chaque partie du modèle peut être traitée séparément et facilite l'implémentation du modèle. De plus, l'existence d'une représentation normale du lien profit facilite l'utilisation de lois normales pour l'inférence bayésienne de ce modèle (Albert et Chib, 1993). Ce modèle peut être préféré au modèle CPG lorsque la séparation des données en présence-absence et des données strictement positives est justifiée (Foster et Bravington, 2012).

6.7 POUR ALLER PLUS LOIN DANS LA MODÉLISATION DE LA DYNAMIQUE D'ESPÈCES INVERTÉBRÉS ÉPIBENTHIQUES

Les modèles dynamiques comme la marche aléatoire ou l'autorégressif d'ordre 1 approchent de manière trop élémentaire un phénomène dynamique. En effet, ces modèles n'intègrent pas les connaissances biologiques et écologiques (p. ex. type de reproduction, dispersion, prédation) pour expliquer la dynamique d'une population.

Un séjour scientifique au Canada (Nouveau-Brunswick et Québec) m'a permis de rencontrer plusieurs spécialistes de la biologie et de l'écologie des invertébrés du golfe du Saint-Laurent, dont les professeurs Heather Hunt et Bernard Sainte-Marie. Cette rencontre me permet aujourd'hui de proposer un modèle parcimonieux qui prend en compte certains aspects importants de la vie des invertébrés épibenthiques afin d'être plus robuste. Ce modèle est construit par convolution discrète comme proposé. Si l'on considère que le modèle des observations est un CPG(λ, a, b) alors, le modèle dynamique latent construit à l'aide d'experts en dynamique de macro-invertébrés est le suivant :

$$\begin{aligned}
\lambda(t,s) &= exp(z(t,s)) \\
z(t,s) &= \mu + \sum_{g=1}^{G} \kappa(s-\omega) x(t,\omega) \\
x(t,\omega) &= \delta x(t-1,\omega) + R(t) + \nu(t) \\
R(t) &= \sum_{l=1}^{G} x(t-d,\omega) f(T(t-d)),
\end{aligned} \tag{6.24}$$

δ est le taux de mortalité, $R(t)$ un terme de recrutement et $f(T(t-d))$ une fonction de variables environnementales influençant le recrutement. Au temps t la biomasse d'oursins au site ω est égale à la biomasse d'oursins au temps $t-1$ au même site ω en tenant compte d'une mortalité δ. Ce taux de mortalité peut être fonction de variables environnementales comme la prédation naturelle P et la pression anthropique C :

$$\delta(t) = exp(-\beta_3 - \beta_4 C(t) - \beta_5 P(t)) \tag{6.25}$$

À cette biomasse restante au site ω s'ajoute un recrutement noté $R(t)$. En effet, les oursins se reproduisent par largage des gamètes mâles et femelles dans l'océan et l'établissement des jeunes oursins s'effectue après un stade larvaire très sensible à la dérive larvaire. Nous supposons donc que tous les oursins contribuent au renouvellement de la population sur l'ensemble du golfe. Ce recrutement à l'échelle du golfe est modélisé par la somme de la biomasse $x(t-d)$ sur l'ensemble du golfe. L'indice d correspond au décalage des individus participant au renouvèlement de la population, il permet également de prendre en compte la croissance des individus. Il est possible d'introduire des variables environnementales dans le recrutement, comme la température Tp et un terme de densité dépendance :

$$f(T(t-d)) = exp(-\beta_0 - \beta_1 Tp - \beta_2 x(t-d)) \tag{6.26}$$

L'approche de modélisation proposée peut être complexifiée en prenant en compte deux échelles spatiales : une échelle régionale et une échelle locale. Ces deux échelles sont prises en compte comme proposé en section 6.1.

$$z(t,s) = \mu + \sum_{l=1}^{G_L} \kappa^L(s-\omega_l) x_L(t,\omega_l) + \sum_{r=1}^{G_R} \kappa^R(s-\omega_r) x_R(t,\omega_r) \tag{6.27}$$

La composante régionale assure le processus dynamique présentée précédemment 6.24, avec les termes de mortalité et de recrutement. La composante locale permet de prendre en compte

des variations à petite échelle qui sont dans cet exemple indépendantes du temps :

$$\begin{aligned} x_R(t,\omega_r) &= \delta x_R(t-1,\omega_r) + R(t,r) + \nu(t) \\ x_L(t,\omega_l) &\overset{i.i.d.}{\sim} Nomal(0,\sigma_l^2) \end{aligned} \tag{6.28}$$

Ce type de modélisation permet d'introduire des connaissances sur la biologie et l'écologie de l'espèce étudiée, mais complexifie notamment l'inférence. Une technique d'inférence prometteuse est proposée en annexe B afin d'estimer de manière astucieuse les paramètres de tels modèles.

7

CONCLUSION

Lors de ces travaux de thèse, je me suis particulièrement intéressé aux données de biomasses d'invertébrés épibenthiques et de poissons de fonds. Les données de biomasses sont couramment utilisées pour construire des indices d'abondances, qui sont eux-mêmes utilisés pour évaluer des stocks à l'aide de modèles complexes (Maunder et Punt, 2004). La qualité des indices d'abondances est par conséquent déterminante dans l'évaluation de ces stocks. Mes travaux de thèse se sont donc portés sur l'amélioration des outils de modélisation utilisés pour étudier ces quantités de biomasses à forte proportion de zéros qui est une caractéristique fondamentale des espèces considérées.

Dans la première partie de mes travaux de thèse, je me suis intéressée à comparer deux méthodes de modélisation adaptées au traitement de données à forte proportion de zéros. La particularité de cette étude est de considérer des quantités de biomasses dont l'effort d'échantillonnage est variable (durée, distance, zone échantillonnée) au cours d'une même campagne d'échantillonnage. Cette variabilité d'effort doit être prise en compte lors de l'analyse des quantités de biomasses et les deux approches de modélisation qui ont été comparées n'incluent pas de la même manière cette variabilité. En effet, le modèle Poisson Composé Gamma (CPG) permet d'introduire l'effort d'échantillonnage de manière naturelle tandis que les modèles delta ont recours à un modèle linéaire généralisé Zuur *et al.* (2009). Ces travaux ont donc permis, grâce à une étude de simulation et à une étude de cas, de mettre en évidence les difficultés des modèles delta à analyser des données dont l'effort d'échantillonnage est variable et de présenter le bon comportement du modèle CPG dans de telles situations. Ces travaux destinés aux praticiens de ces méthodes apportent des recommandations concrètes quant à l'utilisation de ces deux modèles.

La seconde partie de cette thèse utilise le modèle CPG dans sa version hiérarchique bayésienne pour étudier la distribution spatiale des quantités de biomasses d'espèces d'invertébrés épibenthiques dans le sud du golfe du Saint-Laurent (sGSL). La structure spatiale de la répartition de ces espèces est modélisée dans la couche latente par les outils de la géostatistique et plus particulièrement à l'aide d'un variogramme exponentiel. De plus, trois variables environnementales ont été utilisées pour décrire les associations espèces-habitats. Cette approche permet la construction de cartes de répartition de la biomasse sur l'ensemble du sGSL. Ce type de modèles et les prédictions qui lui sont associés sont le point de départ d'étude pour des plans de conservations, des études d'impacts ou encore des analyses de risques (Franklin, 2009). Malheureusement ce type d'approches statiques dans le temps nécessite des précautions quant aux prédictions et conclusions que l'on peut en déduire. Le recours à des modèles plus complexes et notamment à des modèles spatio-temporels est alors nécessaire.

La troisième et dernière partie de mon travail doctoral a consisté à l'élaboration d'un modèle spatio-temporel capable de répondre à ces faiblesses. L'approche développée dans le chapitre 5 doit permettre d'améliorer les prédictions en prenant en compte l'évolution temporelle des quantités de biomasses d'invertébrés épibenthiques récoltées de 1989 à nos jours. Pour ce faire, une grille latente, support de la dépendance spatiale et temporelle, est construite sur la zone d'étude. Les noeuds de cette grille sont des points références qui influencent spatialement et temporellement les points d'observation. Ces outils développés originellement pour l'océanographie et les études de qualité de l'air ont été adaptés dans ce travail à des données de biomasse présentant une forte proportion de zéros. Le modèle CPG est le modèle des observations qui a été utilisé dans cette approche, la grille étant une couche latente de ce modèle modèle bayésien hiérarchique. L'un des principaux intérêts de cette méthode est sa capacité à traiter de grands jeux de données. En effet, l'utilisation d'une grille latente permet à moindre coût de réduire la dimension du problème et de décrire l'évolution spatio-temporelle. Un second avantage à cette approche est sa flexibilité : les extensions proposées au chapitre 6 sont multiples et permettent de tester de nombreuses hypothèses, mais également d'introduire des connaissances sur la biologie ou l'écologie de l'espèce étudiée.

Les approches développées dans cette thèse peuvent être regroupées dans la famille des « Species Distribution Models » et plus particulièrement dans les «correlative models» (Guisan et Zimmermann, 2000; Pearson et Dawson, 2003). Ce type d'approche permet de corréler la présence ou l'abondance d'une espèce avec des variables environnementales dans le but ensuite de prédire sa distribution. Afin d'améliorer ces méthodes, il est à mon avis primordial de développer des modèles dits hybrides, qui permettent d'associer variables environnementales et processus écologiques (p. ex. dispersion, interaction entre espèces). Ces modèles hybrides associent relations espèces-habitats et paramètres de dynamique de population (p. ex. taux de croissance, taux de prédation, *etc.*), qui permettent d'appréhender l'état d'une population. Ce type de modèles stochastiques permettra d'associer au sein d'une même approche de modélisation les outils statistiques novateurs de modélisation spatio-temporelle avec les connaissances des processus écologiques responsable de la distribution des espèces. De plus, je pense qu'il est indispensable de travailler sur l'ajout de la dispersion dans le type d'approches présentées dans ce manuscrit et ceci dans le but de prédire les distributions futures d'espèces face aux scénarios de changements climatiques, mais également la colonisation de territoires par des espèces envahissantes.

BIBLIOGRAPHIE

Albert, J. H. et Chib, S. (1993). Bayesian Analysis of Binary and Polychotomous Response Data. *Journal of the American Statistical Association*, 88(422):669–679.

Allen, J., Brewster, C. et Slone, D. (2001). Spatially explicit ecological models : a spatial convolution approach. *Chaos, Solitons & Fractals*, 12(2):333–347.

Ancelet, S. (2008). *Exploiter l'approche hiérarchique bayésienne pour la modélisation statistique de structures spatiales : application en écologie des populations*. Thèse de doctorat, AgroParisTech.

Ancelet, S., Etienne, M.-P., Benoît, H. et Parent, E. (2009). Modelling spatial zero-inflated continuous data with an exponentially compound Poisson process. *Environmental and Ecological Statistics*, 17(3):347–376.

Andrewartha, H. et Birch, L. (1954). *The distribution and abundance of animals.* The University of Chicago Press, Chicago, IL.

Andrieu, C., Doucet, A. et Holenstein, R. (2010). Particle Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, 72(3):269–342.

Andrieu, C. et Roberts, G. O. (2009). The pseudo-marginal approach for efficient Monte Carlo computations. *The Annals of Statistics*, 37(2):697–725.

Araújo, M. B. et New, M. (2007). Ensemble forecasting of species distributions. *Trends in ecology & evolution*, 22(1):42–7.

Austin, M. (2002). Spatial prediction of species distribution : an interface between ecological theory and statistical modelling. *Ecological Modelling*, 157(2-3):101–118.

Axis-Arroyo, J. et Mateu, J. (2004). Spatio-temporal modeling of benthic biological species. *Journal of environmental management*, 71(1):67–77.

Barry, R. P. et Ver Hoef, J. M. (1996). Blackbox Kriging : Spatial Prediction without Specifying Variogram Models. *Journal of Agricultural, Biological, and Environmental Statistics*, 1(3): 297.

Bell, R. G., Hume, T. M., Dolphin, T. J., Green, M. O. et Walters, R. A. (1997). Characterisation of physical environmental factors on an intertidal sandflat, Manukau Harbour, New Zealand. *Journal of Experimental Marine Biology and Ecology*, 216(1-2):11–31.

Benoît, H. P., Swain, D. P. et Chouinard, G. A. (2009). Using the long-term bottom-trawl survey of the southern Gulf of St. Lawrence to understand marine fish populations and community change. *AZMP Bulletin*, 8:19–27.

Berliner, L. (1996). Hierarchical Bayesian time series models. *In Maximum entropy and Bayesian methods*, pages 15–22. Kluwer Academic Publishers, k. hanson édition.

Berry, D. A. (1987). Logarithmic Transformations in ANOVA. *Biometrics*, 43(2):439.

Besag, J. et Kooperberg, C. (1995). On conditional and intrinsic autoregressions. *Biometrika*, 82(4):733–746.

Besag, J., York, J. et Mollié, A. (1991). Bayesian image restoration, with two applications in spatial statistics. *Annals of the Institute of Statistical Mathematics*, 43(1):1–20.

Box, G., Jenkins, G. et Bacon, D. (1967). Models for forecasting seasonal and non-seasonal times series. Rapport technique, DTIC Document, San Francisco.

Boyce, M. S., Vernier, P. R., Nielsen, S. E. et Schmiegelow, F. K. (2002). Evaluating resource selection functions. *Ecological Modelling*, 157(2-3):281–300.

Brooks, S. et Roberts, G. (1998). Convergence assessment techniques for Markov chain Monte Carlo. *Statistics and Computing*, 8(4):319–335.

Calder, C. (2004). Efficient posterior inference and prediction of space-time processes using dynamic process convolutions. *In Joint Proceedings of the Sixth International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences and the Fifteenth Annual Conference of TIES, The International Environmetrics Society*.

Calder, C. A. (2003). *Exploring Latent Structure in Spatial Temporal Processes Using Process Convolutions*. Duke University.

Calder, C. A. (2007). Dynamic factor process convolution models for multivariate space–time data with application to air quality assessment. *Environmental and Ecological Statistics*, 14(3):229–247.

Candolle, A. D. (1855). *Géographie botanique raisonnée : ou, Exposition des faits principaux et des lois concernant la distribution géographique des plantes de l'époque actuelle*. Masson, Paris.

Candy, S. (2004). Modelling catch and effort data using generalised linear models, the Tweedie distribution, random vessel effects and random stratum-by-year effects. *CCAMLR Science*, 11:59–80.

Chadwick, E., Brodie, W., Colbourne, E., Clark, D., Gascon, D. et Hurlbut, T. (2007). History of annual multi-species trawl surveys on the Atlantic coast of Canada. *Atlantic Zonal Monitoring Program Bulletin*, 6:25–42.

Clark, J. (2007). *Models for ecological data : an introduction*. Princeton University Press.

Clark, J. S. (2004). Why environmental scientists are becoming Bayesians. *Ecology Letters*, 8(1):2–14.

Clark, J. S., Bell, D. M., Hersh, M. H., Kwit, M. C., Moran, E., Salk, C., Stine, A., Valle, D. et Zhu, K. (2011). Individual-scale variation, species-scale differences : inference needed to understand diversity. *Ecology letters*, 14(12):1273–87.

Clarke, K. et Green, R. (1988). Statistical design and analysis for a" biological effects" study. *Mar. Ecol. Prog. Ser.*, 46(1):213–226.

Cook, A., Marion, G., Butler, A. et Gibson, G. (2007). Bayesian inference for the spatio-temporal invasion of alien species. *Bulletin of mathematical biology*, 69(6):2005–25.

Cressie, N. et Wikle, C. (2011). *Statistics for spatio-temporal data*. Wiley.com.

Dauer, D. M. (1993). Biological criteria, environmental health and estuarine macrobenthic community structure. *Marine Pollution Bulletin*, 26(5):249–257.

Dethier, M. et Schoch, G. (2005). The consequences of scale : assessing the distribution of benthic populations in a complex estuarine fjord. *Estuarine, Coastal and Shelf Science*, 62(1-2):253–270.

Diggle, P. J., Menezes, R. et Su, T.-l. (2010). Geostatistical inference under preferential sampling. *Journal of the Royal Statistical Society : Series C (Applied Statistics)*, 59(2):191–232.

Doubleday, W. et Rivard, D. (1981). Bottom trawl surveys. *Canadian Special Publication of Fisheries and Aquatic Sciences*, 58:273.

Edgar, G. J. et Barrett, N. S. (2002). Benthic macrofauna in Tasmanian estuaries : scales of distribution and relationships with environmental variables. *Journal of Experimental Marine Biology and Ecology*, 270(1):1–24.

Elith, J. et Leathwick, J. R. (2009). Species Distribution Models : Ecological Explanation and Prediction Across Space and Time. *Annual Review of Ecology, Evolution, and Systematics*, 40(1):677–697.

Engler, R., Guisan, A. et Rechsteiner, L. (2004). An improved approach for predicting the distribution of rare and endangered species from occurrence and pseudo-absence data. *Journal of Applied Ecology*, 41(2):263–274.

Ferrier, S., Watson, G., Pearce, J. et Drielsma, M. (2002). Extended statistical approaches to modelling spatial pattern in biodiversity in northeast New South Wales. I. Species-level modelling. *Biodiversity & Conservation*, 11(12):2275–2307.

Fitzpatrick, M. C., Weltzin, J. F., Sanders, N. J. et Dunn, R. R. (2007). The biogeography of prediction error : why does the introduced range of the fire ant over-predict its native range ? *Global Ecology and Biogeography*, 16(1):24–33.

Foody, G. (2008). GIS : biodiversity applications. *Progress in Physical Geography*, 32(2):223–235.

Foster, S. D. et Bravington, M. V. (2012). A Poisson–Gamma model for analysis of ecological non-negative continuous data. *Environmental and Ecological Statistics*, 19(4):1–20.

Franklin, J. (2009). *Mapping species distributions : spatial inference and prediction*. Cambridge University Press, New York.

Furrer, R., Nychka, D. et Sain, S. (2012). fields : tools for spatial data. R package version 6.6. 3.

Gelfand, A. E. et Smith, A. F. M. (1990). Sampling-Based Approaches to Calculating Marginal Densities. *Journal of the American Statistical Association*, 85(410):398–409.

Gelman, A., Carlin, J., Stern, H. et Rubin, D. (2004). *Bayesian data analysis*. CRC press.

Gelman, A., Meng, X.-l. et Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica*, 6:733–807.

Gelman, A. et Rubin, D. (1992). Inference from iterative simulation using multiple sequences. *Statistical science*, 7(4):457–472.

Grinnell, J. (1904). The origin and distribution of the chest-nut-backed chickadee. *The Auk*.

Grinnell, J. (1917). Field tests of theories concerning distributional control. *American Naturalist*.

Guisan, A., Lehmann, A., Ferrier, S., Austin, M., Overtnn, J. M. C., Aspinall, R. et Hastie, T. (2006). Making better biogeographical predictions of species' distributions. *Journal of Applied Ecology*, 43(3):386–392.

Guisan, A. et Thuiller, W. (2005). Predicting species distribution : offering more than simple habitat models. *Ecology Letters*, 8(9):993–1009.

Guisan, A. et Zimmermann, N. E. (2000). Predictive habitat distribution models in ecology. *Ecological Modelling*, 135(2-3):147–186.

Hampe, A. (2004). Bioclimate envelope models : what they detect and what they hide. *Global Ecology and Biogeography*, 13(5):469–471.

Hartog, J. R., Hobday, A. J., Matear, R. et Feng, M. (2011). Habitat overlap between southern bluefin tuna and yellowfin tuna in the east coast longline fishery-implications for present and future spatial management. *Deep Sea Research Part II : Topical Studies in Oceanography*, 58(5):746–752.

Hastings, W. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109.

Heikkinen, R. K., Luoto, M., Araujo, M. B., Virkkala, R., Thuiller, W. et Sykes, M. T. (2006). Methods and uncertainties in bioclimatic envelope modelling under climate change. *Progress in Physical Geography*, 30(6):751–777.

Heikkinen, R. K., Luoto, M., Virkkala, R., Pearson, R. G. et Körber, J.-H. (2007). Biotic interactions improve prediction of boreal bird distributions at macro-scales. *Global Ecology and Biogeography*, 16(6):754–763.

Heilbron, D. C. (1994). Zero-Altered and other Regression Models for Count Data with Added Zeros. *Biometrical Journal*, 36(5):531–547.

Higdon, D. (1998). A process-convolution approach to modelling temperatures in the North Atlantic Ocean. *Environmental and Ecological Statistics*, 5(2):173–190.

Higdon, D. (2002). Space and space-time modeling using process convolutions. *In Quantitative methods for current environmental issues*, pages 37–56. Springer Verlag, c. anderso édition.

Himmelman, J. et Dutil, C. (1991). Distribution, population structure and feeding of subtidal seastars in the northern Gulf of St. Lawrence. *Marine ecology progress series. Oldendorf*, 76:61–72.

Hirzel, A. H. et Le Lay, G. (2008). Habitat suitability modelling and niche theory. *Journal of Applied Ecology*, 45(5):1372–1381.

Hobday, A. J. et Hartmann, K. (2006). Near real-time spatial management based on habitat predictions for a longline bycatch species. *Fisheries Management and Ecology*, 13(6):365–380.

Hooten, M. B. et Wikle, C. K. (2007). A hierarchical Bayesian non-linear spatio-temporal model for the spread of invasive species with application to the Eurasian Collared-Dove. *Environmental and Ecological Statistics*, 15(1):59–70.

Hutchinson, G. (1957). Concluding remarks. *Cold Spring Harbor Symposium on Quantitative Biology*, 22:415–427.

Iverson, L. R. et Prasad, A. M. (2002). Potential redistribution of tree species habitat under five climate change scenarios in the eastern US. *Forest Ecology and Management*, 155(1-3):205–222.

Jones, M. M., Olivas Rojas, P., Tuomisto, H. et Clark, D. B. (2007). Environmental and neighbourhood effects on tree fern distributions in a neotropical lowland rain forest. *Journal of Vegetation Science*, 18(1):13–24.

Jorgensen, B. (1987). Exponential dispersion models. *Journal of the Royal Statistical Society. Series B (Methodological)*, 49(2): 127–162.

Jorgensen, B. (1997). *The theory of dispersion models*. Chapman and Hall, London.

Kadane, J. (2011). *Principles of uncertainty*. CRC press.

Kearney, M. et Porter, W. (2009). Mechanistic niche modelling : combining physiological and spatial data to predict species' ranges. *Ecology letters*, 12(4):334–50.

Kendal, W. S. (2004). Taylor's ecological power law as a consequence of scale invariant exponential dispersion models. *Ecological Complexity*, 1(3):193–209.

Kern, J. (2000). *Bayesian process-convolution approaches to specifying spatial dependence structure*. Thèse de doctorat, Duke University.

Kinlan, B. P. et Gaines, S. D. (2003). Propagule dispersal in marine and terrestrial environments : a community perspective. *Ecology*, 84(8):2007–2020.

Krebs, C. J. (1978). Ecology : the experimental analysis of distribution and abundance. *New York : Harper and Row*.

Latimer, A. M., Wu, S., Gelfand, A. E. et Silander, J. a. (2006). Building statistical models to analyze species distributions. *Ecological applications : a publication of the Ecological Society of America*, 16(1):33–50.

Lecomte, J.-B., Benoît, H. P., Ancelet, S., Etienne, M.-P., Bel, L. et Parent, E. (2013a). Compound Poisson-gamma vs. delta-gamma to handle zero-inflated continuous data under a variable sampling volume. *Methods in Ecology and Evolution*, 4(12):1159–1166.

Lecomte, J. B., Benoît, H. P., Etienne, M. P., Bel, L. et Parent, E. (2013b). Modeling the habitat associations and spatial distribution of benthic macroinvertebrates : A hierarchical Bayesian model for zero-inflated biomass data. *Ecological Modelling*, 265:74–84.

Legendre, P. et Legendre, L. (1998). *Numerical ecology*, volume 20. Elsevier Science.

Loring, D. et Nota, D. (1973). *Morphology and sediments of the Gulf of St. Lawrence*. Fisheries and Marine Service, Ottawa.

Lunn, D., Spiegelhalter, D., Thomas, A. et Best, N. (2009). The BUGS project : Evolution, critique and future directions. *Statistics in medicine*, 28(25):3049–67.

MacKenzie, D. (2006). *Occupancy estimation and modeling : inferring patterns and dynamics of species occurrence*. Elsevier, Paris.

Manly, B., McDonald, L. et Thomas, D. (1992). *Resource selection by animals : statistical design and analysis for field studies*. Springer.

Martin, T., Wintle, B., Rhodes, J., Kuhnert, P., Field, S., Low-Choy, S., Tyre, A. et Possingham, H. (2005). Zero tolerance ecology : improving ecological inference by modelling the source of zero observations. *Ecology Letters*, 8(11):1235–1246.

Maunder, M. N. et Punt, A. E. (2004). Standardizing catch and effort data : a review of recent approaches. *Fisheries Research*, 70(2-3):141–159.

McCarthy, M. (2007). *Bayesian methods for ecology*. Cambridge University Press, New York.

Metropolis, N. et Rosenbluth, A. (1953). Equation of state calculations by fast computing machines. *Journal of Chemical Physics*, 21:1087–1092.

Nishizaki, M. T. et Ackerman, J. D. (2006). Juvenile–adult associations in sea urchins (Strongylocentrotus franciscanus and S. droebachiensis) : protection from predation and hydrodynamics in S. franciscanus. *Marine Biology*, 151(1):135–145.

Ntzoufras, I. (2011). *Bayesian modeling using WinBUGS*, volume 698. Wiley, Hoboken, NJ.

Olsen, N., Workman, G., Stanley, R. et Station, P. (2007). Queen Charlotte Sound groundfish bottom trawl survey, July 5th to August 9th, 2004. Rapport technique, Science Branch, Pacific Region Fisheries and Oceans Canada, Pacific Biological Station, Nanaimo, BC.

Parent, E. et Rivot, E. (2012). *Introduction to hierarchical Bayesian modeling for ecological data*. CRC press.

Pearson, R., Dawson, T., Berry, P. et Harrison, P. (2002). SPECIES : A Spatial Evaluation of Climate Impact on the Envelope of Species. *Ecological Modelling*, 154(3):289–300.

Pearson, R. G. et Dawson, T. P. (2003). Predicting the impacts of climate change on the distribution of species : are bioclimate envelope models useful ? *Global Ecology and Biogeography*, 12(5):361–371.

Pearson, R. G., Dawson, T. P. et Liu, C. (2004). Modelling species distributions in Britain : a hierarchical integration of climate and land cover data. *Ecography*, 27(3):285–298.

Peel, D., Bravington, M. V., Kelly, N., Wood, S. N. et Knuckey, I. (2012). A Model-Based Approach to Designing a Fishery-Independent Survey. *Journal of Agricultural, Biological, and Environmental Statistics*, 18(1):1–21.

Perry, R. I. et Smith, S. J. (1994). Identifying Habitat Associations of Marine Fishes Using Survey Data : An Application to the Northwest Atlantic. *Canadian Journal of Fisheries and Aquatic Sciences*, 51(3):589–602.

Peterson, A. T. et Holt, R. D. (2003). Niche differentiation in Mexican birds : using point occurrences to detect ecological innovation. *Ecology Letters*, 6(8):774–782.

Peterson, A. T., Ortega-Huerta, M. A., Bartley, J., Sánchez-Cordero, V., Soberón, J., Buddemeier, R. H. et Stockwell, D. R. B. (2002). Future projections for Mexican faunas under global climate change scenarios. *Nature*, 416(6881):626–9.

Pinkus, A. (1997). *Fourier series and integral transforms*. Cambridge University Press.

Porch, C. et Scott, G. (1994). A numerical evaluation of GLM methods for estimating indices of abundance from West Atlantic bluefin tuna catch per trip data when a high proportion of the trips. *ICCAT Col. Vol. Sci. Pap*, 42:241–245.

Redfern, J., Ferguson, M., Becker, E., Hyrenbach, K., Good, C., Barlow, J., Kaschner, K., Baumgartner, M., Forney, K., Ballance, L., Fauchald, P., Halpin, P., Hamazaki, T., Pershing, A., Qian, S., Read, A., Reilly, S., Torres, L. et Werne, F. (2006). Tehniques for cetaceans habitat modelling. *Marine Ecology Progress Series*, 310:271–295.

Robert, C. et Casella, G. (2004). *Monte Carlo statistical methods*. Springer, New York.

Robinson, L. M., Elith, J., Hobday, A. J., Pearson, R. G., Kendall, B. E., Possingham, H. P. et Richardson, A. J. (2011). Pushing the limits in marine species distribution modelling : lessons from the land present challenges and opportunities. *Global Ecology and Biogeography*, 20(6):789–802.

Rotenberry, J. T., Preston, K. L. et Knick, S. T. (2006). GIS-Based Niche Modeling for Mapping Species Habitat. *Ecology*, 87(6):1458–1464.

Royle, J. et Dorazio, R. (2008). *Hierarchical modeling and inference in ecology : the analysis of data from populations, metapopulations and communities*. Academic Press.

Rue, H. et Held, L. (2005). *Gaussian Markov random fields : theory and applications*. CRC Press.

Scheffer, M. et Carpenter, S. R. (2003). Catastrophic regime shifts in ecosystems : linking theory to observation. *Trends in Ecology & Evolution*, 18(12):648–656.

Schneider, D. (1994). *Quantitative ecology : spatial and temporal scaling*. Academic Press.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464.

Scott, J. (2002). *Predicting species occurrences : issues of accuracy and scale*. Island Press, Washington.

Shono, H. (2008). Application of the Tweedie distribution to zero-catch data in CPUE analysis. *Fisheries Research*, 93(1-2):154–162.

Sinclair, A., Schnute, J., Haigh, R., Starr, P., Stanley, R., Fargo, J. et Workman, G. (2003). Feasibility of multispecies groundfish bottom trawl surveys on the BC coast. Rapport technique, Science Branch, Pacific Region Fisheries and Oceans Canada, Pacific Biological Station, Nanaimo, BC.

Smyth, G. (1996). Regression analysis of quantity data with exact zeros. *Proceedings of the Second Australia–Japan Workshop on Stochastic Models in Engineering, Technology and Management. Technology Management Centre, University of Queensland*, pages 572–580.

Snelgrove, P. (1999). Getting to the bottom of marine biodiversity : Sedimentary habitats : Ocean bottoms are the most widespread habitat on earth and support high biodiversity and key ecosystem services. *BioScience*, 49(2):129—-138.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P. et van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, 64(4):583–639.

Stanley, R., Starr, P., Olsen, N., Rutherford, K. et Wallace, S. (2005). Status Report on Canary Rockfish Sebastes Pinniger. Rapport technique, Science Branch, Pacific Region Fisheries and Oceans Canada, Pacific Biological Station, Nanaimo, BC.

Stefansson, G. (1996). Analysis of groundfish survey abundance data : combining the GLM and delta approaches. *ICES Journal of Marine Science*, 53(3):577–588.

Tascheri, R., Saavedra-Nievas, J. et Roa-Ureta, R. (2010). Statistical models to standardize catch rates in the multi-species trawl fishery for Patagonian grenadier (Macruronus magellanicus) off Southern Chile. *Fisheries Research*, 105(3):200–214.

Taylor, L. (1961). Aggregation, variance and the mean. *Nature*, 189:732–735.

Thuiller, W., Richardson, D. M., Pysek, P., Midgley, G. F., Hugues, G. O. et Rouget, M. (2005). Niche-based modelling as a tool for predicting the risk of alien plant invasions at a global scale. *Global Change Biology*, 11(12):2234–2250.

Torres, L. G., Read, A. J. et Halpin, P. (2008). Fine-scale habitat modeling of a top marine predator : do prey data improve predictive capacity. *Ecological Applications*, 18(7):1702–1717.

Tu, W. (2002). Zero-Inflated Data. *Encyclopedia of environmetrics*.

Tyre, A. J., Tenhumberg, B., Field, S. A., Niejalke, D., Parris, K. et Possingham, H. P. (2003). Improving precision and reducing bias in biological surveys : estimating false-negative error rates. *Ecological Applications*, 13(6):1790–1801.

Ver Hoef, J. M. et Jansen, J. K. (2007). Space—time zero-inflated count models of Harbor seals. *Environmetrics*, 18(7):697–712.

von Humboldt, A. et Bonpland, A. (1807). *Essai sur la géographie des plantes*. Paris.

Walsh, S. (1997). Efficiency of bottom sampling trawls in deriving survey abundance indices. *Oceanographic Literature Review*, 44(7):748–748.

Wikle, C. (1999). A dimension-reduced approach to space-time Kalman filtering. *Biometrika*, 86(4):815–829.

Wikle, C., Berliner, L. et Cressie, N. (1998). Hierarchical Bayesian space-time models. *Environmental and Ecological Statistics*, 5(2):117–154.

Wikle, C. K. (2003). Hierarchical Bayesian Models for Predicting the Spread of Ecological Processes. *Ecology*, 84(6):1382–1394.

Williams, A. et Bax, N. (2001). Delineating fish-habitat associations for spatially based management : an example from the south-eastern Australian continental shelf. *Marine and Freshwater Research*, 52(4):513 – 536.

Wolpert, R. (1998). Poisson/gamma random field models for spatial statistics. *Biometrika*, 85(2):251–267.

Ysebaert, T. et Herman, P. (2002). Spatial and temporal variation in benthic macrofauna and relationships with environmental variables in an estuarine, intertidal soft-sediment environment. *Marine Ecology Progress Series*, 244:105–124.

Zhu, Z. et Wu, Y. (2010). Estimation and Prediction of a Class of Convolution-Based Spatial Nonstationary Models for Large Spatial Data. *Journal of Computational and Graphical Statistics*, 19(1):74–95.

Zuur, A. F., Ieno, E. N., Walker, N., Saveliev, A. A. et Smith, G. M. (2009). *Mixed effects models and extensions in ecology with R*. Statistics for Biology and Health. Springer, New York.

A

INFÉRENCE : MONTE-CARLO PAR CHAINES DE MARKOV

Les modèles hiérarchiques bayésiens sont généralement complexes et peuvent être composés d'un grand nombre de paramètres inconnus. Dans ce cas, il est difficile d'écrire analytiquement la fonction de densité *a posteriori* $[\theta|y]$. En effet, il est souvent impossible de calculer la constante de normalisation $[y]$ car elle nécessite une intégration àn dimension, n étant le nombre de paramètres. La densité *a posteriori* est donc connue à cette constante de normalisation près :

$$[\theta|y] \propto [y|\theta][\theta] \tag{A.1}$$

L'inférence de modèles bayésiens complexes a été possible par l'accessibilité à de grandes capacités de calculs associés au développement d'algorithmes connus sous le nom de méthode de Monte-Carlo par Chaines de Markov (MCMC) (Robert et Casella, 2004). L'idée générale de ce type d'algorithmes est de tirer un grand nombre d'échantillons de la loi *a posteriori* $[\theta|y]$ sans se préoccuper de la constante $[y]$. Pour ce faire, une chaine de Markov dont la distribution stationnaire est $[\theta|y]$ est construite. Elle permet d'obtenir un échantillon aléatoire $\{\theta^1, \theta^2, ..., \theta^I\}$ qui (lorsqu'il est suffisamment grand) donne une approximation de la distribution *a posteriori*. Les quantités d'intérêt (p. ex. moyenne, variance) peuvent alors être calculées empiriquement à partir de cet échantillon. Cette méthode ne requiert donc pas l'expression analytique de la loi *a posteriori* $[\theta|y]$.

L'algorithme de Metropolis-Hasting et l'algorithme de Gibbs sont les deux algorithmes MCMC les plus utilisés (Robert et Casella, 2004). Leur fonctionnement est décrit dans les sections suivantes A.1 et A.2.

A.1 ALGORITHME DE METROPLIS-HASTINGS

L'algorithme de Metroplis-Hastings a été développé tout d'abord par (Metropolis et Rosenbluth, 1953) pour la Physique nucléaire puis par (Hastings, 1970) pour la Statistique. Cet algorithme est adapté d'une marche aléatoire, mais utilise une règle d'acceptation-rejet pour converger vers la distribution cible $\pi(\theta) = [\theta|y] = K \times [y|\theta][\theta]$ connue à une constante près. Les différentes étapes se déroulent comme suit :

1. initialisation avec une valeur $\theta^{(0)}$
2. A l'itération i, mettre à jour $\theta^{(i)}$ par $\theta^{(i+1)}$ selon :
   a) Générer le candidat θ^* a partir de la valeur courante $\theta^{(i)}$ par un tirage dans la loi $q(.|\theta^{(i)})$,
   b) Calculer le ratio

$$\rho = \frac{\pi(\theta^*)q(\theta^{(i)}|\theta^*)}{\pi(\theta^{(i)})q(\theta^*|\theta^{(i)})} \tag{A.2}$$

   On note que pour calculer ce ratio, il n est pas nécessaire de connaitre la constante de normalisation de $\pi(\theta)$ qui se simplifie.
   c) Prendre

$$\theta^{(i+1)} = \begin{cases} \theta^* \text{ avec probabilité } min(1, \rho) \\ \theta^{(i)} \text{ sinon.} \end{cases} \tag{A.3}$$

La loi de densité $q(.|\theta^{(i)})$ est appelée loi de proposition ou loi instrumentale. On prend souvent une loi uniforme ou normale centrée sur $\theta^{(i)}$ de variance à régler. Le choix de cette loi de proposition q est déterminant. En effet, il conditionne la vitesse de convergence de l'algorithme. Si la loi instrumentale est trop dispersée, les candidats θ^* générés seront souvent refusés ralentissant ainsi la convergence. Inversement, si la loi de proposition ne permet pas de générer des candidats relativement différents la chaine peut rester bloquée longtemps avant de converger vers la loi cible.

A.2 ALGORITHME DE GIBBS

Prenons comme exemple un modèle à trois paramètres $(\theta_1, \theta_2, \theta_3)$. Nous souhaitons obtenir un large échantillon du posterior $[\theta_1, \theta_2, \theta_3|y]$, ou y sont les données dont nous disposons. L'algorithme de Gibbs utilise les distributions conditionnelles complètes pour obtenir une distribution *a posteriori*. Supposons que ces distributions conditionnelles complètes soient facilement calculables, l'algorithme de Gibbs procède ainsi :

1. initialisation avec une valeur $(\theta_1^{(0)}, \theta_2^{(0)}, \theta_3^{(0)})$
2. Générer $\theta_1^{(i+1)}$ selon $[\theta_1|\theta_2^{(i)}, \theta_3^{(i)}y]$
3. Générer $\theta_2^{(i+1)}$ selon $[\theta_2|\theta_1^{(i+1)}, \theta_3^{(i)}y]$
4. Générer $\theta_3^{(i+1)}$ selon $[\theta_3|\theta_1^{(i+1)}, \theta_2^{(i+1)}y]$

Cet algorithme converge vers la loi jointe *a posteriori* $[\theta_1, \theta_2, \theta_3|y]$. Contrairement à l'algorithme de Metropolis-Hastings, l'algorithme de Gibbs ne requiert pas de loi de proposition ni d'étape d'acceptation-rejet. La limite de cet algorithme réside dans la capacité à simuler les distributions conditionnelles complètes des éléments du vecteur de paramètres θ.

Il est possible de construire des algorithmes hybrides faisant appel aux deux méthodes (*Metropolis-Hastings within Gibbs*). Cette méthode est principalement utilisée lorsque certaines distributions conditionnelles complètes sont difficiles à échantillonner avec l'échantillonneur de Gibbs. Un algorithme de Metropolis-Hastings est alors utilisé à l'intérieur de l'algorithme de Gibbs.

A.3 DIAGNOSTICS DE CONVERGENCE

Afin d'obtenir un échantillon représentatif de la loi *a posteriori*, il est déterminant de contrôler la convergence des algorithmes MCMC utilisés. L'utilisation de plusieurs chaines MCMC indépendantes, dont les valeurs initiales étaient différentes, permet l'utilisation d'outils statistiques pour contrôler leur convergence (Gelman et Rubin, 1992; Brooks et Roberts, 1998). Ces diagnostics consistent à vérifier qu'après une certaine période de chauffe, les chaines ont convergé vers le même état stationnaire.

En pratique, trois chaines indépendantes sont généralement utilisées. Une première inspection visuelle permet de dire si les chaines se mélangent correctement. Si graphiquement les chaines sont mélangées, alors la statistique de Gelman-Rubin peut être utilisée. Avant d'utiliser le diagnostic de convergence, il est préférable d'écarter les premières itérations de chaque chaine correspondant à une période de chauffe qui ne représente pas la distribution stationnaire. Soit m le nombre de chaines, n la longueur de chaque chaine après avoir écarté les premières itérations, pour calculer la statistique de Gelman-Rubin, il est tout d'abord nécessaire de calculer la variance intra chaines :

$$W = \frac{1}{m}\sum_{j=1}^{m} s_j^2$$

avec

$$s_j^2 = \frac{1}{n-1}\sum_{i=1}^{n}(\theta_{ij} - \bar{\theta}_j)^2 \tag{A.4}$$

s_j^2 est la variance de la chaine j et W est la moyenne des variances de chacune des m chaines. W sous-estime généralement la *vraie* variance de la distribution stationnaire. En effet, il est probable que les chaines n'aient pas toutes atteint la distribution stationnaire. La statistique de Gelman-Rubin nécessite également le calcul de la variance entre chaines :

$$B = \frac{n}{m-1}\sum_{j=1}^{m}(\bar{\theta}_j - \bar{\bar{\theta}})^2$$

avec

$$\bar{\bar{\theta}} = \frac{1}{m}\sum_{j=1}^{m}(\bar{\theta}_j)^2 \tag{A.5}$$

B est la variance des moyennes de chaque chaine multipliée par n, car les chaines sont composées de n tirages. Avec W et B il est possible de calculer la variance de la distribution stationnaire :

$$\hat{\mathbb{V}ar}(\theta) = (1 - \frac{1}{n})W + \frac{1}{n}B \tag{A.6}$$

Si les premières itérations ont été écartées (période de chauffe) alors $\hat{\mathbb{V}ar}(\theta)$ est un estimateur non biaisé de la variance de la distribution stationnaire. Il est donc possible de calculer la statistique de Gelman-Rubin :

$$\hat{R} = \sqrt{\frac{\hat{\mathbb{V}ar}(\theta)}{W}} \tag{A.7}$$

En pratique, un score $\hat{R}$ proche de 1 signifie que les chaines ont convergé. Cependant, si $\hat{R} > 1.2$ alors les chaines n'ont pas atteint leur distribution stationnaire, il est donc nécessaire de poursuivre la procédure d'échantillonnage de la chaine jusqu'à obtenir un score $\hat{R} < 1.2$. Lorsque le score de R est satisfaisant, les chaines indépendantes peuvent être fusionnées en une seule de dimensions nm.

B

VERS UN NOUVEL ALGORITHME D'INFÉRENCE ADAPTÉ AUX MODÈLES HIÉRARCHIQUES SUR GRILLES LATENTES NORMALES

B.1 MOTIVATIONS

1. S'appuyer sur des outils génériques d'inférence MCMC tels que JAGS ou OpenBUGS permet de s'affranchir du développement d'algorithme d'inférence. Néanmoins, l'estimation des paramètres des modèles les plus complexes développés dans ce travail doctoral peut durer plusieurs jours. De ce fait, développer ses propres algorithmes permet de contrôler toutes les étapes de l'inférence et notamment d'améliorer les temps de calcul. Ce chapitre annexe présente les premières tentatives de recherches en la matière, bien qu'elles n'aient pas été menées jusqu'à un terme très avancé.

2. Pour commencer, les données observées sont supposées être distribuées selon une loi normale. L'inférence d'un modèle à convolution discrète est alors très simple. En effet, puisqu'il s'agit simplement d'un modèle à effets aléatoires normaux (Higdon, 2002). Bien entendu, la couche d'observations considérée dans cette thèse est un modèle plus compliqué, c'est une loi CPG. Cependant, commet-on une grande erreur si, pour l'inférence, on remplace en chaque point, la distribution CPG par une loi normale dont la moyenne et la variance ont été estimées préalablement ? Après cette première inférence peu couteuse, car s'appuyant sur la conjugaison normale-normale, on dispose d'un simulateur de champs latents. Ce simulateur peut alors être considéré comme une fonction auxiliaire efficace dans une procédure de type *importance sampling*.

3. Pour réaliser l'inférence des paramètres du modèle, il est ensuite proposé de combiner importance sampling (pour le champlatent) et Métropolis-Hasting (pour les paramètres). L'algorithme de Metropolis-Hasting pseudo-marginal de Andrieu et Roberts (2009), donné à la section C.2 de l'annexe correspondante, permet de faire le mariage entre l'importance sampling qui fournit un estimateur sans biais de la constante de normalisation et des pas de Metropolis-Hasting pour les paramètres.

Ce chapitre rappelle d'abord les commodités de la conjugaison dans le cas normal puis, pour le cas qui nous intéresse (modèle normal latent sur grille, mais observations de type CPG), détaille l'algorithme sur lequel nous avons commencé à développer nos essais. Nous notons les distributions de probabilité par des crochets $[.]$ quand ces lois font référence à la structure de modélisation que nous avons adoptée et par des lettres telles que $\pi(.,.)$ ou $p(.|.)$ quand il s'agit de lois auxiliaires utiles aux algorithmes d'inférence (ces lois auxiliaires présupposent que les données sont connues).

B.2 Inférence d'un modèle à convolution discrète normale

B.2.1 *Notations*

Fixons les notations de ce premier modèle avec θ le paramètre, Z le vecteur latent ou de façon équivalente X les composantes de grille de la convolution discrète. On numérote par l'indice s les points de mesure (de 1 à S) et les points de grille par le numéro courant g (de 1 à G).

Observations normales :

$$\begin{aligned} Y_s &= Z_s + \varepsilon_s \\ \varepsilon_s &\sim N(h_s, \sigma_s^2) \end{aligned} \tag{B.1}$$

Champ latent construit par convolution discrète :

$$Z_s = \sum_g K_\phi(s,g) \times X_g$$

Grille de points $g = 1, ..., G$ où se manifestent les intensités des composantes locales indépendantes X_g :

$$X_g \sim N(0, \sigma_x^2)$$
$$K_\phi(s, g) = \exp -\frac{d(s, g)}{\phi}$$

On note $\theta = (\phi, \sigma_x^2, (h_s, \sigma_s^2)_{s=1:S})$ le vecteur des paramètres incluant les biais et variances de mesures en chaque point s.

Notons $Z = \begin{pmatrix} Z_1 \\ ... \\ Z_s \\ ... \\ Z_S \end{pmatrix}$ le vecteur-colonne des S valeurs du champ latent et $X = \begin{pmatrix} X_1 \\ ... \\ X_g \\ ... \\ X_G \end{pmatrix}$ le vecteur colonne des valeurs de grille. La matrice $K = \begin{pmatrix} K(1,1) & ... & K(1,G) \\ ... & ... & ... \\ ... & K(s,g) & ... \\ ... & ... & ... \\ K(S,1) & ... & K(S,G) \end{pmatrix}$ a donc S lignes et G colonnes.

La variable $Y|X$ est normale de moyenne $\mathbb{E}(Y|X) = h + KX$ et de précision $P = \Sigma^{-1}$ diagonale de terme générique $P_{ss} = \sigma_s^{-2}$. Supposons que l'on ait observé $Y = y$. On sait que X est marginalement normal de moyenne nulle et de matrice de variance covariance $V_0 = \sigma_x^2 \times 1_{GG}$.

B.2.2 *Une loi conditionnelle explicite*

Le calcul qui suit est un calcul classique de la conjugaison normale-normale :

$$\begin{aligned}\log[X|y,\theta] &= \log[Y=y|X,\theta] + \log[X|\theta] + Cste \\ &= -\frac{1}{2}\left(y - X'K' - h'\right) P \left(y - KX - h\right) + \\ &- \frac{1}{2}\left(X'\right) V_0^{-1} \left(X\right) + Cste\end{aligned}$$

En réarrangeant les termes dépendants de X, on voit que :

$$\begin{aligned}\log[X|y,\theta] = &-\frac{1}{2}(X'K'PKX + X'V_0^{-1}X) \\ &- \frac{1}{2}X'\left(-K'Py - K'Ph\right) \\ &- \frac{1}{2}\left(-K'Py - KPh\right)' X + + Cste\end{aligned}$$

$X|y$ est donc la loi normale multivariée de moyenne $h_1(y,\theta)$ et de matrice de variance-covariance $V_1(y,\theta)$ telles que :

$$\begin{aligned}V_1^{-1} &= V_0^{-1} + K'PK \\ h_1 &= V_1 K' P \left(y - h\right)\end{aligned}$$

C'est à dire avec nos notations pour la convolution discrète :

$$\begin{aligned}\mathbb{V}ar(X|y,\theta)^{-1} = V_1^{-1} &= \left(\sigma_x^{-2} 1_{GG} + K'PK\right) \\ \mathbb{E}(X|y,\theta) = h_1 &= V_1 K' P \left(y - h\right)\end{aligned} \tag{B.2}$$

B.2.3 *Intérêts*

On note l'effet de réduction de dimension du à la grille : compte tenu de la forme particulière de V_1^{-1} , son inverse de dimension $G \times G$ est plus facile à calculer que celui de la matrice covariance de Z (de dimension $S \times S$).

En appelant $[\theta]$ la loi *a priori* des paramètres de ce modèle, $[x|y,\theta]$ est la loi normale de caractéristiques données par l'équation B.2. La meilleure façon d'échantillonner dans la loi *a posteriori* $[\theta|y] = \frac{1}{[y]} \times \left(\frac{[y|x,\theta][x|\theta][\theta]}{[x|y,\theta]}\right)$ est surement la façon traditionnelle par un algorithme de type Metropolis-Hasting puis-

qu'elle est proportionnelle à une fonction connue (où X est une variable muette qu'on pourra prendre égale à 0).

Une autre façon de procéder, plus longue et bien sûr inutilement compliquée pour ce cas normal (les X sont muets dans l'expression $\left(\frac{[y|x,\theta][x|\theta][\theta]}{[x|y,\theta]}\right)$) serait de mettre en place l'algorithme $MH + IS$ décrit à la section C.2 en annexe à la thèse, en appelant $\pi(x,\theta) = \frac{1}{[y]} \times [y|x,\theta][x|\theta][\theta]$ et $p(x|\theta) = [x|y,\theta]$ la loi normale de caractéristiques données par l'équation B.2, pour laquelle il est néanmoins très facile de tirer les nombreux réplicats nécessaires au fonctionnement de l'algorithme. C'est pourtant cette idée de commodité de génération par une loi auxiliaire normale que nous souhaitons conserver par la suite.

B.3 Idées de base pour la mise en oeuvre d'un algorithme d'inférence type $MH + IS$

Considérons maintenant le cas où l'équation B.1 est remplacée par une loi d'observation non normale, de type CPG par exemple. Pour ne pas trop compliquer les choses, prenons le cas d'un processus de Poisson à marques exponentielles (Ancelet *et al.*, 2009) qui donne des zéros avec une probabilité $\exp(-\lambda(s))$ et pour lequel on connait la vraisemblance quand les observations ne sont pas nulles :

$$\begin{aligned}[Y(s) = y(s)|\lambda(s),\rho] \\ = \lambda(s)\rho\exp(-\lambda(s) - \rho y(s))\frac{I_1(2\sqrt{\lambda(s)\rho y(s)})}{\sqrt{\lambda(s)\rho y(s)}}\end{aligned} \tag{B.3}$$

où I_1 est la fonction de Bessel de deuxième espèce modifiée d'ordre 1, avec $\lambda(s) = \exp(Z(s) + \mu(s))$, $\mu(s)$ représentant l'effet des covariables au point s, généralement sous la forme d'une combinaison linéaire à coefficients inconnus (équation 5.11). La figure B.1.a où les triangles indiquent les zéros montre par exemple une réalisation de ce modèle hiérarchique obtenue à partir de $\rho = 1, \sigma_x = 9, \phi = 5.23, m = 0.5$ dans une configuration d'espace linéaire à 90 points de mesures équirépartis et 21 noeuds réguliers sur le segment allant de l'abscisse 1 à 100 .

Peut-on se rapprocher du cas simple précédent (équation B.1) ? Imaginons que nous disposions d'un estimateur $T_s(Y)$ de $Z(s)$. Sachant θ, mais également $\mu(s)$, nous pouvons générer un grand nombre de jeux de données respectant la structure du modèle, afin d'évaluer son biais $\hat{h}_s(\theta)$ et sa variance $\hat{\sigma}_s^2(\theta)$. Posant $\hat{y}(s) = T_s(y)$, nous écrivons l'approximation :

$$\begin{aligned} \hat{y}(s) &\approx Z_s + \varepsilon_s \\ \varepsilon_s &\sim N(\hat{h}_s, \hat{\sigma}_s^2) \end{aligned} \tag{B.4}$$

La figure B.1.b montre l'estimateur construit (en ligne rouge). Cet estimateur a été conçu de la manière suivante : nous appuyant sur $\mathbb{E}(Y(s)|\theta) = \frac{\mu(s)}{\rho}$, nous avons d'abord évalué les valeurs $\hat{Z}(s) = (\log(Y(s) - \log(\rho) - \mu(s))$ pour les $Y(s)$ non nuls puis, fort de la connaissance de $[Y(s) = 0|\theta] = \exp(-\mu(s))$, nous calculons $f_s(Y)$ la fréquence de valeurs nulles dans un voisinage du point s où $Y(s) = 0$, et $\hat{Z}(s) = (\log(-\log(f_s(Y))) - \mu(s))$. L'estimateur $T_s(Y)$ est obtenu en lissant les valeurs $\hat{Z}(s)$ ainsi construites à partir de l'échantillon. Par simulation sur des échantillons virtuels, on évalue son biais $\hat{h}_s$ et sa variance $\hat{\sigma}_s^2$ pour les valeurs nulles et non nulles de Y. Sur la figure B.1c apparait également l'estimateur débiaisé (en ligne bleue), reconnaissable à sa trajectoire plus heurtée.

Remplaçant dans la formule B.2 y, h_s, σ_s par $\hat{y}, \hat{h}, \hat{\sigma}_s^2$, on dispose maintenant d'un générateur $p(X|\theta)$ de champs latents auxiliaires.

La figure B.1.d semble indiquer que le générateur fournit des champs latents candidats proches de la configuration initiale (bien sûr inconnue en situation réelle, mais qui sont ici représentés par quatre panels) qui nous avait servi à générer les données et qui ont donc de très bonnes chances d'être acceptés par l'algorithme $MH + IS$ (voir section C.2).

En résumé, pour faire fonctionner l'algorithme d'inférence $MH + IS$ dont la structure est donnée en annexe C, il faut donc les trois éléments suivants à coder sous forme de trois sous-programmes.

1. Une loi de proposition $q(\theta^*|\theta)$ pour le Metropolis sur les paramètres, par exemple une marche aléatoire, avec inno-

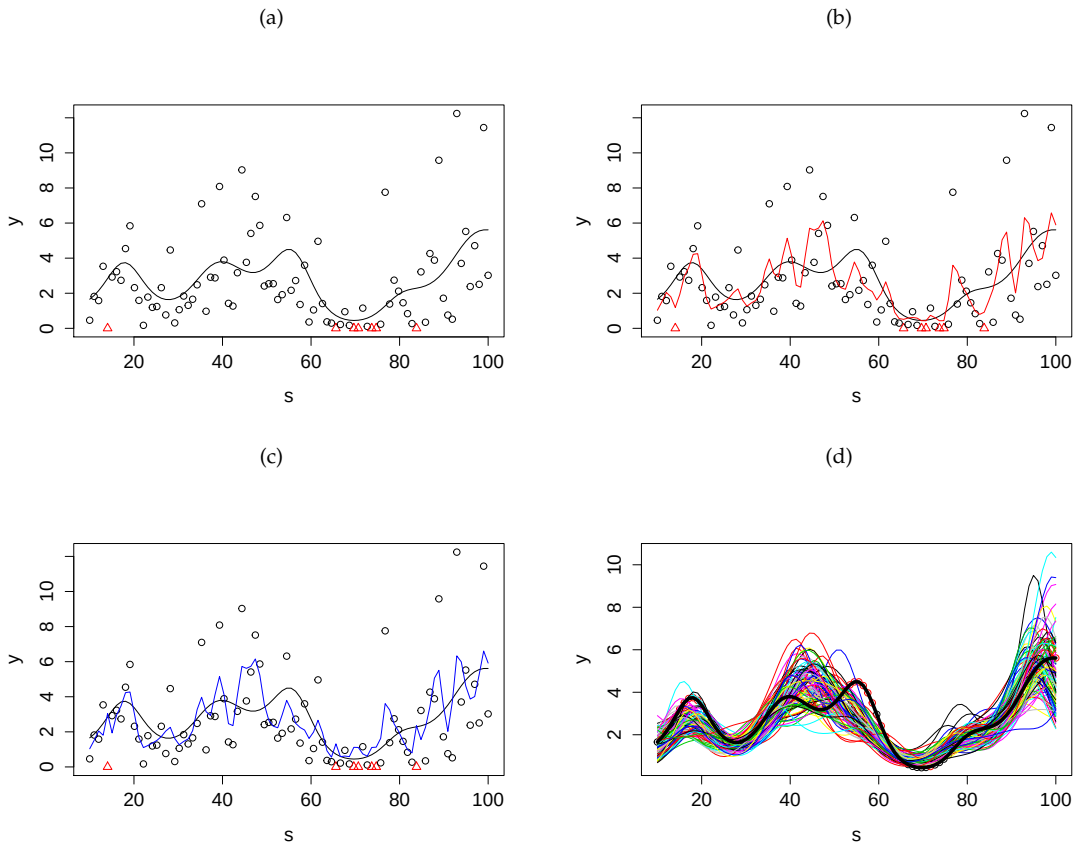

FIGURE B.1 : Visualisation d'une configuration de grille linéaire. (a) Couche latente et données. (b) Estimateur obtenu par lissage. (c) Estimateur débiaisé. (d) 100 Champs latents auxiliaires.

vation normale ou uniforme dont il faut régler la variance d'exploration.

2. Ayant les données y, sur les S sites, on doit pouvoir disposer d'une routine pour calculer $\pi(\theta, X) = [\theta, X|, y]$ à une constante près, ce que donne la loi conjointe :

$$[\theta, X, y] = \left(\prod_{s=1}^{S} [y(s)|z(s), \theta] \right) \times \left(\prod_{g=1}^{G} N(x(g), 0, \sigma_x^2 \right) \times [\theta]$$

où les lois $[y(s)|z(s), \theta]$ proviennent du processus d'observation avec $Z = KX$.

3. Une dernière routine calcule biais et variance de l'estimateur $T_s(Y)$, puis génère R tirages multivectoriels de la

loi normale $p(.|\theta)$ de caractéristiques données par l'équation B.2.

B.4 PREMIÈRES APPLICATIONS DE L'ALGORITHME $MH + IS$

Pour mettre au point l'algorithme, nous avons commencé par fixer dans le vecteur des paramètres θ toutes les composantes à une valeur connue, sauf une : μ (puis σ_x). La simulation de données distribuée selon une loi CPG est effectuée selon la procédure présentée en section 5.2.1 avec un nombre de points réduit ($S = 150$). La structure ainsi obtenue est similaire à celle présentée en figure 5.1.

L' inférence bayésienne partant de données générées avec une moyenne nulle $(\mu = 0)$, une moyenne positive ($\mu = 1$) ou une moyenne négative ($\mu = -1$) donne à chaque fois des résultats concluants du type de la figure B.2, obtenue dans le cas $\mu = 0$ avec une loi *a priori* peu informative pour ce paramètre (normale centrée d'écart type 10). L'algorithme $MH + IS$ a été configuré avec 5000 itérations Metropolis-Hasting (selon une marche aléatoire normale partant de $\mu = 10$ avec une variance d'exploration de 1) au sein de chacune desquelles on a effectué $R = 100$ réplicats de valeurs aux noeuds de grille.

De la même façon, on peut mener l'inférence de σ_x. Partant d'un prior assez peu informatif (inverse-gamma(1,1)), la figure B.3 montre une rassurante allure de la loi *a posteriori* obtenue par le même algorithme $MH + IS$ pour le paramètre de variance.

B.5 CONCLUSIONS

Dans les résultats partiels, mais encourageants, présentés ici, seul un paramètre à la fois est estimé par l'algorithme. Les prochaines étapes, non mises en oeuvre dans cette thèse faute de temps, seraient d'estimer simultanément tous les paramètres du modèle, c'est à dire conjointement la moyenne du champ latent ainsi que les paramètres qui contrôlent la quantité de biomasses présente dans un patch. Bien que nos premiers essais se révèlent prometteurs, un grand nombre de questions de réglage de l'algorithme $MH + IS$ restent aujourd'hui ouvertes :

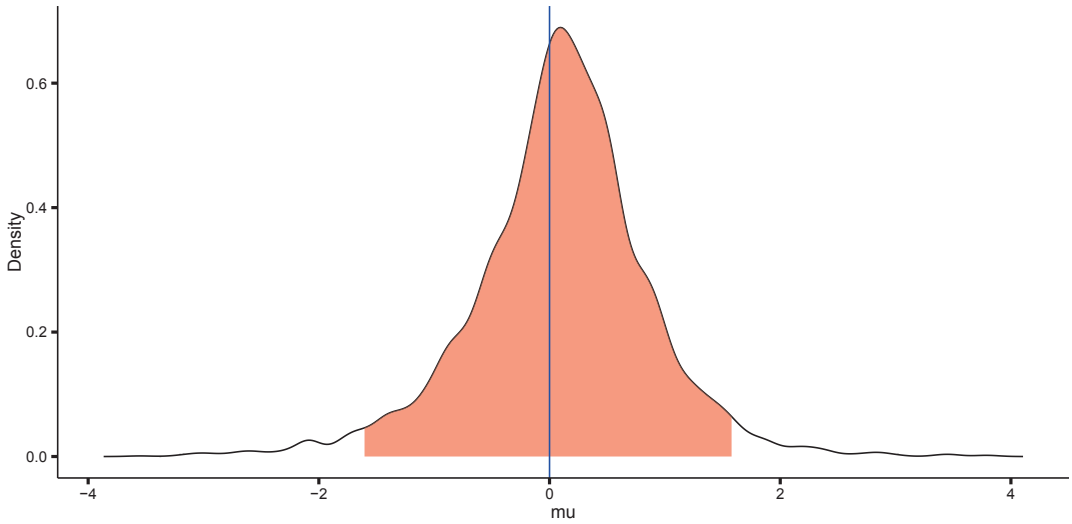

FIGURE B.2 : Distribution *a posteriori* du paramètre μ. La zone rouge représente l'intervalle à 95% *a posteriori*. La valeur initiale du paramètre μ est représentée par la ligne bleue.

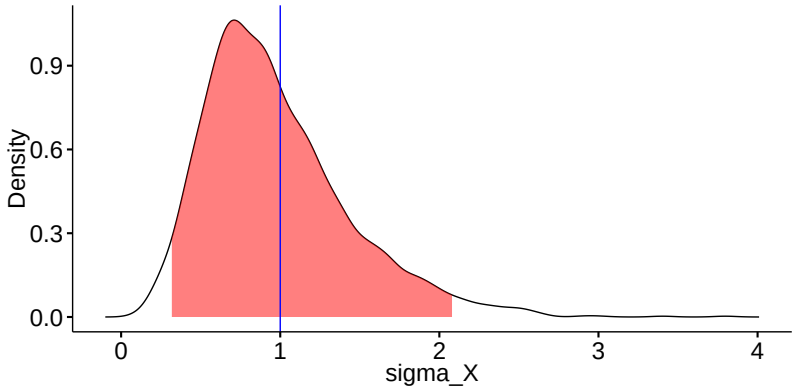

FIGURE B.3 : Distribution *a posteriori* du paramètre σ_x^2, variance des points de grille. La zone rouge représente l'intervalle à 95% *a posteriori*. La valeur simulée du paramètre σ_x^2 est représentée par la ligne bleue.

1. en théorie, on pourrait choisir $R = 1$ (une seule génération du champ latent), mais l'estimateur $\widehat{[\theta|y]}$ ainsi construit sera pauvre quoique sans biais et la convergence (asymptotique) reste garantie. En pratique, dans les réglages de l'algorithme, le choix de R, le nombre de tirages d'importance de la couche latente à chaque itération du Metropolis joue un rôle crucial : plus on l'augmente, meilleur sera l'estimation de $[\theta|y]$ ainsi que le rapport d'acceptation de Metropolis. Cependant, l'augmentation du nombre de tirages peut ralentir l'algorithme et augmenter les exigences d'espaces de stockage.

2. Rien ne dit que la technique de production de la fonction d'importance $p(X|\theta)$ par conjugaison normale que nous avons mis en oeuvre soit la meilleure ; ignorer la structure spatiale des Z peut simplifier cette fonction, plus de variabilité peut également être introduite (en perturbant par un bruit additionnel autour des $\hat{z}$) afin de plus faire *respirer* plus la fonction d'importance et de mieux explorer mieux les possibles champs latents ?

3. Enfin, malgré le caractère très générique de la méthode, l'estimateur $T(Y)$ doit être adapté aux spécificités de la loi d'observation. Pour mettre en place une procédure computationnelle bayésienne $MH + IS$ la plus efficace possible, il est donc nécessaire de se rapprocher de la construction d'estimateurs très utilisée en statistique fréquentiste.

C

MÉTHODES AVANCÉES D'INFÉRENCE

C.1 L'ALGORITHME DE METROPOLIS-HASTINGS PSEUDO MARGINAL D'ANDRIEU, DOUCET ET HOLENSTEIN

Imaginons que, comme à la section A.1, nous voulions échantillonner une distribution cible $\pi(\theta)$ grâce à un algorithme de Metropolis selon la loi de proposition $q(\theta^*|\theta)$, mais que nous n'ayons accès qu'à une estimation $\hat{\pi}$ de la densité de π, par exemple obtenue grâce à un algorithme d'échantillonnage préférentiel (*importance sampling* comme dans les systèmes particulaires de Andrieu *et al.* (2010)). La condition de non biais $\mathbb{E}(\hat{\pi}) = \pi$ est la condition clé pour montrer que la substitution brutale de π par la densité bruitée $\hat{\pi}$ dans le rapport d'acceptation de Métropolis Hasting $\rho(\theta, \theta^*) = \frac{\pi(\theta^*)q(\theta|\theta^*)}{\pi(\theta^{(i)})q(\theta^*|\theta)}$ permet d'obtenir un échantillon dont la loi est *exactement* π. Voici schématiquement les éléments de preuve.

Appelons η la variable aléatoire $\frac{\hat{\pi}(\theta)}{\pi(\theta)}$ représentant le bruit de cette estimation et appelons $f(\eta|\theta)$ sa densité (qui est conditionnelle à θ : car de fait, à chaque itération de l'algorithme MH on propose à la fois θ et η, même si η reste invisible). Considérons l'algorithme de MH étendu sur l'espace (θ, η). Chaque itération de l'algorithme génère alors *ipso facto* un couple (θ^*, η^*) selon la loi de proposition $q(\theta^*|\theta) \times f(\eta^{*\prime}|\theta^*)$.

Le rapport d'acceptation avec plug-in brutal est $\hat{\rho}(\theta, \theta^*) = \frac{\hat{\pi}(\theta^*)\times q(\theta|\theta^*)}{\hat{\pi}(\theta)\times q(\theta^*|\theta)} = \frac{\eta^*\times q(\theta|\theta^*)\times\pi(\theta^*)}{\eta\times q(\theta^*|\theta)\times\pi(\theta)}$. Si on le réécrit pour faire apparaitre la proposition sur l'espace augmenté en η, il vient :

$$\hat{\rho}(\theta, \theta^*) = \frac{\{q(\theta|\theta^*) \times f(\eta|\theta)\} \times \pi(\theta^*) \times \eta^* \times f(\eta^*|\theta^*)}{\{q(\theta^*|\theta) \times f(\eta^*|\theta^*)\} \times \pi(\theta) \times \eta \times f(\eta|\theta)}$$

On voit que la loi cible en (θ, η) est $[\theta, \eta] \propto \pi(\theta) \times \eta \times f(\eta|\theta)$.

Or la loi marginale en θ de cette loi cible est $\int [\theta, \eta] d\eta \propto \pi(\theta) \times \int \eta \times f(\eta|\theta) d\eta = \pi(\theta)$ puisque $\mathbb{E}(\eta) = 1$ (il suffit d'ailleurs que ce soit une constante).

Ainsi, cet algorithme donne bien un échantillonnage de θ selon π de façon complètement *exacte* malgré (ou plutôt grâce) à des tirages dans $\hat{\pi}$, pour peu que $\hat{\pi}$ soit un estimateur sans biais de π.

C.2 L'IMPORTANCE SAMPLING MARGINALISÉE D'ANDIEU ET DOUCET

On dispose d'une loi conjointe $\pi(\theta, X)$ et notre objectif est d'évaluer la loi marginale $\pi(\theta)$. Soit une loi auxiliaire paramétrée par θ, $p(X|\theta)$ avec laquelle nous engendrons le vecteur X à R composantes (réplicats *iid*) $(X^r)_{r=1..R} \underset{iid}{\sim} p(X|\theta)$. Considérons la quantité :

$$\tilde{\pi}(\theta, \mathbf{X}) = \frac{1}{R} \sum_{r=1}^{R} \frac{\pi(\theta, X^r)}{p(X^r|\theta)}$$

Vue comme une fonction de θ, cette quantité est la somme des poids non normalisée qui estime (sans biais, voir section C.1) la constante d'intégration ($\pi(\theta) = \int_x \frac{\pi(\theta,x)}{p(x|\theta)} p(x|\theta) dx$) quand on génère $X|\theta$ par tirage d'importance avec l'instrumentale p.

Regardons l'algorithme markovien introduisant un calcul d'importance sampling au sein d'une procédure de Metropolis-Hasting $(MH + IS)$ suivant :

- proposer le vecteur candidat : $\theta^* \sim q(\cdot|\theta)$ et $X^* \underset{iid}{\sim} \prod_{r=1}^{R} p(X_r|\theta^*)$
- accepter (ou doublonner) selon

$$\rho((\mathbf{X}, \theta) \to (\mathbf{X}^*, \theta^*)) = \frac{\tilde{\pi}(\theta^*, \mathbf{X}^*)}{\tilde{\pi}(\theta, \mathbf{X})} \times \frac{q(\theta|\theta^*)}{q(\theta^*|\theta)}$$

Cet algorithme markovien s'appuie sur la quantité $\tilde{\pi}(\theta, X) = \frac{1}{R}\sum_{r=1}^{R}\frac{\pi(\theta,X^r)}{p(X^r|\theta)}$; Andrieu et Roberts (2009) montrent qu'il a pour loi invariante :

$$\pi^{\infty}(\theta, \mathbf{X}) = \frac{1}{R}\sum_{r=1}^{R}\pi(\theta, X^r)\prod_{l\neq r}p(X^l|\theta)$$

La loi marginale de la chaine engendrée après tirage MCMC est $\pi(\theta)$.

Il n'y a donc pas de biais dans l'approximation Monte-Carlo π^{∞}pour π.

RÉSUMÉ

Comprendre et étudier la distribution spatiale et temporelle des espèces est un enjeu margeur pour leur gestion. De ce fait, de nombreuses approches statistiques ont été développées pour étudier leur répartition spatio-temporelles. Cependant, modéliser la distribution d'une population pose deux difficultés majeures : (i) les jeux de données disponibles présentent souvent une forte quantité de zéros. Cet excès de zéros implique l'utilisation d'outils statistiques spécifiques, (ii) les valeurs de variables prélevées en des sites voisins sont spatialement corrélées. Ignorer cette autocorrélation peut amener à de fausses conclusions. Deux approches, adaptées aux données continues à forte proportion de zéros, sont comparées lorsque l'effort d'échantillonnage (p. ex. durée, distance) est variable. La comparaison est effectuée par simulations et avec des données de pêches commerciales de poissons de fond au large de Vancouver. L'approche delta, la plus utilisée en écologie, ne permet pas de prendre en compte cet effort variable correctement, tandis que la seconde approche, un Poisson composé à marques gamma (CPG), est robuste à cette variabilité. La seconde partie de ces travaux s'est intéressée à la modélisation spatiale de quantités de biomasses d'invertébrés épibenthiques dans le golfe du Saint-Laurent. Chaque année depuis 1989, Pêche et Océan Canada organise une campagne de pêche scientifique dans cette région. De telles données sont idéales pour comprendre les changements spatio-temporels d'abondance de populations. Un modèle hiérarchique bayésien (HBM) basé sur le modèle CPG est proposé pour étudier la répartition spatiale de ces espèces d'invertébrés. La structure spatiale de la biomasse est modélisée grâce à un variogramme exponentiel associé à trois variables environnementales (sédiment, profondeur et température) dans la couche latente du HBM. Cette méthode produit des cartes de quantités d'intérêt (p. ex. probabilité de présence, biomasse moyenne) en tenant compte de l'incertitude des paramètres estimés ainsi que les erreurs associées aux observations. Ce HBM fournit des outils utiles pour la gestion spatiale de populations. Enfin, une approche de modélisation spatio-temporelle des quantités de biomasses d'invertébré est proposée. Le HBM basé sur le modèle CPG est une nouvelle fois utilisé, mais l'utilisation du variogramme est remplacée par une grille latente, support de la dépendance spatio-temporelle. La structure spatiale est approchée par convolution discrète d'un bruit blanc à l'aide d'un noyau exponentiel sur la grille latente. Cette même grille contrôle également la dépendance temporelle qui est modélisée par un processus autorégressif d'ordre 1. Ce modèle permet d'étudier l'évolution des quantités de biomasses des espèces d'invertébrés du golfe. L'approche spatio-temporelle développée, flexible et performante, ouvre beaucoup de perspectives quant à la modélisation spatio-temporelle de distribution d'espèce.

ABSTRACT

Understanding spatial and spatio-temporal species distribution is a key aspect in a management context. Thereby, a large number of statistical approaches have been developed to study their spatio-temporal distributions. However, the modeling of species distribution raises two main difficulties: (i) typical data present an excess of zero. This large proportion of zero requires specific zero-inflated modeling approaches. (ii) variables sampled in the same area often present spatial autocorrelation. Ignoring this autocorrelation may produce to false conclusion. Two approaches, which can handle zero-inflated data are compared in terms of response to variable sampling effort. This comparison is performed by simulations and using fishery dependent data of groundfish species. Delta-distribution approaches, which are the classical tool to model such zero-inflated data in ecology, can lead to poor estimates when the variability in the sampling effort is important. On the opposite, a compound Poisson with gamma marks (CPG) is much more robust to variable sampling conditions. The second part of this work aims at developing spatial model for representing the biomass distribution of epibenthic invertebrates sampled in the southern Gulf of Saint Lawrence (sGSL) since 1989 by Fisheries and Ocean Canada. Data from such broad scale surveys are ideal for understanding spatial and temporal changes in population abundance and community structure. A Bayesian hierarchical model (HBM) based on the CPG is proposed to study the distribution of these species. The latent spatial process is modeled with an exponential variogram associated with three environmental variables (sediment, depth and temperature). Maps of quantities of interest (e.g. probability of presence, quantity of biomass) are produced, taking into account the uncertainty of the estimated parameters and observation errors. This hierarchical Bayesian modeling approach provides a useful tool for spatial management of human activities that may affect living resources. Finally, a spatio-temporal model is presented to study the evolution of the biomass of invertebrates in the sGSL. The HBM proposed previously is used again, but the variogram is replaced by a latent grid, which accounts for the spatial and temporal dependencies. The spatial structure is modeled by convolving an independent process with an exponential kernel on the grid. This grid is also controlling the temporal dependency, which is modeled with an autoregressive model of order 1. The approach allows to study the evolution of biomass quantities of the epibenthic invertebrates of the sGSL. The spatio-temporel model developed is flexible and efficient and open a whole range of perspectives.