# Modèles statistiques spatialement explicites de données de comptage : analyse bibliographique et comparaison de différentes approches

Yannick Saas



Master 2 Mathématiques et Applications

Spécialité Statistique

Université de Strasbourg

# Modèles statistiques spatialement explicites de données de comptage : analyse bibliographique et comparaison de différentes approches

*Rapport de stage*


| | |
|---|---|
| Auteur : | Yannick SAAS |
| Maître de Stage : | Frédéric GOSSELIN |
| Responsable du Master : | Armelle GUILLOU |
| Référent à l'Université : | Nicolas POULIN |


IRSTEA, 13/02/2012 - 27/07/2012

Université de Strasbourg
UFR de Mathématique et d'Informatique
7, rue René Descartes
67084 - Strasbourg
Tél : 03 68 85 50 00
Fax : 03 68 85 03 28
www.unistra.fr

IRSTEA - Centre de Nogent-sur-Vernisson
Domaine des Barres
45290 - Nogent-sur-Vernisson
Tél : 02 38 95 03 30
Fax : 02 38 95 03 59
www.irstea.fr

*Everything is related to everything else,*
*but closer things more so.*
Waldo Tobler

# Résumé

Les jeux de données écologiques présentent fréquemment des dépendances de type spatiales, consistant en ce que les observations géographiquement proches les unes des autres ont plus tendance à se ressembler que les observations plus distantes. Cet état de dépendance, appelé autocorrélation spatiale, s'il n'est pas correctement pris en compte dans la modélisation statistique, est connu pour produire des tests non fiables ainsi que des estimations imprécises des effets fixes.

Bien que les chercheurs disposent de nombreuses méthodes permettant de modéliser l'autocorrélation spatiale, ils restent souvent perplexes face aux nombreux débats présents dans la littérature d'écologie statistique, où s'opposent les partisans des méthodes spatiales à ceux qui s'en méfient. Beale et al. (2010) apportent des éclaircissements dans le contexte de données continues réparties sur grille, mais les chercheurs ont besoin de conseils supplémentaires pour d'autres types de données dans des contextes spatiaux plus flexibles.

Dans cette étude, nous étendons les travaux de Beale et al. (2010) au cas de données de comptage irrégulièrement réparties dans l'espace. Nous avons sélectionnné quatre méthodes spatiales faciles à implémenter dans la famille des modèles linéaires généralisés mixtes et des modèles additifs généralisés mixtes, incorporant des effets aléatoires spatialement structurés qui permettent de modéliser les dépendances spatiales, soit à partir d'une notion de distance, soit par la spécification de relations de voisinages. A l'aide d'une approche par simulation, on évalue les performances relatives des méthodes testées en termes d'erreur de type I et de précision réelle des estimations. On s'intéresse, de plus, à leur robustesse face à des phénomènes de non-stationnarité spatiale et de sur-dispersion, qui sont des propriétés typiques des données de comptage spatialement autocorrélées en écologie.

Parmi les méthodes fréquentistes et bayésiennes implémentées, les méthodes bayésiennes spatialement structurées estiment les effets fixes de la manière la plus précise avec des erreurs de type I correctes. La méthode spatiale basée sur une inférence MCMC avec mises à jours de Langevin-Hastings s'adapte le mieux à toutes les complexités simulées.

# Remerciements

Je tiens tout d'abord à remercier mon maître de stage, Frédéric Gosselin, de m'avoir proposé un sujet d'étude riche et ambitieux, et de m'avoir accordé sa confiance tout au long du stage en me laissant libre dans mes choix. Merci à lui pour les vives discussions, très constructives, que nous avons eues ensemble.

Je remercie ensuite Christian Ginisty, chef de l'unité de recherche "Ecosystèmes Forestiers", et Frédéric Archaux, chef de l'équipe "Biodiversité", de m'avoir donné l'opportunité de rejoindre les équipes de recherche le temps d'un stage.

Je veux maintenant remercier l'ensemble du personnel de l'unité de recherche de m'avoir accueilli, dans une ambiance amicale et détendue. Je remercie, et ce dans un ordre aléatoire : Gilles, Philippe, Yoan, Emmanuelle, Carl, Marie-Charlotte, Jean-Pascal, Agnès, Dominique, Patrick, Marion, Benoît, Nicolas, Julien, Christophe, Richard, Liping, Vicky, Vincent, etc. Je remercie ensuite plus particulièrement Coryse et Aminata, mes deux collègues de bureau, pour leur bonne humeur.

Merci aussi à Nicolas Poulin, ingénieur de recherche au Centre de Statistique de Strasbourg (CeStatS), pour ses conseils concernant la rédaction du présent rapport.

Pour terminer, je tiens à remercier chaudement les personnes, stagiaires et thésards, avec lesquelles j'ai vécues en colocation pendant plusieurs mois sur le domaine des Barres, à savoir : Céline, Mathilde-Heloïse, Emilie, Guillem, Clément, Christophe et enfin Léonie.

# Table des matières

# Liste des tableaux

# Liste des figures

# Liste des sigles et abréviations

| | |
|---|---|
| **IRSTEA** | Institut national de Recherche en Sciences et Technologies pour l'Environnement et l'Agriculture |
| **CEMAGREF** | CEntre national du Machinisme Agricole, du Génie Rural, des Eaux et des Forêts |
| **ONF** | Office National des Forêts |
| **GIP-ECOFOR** | Groupe d'Intérêt Public - ECOsystèmes FORestiers |
| **CNRS** | Centre National de la Recherche Scientifique |
| **INRA** | Institut National de la Recherche Agronomique |
| **GNB** | Gestion forestière, Naturalité et Biodiversité |
| **RBI** | Réserve Biologique Intégrale |
| **i.i.d.** | indépendantes et identiquement distribuées |
| | |
| **GLM** | Generalized Linear Model |
| **GLMM** | Generalized Linear Mixed Models |
| **GAM** | Generalized Additive Model |
| **GAMM** | Generalized Additive Mixed Model |
| **GLS** | Generalized Least Squares |
| **GWR** | Geographically Weighted Regression |
| **WRM** | Wavelet Revised Method |
| **MCMC** | Markov Chain Monte Carlo |
| **MCMCLH** | Markov Chain Monte Carlo with Langevin-Hastings updates |
| **INLA** | Iterative Nested Laplace Approximation |
| **CAR** | Conditional Auto Regressive |
| **ICAR** | Intrinsic Conditional Auto Regressive |
| **SAR** | Simultaneous Auto Regressive |
| **ARMA** | Auto Regressive Moving Average models |
| **PQL** | Penalized Quasi-Likelihood |
| **QML** | Quasi-Maximum Likelihood |
| **MSE** | Mean Squared Error |
| **RMSE** | Root Mean Squared Error |

# Chapitre 1

# Introduction

## 1.1 IRSTEA, EFNO et l'équipe BIODIVERSITE

J'ai réalisé, du mois de Février au mois de Juillet 2012, mon stage de fin d'études au sein de l'IRSTEA (Institut national de Recherche en Sciences et Technologies pour l'Environnement et l'Agriculture), anciennement CEMAGREF (CEntre national du Machinisme Agricole, du Génie Rural, des Eaux et des Forêts), sous la direction de Frédéric GOSSELIN, ingénieur-chercheur.

IRSTEA est un institut de recherche scientifique, qui se trouve sous la double tutelle du ministère de l'écologie et de l'agriculture. Il mène des recherches à l'échelle du territoire, dans une optique d'aide à la décision publique, sur des enjeux liés au développement durable, tels que les ressources en eau, les systèmes écologiques aquatiques et terrestres et les technologies pour l'environnement et l'agriculture. L'institut est financé pour la majorité de ses projets par des fonds publics (ministères, collectivités territoriales, etc.) et, en moindre partie, par des fonds européens et internationaux.

J'ai intégré l'un des neufs centres régionaux de IRSTEA, le centre de Nogent-sur-Vernisson sur le domaine des Barres, constitué d'une unique unité de recherche : l'unité EFNO - Ecosystèmes forestiers. Les recherches de l'unité portent sur les écosystèmes forestiers de plaine et les pratiques de gestion sylvicole favorables à la préservation de la biodiversité forestière. Les études réalisées conduisent à des modèles théoriques et des publications scientifiques, mais l'unité se veut aussi tournée vers l'appui aux décideurs publics et aux gestionnaires d'espaces forestiers.

Plusieurs équipes de recherche sont présentes au sein de l'unité EFNO, dont l'équipe "Biodiversité" à laquelle j'ai été intégré. L'équipe "Biodiversité" centre ses activités de recherche sur l'étude des composantes de la biodiversité dans les massifs forestiers. L'objectif est de mieux comprendre la biodiversité ainsi que les pressions qui pèsent dessus, afin de proposer des recommandations de gestion forestière et d'aménagement du territoire préservant la biodiversité. Plusieurs projets sont actuellement menés par l'équipe, et c'est dans l'un d'entre eux, le projet GNB, que mon stage s'insère.

## 1.2 Le projet GNB

Le projet GNB (Gestion forestière, Naturalité et Biodiversité) est un projet écologique d'envergure national mené depuis 2008 par l'équipe "Biodiversité" sous la responsabilité scientifique de Frédéric Gosselin. Le projet GNB fait partie du programme BGF (Biodiversité et Gestion Forestière), centré de manière spécifique sur l'étude des relations entre biodiversité et gestion

forestière. Le projet BGF est financé, avec le soutien du ministère de l'écologie, par un groupement d'intérêt public, le GIP-ECOFOR (Groupe d'Intérêt Public - ECOsystèmes FORestiers), qui a pour objectif de permettre à plusieurs organismes publics, notamment l'IRSTEA, l'ONF (Office National des Forêts), l'INRA (Institut National de la Recherche Agronomique) et le CNRS (Centre National de la Recherche Scientifique), de mettre en commun leurs moyens pour la réalisation de projets de grande envergure, tel que le projet GNB.

Le but du projet GNB est de mieux comprendre et de quantifier l'impact de l'exploitation forestière sur la biodiversité, et ainsi d'apporter une aide à la décision politique publique, concernant la préservation de la biodiversité dans les massifs forestiers français. Pour cela, il s'agit de comparer des parcelles forestières exploitées à des parcelles forestières non-exploitées. Les parcelles non-exploitées sont choisies dans des parties intégrales de réserves naturelles ou des Réserves Biologiques Intégrales (RBI). Alors que les réserves naturelles, de manière globale, sont des zones protégées, autorisant néanmoins une exploitation forestière encadrée préservant la biodiversité, les parties intégrales de réserves naturelles, de même que les RBI, sont des zones où l'exploitation forestière est totalement proscrite, et où la forêt est rendue à une évolution naturelle. L'étude repose sur l'échantillonnage de sept groupes taxonomiques : plantes vasculaires, mousses, champignons, chauve-souris, oiseaux, insectes coléoptères carabiques (insectes carnivores très répandus) et insectes coléoptères saproxyliques (insectes dépendants du bois).

Pour mener à bien cette étude, des données, liées à la fois à des variables indicatrices de la biodiversité mais aussi à des variables environnementales, ont été recueillies dans 15 massifs forestiers français, dans 213 placettes d'études en tout. Les massifs choisis sont représentatifs de l'ensemble des massifs forestiers français, et présentent des hétérogénéités, notamment en termes d'altitude (massifs de plaine, massifs de montagne), de type de substrat (massifs acides, massifs calcaires) et d'étendue géographique. Quant aux placettes, elles ont été choisies afin d'obtenir une réelle représentativité des sites, via un tirage aléatoire sous contraintes. En effet, tout d'abord, les placettes en zones non-exploitées sont tirées au sort dans un ensemble de points d'échantillonages. Puis, afin d'obtenir le même nombre de placettes en zones exploitées qu'en zones non-exploitées, on associe à chaque placette tirée au sort dans la zone non-exploitée une placette dans la zone exploitée, en s'assurant de leur complémentarité notamment en termes de type de sol et d'exposition, pour éviter des biais stationnels liés à des différences que l'on ne veut pas tester. Notons bien que la placette est l'unité spatiale de base dans l'étude ; toutes les variables dont on dispose, qu'elles soient environnementales ou indicatrices de la biodiversité, sont mesurées à l'échelle de la placette, générant une observation par placette.

Afin de mesurer de manière quantitative l'effet de l'exploitation forestière sur la biodiversité, une approche statistique est mise au point. Elle doit permettre de tester la significativité de l'influence de la gestion forestière sur les différentes composantes de la biodiversité et, si cet effet est significatif, de mettre en lumière les variables qui expliquent le mieux les variations de biodiversité entre zones exploitées et zones non-exploitées. C'est précisément au niveau de la modélisation statistique que mon stage entre en jeu ; il s'agit d'améliorer la modélisation en tenant compte d'une caractéristique importante des données, à savoir l'autocorrélation spatiale.

## 1.3 Problématique statistique

La biodiversité forestière est étudiée principalement à travers deux grandeurs : l'abondance et la richesse. L'abondance représente, pour une espèce donnée et une placette donnée, le nombre d'individus de l'espèce présents sur cette placette. Pour un groupe taxonomique donné et une

placette donnée, la richesse représente le nombre d'espèces différentes appartenant au groupe considéré sur la placette donnée. Les données d'abondance et de richesse sont des grandeurs discrètes et positives ; il s'agit de données de comptage. Il s'agira bien sûr de tenir compte de la nature discrète positive des données, que l'on modélisera à l'aide de distributions de probabilités appropriées, car les modèles normaux sont inadaptés.

Le point de départ de mon stage est le suivant : les données d'abondance et de richesse présentent fréquemment des dépendances de type spatiales ; deux observations géographiquement proches ont plus tendance à se ressembler que deux observations géographiquement éloignées. Ce phénomène, appelé autocorrélation spatiale, sera présenté en détails dans le chapitre 2. Ainsi, au sein d'un même massif, les observations d'abondance ou de richesse issues de deux placettes spatialement proches sont plus "corrélées" que celles issues de deux placettes éloignées. La question se pose alors de la nécessité de tenir compte de la nature spatiale des données dans la modélisation statistique.

L'objet premier de mon stage consiste à montrer dans quelle mesure il est crucial de tenir compte du caractère spatial des données, par la modélisation des corrélations inter-placettes à l'aide de modèles géostatistiques appropriés. L'intérêt d'une telle modélisation ainsi que la mise en oeuvre de modèles spatiaux explicites adaptés à des données de comptage est l'objet du * chapitre 3. Dans cette étude, on s'intéresse de manière spécifique à la mise en oeuvre de modèles linéaires généralisés mixtes (GLMM) et de modèles additifs généralisés mixtes (GAMM), qui modélisent les dépendances spatiales à l'aide d'effets aléatoires corrélés ; on parle d'effets aléatoires spatialement structurés. L'object second du stage est alors de comparer les performances relatives des différentes méthodes spatiales implémentées, lorsqu'elles sont confrontées à des jeux de données présentant des caractéristiques spatiales de plus en plus complexes. On cherche, par ailleurs, à tester la robustesse des méthodes les plus performantes à des phénomènes de non-stationnarité spatiale, et de dispersion extra-spatiale (sur-dispersion et sous-dispersion).

## 1.4 Stratégie d'attaque

Afin de tester et de comparer les différents modèles implémentés, une approche par simulation est mise en oeuvre chapitre 4. Plusieurs scénarios spatiaux sont envisagés. On commence par générer des données de comptage régulièrement espacées sur une grille 20x20, en simulant des propriétés spatiales de plus en plus complexes. On génère ensuite des observations irrégulièrement réparties dans l'espace, à l'aide des coordonnées géographiques réelles des placettes d'étude du projet GNB ; l'objectif est de simuler des données qui présentent des propriétés spatiales réelles.

L'intérêt principal d'une telle approche par simulation repose sur le fait que le paramètre vectoriel de régression $\beta$ est fixé et connu a priori, ce qui n'est jamais le cas dans l'étude de données réelles. Cela permet de comparer la qualité de l'inférence des différents modèles, en se concentrant sur l'estimation des effets fixes, par la mise en oeuvre d'indicateurs de performances classiques tels que l'erreur de type I et le Root Mean Squared Error (RMSE).

Une étude similaire a été réalisée par Beale et al. (2010) dans un cadre gaussien de données quantitatives continues. Malheureusement, cette étude souffre de plusieurs limitations à différents niveaux, concernant notamment la qualité des méthodes spatiales testées et le réalisme des scénarios spatiaux simulés. Dans cette étude, on essaye d'aller au-delà de ces limitations en implémentant des modèles spatiaux pertinents, notamment via des algorithmes bayésiens évolués,

en explorant différentes manières de modéliser l'information spatiale (voir chapitre 3). Enfin, un grand soin a été apporté à l'élaboration de scénarios spatiaux réalistes (voir chapitre 4) afin de permettre une extrapolation justifiée de nos résulats à l'étude de données réelles d'écologie.

# Chapitre 2

# L'autocorrélation spatiale

Ce premier chapitre est consacré à la notion d'autocorrélation spatiale, que l'on définit dans un premier temps avant d'en présenter les principales caractéristiques. On évoque ensuite quelques diagnostics permettant de la détecter et de la quantifier.

## 2.1 Introduction

### 2.1.1 Définition

L'autocorrélation spatiale est définie par l'absence d'indépendance entre observations géographiques. En présence d'autocorrélation spatiale, la répartition spatiale des valeurs n'est pas aléatoire ; il y a un phénomène de ressemblance des valeurs en fonction de la localisation géographique. On parle d'autocorrélation spatiale *positive* lorsque les observations proches ont plus tendance à se ressembler que les observations plus éloignées, et d'autocorrélation spatiale *négative* dans le cas inverse de dissemblance des observations proches. Dans cette étude, on s'intéresse uniquement à l'autocorrélation spatiale positive qui est de loin la plus fréquente en écologie. Concernant le vocabulaire, on parlera de données *spatialement autocorrélées* ou *spatialement structurées*, ou sinon simplement de données *spatiales*.

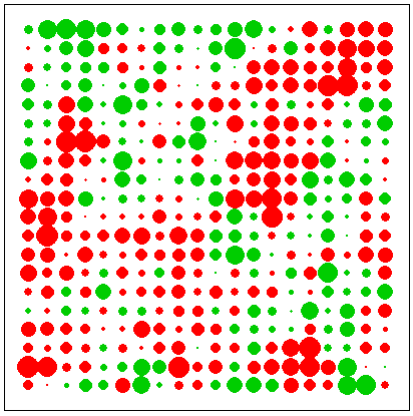
Cette notion a été introduite pour la première fois par Tobler (1970), dans son énoncé de la première loi de la géographie :

*Everything is related to everything else, but closer things more so.*

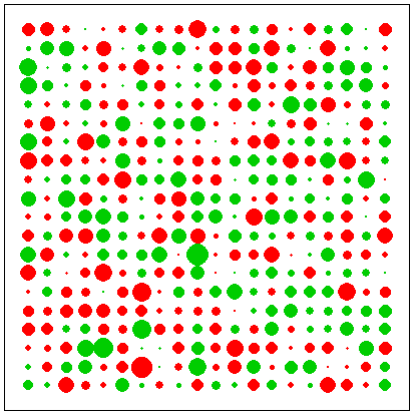
Une illustration graphique du phénomène d'autocorrélation spatiale est donnée figure 2.1. Deux types de données ont été simulés, à savoir des données quantitatives continues simulées à partir d'une loi normale, et des données de comptage simulées à partir d'une loi de Poisson. Pour chaque type de données, on a simulé sur une grille 20x20 des observations spatialement structurées ainsi que des observations aléatoirement réparties dans l'espace ; le chapitre 4 détaille précisément la manière dont de telles simulations ont été mises en oeuvres. Concernant l'intepprétation de tels graphes, précisons que la couleur des points correspond au signe de l'observation, à savoir rouge pour les valeurs négatives et vert pour les valeurs positives. Quant à la taille des points, elle correspond à la valeur absolue de l'observation. Les graphes (a) et (c) illustrent de manière explicite le phénomène d'autocorrélation spatiale, où l'on observe des dépendances spatiales marquées qui se manifestent par la formation de paquets, ce qui n'est pas le cas pour les observations aléatoirement réparties dans l'espace, représentées graphes (b) et (d).

FIGURE 2.1 – Exemple de données spatialement structurées et de données aléatoirement réparties dans l'espace

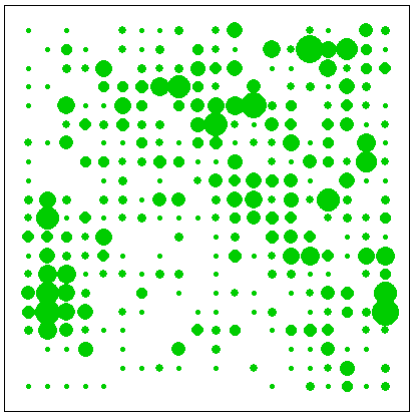
(a) Données continues spatiales



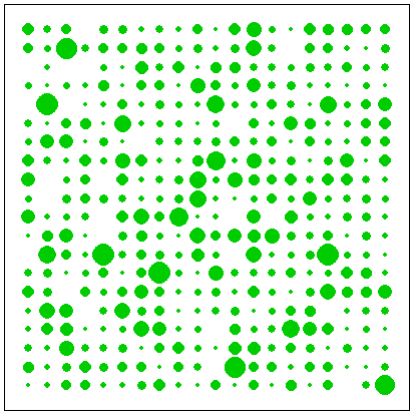
(b) Données continues non-spatiales



(c) Données de comptage spatiales



(d) Données de comptage non-spatiales



### 2.1.2 Causes

L'autocorrélation spatiale trouve son origine, de manière générale, dans le fait que les données sont affectées par des processus qui relient des lieux différents, mettant en jeu des interactions spatiales sous-jacentes : on parle d'*environnement aléatoire corrélé dans l'espace*. Ces processus génèrent une organisation spatiale particulière des données, qui est dûe principalement à un phénomène de diffusion. Celui-ci consiste en effet à observer que l'intensité d'un phénomène dépend de la distance à l'origine. Ainsi, si deux observations sont spatialement proches et ont donc des distances à l'origine proches, elles auront des valeurs similaires. Pour illustrer cette idée, considérons par exemple la répartition spatiale du nombre de fourmis dans une fourmillière. D'un point de vue spatial, la fourmillière consiste en un phénomène de diffusion du nombre de fourmis à partir d'un point central, le centre de la fourmillière. Il se produit un phénomène de diffusion à partir de ce point central ; on observe en pratique de moins en moins de fourmis à mesure que l'on s'éloigne du centre, ainsi que des dépendances spatiales entre observations géographiques proches.

Dans le contexte de données de répartition d'espèces, l'autocorrélation spatiale traduit simplement le fait que les espèces ne sont pas réparties de manière aléatoire dans l'espace. Cela

s'explique par des principes dynamiques d'écologie, tels que la spéciation (le principe de génération des espèces) et la dispersion (le principe de mobilité des espèces) (Legendre and Fortin, 1989; Legendre, 1993). Par ailleurs, des variables environnementales spatialement structurées telles que la température, le type de sol ou le niveau d'acidité, formant un environnement aléatoire spatialement corrélé dans l'espace, influent directement sur les phénomènes de spéciation et de dispersion, et donc sur la structuration spatiale des données (Legendre, 1993; Fortin and Dale, 2005).

## 2.2 Caractéristiques des structures spatiales

### 2.2.1 Portée et intensité

L'autocorrélation spatiale peut être caractérisée principalement par deux grandeurs : la portée et l'intensité. La portée correspond à la distance sur laquelle vont s'étaler les dépendances spatiales ; elle peut donc être vue comme la distance à partir de laquelle on peut considérer que les observations sont spatialement indépendantes. L'intensité, encore appelée force, correspond au degré d'intensité des dépendances spatiales.

Les notions de portée et d'intensité sont illustrées figure 2.2, qui représente quatre jeux de données continues et spatialement structurées, avec des paramètres de portée et d'intensité qui varient. Les graphes (a) et (c) illustrent l'effet d'une différence de portée spatiale, à intensité fixe : la distance sur laquelle s'étalent les dépendances spatiales augmente lorsque la portée est plus élevée. Les graphes (a) et (b) illustrent l'effet d'une différence d'intensité, à portée fixe : le degré de variation des valeurs associées aux observations spatialement dépendantes est plus élevé lorsque l'intensité est plus forte.

### 2.2.2 Stationnarité et non-stationnarité spatiale

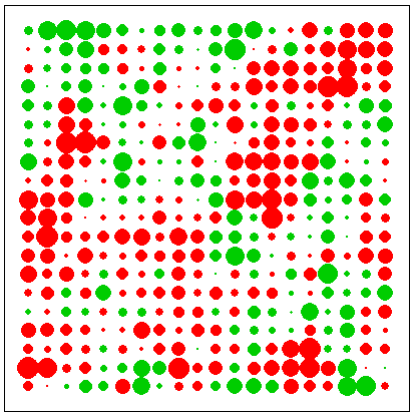
Les jeux de données spatialement structurées vus jusqu'à présent présentent tous la caractéristique de stationnarité spatiale : l'autocorrélation ne varie pas selon la région de l'espace. L'autocorrélation spatiale entre deux points dépend uniquement de la distance qui sépare les deux points, et non de la région de l'espace dans laquelle ils se situent.

Cependant, il peut arriver en pratique que ce ne soit pas le cas ; la manière dont les observations dépendent spatialement les unes des autres n'est pas constante dans l'espace. Cette non-stationnarité spatiale trouve sa cause dans les mécanismes sous-jacents à l'origine de l'autocorrélation spatiale, qui peuvent tout à fait ne pas agir de la même manière selon les régions. Par exemple, il est bien connu que la mobilité des espèces (dispersion), qui est l'un des mécanismes à l'origine de l'autocorrélation spatiale, tend à être plus faible en montagne qu'en plaine, générant ainsi une autocorrélation spatiale de plus faible portée en montagne qu'en plaine.

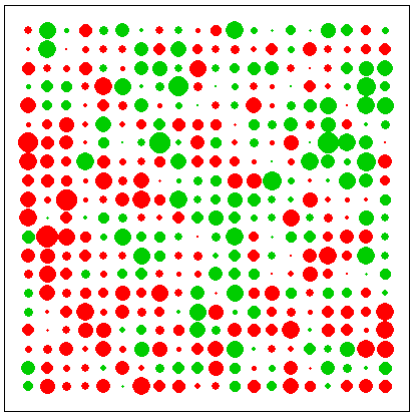
La figure 2.3 représente deux jeux de données spatialement structurées, incorporant des composantes de non-stationnarité spatiale. On a simulé, tout d'abord sur le graphe (a), un phénomène de non-stationnarité de la portée spatiale : la portée de l'autocorrélation spatiale varie selon les régions. Sur notre simulation, la portée est faible du côté gauche, et augmente au fur et à mesure que l'on va vers la droite. Un autre type de non-stationnarité spatiale a été exploré sur le graphe (b) ; on a simulé des tendances spatiales continues et linéaires d'une région à l'autre de l'espace, à l'aide d'un gradient linéaire d'intensité spatiale, de gauche à droite sur notre simulation. On renvoit le lecteur au chapitre 4 pour des détails concernant la manière dont ces champs non-stationnaires ont été générés.

FIGURE 2.2 – Exemple de données spatialement structurées avec différents paramètres de portée et d'intensité

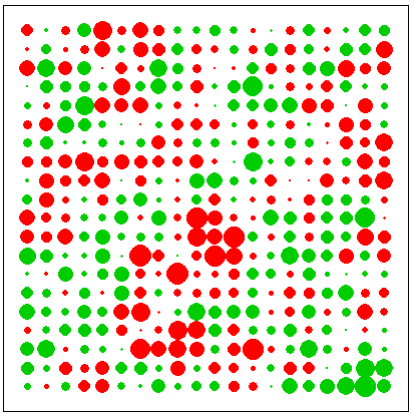
(a) Portée élevée, intensité élevée



(b) Portée élevée, intensité faible



(c) Portée faible, intensité élevée



(d) Portée faible, intensité faible

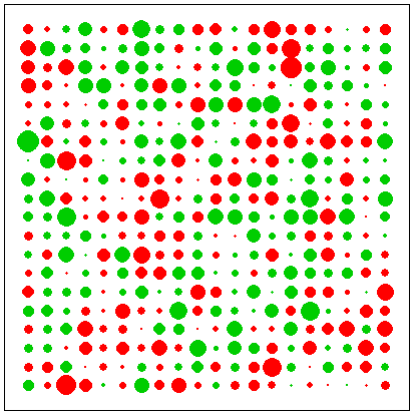
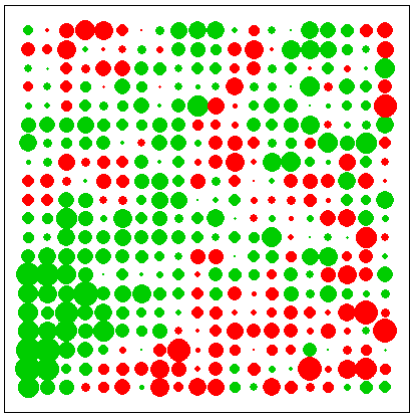
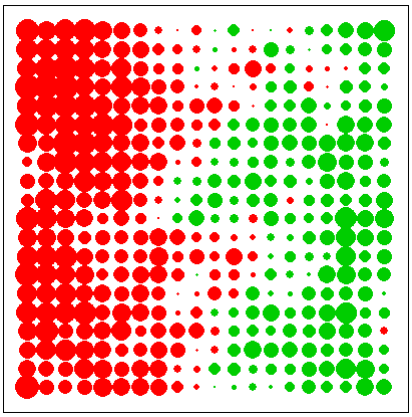


FIGURE 2.3 – Illustration de deux types de non-stationnarité spatiale

(a) Portée non-stationnaire



(b) Tendances spatiales linéaires



## 2.3 Diagnostics d'autocorrélation spatiale

### 2.3.1 Introduction

Dans le cadre de l'étude de données réelles, on aimerait savoir si une série de données présente ou non de l'autocorrélation spatiale, et si une modélisation spatiale explicite, telle que décrite au chapitre 3, présente un intérêt. Cela est possible grâce à une analyse spatiale exploratoire des données, qui permet, avant toute modélisation statistique, de détecter et de quantifier le degré de structuration spatiale présen dans les données. Il existe de nombreuses analyses spatiales exploratoires, telles que la détection de l'autocorrélation spatiale, la détection de tendances spatiales non-stationnaires, ou encore la détection d'individus atypiques du point de vue spatial (R.P., 2003).

Dans cette étude, on s'intéresse de manière spécifique au diagnostic de la présence d'autocorrélation spatiale, et à la représentation graphiques des dépendances spatiales. Deux grandes familles de méthodes ont été mises au point pour cela : les variogrammes et les indices d'autocorrélation (Cressie, 1993; S., 2010). L'approche par variogramme repose sur l'étude de la variance des différences des variables aléatoires considérées en différents points de l'espace, par analogie avec les fonctions d'autocorrélation introduites pour l'analyse des dépendances temporelles dans la théorie des séries chronologiques. L'approche par variogramme n'est pas présentée dans ce rapport ; on s'intéresse uniquement à l'approche par indice d'autocorrélation.

Les deux indices d'autocorrélation les plus connus sont le le $C$ de Geary (Geary, 1954) et le $I$ de Moran (Moran, 1950), auquel on va s'intéresser de manière spécifique. En effet, le $I$ de Moran est réputé pour être le plus robuste (Cressie, 1993), et la pratique a montré que le test statistique d'autocorrélation spatiale basé sur le $I$ de Moran est plus puissant que celui basé sur le $C$ de Geary. Par ailleurs, comme on le verra dans la suite, le $I$ de Moran s'interprète facilement et offre des possibilités graphiques intéressantes.

Comme nous le verrons dans le chapitre 4, on s'intéressera surtout dans le cadre de notre approche par simulation, à la détection d'autocorrélation spatiale dans les résidus des modèles ajustés, et non sur les observations spatiales elles-mêmes.

### 2.3.2 Le $I$ de Moran

Le $I$ de Moran est un indice permettant de quantifier le degré de structuration spatiale d'une série d'observation. Il repose sur la spécification a priori de relations de voisinage entre observations. Avant d'introduire le $I$ de Moran dans son approche globale et locale, introduisons donc tout d'abord la notion de voisinage.

#### La spécification des voisinages

La notion de voisinage, qui représente la façon dont sont spatialement connectées les observations, est une notion centrale en statistiques spatiales, qui va intervenir à la fois dans la mise en oeuvre d'indices d'autocorrélation tels que le $I$ de Moran, mais aussi dans la mise en oeuvre pratique de certains modèles spatiaux explicites (voir chapitre 3).

Les relations de voisinages sont spécifiées de manière plus ou moins complexe via une matrice carrée de dimension le nombre d'observations, appelée matrice de voisinage ou matrice de poids. Dans cette étude, on considère un cas simple non-pondéré de matrice de poids : la matrice de contiguïté, notée $\mathbf{W}$. Elle est définie de la manière suivante :

$$\mathbf{W} = ((w_{ij})) = \begin{cases} 1 & \text{si i et j sont voisins} \\ 0 & \text{sinon} \end{cases}$$

Pour spécifier de manière concrète les relations de voisinages dans le contexte d'observations irrégulièrement espacées, on utilise dans cette étude la notion de distance limite, en considèrant que deux observations sont voisines si la distance qui les sépare est inférieure à une distance limite fixée a priori. On renvoie le lecteur au chapitre 4 pour la spécification pratique des voisinages, selon le contexte spatial des simulations.

**Le $I$ de Moran : approche globale**

Notons $I_M$ l'indice de Moran global pour la série d'observations $(z_1, \ldots, z_n)$. $I_M$ est défini par la formule suivante :

$$I_M = \frac{\frac{1}{m}\sum_{i,j} w_{ij}(z_i - \bar{z})(z_j - \bar{z})}{\frac{1}{n}\sum_{i}(z_i - \bar{z})} \tag{2.1}$$

Avec :
$\mathbf{W} = ((w_{ij}))$ : la matrice de contiguïté spatiale décrite précédemment
$\bar{z}$ : la moyenne des observations
$m = \sum_{i,j} w_{ij}$ : le nombre total de paires de voisins

Le $I$ de Moran s'interprète comme le rapport d'une covariance entre observations voisines sur la variance totale observée. Il mesure donc la covariation d'un point et de ses voisins, en ramenant le résultat à la variance de l'ensemble des points. Ainsi, le $I$ de Moran, qui peut être interprété comme la part de variance explicable par les relations de voisinage spatiales, est donc analogue à un coefficient de corrélation. Il est d'autant plus grand que des valeurs semblables apparaissent entre voisins, et d'autant plus petit que des valeurs dissemblables apparaissent entre voisins ; il prend des valeurs proches de 1 en cas d'autocorrélation spatiale positive, de -1 en cas d'autocorrélation spatiale négative et de 0 en cas d'absence d'autocorrélation spatiale. Néanmois, en pratique, contrairement au coefficient de corrélation de Pearson, il peut prendre des valeurs légèrement inférieures à -1 ou supérieures à 1. Par ailleurs, en absence d'autocorrélation spatiale, la valeur théorique attendue du $I$ de Moran est :

$$\mathbb{E}(I_M) = -\frac{1}{n-1}, \text{ qui est proche de 0 quand n devient grand.}$$

Dans le cas où les observations sont gaussiennes, il existe une expression asymptotique de la variance de $I_M$ (voir Cressie (1993) pour plus de détails) :

$$Var(I_M) = \frac{s_1}{m^2}$$

Avec $m = \sum_{i,j} w_{ij}$ et $s_1 = \sum_{i,j}(w_{ij}^2 + w_{ij}w_{ji})$.

Par ailleurs, Cliff and Ord (1981) ont montré que $I_M$ est asymptotiquement gaussien. Cela permet de construire aisément un test testant la significativité statistique de la présence d'autocorrélation spatiale. Il permet de juger si la ressemblance entre points voisins est significativement plus grande que celle attendue dans le cas d'indépendance. Plus précisément, considérons le test d'hypothèse nulle l'hypothèse d'indépendance spatiale. Considérons la statitistique de test suivante :

$$T_M = \frac{I_M - \mathbb{E}(I_M)}{\sqrt{Var(I_M)}}$$

D'après la normalité asymptotique de $I_M$, $T_M$ suit une loi normale centrée-réduite. On peut alors mettre en oeuvre une procédure de décision classique, en fixant l'erreur de première espèce $\alpha$ (typiquement à 5%) et en considérant $q^{1-\frac{\alpha}{2}}$ le quantile à l'ordre $1 - \frac{\alpha}{2}$ de la loi normale centrée-réduite.

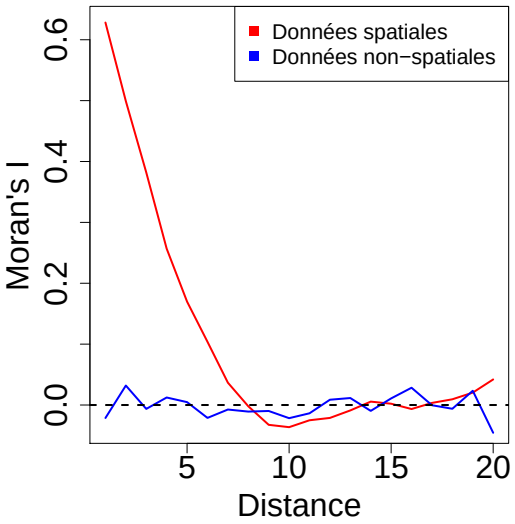
### Le $I$ de Moran : approche locale

L'approche qui vient d'être vue est l'approche globale du $I$ de Moran, qui permet de quantifier le degré d'autocorrélation globale. Cependant, l'indicateur global n'est pas assez fin pour détecter d'éventuelles spatialités locales marginales. Dans l'approche locale du $I$ de Moran, l'idée est de faire varier les relations de voisinages, des plus étroites aux plus larges, en spécifiant des distances limites, à l'origine de la spécification pratique des voisinages, de plus en plus grandes.

Pour une distance limite donnée, et donc une matrice de voisinage $\mathbf{W}$ donnée, on calcule le $I$ de Moran à l'aide de la formule 2.1. En faisant varier ces distances limites, on obtient ainsi un graphe, appelé le *corrélogramme de Moran*, qui représente la valeur de l'indice en fonction de la distance limite ; chaque point traduit donc le degré d'autocorrélation spatiale pour un voisinage donné. Un tel corrélogramme, appliqué à des données spatialement structurées, décroît typiquement d'un certain niveau d'autocorrélation à une valeur proche de 0, indiquant une absence d'autocorrélation spatiale à partir d'une certaine distance entre observations.

La figure 2.4 présente deux corrélogrammes de Moran. Le premier, en rouge, appliqué à des données spatialement structurées, présente une décroissance d'un certain seuil d'autocorrélation à une valeur proche de 0. Le second, en bleu, appliqué à des données non-spatiales, est constant autour de la valeur 0, indiquant une absence d'autocorrélation spatiale à toutes les échelles.

FIGURE 2.4 – Corrélogrammes de Moran appliqués à des données spatiales et des données non-spatiales



# Chapitre 3

# La modélisation de l'autocorrélation spatiale dans un cadre de régression

On introduit dans ce chapitre la modélisation spatiale explicite de données de comptage spatialement autocorrélées, en présentant tout d'abord l'intérêt d'une telle modélisation dans un cadre de régression. On s'intéresse ensuite, de manière spécifique, à la définition des modèles spatiaux ainsi qu'aux problématiques d'estimation numérique de ces modèles.

## 3.1 Intérêt de la modélisation

Dans un cadre de régression où la variable réponse et les covariables sont spatialement structurées, on peut se demander si les modèles non-spatiaux classiques sont satisfaisants, ou s'il est nécessaire de modéliser de manière explicite l'autocorrélation spatiale.

Théoriquement parlant, si toute la variabilité spatiale de la réponse peut être expliquée par la variabilité spatiale des covariables introduites dans le modèle, l'autocorrélation spatiale est bien modélisée ; on mettra en évidence dans cette étude que dans ce cas les modèles non-spatiaux classiques s'ajustent bien et sont performants (voir chapitre 5). Cependant, en pratique, dans le cadre d'une modélisation de données réelles, on ne peut jamais être sur d'avoir introduit dans le modèles toutes les variables influentes ; si l'une d'entre elles est spatialement structurée, une partie de la variabilité spatiale de la variable réponse n'est alors pas correctement modélisée, et se retrouve dans les résidus du modèle, générant ainsi de l'autocorrélation spatiale résiduelle. Les méthodes non-spatiales classiques ne permettent pas de modéliser cette autocorrélation spatiale résiduelle, qui est toujours potentiellement présente, et qui viole l'une des hypothèses fondamentales du modèle linéaire, à savoir l'indépendance des résidus. Les observations corrélées sont traitées comme indépendantes, ce qui est à l'origine d'une forme de pseudoréplication (Hurlbert, 1984; Cressie, 1993; Fortin and Dale, 2005), qui affecte négativement la qualité de l'inférence des effets fixes. Plus précisément, on note deux conséquences problématiques de l'application de modèles non-spatiaux sur données spatialement autocorrélées :

1) La sous-estimation des écarts-types des estimateurs des effets fixes

La variance des estimateurs des effets fixes est sous-estimée ; il y a un biais dans l'estimation. L'intervalle de confiance fourni par le modèle pour le paramètre de régression $\beta$ est trop serré autour de la valeur estimée $\hat{\beta}$, et la probabilité qu'il contienne la vraie valeur $\beta_{vrai}$ est trop faible. En terme de test, en considérant le test d'adéquation d'hypothèse nulle $\beta = \beta_{vrai}$, cela revient à dire que l'erreur de type I associée à ce test (probabilité de rejeter à tort l'hypothèse nulle)

est bien plus élevée que le taux nominal de 5%. Dans un contexte de tests de significativité, où l'on teste les différentes hypothèses nulles $\beta = 0$, comme les intervalles de confiance vont être trop serrés autour des valeurs estimées et vont donc contenir la valeur 0 avec une probabilité trop faible, les erreurs de type I associées vont elles aussi être trop élevées. Les p-valeurs sont sous-estimées, et les procédures de sélection de modèle ont alors tendance à accepter trop de variables, aboutissant à des choix de modèles trop peu parcimonieux (Lennon, 2000).

Une telle sous-estimation des écarts-types peut s'expliquer de manière intuitive. En effet, les modèles non-spatiaux ne sont pas capables de séparer la variabilité non-spatiale liée au effets fixes à la variabilité liée purement aux dépendances spatiales. Toute la variabilité de la réponse va donc être expliquée par les covariables, dont la significativité va donc être sur-estimée, aboutissant typiquement à une sous-estimation des écarts-types. C'est l'une des conséquences du problème général de pseudoréplication introduit par Hurlbert (1984), consistant à traiter des observations corrélées comme indépendantes. Le phénomène d'autocorrélation spatiale (dépendances spatiales), comme celui des mesures répétées (dépendances temporelles), sont des exemples typiques de pseudoréplication.

Cressie (1993) a introduit quelques développement mathématiques permettant de justifier théoriquement la sous-estimation des écart-types des estimateurs dans un cadre de régression multiple. Nous sommes parvenus à apporter une autre justification mathématique à ce phénomène, dans un cas de dépendances spatiales simples, en montrant que la variance d'une suite de variables aléatoires spatialement structurées est sous-estimée (voir annexe A).

2) Une estimation imprécise des effets fixes

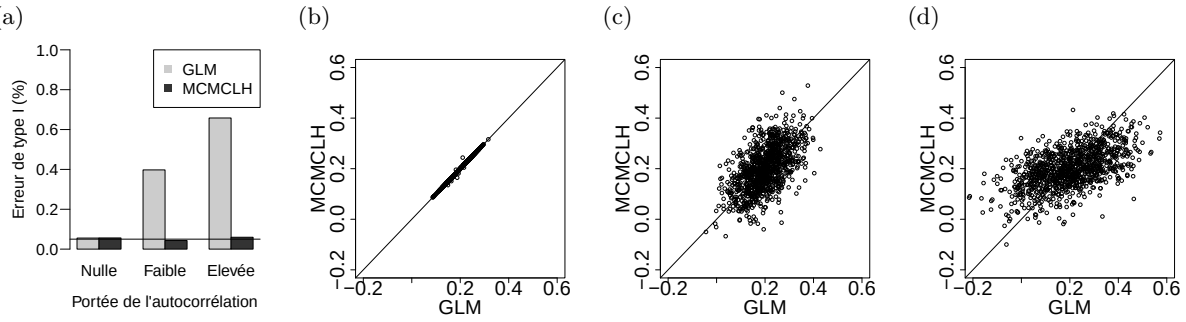
Bien que les estimateurs soient statistiquement non-biaisés, les coefficients de régression des effets fixes sont estimés de manière imprécise. Bien que la mesure de variabilité (écart-type) fournie par le modèle soit sous-estimée, comme décrit précédemment, les approches par simulation permettent d'exhiber des erreurs aléatoires réelles élevées (Beale et al., 2007, 2010). Les estimations deviennent de moins en moins précises à mesure que le degré d'autocorrélation spatiale augmente. C'est une conséquence moins connu de l'application des méthodes non-spatiales, qui ne peut être décelée qu'à travers une approche par simulation, grâce à laquelle on a accès à la variabilité réelle des estimations, en opposition à la variabilité estimée par le modèle. On renvoie le lecteur au chapitre 4 pour plus de précision sur ces divergences de variablité.

Ces deux phénomènes seront explicitement mis en évidence dans cette étude (voir chapitre 5). Cependant, en guise d'illustration, une petite simulation préliminaire a été mise en oeuvre 3.1. On a généré 3 scénarios de 1000 jeu de données sur une grille rectangulaire 20x20, en simulant une variable réponse spatialement structurée ainsi que deux covariables indépendantes spatialement structurées, de telle manière à obtenir selon les scénarios un degré d'autocorrélation résiduelle soit nul, soit faible, soit importante. Pour plus d'informations concernant le processus de simulation, voir chapitre 4. Sur chaque jeu de données de chaque scénario, un modèle non-spatial classique (GLM, en gris) et un modèle spatial performant (MCMCLH, en noir) sont ajustés et comparés en termes d'erreur de type I et de précision des estimations. Pour des informations sur les modèles et la comparaison de modèles, voir chapitre 3 et chapitre 4. On obtient les résultats suivants : (a) En terme d'erreur de type I, qui évalue la qualité globale de l'inférence des effets fixes, alors qu'elle est stable sur tout les scénarios à sa valeur de référence 5% pour le modèle spatial MCMCLH, elle augmente dramatiquement bien au-delà du taux de 5% avec l'augmentation de le portée de l'autocorrélation spatiale, pour le modèle non-spatiale

GLM. Les graphes (b), (c) et (d) illustrent les différences entre les estimations fournies par les deux modèles, lorsque la portée de l'autocorrélation résiduelle augmente de (b) à (d). Alors que les estimations fournies par GLM deviennent de plus en plus éparpillées autour de la vrai valeur 0.2, les estimations fournies par MCMCLH sont bien plus précises.

Précisons enfin que de nombreuses polémiques agitent la littérature d'écologie statistique, concernant l'origine réel de ces problèmes d'inférence ; il n'est pas clair si ces problèmes résultent uniquement ou non de la présence d'autocorrélation spatiale dans les résidus. On va tenter, dans cette étude, d'investiguer ce problème (voir chapitre 4), en essayant de faire le lien entre la présence d'autocorrélation dans les résidus et de tels problèmes d'estimation des effets fixes.

FIGURE 3.1 – Comparaison d'un modèle non-spatial (GLM) et d'un modèle spatial (MCMCLH) en terme d'inférence des effets fixes




## 3.2 La modélisation de données de comptage spatialement autocorrélées

Afin de palier à ces problèmes qui rendent l'inférence statistique peu fiable, des modèles spatiaux explicites permettant de modéliser l'autocorrélation spatiale résiduelle ont été développés. Dans le contexte de données quantitatives continues, de nombreux modèles existent, tels que les modèles auto-régressifs à moyennne mobile (ARMA), les auto-régressions simultanées (SAR) et les auto-régressions conditionnelles (CAR). Dans le contexte de données de comptage qui est celui qui nous intéresse, de tels modèles, basés sur l'hypothèse de normalité, sont inadaptés. Dans cette étude, on s'intéresse alors à des régressions de Poisson, dans la famille des modèles linéaires généralisés mixtes (GLMM) (Breslown and Clayton, 1993; Venables and Ripley, 2002) et des modèles additifs généralisés mixtes (GAMM) (Hastie and Tibshirani, 1986; Wood, 2006), dans un contexte fréquentiste et bayésien.

Les modèles mixtes sont des généralisations des modèles à effets fixes, qui permettent de modéliser des sources de variabilité additionnelles à travers des effets aléatoires non observés. Dans le cas qui nous intéresse, il s'agit de modéliser l'autocorrélation spatiale résiduelle. Ces dépendances spatiales résiduelles seront prises en compte, comme on le verra, par la structuration spatiale de la matrice de variance-covariance des effets aléatoires. Les GAMM, contrairement aux GLMM, permettent d'introduire en plus des variables spatiales de coordonnées (latitude et longitude, par exemple) en tant que covariables supplémentaires, à l'aide de fonctions de lissages non-linéaires, appelées fonction splines (Wood, 2006). Ils permettent ainsi de modéliser les dépendances spatiales à la fois dans les effets fixes et les effets aléatoires.

Notons qu'il existe une autre approche permettant de tenir compte de l'autocorrélation spatiale, non pas de manière explicite au coeur de la modélisation, mais en adaptant les tests statistiques usuels qui sont caduques, en se concentrant sur la correction des écarts-types des estimateurs (Dutilleul, 1993; Fortin and Dale, 2005; Dale and Fortin, 2009). Des modèles non-spatiaux classiques sont ajustés sur les données, et les tests de significativité de l'influence des régresseurs sont adaptés de sorte à prendre en compte l'autocorrélation spatiale à ce niveau. L'idée est de recalculer les degrés de liberté, qui mesurent la quantité d'information apportée par les observations, en considérant le fait que des observations spatialement autocoréllées fournissent moins d'information que des observations indépendantes.

Avant de présenter les modèles spatiaux explicites en détails, attardons-nous tout d'abord sur la définition des modèles non-spatiaux et des modèles spatiaux indépendants. Les différents modèles présentés ci-après ont tous été implémentés à l'aide du logiciel R (R Development Core Team, 2011). Une synthèse globale des méthodes est disponible table 3.1.

## 3.3 Les modèles non-spatiaux et les modèles spatiaux indépendants

### 3.3.1 Les modèles non-spatiaux

Dans le contexte de données de comptage, le modèle non-spatiale canonique est le modèle linéaire généralisé de Poisson (GLM de Poisson) avec lien log. C'est un modèle classique, faisant intervenir uniquement des effets fixes ; les dépendances spatiales ne sont ainsi modélisées d'aucune manière.

Considérons les notations suivantes :

$\mathcal{P}(\lambda)$ : la loi de Poisson de paramètre $\lambda$
$(Y_1, \ldots, Y_n)$ : le vecteur des variables aléatoires réponses associées à des données de comptage spatialement structurées
$\mathbb{X} = ((X_i^p))$ : la matrice des covariables de dimension $n$ x $P$
$\boldsymbol{\beta}$ : le vecteur des paramètres de la régression
$\mu_i = \mathbb{E}(Y_i|\mathbb{X}_i)$ : la moyenne conditionnelle de la variable réponse au point $i$

Le GLM de Poisson avec lien log est alors défini par, $\forall i \in \{1, \ldots, n\}$ :

$$\begin{cases} \mathcal{L}(Y_i|X_i^1, \ldots X_i^P) = \mathcal{P}(\mu_i) \\ \log(\mu_i) = (\mathbb{X}\boldsymbol{\beta})_i \\ (Y_i|X_i^1, \ldots, X_i^P) \text{ indépendants } \forall i \in \{1, \ldots, n\} \end{cases}$$

Un GLM avec une distribution de Quasi-Poisson, qui permet de tenir compte d'une variabilité supplémentaire non observée à travers un paramètre de sur-dispersion, est aussi implémenté.

### 3.3.2 Les modèles spatiaux indépendants

Un premier pas pour modéliser l'autocorrélation spatiale résiduelle est d'introduire des effets aléatoires indépendants et identiquement distribués (i.i.d.) dans le GLM de Poisson, créeant ainsi un modèle linéaire généralisé mixte de Poisson(GLMM de Poisson). Cela permet de tenir

TABLE 3.1 – Synthèse des méthodes de régression implémentées

| Abréviation | Description | Inférence | Classification | package R |
|---|---|---|---|---|
| GLMpoiss | Modèle linéaire généralisé (GLM) de Poisson | fréquentiste (Newton-Raphson) | non-spatial | stats |
| GLMquasipoiss | GLM de Quasi-Poisson | fréquentiste (Quasi-Maximum Likelihood - QML) | non-spatial | stats |
| GLMMPQLiid | Modèle linéaire généralisé mixte (GLMM) de Poisson avec effets aléatoires indépendants | fréquentiste (Penalized Quasi-Likelihood - PQL) | spatial indépendant | MASS |
| GAMMiid | Modèle additif généralisé mixte (GAMM) de Poisson avec covariables spatiales et effets aléatoires indépendants | fréquentiste (PQL) | spatial indépendant | mgcv |
| INLAiid | GLMM de Poisson avec effets aléatoires indépendants | bayésienne (Iterative Nested Laplace Approximation - INLA) | spatial indépendant | INLA |
| GLMMPQL | GLMM de Poisson avec effets aléatoires corrélés à partir des distances | fréquentiste (PQL) | spatial structuré | MASS |
| GAMM | GAMM de Poisson avec covariables spatiales et effets aléatoires corrélés à partir des distances | fréquentiste (PQL) | spatial structuré | mgcv |
| INLA | GLMM de Poisson avec effets aléatoires corrélés à partir des voisinages | bayésienne (INLA) | spatial structuré | INLA |
| MCMCLH | GLMM de Poisson avec effets aléatoires corrélés à partir des distances | bayésienne (MCMC avec mises à jour de Langevin-Hastings - MCMCLH) | spatial structuré | geoRglm |

compte de sources de variabilités indépendantes supplémentaires par rapport 7 aux effets fixes, et de modéliser ainsi de manière qualitative la nature spatiale des données. C'est un modèle intermédiaire, qui ne modélise pas de manière explicite les dépendances spatiales, contrairement aux modèles spatiaux structurés avec effets aléatoires corrélés. Notons que le modèle spatial indépendant prend tout son sens lorsqu'il y a plusieurs observations par zone spatiale.

On modélise en pratique les effets aléatoires à l'aide d'un champ gaussien centré et indépen-

dant, en considérant les notations additionnelles suivantes :

$\mathbf{A} = (A_1, \ldots, A_n)$ : le vecteur des effets aléatoires spatiaux i.i.d.
$\sigma^2$ : le paramètre de variance associée aux effets aléatoires homoscédastiques
$\mathcal{N}(\mu, \sigma^2)$ : la loi normale de moyenne $\mu$ et de variance $\sigma^2$
$\mu_i = \mathbb{E}(Y_i | X_i^1, \ldots, X_i^P, \mathbf{A})$ : la moyenne de la variable réponse au point $i$ conditionnellement aux effets aléatoires

Le modèle spatial indépendant est alors défini par, $\forall i \in \{1, \ldots, n\}$ :

$$\begin{cases} \mathcal{L}(Y_i | X_i^1, \ldots, X_i^P, \mathbf{A}) = \mathcal{P}(\mu_i) \\ \log(\mu_i) = (\mathbb{X}\boldsymbol{\beta})_i + A_i \\ \mathcal{L}(\mathbf{A}) = \mathcal{N}(0, \sigma^2 Id) \\ (Y_i | X_i^1, \ldots, X_i^P) \text{ indépendants } \forall i \in \{1, \ldots, n\} \text{ conditionnellement à } \mathbf{A} \end{cases}$$

On a implémenté, en pratique, 3 méthodes spatiales indépendantes dont 2 GLMM et 1 GAMM (voir table 3.1).

## 3.4 Les modèles spatialement structurés

Les modèles spatialement structurés, encore appelés modèles spatiaux corrélés ou modèles spatiaux explicites, permettent de prendre en compte de manière quantitative les dépendances spatiales inter-sites, et donc de modéliser correctement l'autocorrélation spatiale. Les dépendances spatiales sont modélisées par des effets aléatoires corrélés ; concrètement, il s'agit de structurer spatialement la matrice de variance-covariance $\Sigma$ du vecteur des effets aléatoires.

En considérant les notations précédentes, et en notant $C$ la matrice de corrélation du vecteur des effets aléatoires, de telle sorte que $\Sigma = \sigma^2 C$, les modèles spatialement structurés sont définis par, $\forall i \in \{1, \ldots, n\}$ :

$$\begin{cases} \mathcal{L}(Y_i | X_i^1, \ldots, X_i^P, \mathbf{A}) = \mathcal{P}(\mu_i) \\ \log(\mu_i) = (\mathbb{X}\boldsymbol{\beta})_i + A_i \\ \mathcal{L}(\mathbf{A}) = \mathcal{N}(0, \sigma^2 C) \\ (Y_i | X_i^1, \ldots, X_i^P) \text{ indépendants } \forall i \in \{1, \ldots, n\} \text{ conditionnellement à } \mathbf{A} \end{cases}$$

Dans le cas des modèles spatiaux indépendents, on a simplement $C = I_d$. Dans le cas des modèles spatiaux structurés, on a exploré deux manières différentes de structurer spatialement la matrice de corrélation $C$ pour rendre compte des dépendances spatiales. $C$ peut être structuré de manière spatiale soit à partir d'une notion de distance et d'une fonction de corrélation paramétrique de la distance, soit à partir de la spécification de relations de voisinages entre observations. Ces deux approches vont être présentées successivement dans les prochaines sections.

En tout, 3 GLMM et 1 GAMM spatialement structurés ont été implémentés. Le code de l'implémentation de ces quatre méthodes dans le logiciel R (R Development Core Team, 2011) est fourni annexe B.

### 3.4.1 Le modèle basé sur les distances

Les méthodes basées sur les distances modélisent la décroissance des dépendances spatiales entre deux observations à partir de la distance qui les sépare, à l'aide d'une fonction de corrélation stationnaire, considérant que les corrélations dépendent uniquement de la distance entre les points. Dans les méthodes implémentées, on considère la fonction de corrélation exponentielle à un paramètre. Le vecteur des effets aléatoires est alors modélisé par un champ aléatoire gaussien de matrice de corrélation $C$ donnée par :

$$C_{ij} = \exp\left(\frac{-d_{ij}}{\phi}\right)$$

Avec :
$d_{ij}$ : la distance euclidienne entre les points $i$ et $j$
$\phi$ : le paramètre de portée spatiale

Lorsque $\phi = 0$, il n'y a pas d'autocorrélation spatiale résiduelle, et celle-ci augmente à mesure que $\phi$ augmente. Si deux observations sont éloignées d'une distance supérieure à $\phi$, on peut alors considérer qu'elles sont spatialement indépendantes.

Il existe une écriture conditionnelle du modèle basé sur les distances, mais dont la formule explicite n'est pas évidente (voir Cressie (1993)). Si l'on note, $\forall i \in \{1, \ldots, n\}$, $\mathbf{A_{-i}} = (A_1, \ldots, A_{i-1}, A_{i+1}, \ldots, A_n)$ le vecteur des effets aléatoires dont on a enlevé la $i$-ième composante, on peut voir que l'espérance conditionnelle de $A_i$ sachant $\mathbf{A_{-i}}$ s'exprime comme une somme pondérée des composantes de $\mathbf{A_{-i}}$, qui fait intervenir les distances $d_{ij}$.

En tout, 3 méthodes spatiales basées sur les distances ont été implémentées, dont 2 GLMM et 1 GAMM (voir table 3.1).

### 3.4.2 Le modèle basé sur les voisinages

Les méthodes spatiales basées sur les voisinages modélisent les dépendances spatiales par la spécification a priori de relations de voisinages, avec l'idée que les observations voisines partagent les même structures de dépendances spatiales. La spécification des voisinages, introduite dans le chapitre 2, dépend en pratique du contexte spatial dans lequel on se situe, et est basée, dans cette étude, sur une notion de distance limite. On rappelle ici la forme de la matrice de contiguïté spatiale $\mathbf{W} = ((w_{ij}))$ :

$$\mathbf{W} = ((w_{ij})) = \begin{cases} 1 & \text{si i et j sont voisins} \\ 0 & \text{sinon} \end{cases}$$

#### Le modèle ICAR

On s'intéresse ici à un modèle spatial très connu, basé sur les voisinages : le modèle conditionnel autorégressif (CAR), introduit par Besag (1974). Ce type de modèle est fréquemment utilisé dans un contexte gaussien pour modéliser des dépendances spatiales, en mettant en place une structure conditionnelle de dépendances sur la variable réponse. Dans le contexte de données de comptage auxquelles on s'intéresse, on se sert des structures conditionnelles autorégressives pour structurer spatialement les effets aléatoires d'un GLMM de Poisson.

Considérons une suite de variables aléatoires centrées $(A_1, \ldots, A_n)$, représentant les effets aléatoires. Il existe plusieurs types de modèles CAR, dont le plus connu est le modèle conditionnel

autorégressif intrinsèque (ICAR) introduit par Besag and Green (1993). Ce modèle est défini, dans sa version homoscédastique, par la formule conditionnelle suivante, $\forall i \in \{1, \ldots, n\}$ :

$$A_i | \mathbf{A_{-i}} \sim \mathcal{N}\left(\sum_{j \neq i} w_{ij} A_j, \sigma^2\right)$$

Avec $\mathbf{A_{-i}} = (A_1, \ldots, A_{i-1}, A_{i+1}, \ldots, A_n)$.

Ce modèle s'interprète simplement, en observant que l'effet aléatoire associé à une région spatiale donnée prend la valeur de la somme des effets aléatoires associés aux régions avec lesquelles elle est en relation de voisinage. Ce mécanisme conditionnelle permet ainsi de rendre compte des dépendances spatiales.

Il existe une écriture condensée de la loi du vecteur des effets aléatoires (Cressie, 1993), qui permet de voir la manière dont le modèle ICAR structure la matrice de corrélation des effets aléatoires :

$$\mathcal{L}(\mathbf{A}) = \mathcal{N}(0, \sigma^2 (I_n - W)^{-1})$$

**Analogie avec les séries temporelles**

Le modèle CAR peut se comprendre le mieux par analogie avec la modélisation des corrélations temporelles dans la théorie des séries chronologiques. En effet, considérons $(X_t)$ un processus temporel autorégressif stationnaire centré d'ordre $p$. $(X_t)$ est défini par la formule de récurrence suivante :

$$X_t = \sum_{i=1}^{p} \phi_i X_{t-i} + \epsilon_t$$

Avec $\epsilon_t$ un bruit blanc gaussien de variance $\sigma^2$.

Si l'on note $\mathbf{X_{-t}} = (X_1, \ldots, X_{t-1})$, on obtient la formule conditionnelle suivante :

$$X_t | \mathbf{X_{-t}} \sim \mathcal{N}\left(\sum_{i=1}^{p} \phi^i X_{t-i}, \sigma^2\right)$$

Cette formule signifie que les corrélations temporelles sont modélisées par la considération qu'en moyenne l'observation à un temps t prend la valeur d'une somme pondérée des p observations antérieures. Si l'on applique ce principe à la modélisation des dépendances spatiales, en tenant compte de leur nature multidimensionnelle et multidirectionnelle, on retrouve la formule conditionnelle définissant le modèle ICAR. Ainsi, les modèle CAR sont une généralisation des modèles temporels autorégressifs au cas de dépendances spatiales multidimensionnelles.

**Une version *propre* du modèle ICAR**

Le modèle ICAR, dans son implémentation bayésienne théoriquement étudiée et mise en oeuvre par Besag et al. (1991), souffre de problèmes numériques de convergence, car le vecteur des effets aléatoires $\mathbf{A}$ possède une loi de probabilité impropre Pettitt et al. (2002). Par ailleurs, dans le modèle ICAR, il n'y a pas de paramètre permettant d'évaluer l'intensité des dépendances spatiales. C'est pourquoi nous introduisons ici une version *propre* du modèle CAR, telle que présentée par Pettitt et al. (2002). Elle fait intervenir un paramètre spatial $\phi$ qui, en plus

d'assurer le caractère *propre* de la loi, quantifie l'intensité des dépendances spatiales.

Ce modèle est implémenté en pratique dans une version hétéroscédastique (la variance dépend des observations), définie par la formule conditionnelle suivante, $\forall i \in \{1, \ldots, n\}$ :

$$A_i|\mathbf{A_{-i}} \sim \mathcal{N}\left(\frac{\phi}{1+|\phi|n_i}\sum_{j\sim i} Y_j, \frac{1}{1+|\phi|n_i}\right)$$

Avec :
$n_i$ : le nombre de voisins de l'observation $i$
$\phi$ : le paramètre d'intensité spatiale

$\phi$ est égal à 0 lorsque toutes les régions sont spatialement indépendantes et, lorsque $\phi$ tend vers l'infini, on voit, d'après la formule conditionnelle suivante, que l'influence du voisinage augmente avec des valeurs croissantes de $\phi$, l'effet aléatoire prenant des valeurs de plus en plus proches de la valeur moyenne de ses voisins, et ce de manière de plus en plus précise :

$$A_i|\mathbf{A_{-i}} \overset{\phi\to\infty}{\sim} \mathcal{N}\left(\frac{1}{n_i}\sum_{j\sim i} A_j, 0\right)$$

Notons enfin qu'il existe une écriture explicite de ce modèle en terme de matrice de précision (inverse de la matrice de corrélation) :

$$C_{ij}^{-1} = \begin{cases} 1+|\phi|\cdot n_i & \text{si } i = j \\ -\phi & \text{si } i \text{ et } j \text{ sont voisins et } i \neq j \\ 0 & \text{sinon} \end{cases}$$

Un unique modèle CAR *propre* ainsi défini est implémenté en pratique (voir table 3.1).

## 3.5 Estimation des modèles spatiaux

Concernant les modèles non-spatiaux, l'estimation est réalisée via une inférence fréquentiste classique, à savoir l'algorithme de Newton-Raphson pour le GLM de Poisson et la technique du Quasi-Maximum Likelihood (QML) pour le modèle de Quasi-Poisson.

Les modèles spatiaux indépendants et spatialement structurés, pour lesquels on veut estimer les effets aléatoires et les paramètres spatiaux en plus des effets fixes, sont trop compliqués pour être estimés par maximum de vraisemblance ; les algorithmes d'estimation simultanée de l'ensemble des paramètres ne convergent pas en pratique. On a alors recours à des techniques d'approximation, soit de type fréquentiste via la technique du Penalized Quasi-Likelihood (PQL) introduite par Breslown and Clayton (1993), soit de type bayésienne par la mise en oeuvre de méthodes de Monte Carlo par Chaînes de Markov (MCMC).

Dans ce rapport, on s'intéresse de manière spécifique à la présentation de l'inférence bayésienne MCMC, qui sera introduite dans un exposé synthétique, et appliqué, en guise d'exemple, à l'estimation du modèle spatialement structuré basé sur les distances, en se basant sur les recommandations de Diggle et al. (1998), qui a étudié le premier l'implémentation bayésienne de modèles spatiaux basés sur les distances, à l'aide de méthodes MCMC.

En pratique, les modèles basés sur les distances et les modèles basés sur les voisinages ont été implémentés à l'aide d'algorithmes MCMC adaptatifs, que l'on présentera dans un second temps.

### 3.5.1 Inférence bayésienne et méthodes de Monte Carlo par Chaînes de Markov

#### Généralités

L'objectif de cette section et de fournir au lecteur les notions de base de l'inférence bayésienne via les méthodes MCMC. On reporte le lecteur au livre de Gelman et al. (2004) pour un exposé exhaustif de ces concepts.

Supposons, dans un cadre de régression, que l'on cherche à estimer le vecteur de paramètres $\boldsymbol{\theta}$. L'inférence bayésienne consiste à considérer $\boldsymbol{\theta}$ comme un vecteur de variables aléatoires auxquelles on associe des lois de probabilités *a priori*, permettant de modéliser un degré d'incertitude sur les paramètres. L'idée est ensuite d'estimer la distribution *a posteriori* de $\boldsymbol{\theta}$ conditionnellement au vecteur des observations $\mathbf{y}$ ; cette distribution conditionnelle résume l'information disponible sur $\boldsymbol{\theta}$ une fois $\mathbf{y}$ observé. L'estimation de la distribution a posteriori $p(\mathbf{y}|\boldsymbol{\theta})$ est réalisée grâce à la règle de Bayes, que l'on présente sous sa forme proportionnelle dans le cas continu, en notant $\propto$ le symbole *proportionnel à* :

$$p(\boldsymbol{\theta}|\mathbf{y}) \propto p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})$$

Avec :
$p(\mathbf{y}|\boldsymbol{\theta})$ : la vraisemblance du modèle
$p(\boldsymbol{\theta})$ : la loi a priori sur le vecteur de paramètres $\boldsymbol{\theta}$

Il s'agit donc d'estimer la loi a posteriori $p(\boldsymbol{\theta}|\mathbf{y})$, puis d'en prendre l'espérance $\mathbb{E}(\boldsymbol{\theta}|\mathbf{y})$, qui constitue l'estimateur bayésien du paramètre $\boldsymbol{\theta}$.

Cependant, il est rare de pouvoir obtenir une expression explicite de la loi a posteriori, notamment lorsque la dimension de $\boldsymbol{\theta}$ est élevée, ce qui est le cas pour nos modèles spatiaux. Des outils de simulations, tels que les méthodes MCMC, ont donc été développés pour obtenir une estimation numérique de la loi a posteriori, et donc de son espérance. Le principe de base des méthodes MCMC, dans le contexte de l'inférence bayésienne, est de générer une chaîne de Markov $(\boldsymbol{\theta}^t; t = 1, \ldots, T)$ de loi stationnaire la distribution a posteriori $p(\boldsymbol{\theta}|\mathbf{y})$, puis d'approcher $\mathbb{E}(\boldsymbol{\theta}|\mathbf{y})$ par la loi forte des grands nombres, de la manière suivante :

$$\mathbb{E}(\boldsymbol{\theta}|\mathbf{y}) \approx \frac{1}{T}\sum_{t=1}^{T} \boldsymbol{\theta}^t$$

Il existe plusieurs algorithmes permettant de générer une telle chaîne ; les deux plus connus sont l'algorithme de Gibbs et et l'algorithme de Metropolis-Hastings. Dans ce rapport, seul l'algorithme de Gibbs est présenté. L'implémentation standard de l'algorithme de Gibbs consiste à fournir des réalisations des distributions conditionnelles unidimensionnelles des paramètres les uns par rapport aux autres.

Notons à nouveau $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_P)$ le vecteur des paramètres à estimer, où $P$ est la dimension de $\boldsymbol{\theta}$. Considérons une valeur initiale arbitraire, et supposons que l'on dipose à l'étape $t$ du vecteur

des réalisations $\boldsymbol{\theta}^t = (\theta_1^t, \ldots, \theta_P^t)$. On obtient alors une réalisation $\boldsymbol{\theta}^{t+1} = (\theta_1^{t+1}, \ldots, \theta_P^{t+1})$ du vecteur aléatoire $\boldsymbol{\Theta}^{t+1} = (\Theta_1^{t+1}, \ldots, \Theta_P^{t+1})$ en simulant selon les lois conditionnelles suivantes :

$$\begin{aligned}
&\Theta_1^{t+1} \sim \mathcal{L}\left(\Theta_1 | \Theta_2 = \theta_2^t, \ldots, \Theta_P = \theta_P^t\right) \\
&\Theta_p^{t+1} \sim \mathcal{L}\left(\Theta_p | \Theta_1 = \theta_1^{t+1}, \ldots, \Theta_{p-1} = \theta_{p-1}^{t+1}, \Theta_{p+1} = \theta_{p+1}^t, \Theta_P = \theta_P^t\right) \\
&\Theta_P^{t+1} \sim \mathcal{L}\left(\Theta_P | \Theta_1 = \theta_1^{t+1}, \ldots, \Theta_{P-1} = \theta_{P-1}^{t+1}\right)
\end{aligned}$$

Avant d'appliquer ces procédures bayésiennes au cas du modèle basé sur les distances, on aimerait insister sur quelques notions importantes concernant la convergence pratique des chaînes de Markov. Il faut être très précautionneux en inférence bayésienne, et toujours vérifier en pratique la bonne convergence des chaînes de Markov ; il s'agit d'obtenir une chaîne sans tendance ni trop de variabilité. En pratique, dans cette étude, nous nous limiterons à un examen visuel des chaînes de Markov pour nous assurer de leur bonne convergence, qui peut être obtenue en contrôlant les paramètres suivants :

- La période de chauffe (burn-in) $B$ : c'est le nombre d'itérations nécessaires pour que la chaîne atteigne sa distribution statinnaire.
- Le nombre d'itérations $I$ : il doit être suffisamment grand pour assurer une bonne convergence de la chaîne de Markov.
- La période d'échantillonage (thin) $t$ : les variables issues de la simulation des chaînes de Markov sont généralement corrélées, on le considère donc pas toutes les réalisations, mais seulement une toutes les $t$ réalisations. La taille de la chaîne de Markov obtenue est alors égale à $\frac{I-B}{t}$.

**Application au modèle basé sur les distances**

Appliquons à présent ces procédures d'inférences bayésiennes au cas du GLMM de Poisson basé sur les distances. Considérons, pour cela, les notations suivantes :

$\boldsymbol{\theta} = (\boldsymbol{\beta}, \mathbf{A}, \sigma^2, \phi)$ : le vecteur de paramètres que l'on cherche à estimer
$\Sigma = \sigma^2 C$ : la matrice de variance-covariance des effets aléatoires, qui spatialement structurée à partir des distances et dépend des paramètres $\sigma^2$ et $\phi$
$\mathcal{U}[a, b]$ : la loi uniforme sur le segment $[a, b]$

On considère ensuite les lois a priori suivantes, indépendantes deux à deux, sur les paramètres du modèle :

$$\begin{cases}
\boldsymbol{\beta} \sim \mathcal{N}(0, 1) \\
(\mathbf{A} | \sigma^2, \phi) \sim \mathcal{N}(0, \Sigma) \\
\sigma^2 \sim \mathcal{U}[a, b] \\
\phi \sim \mathcal{U}[0, c]
\end{cases}$$

Où les valeurs $a$, $b$ et $c$ des bornes supérieures et inférieures des lois uniformes dépendent du contexte spatial dans lequel on se trouve.

Ecrivons à présent les dénsités de probabilités associées à ces lois a priori :

$$\begin{cases} p(\boldsymbol{\beta}) \propto \exp(-\frac{1}{2}\boldsymbol{\beta}'\boldsymbol{\beta}) \\ p(\mathbf{A}|\sigma^2,\phi) \propto |\Sigma|^{-\frac{1}{2}} \exp(-\frac{1}{2}\mathbf{A}'\Sigma^{-1}\mathbf{A}) \\ p(\sigma^2) \propto 1 \\ p(\phi) \propto 1 \end{cases}$$

En appliquant la règle de Bayes de manière successive, ainsi que l'indépendance des lois a priori, on peut calculer la loi a posteriori du vecteur $\boldsymbol{\theta} = (\boldsymbol{\beta}, \mathbf{A}, \sigma^2, \phi)$ conditionnellement aux observations $\mathbf{Y}$ :

$$\begin{aligned} p(\boldsymbol{\beta}, \mathbf{A}, \sigma^2, \phi|\mathbf{Y}) &\propto p(\mathbf{Y}|\boldsymbol{\beta}, \mathbf{A}, \sigma^2, \phi)p(\boldsymbol{\beta}, \mathbf{A}, \sigma^2, \phi) \\ &\propto p(\mathbf{Y}|\boldsymbol{\beta}, \mathbf{A})p(\boldsymbol{\beta})p(\mathbf{A}, \sigma^2, \phi) \\ &\propto p(\mathbf{Y}|\boldsymbol{\beta}, \mathbf{A})p(\boldsymbol{\beta})p(\mathbf{A}|\sigma^2, \phi)p(\sigma^2, \phi) \\ &\propto p(\mathbf{Y}|\boldsymbol{\beta}, \mathbf{A})p(\mathbf{A}|\sigma^2, \phi)p(\boldsymbol{\beta})p(\sigma^2)p(\phi) \\ &\propto p(\mathbf{Y}|\boldsymbol{\beta}, \mathbf{A})p(\mathbf{A}|\sigma^2, \phi)p(\boldsymbol{\beta}) \end{aligned}$$

Or, le premier terme $p(\mathbf{Y}|\beta, \mathbf{A})$ est la vraisemblance du modèle de Poisson, que l'on obtient à l'aide de l'hypothèse d'indépendance conditionnelle :

$$p(\mathbf{Y}|\boldsymbol{\beta}, \mathbf{A}) = \prod_{i=1}^{n} exp(-\mu_i)\frac{\mu_i^{Y_i}}{Y_i!}$$

Avec $\mu_i = \exp\left((\mathbb{X}\boldsymbol{\beta})_i + A_i\right)$.

Il s'agit de multiplier cette vraisemblance par deux lois normales $p(\mathbf{A}|\sigma^2, \phi)$ et $p(\boldsymbol{\beta})$. Cela aboutit à une expression non-triviale, à partir laquelle on ne peut identifier de loi explicite.

Appliquons donc l'algorithme de Gibbs, qui nécessite la connaissance des densités conditionnelles suivantes :

$$\begin{cases} p(\beta_p|\boldsymbol{\beta_{-p}}, \mathbf{A}, \mathbf{Y}) \\ p(A_i|\mathbf{A_{-i}}, \boldsymbol{\beta}, \sigma^2, \phi, \mathbf{Y}) \\ p(\sigma^2|\mathbf{A}, \phi, \mathbf{Y}) \\ p(\phi|\mathbf{A}, \sigma^2, \mathbf{Y}) \end{cases}$$

Avec :
$\boldsymbol{\beta_{-p}} = (\beta_1, \ldots, \beta_{p-1}, \beta_{p+1}, \ldots, \beta_P)$
$\mathbf{A_{-i}} = (A_1, \ldots, A_{i-1}, A_{i+1}, \ldots, A_n)$

En appliquant de manière successive la règle de Bayes ainsi que les propriétés d'indépendance entre les paramètres du modèle, on obtient aisément l'expression des densités des deux dernières lois conditionnelles :

$$\begin{aligned} p(\sigma^2|\mathbf{A}, \phi, \mathbf{Y}) &\propto p(\mathbf{A}|\sigma^2, \phi)p(\sigma^2)p(\phi) \\ &\propto exp\left(-\frac{1}{2}\mathbf{A}'\Sigma^{-1}\mathbf{A}\right) \end{aligned}$$

$$\begin{aligned} p(\phi|\mathbf{A},\sigma^2,\mathbf{Y}) &\propto p(\mathbf{A}|\sigma^2,\phi)p(\sigma^2)p(\phi) \\ &\propto exp\left(-\frac{1}{2}\mathbf{A}'\Sigma^{-1}\mathbf{A}\right) \end{aligned}$$

On identifie deux lois normales, à partir desquelles on peut alors aisément simuler.

Par contre, pour les deux premières lois conditionnelles, on obtient :

$$\begin{cases} p(\beta_p|\boldsymbol{\beta_{-p}},\mathbf{A},\mathbf{Y}) \propto p(Y_i|\boldsymbol{\beta},\mathbf{A})p(\beta_p|\boldsymbol{\beta_{-p}}) \\ p(A_i|\boldsymbol{A_{-i}},\boldsymbol{\beta},\sigma^2,\phi,\mathbf{Y}) \propto p(Y_i|A_i,\boldsymbol{\beta})p(A_i|\mathbf{A_{-i}},\sigma^2,\phi) \end{cases}$$

On retrouve ainsi encore une fois, dans les deux cas, le produit d'une vraisemblance de Poisson avec une loi normale (Gemperli and Vounatsou, 2003). L'expression ainsi obtenue ne permet pas donc pas d'exhiber une loi de probabilité connue, dont il n'est donc pas possible de simuler des réalisations. Il est alors nécessaire, arrivé là, de mettre en oeuvre un algorithme de Metropolis-Hastings, qu'on ne détaillera dans le présent rapport (voir Gelman et al. (2004)).

### 3.5.2 Algorithmes MCMC adaptatifs et implémentation pratique

La mise en oeuvre de procédures numériques MCMC engendre des temps de calcul très longs, notamment dans les problèmes complexes à grandes dimensions, de par la nécessité d'inverser de manière répétée la matrice de variance-covariance des effets aléatoires. En effet, une inférence bayésienne MCMC peut nécessiter plusieurs heures de calcul. Dans l'optique d'une mise en oeuvre d'une approche par simulation comprenant des milliers de jeux de données, le temps de calcul devient alors déraisonnable. C'est pour cela que l'on a recours à des versions modifiées ou approximées des méthodes MCMC, permettant d'accélérer la convergence. Dans cette étude, le modèle basé sur les distances est estimé à l'aide d'une modification de l'algorithme MCMC : l'algorithme MCMC avec mises à jours de Langevin-Hastings (Markov Chain Monte Carlo with Langevin-Hastings updates - MCMCLH). Quant au modèle basé sur les voisinages, il est estimé à l'aide d'une approximation de l'algorithme MCMC appelée Iterative Nested Laplace Approximation (INLA). Quelques secondes à quelques minutes suffisent alors pour mener à bien l'inférence.

#### Markov Chain Monte Carlo with Langevin-Hastings updates

L'algorithme MCMC avec mises à jours de Langevin-Hastings (MCMCLH), tel que décrit par Christensen and Waagepetersen (2002), est un algorithme hybride, basé sur celui de Metropolis-Hastings, mettant en oeuvre une marche aléatoire gaussienne où ,à chaque itération, les effets aléatoires sont simultanément mises à jour avec les effets fixes. En plus d'être beaucoup moins coûteux en temps qu'une procédure MCMC classique, il réduit les erreurs de sampling. Pour plus de détails concernant le socle théorique de cette méthode, voir Christensen et al. (2001).

L'implémentation pratique de l'algorithme dans le logiciel R a été réalisée à l'aide du package geoRglm (Christensen and Ribeiro Jr, 2002). Concernant le calibrage de l'inférence bayésienne, permettant de s'assurer de la bonne convergence de la chaîne de Markov, on spécifie un nombre d'itérations égal à 70000, une période de burn-in égale à 30000, pour obtenir une chaîne de Markov de taille 7000 en ne considérant qu'une réalisation sur 10 ($thin = 10$). L'implémentation pratique de l'algorithme MCMCLH necessite par ailleurs de spécifier des paramètres supplémentaires liés aux probabilités d'acceptation des nouvelles valeurs dans l'algorithme. Voir Christensen and Waagepetersen (2002) pour plus de détails.

**Iterative Nested Laplace Approximation**

Contrairement à l'algorithme MCMCLH, l'algorithme INLA (Iterative Nested Laplace Approximation) est une approximation de l'algorithme MCMC. Elle consiste à approximer des lois marginales unidimensionnelles a posteriori des paramètres du modèle. Elle repose sur la notion de champ de Markov gaussien étudié par Rue and Held (2005).

L'implémentation pratique de cet algorithme se fait à l'aide du package INLA (Rue et al., 2009). Concernant les lois a priori des paramètres, des lois inverse-gamma sont proposés par défaut pour les paramètres $\sigma^2$ et $\phi$, et des lois a priori plates sont utilisées pour les effets fixes. On reporte le lecteur à Rue et al. (2009) ainsi qu'au site internet `www.r-inla.org` pour plus de détails à ce sujet. Enifn, les paramètres de réglage liés à la convergence de la chaîne de Markov ont les mêmes valeurs que pour la méthode MCMCLH.

# Chapitre 4

# L'approche par simulation

Ce chapitre est dédié à la présentation de l'approche par simulation. Après avoir explicité la mise en oeuvre des différents scénarios de simulation, on se penchera sur la présentation de la comparaison de modèles.

## 4.1 Introduction

Afin de comparer les performances relatives des méthodes implémentées et d'exhiber leurs forces et leurs faiblesses dans différents contextes spatiaux, une approche par simulation est envisagée, mettant en oeuvre des données de comptage spatialement autocorrélées. Une approche par simulation offre un cadre adéquat pour comparer des modèles, qui bien plus fiable que de comparer des modèles à partir de jeux de données isolés (voir chapitre 6 pour une discussion sur le sujet).

Cette étude s'inscrit à la suite des travaux de Beale et al. (2010), dont les objectifs sont globalement les mêmes que ceux de cette étude : mettre en évidence la faillite des modèles non-spatiaux lorsqu'ils sont ajustés sur des données spatiales, et comparer différentes méthodes spatiales entre elles. Malheureusement, l'étude de Beale et al. (2010) souffre de plusieurs limitations à différents niveaux, concernant la qualité des méthodes spatiales testées et le réalisme des scénarios spatiaux simulés. Dans cette étude, on essaye d'aller au-delà de ces limitations en implémentant des modèles spatiaux a priori pertinents, notamment via des algorithmes bayésiens évolués, avec différentes manières de modéliser l'information spatiale (voir chapitre 3). Par ailleurs, un grand soin a été apporté à l'élaboration de simulations réalistes, afin de permettre une extrapolation justifiée de nos résultats à l'étude de jeux de données réelles. Cela a été réalisé en se concentrant sur les points suivants :

1) Beale et al. (2010) s'intéressent uniquement à des données quantitatives continues gaussiennes, ajustées à l'aide de modèles normaux. Or, il ne s'agit pas d'un type de données très répandus en écologie, contrairement aux données de comptage qui sont très fréquents notamment dans le cadre de l'étude de la répartition des espèces, à travers des données d'abondance et de richesse. C'est pour cela que dans cette étude, on tente de généraliser les résultats de Beale et al. (2010) aux données de comptage, que l'on ajuste à l'aide de modèles de Poisson (voir chapitre 3).

2) Les scénarios spatiaux simulés par Beale et al. (2010) ne reflètent pas la complexité spatiale des jeux de données réelles, car toutes les données ont été simulées uniquement sur grille. Dans cette étude, après avoir comme Beale et al. (2010) mis en place des scénarios spatiaux sur grille et généré ainsi des observations régulièrement espacées, on génère des observations

irrégulièrement espacées, en se basant sur les coordonnées géographiques réelles des placettes du projet GNB, afin de voir si les modèles spatiaux performants sur grille s'adaptent de manière satisfaisante dans des contextes spatiaux plus flexibles.

3) On ne se contente pas, dans cette étude, de comparer la performance des modèles sur la base de scénarios spatiaux simples. On incorpore en effet, tout d'abord, dans les scénarios sur grilles et les scénarios GNB, des caractéristiques de non-stationnarité spatiale, telles qu'introduites dans le chapitre 2. De plus, on met en oeuvre des simulations additionnelles dont le but est de tester la robustesse des méthodes implémentées à la violation de certaines de leurs conditions d'application, en envisageant notamment les problématiques de dispersion non-spatiale, à savoir la sur-dispersion et la sous-dispersion, que présentent souvent les données de comptage. On teste enfin si le modèle basé sur les distances dans son implémentation bayésienne, à savoir MCMCLH (voir tableau 3.1), est robuste à une mauvaise spécification de sa fonction de corrélation spatiale.

Présentons donc plus en détails les scénarios de base, tout d'abord sur grille puis sur points irrégulièrement espacés, avant de décrire les simulations additionnelles.

## 4.2 Mise en oeuvre des simulations de base

On simule des jeux de données de comptage spatialement autocorrélées, qui représentent des distributions virtuelles d'espèces ainsi que des covariables environnementales spatialement structurées. On met en oeuvre en tout 16 scénarios de base : 10 scénarios sur grille, générant des observations régulièrement espacées, et 6 scénarios sur placettes GNB, générant des observations irrégulièrement espacées. Une synthèse globale des scénarios de base peut être trouvée 4.1. Notons bien que l'on simule, dans tout les cas, une observation par unité spatiale ; l'unité spatiale des simulations sur grille est la cellule, et celle des simulations sur points irrégulièrement espacés est la placette GNB.

De manière générale, en pratique, on génère une autocorrélation spatiale résiduelle dans un terme d'erreur, modélisé par un champ gaussien centré et spatialement structuré, que l'on ajoute au champ des effets fixes. Toutes les covariables simulées, qui sont quantitatives continues et distribuées normalement, sont indépendantes et toutes spatialement structurées.

Plus précisément, notons $n$ le nombre d'observations et $P$ le nombre de covariables. On veut simuler des données de comptage $\mathbf{Y} = (Y_1, \ldots, Y_n)$ de sorte à introduire de l'autocorrélation spatiale résiduelle dans la modélisation. Notons $\mathbb{X}$ la matrice des covariables simulées, et $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_P)$ le vecteur des coefficients de régression fixés a priori. On simule un terme d'erreur $\boldsymbol{\epsilon} = (\epsilon_1, ...\epsilon_n)$ à l'aide d'une loi normale centrée multidimensionnelle, de matrice de variance-covariance $\Sigma$. Afin d'obtenir un terme d'erreur spatialement autocorrélé, on structure spatialement sa matrice de corrélation $C$, définie par l'expression $\Sigma = \sigma^2 C$, de telle sorte à obtenir des corrélations spatiales de plus en plus importantes à mesure que la distance entre les observations diminue. On simule cet état de dépendance à l'aide de la fonction de corrélation stationnaire exponentielle à un paramètre, en écrivant :

$$C_{ij} = \exp\left(\frac{-d_{ij}}{\phi}\right)$$

Avec :
$d_{ij}$ : la distance séparant les observations $i$ et $j$

$\phi$ : le paramètre de portée spatial

On construit ensuite le prédicteur linéaire $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_n)$ en additionnant le champ des effets fixes à celui des erreurs, de la manière suivante :

$$\boldsymbol{\eta} = \mathbb{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

On simule enfin selon une loi de Poisson avec lien log pour obtenir le vecteur réponse $\mathbf{Y} = (Y_1, \ldots, Y_n)$ :

$$\forall i \in \{1, \ldots, n\} : Y_i \sim \mathcal{P}(\exp(\eta_i))$$

Pour l'ensemble des scénarios, la valeur de $\sigma$ est fixée à 0.7. La valeur $\phi$ de la portée spatiale résiduelle , quant à elle, varie selon les scénarios.

Présentons à présents les spécificités de la mise en oeuvre des scénarios sur grille et des scénarios GNB, dont on peut trouver un résumé synthéthique table 4.1.

### 4.2.1 Scénarios sur grille

On génère 10 scénarios, en simulant pour chacun d'eux 1000 jeux de données de 400 observations sur une grille 20x20.

Les covariables introduits présentent une autocorrélation spatiale qui varie à la fois en termes de portée (portée faible de valeur 0.7 et portée élevée de valeur 3) et d'intensité relative (intensité faible et intensité élevée). Des intensités faibles ont été générées en ajoutant au terme d'erreur spatialement structuré $\boldsymbol{\epsilon}$ un bruit blanc gaussien de variance égale à 1.

Le premier scénario GRID.1 est un scénario de référence qui n'incorpore pas de variabilité spatiale supplémentaire (pas de terme $\epsilon$) et donc pas d'autocorrélation spatiale résiduelle. Les scénarios GRID.2-7 sont des scénarios incorporant une autocorrélation spatiale stationnaire, qui varie selon les scénarios en termes de portée (portée faible de valeur 0.7, portée intermédiaire de valeur 3 et portée élevée de valeur 5) et d'intensité relative (intensité faible et intensité élevée, comme pour les covariables).

Les scénarios GRID.8-10 incorporent différentes formes de non-stationnarité spatiale, qui ont été inspirées des travaux de Beale et al. (2010). Plus précisément, on a exploré deux types de non-stationnarité spatiale, tels que décrits et illustrés graphiquement figure 2.3 :

1) La non-stationnarité de la portée de l'autocorrélation spatiale. Concrètement, on a simulé deux champs aléatoires gaussiens spatiaux respectivement de faible portée ($\phi = 0.7$) et de forte portée ($\phi = 3$). Afin d'obtenir un champ aléatoire de portée forte sur la partie gauche de la grille et de portée faible sur la partie droite, on multiplie chacun des champs aléatoires par des matrices gradientes de valeurs comprises entre 0 et 1, respectivement de gauche à droite et de droite à gauche, et on somme les deux champs obtenus.

2) Les tendances spatiales linéaires. On obtient un champ aléatoire gaussien de valeurs croissantes de gauche à droite de la grille, en ajoutant à un champ aléatoire gaussien spatialement structuré $\mathcal{G}$ un champ de valeurs, de gauche à droite de la grille, prenant des valeurs croissantes entre $\min \mathcal{G}$ et $\max \mathcal{G}$.

Ces deux types de non-stationnarité spatiale sont appliqués soit dans la structure spatiale des covariables (scénario GRID.8), soit dans la structure spatiale du terme d'erreur (scénario GRID.9), soit dans les deux structures simultanément (scénario GRID.10). Voir table 4.1 pour plus de précisions.

### 4.2.2 Scénarios sur points irrégulièrement espacés

On génère 6 scénarios, appelés *scénarios GNB*, en simulant pour chacun d'eux 1000 jeux de données de 197 observations, à partir des coordonnées géographiques réelles des placettes du projet GNB, présenté chapitre 1.

Les 15 massifs forestiers du projet GNB sont suffisamment éloignés les uns des autres pour que l'on puisse considérer que deux placettes appartenant à deux massifs différents sont spatialement indépendantes. Cependant, au sein des massifs, il existe de potentielles dépendances spatiales entre les placettes ; il s'agit d'une autocorrélation spatiale intra-massifs inter-placettes.

On a simulé trois covariables quantitatives continues distribuées normalement, indépendantes et spatialement structurées, avec différents paramètres de portée spatiale (0.7km et 3km). Une variable binaire, la variable *Gestion*, indiquant quelles placettes sont en zones exploitées et quelles placettes sont en zones non-exploitées, est aussi introduite dans les simulations.

Alors que le scénario GNB.1 est un scénario non-spatial de référence, ne présentant aucune autocorrélation spatiale résiduelle, les scénarios GNB.2-4 sont des scénarios spatiaux incorportant un terme d'erreur spatial avec une portée qui varie selon les scénarios (0.7km pour GRID.2, 1.5km pour GRID.3 et 3km pour GRID.4).

Quant aux scénarios spatiaux GNB.5-6, ils incorporent des caractéristiques de non-stationnarité spatiale, qui sont spécifiques à la configuration spatiale des placettes GNB, donc différentes des composantes stationnaires implémentées sur grille. Plus précisément, le scénario GNB.5 présente une caractéristique de non-stationnarité spatiale inter-massifs : différentes portées spatiales ont été choisies pour les différents massifs, par tirage au sort dans une loi uniforme de paramètres 0.2 et 3. Cependant, au sein de chaque massif, la portée de l'autocorrélation spatiale est bien fixe. Cela n'est pas le cas dans le le scénario GNB.6, qui présente une caractéristique de non-stationnarité intra-massif. En effet, on a superposé au sein des massifs trois champs gaussiens spatialement structurés de portées spatiales respectives 0.7km, 1.5km et 3km, modélisant différents phénomènes écologiques à l'origine de l'autocorrélation spatiale et agissant à différentes échelles. Le terme d'erreur est alors crée concrètement à l'aide d'un mélange de trois lois gaussiennes multivariées.

### 4.2.3 Spécification pratique des voisinages

La notion de voisinage, introduite au chapitre 2, intervient à la fois dans la modélisation spatiale (voir chapitre 3) mais aussi dans la mise en oeuvre de diagnostics d'autocorrélation spatiale tels que le $I$ de Moran, présenté dans le chapitre 2.

Il s'agit d'un problème délicat qu'il faut considérer avec soin car, comme on le verra dans les résultats de l'étude chapitre 5, une mauvaise spécification des voisinages, soit par la création de relations de voisinages artificielles, soit par l'omission de relations de voisinages réelles, peut

avoir des conséquences néfastes sur la qualité de l'estimation des modèles basés sur les voisinages.

La spécification des voisinages sur grille n'est pas problématique ; on suit simplement la règle classique des voisins non-diagonaux consistant à considérer deux cellules voisines si elles partagent un côté en commun.

Concernant la spécification des voisinages sur points irrégulièrement espacés, elle est basée, comme on l'a déjà rappelé, sur une notion de distance limite, qu'il s'agit de choisir a priori ; deux placettes sont considérées voisines si la distance qui les sépare est inférieure à la distance limite choisie. Dans cette étude, on choisit deux distances limites différentes, respectivement 1.5km et 3km, qui génèrent donc deux manières différentes d'organiser les placettes en voisinages, et donc deux modèles CAR (voir chapitre 3) bien distincts, appelés respectivement INLA1 et INLA2. Pour avoir une idée du caractère approprié de ces choix de distance, on peut considérer les repères suivants : en moyenne, dans l'ensemble des massifs, les placettes GNB sont séparées d'une distance minimale d'environ 200m, d'une distance maximale de 5km, avec une distance moyenne globale d'environ 3km.

## 4.3 Simulations additionnelles

Des simulations additionnelles sont mises en oeuvre afin de tester la robustesse des méthodes implémentées à la violation de certaines de leurs conditions d'application, en envisageant tout d'abord le cas d'une inadéquation au modèle de Poisson (inégalité de l'espérance et de la variance) à l'origine des phénomènes de sur-dispersion et de sous-dispersion des données de comptage. On cherche ensuite à tester la robustesse de la méthode basée sur les distances la plus performante, qui est comme on le verra la méthode $MCMCLH$, à une mauvaise spécification de sa fonction de corrélation spatiale.

### 4.3.1 Simulation de données de comptage sur-dispersées et sous-dispersées

Les phénomènes de sur-dispersion (variance plus grande que l'espérance) et en moindre mesure de sous-dispersion (variance plus petite que l'espérance) de données de comptage sont très fréquents en écologie. Il s'agit de propriétés de dispersion des données, qui ne sont pas liées à leur caractère spatial, mais qui s'y ajoutent fréquemment.

La sur-dispersion correspond, dans ce cadre, à une variabilité supplémentaire non-spatiale non-observée, qui peut trouver sa cause dans l'omission dans le modèle d'une variable influente non-spatialement structurée ou dans la présence de corrélations positives non-spatiales. Quant à la sous-dispersion, elle est principalement liée à la présence de corrélations négatives non-spatiales entre observations.

Si l'on ne tient pas compte de ces phénomènes de dispersion dans la modélisation, bien que les estimateurs des effets fixes ne soient pas biaisés, ceux des variances associées sont sous-estimés dans le cas de la sur-dispersion et sur-estimés dans le cas de la sous-dispersion. On tient typiquement compte de ces phénomènes de dispersion soit à l'aide de distributions de probabilités appropriées, telles que les distributions de Quasi-Poisson, les distribution binomiales négatives ou les distributions de Poisson généralisées, soit à l'aide de modèles mixtes incorporant des effets aléatoires i.i.d. au niveau observation. La distribution binomiale négative ainsi que la distribution de Poisson généralisée sont présentées annexe D.

Dans le cadre de la modélisation spatiale, la question se pose si ce type de dispersion, non lié à l'information spatiale, doit être modélisée de manière spécifique, ou si les effets aléatoires spatiaux introduits jusqu'alors sont capables d'absorber de manière satisfaisante cette variabilité additionnelle.

Pour répondre à cette question, on simule des données de comptage spatialement autocorrélés incorporant différents degrés de sur-dispersion et de sous-dispersion, sur la base du scénario spatial GNB.4, avec la même configuration de simulation décrite précédemment. Tandis que les données sur-dispersées sont générées en tirant en sort dans une distribution binomiale négative, en fixant le paramètres $r$ (voir annexe D) aux valeurs 10, 5 et 2, de la sous-dispersion est générée à l'aide d'une distribution de Poisson généralisée, en fixant le paramètre de sur-dispersion $\phi$ (voir annexe D) aux valeurs 0.8, 0.6 et 0.4.

### 4.3.2 Simulation d'une mauvaise spécification de la fonction de corrélation spatiale du modèle MCMCLH

Dans le cadre d'une étude réelle, dans le cas des modèles basés sur les distances, on ne peut savoir a priori quelle fonction de corrélation spatiale est la plus adaptée pour modéliser les dépendances spatiales. Il existe en effet plusieurs fonctions de corrélation disponibles et implémentables avec le logiciel R, telles que la fonction de corrélation exponentielle, sphérique, gaussienne, linéaire etc.

Afin de tester la robustesse de la méthode MCMCLH, qui est comme on le verra la méthode basée sur les distances la plus performante, à une mauvaise spécification de la fonction de corrélation spatiale, on met en oeuvre des scénarios additionels de simulations, basés sur les scénarios spatiaux GNB.2-5.

On simule, pour chaque scénario, 1000 jeux de données spatialement structurées à l'aide de la fonction de corrélation exponentielle et, sur chaque jeu de données, on ajuste deux modèles MCMCLH différents : le modèle MCMCLHexp implémenté avec la "bonne" fonction de corrélation, à savoir la fonction exponentielle, et le modèle MCMCLHspher implémenté avec la fonction de corrélation sphérique. L'objectif est de voir si le modèle MCMCLHspher est aussi performant en termes d'estimation des effets fixes que le modèle MCMCLHexp. Les matrices de corrélation associées aux modèles MCMCLHexp et MCMCLHspher sont données par :

$$\begin{cases} C_{ij}^{exp} = \exp\left(-\frac{d_{ij}}{\phi}\right) \\ C_{ij}^{spher} = 1 - 1.5\left(\frac{d_{ij}}{\phi}\right)^2 + 0.5\left(\frac{d_{ij}}{\phi}\right)^3 \end{cases}$$

Où $d_{ij}$ représente la distance qui sépare les observations $i$ et $j$.

## 4.4 Comparaison de modèles

Tout l'intérêt d'une approche par simulation est de pouvoir comparer les performances des modèles implémentés. En effet, le coefficient de régression $\boldsymbol{\beta}$ est fixé et connu a priori, ce qui n'est jamais le cas dans les études de données réelles. Cela nous permet de comparer les performances intrinsèques des modèles en termes d'inférence des effets fixes, à l'aide d'indicateurs classiques tels que l'erreur de type I et le Root Moot Square Error (RMSE).

On cherche aussi, dans un deuxième temps, à comparer les performances des méthodes de manière relative, en se concentrant sur les différences, appelées *shifts*, entre les estimations fournis par les différentes modèles.

Il est important de préciser que, dans l'approche de simulation mise en oeuvre, aucune procédure de validation de modèle n'est effectuée ; on s'intéresse à la qualité brute de l'inférence des effets fixes, et les problématiques de validation de modèle sont occultées. On sait en effet par avance que dans de nombreux modèles tels que les modèles non-spatiaux, l'hypothèse d'indépendance des résidus sera violée en présence d'autocorrélation spatiale ; l'un des objectifs est justement de quantifier l'erreur commise par ces modèles, lorsque l'hypothèse d'indépendance est violée.

On testera néanmoins, dans un deuxième temps, la présence d'autocorrélation spatiale dans les résidus des modèles ajustés, à l'aide du $I$ de Moran présenté chapitre 2, afin de voir quels modèles ont été capables de supprimer l'autocorrélation spatiale dans résidus. On va ainsi essayer de faire un lien entre la performance des modèles et la présence d'autocorrélation dans les résidus.

### 4.4.1 Critères de peformances intrinsèques

Une bonne inférence des effets fixes repose sur une estimation juste et précise des coefficients de régression. En termes statistiques, il s'agit de s'intéresser en premier lieu à l'erreur systématique, dont la mesure la plus connue est le biais, défini par la différence entre la moyenne des estimations et la vraie valeur. En ce qui concerne la précision de l'estimation, il s'agit de s'intéresser à une notion d'erreur aléatoire, c'est à dire à une mesure de la variabilité des estimations autour de leur moyenne, telle que la variance. On construit ainsi deux indicateurs, l'erreur de type I et le RMSE, que l'on présentera en détails dans la suite, qui permettent de tenir compte de cette double notions d'erreur.

Il y a cependant une différence fondamentale pour notre étude entre ces deux critères, qui se situe au niveau de la nature de la variabilité aléatoire prise en compte. En effet, l'erreur de type I, qui mesure le degré de fiabilité globale de l'inférence, est calculée à partir d'une variabilité *estimée* par le modèle lui-même, à savoir l'écart-type de l'estimateur qui, comme on l'a déjà vue et comme on le verra encore, est sous-estimée par les modèles non-spatiaux. Quant au RMSE, en absence de biais, il rend compte de la variabilité *réelle* de l'estimation, à laquelle on a accès grâce à l'approche par simulation ; il n'est pas calculé à partir d'une mesure de variabilité estimée par le modèle. Ces deux notions de variabilités, variabilité estimée et variabilité réelle, peuvent diverger en pratique notamment si l'on ajuste des modèles non-spatiaux sur des données spatiales, à cause de la sous-estimation des écarts-types des estimateurs. Cette problématique de divergence de variabilités sera abordée plus en détails dans la discussion chapitre 6.

#### Erreur de type I

Pour un modèle fixé, considérons à présent les notations suivantes :

$P$ : le nombre de covariables
$B$ : le nombre de simulations pour lesquelles le modèle a convergé
$\boldsymbol{\beta}^{vrai} = (\beta_1^{vrai}, \dots, \beta_P^{vrai})$ : le vecteur des vrais coefficients de régression fixés a priori
$\forall (p,b) \in \{1,\dots,P\} \times \{1,\dots,B\}, \hat{\beta}_p^b$ : l'estimation du paramètre $\beta_p$ dans la $b$-ième simulation
$\forall (p,b) \in \{1,\dots,P\} \times \{1,\dots,B\}, \hat{\sigma}_p^b$ : l'estimation de l'écart-type $\sigma_p$ de l'estimation de $\beta_p$ dans

la $b$-ième simulation

Fixons à présent $p \in \{1, \ldots, P\}$, et considérons le test d'adéquation suivant :

$$\mathcal{H}_0 : \beta_p = \beta_p^{vrai} \text{ contre } \mathcal{H}_1 : \beta_p \neq \beta_p^{vrai}$$

L'erreur de type I associée au test d'adéquation est la probabilité de rejeter à tort l'hypothèse nulle :

$$E_I^p = \mathbb{P}(\text{rejeter } \mathcal{H}_0 | \mathcal{H}_0 estvraie)$$

Afin d'estimer et d'interpréter aisément de l'erreur de type I, il faut se rappeler qu'une approche par test est toujours équivalente à une approche par intervalle de confiance. En effet, l'erreur de type I correspond à la probabilité que l'intervalle de confiance fourni par le modèle pour le paramètre $\beta_p$ ne contienne pas la vraie valeur $\beta_p^{vrai}$. Dans une approche par simulation, l'erreur de type I est alors estimée par le pourcentage d'intervalles de confiance fournis par le modèle qui ne contiennent pas la bonne valeur :

$$\hat{E}_I^p = \frac{1}{B} \sum_{b=1}^{B} \mathbb{1}(\beta_p \notin ]\hat{\beta}_{p,inf}^b, \hat{\beta}_{p,sup}^b[)$$

Où $]\hat{\beta}_{p,inf}^b, \hat{\beta}_{p,sup}^b[$ est l'intervalle de confiance fourni par le modèle pour le paramètre $\beta_p$ à la $b$-ième simulation.

Afin de fournir un indicateur de performance pour l'ensemble des $P$ paramètres, on prend la moyenne des erreurs de type I associées à chacun des paramètres :

$$\hat{E}_I = \frac{1}{P} \sum_{p=1}^{P} \hat{E}_I^p$$

On reportera sur les graphes des erreurs de type I les valeur minimum et maximum des $\hat{E}_I^p$, afin d'avoir une idée de la variabilité des erreurs de type I entre paramètres.

Les intervalles de confiance ne se calculent pas de la même manière selon que l'on se trouve dans un cadre fréquentiste ou bayésien. Pour les modèles fréquentistes, on fait l'hypothèse classique de normalité asymptotique des estimateurs (qui sera vérifiée a posteriori à l'aide d'un test de normalité). En fixant alors le seuil de significativité $\alpha$ (typiquement à 5%), et en notant $q$ le quantile d'ordre $1 - \frac{\alpha}{2}$ de la loi normale centrée-réduite, on peut écrire les bornes de l'intervalle de confiance asymptotique du paramètre $\beta_p$ au niveau de confiance $1 - \alpha$ :

$$\begin{cases} \hat{\beta}_{p,inf} = \hat{\beta}_p - q \cdot \hat{\sigma}_p \\ \hat{\beta}_{p,sup} = \hat{\beta}_p + q \cdot \hat{\sigma}_p \end{cases}$$

Cela revient simplement à considérer que la statistique de test $T_p = \frac{\hat{\beta}_p - \beta_p^{vrai}}{\hat{\sigma}_p}$ suit sous $\mathcal{H}_0$ une loi normale centrée-réduite.

En ce qui concerne les intervalles de confiance associés à des modèles bayésiens, ils sont basés sur les quantiles de la loi estimée a posteriori des estimateurs (Gelman et al., 2004). En notant respectivement $\hat{\beta}_{p,inf}$ et $\hat{\beta}_{p,sup}$ les quantiles d'ordre respectif $1 - \frac{\alpha}{2}$ et $\frac{\alpha}{2}$ de la distribution, l'intervalle de confiance bayésien du paramètre $\beta_p$ au niveau de confiance $1 - \alpha$ est donné par

$]\hat{\beta}_{p,inf}, \hat{\beta}_{p,sup}[$.

La qualité de l'inférence globale de effets fixes d'un modèle sera jugée bonne si l'erreur de type I est égale au seuil de significativité $\alpha = 5\%$. Si l'erreur de type I est supérieure ou inférieure à ce taux, cela révèle un problème dans l'estimation.

**Root Mean Squared Error**

Le Root Mean Squared Error (RMSE) est un critère classique de performance, qui mesure la variabilité des estimations autour des vraies valeurs. C'est la racine carrée du Mean Squared Error (MSE) qui est défini, en considérant les notations précédentes, pour un paramètre $p$ fixé, par :

$$MSE_p = \mathbb{E}\left((\hat{\beta}_p - \beta_p)^2\right)$$

Il est estimé à l'aide de l'approche par simulation par :

$$\widehat{MSE}_p = \frac{1}{B}\sum_{b=1}^{B}\left(\hat{\beta}_p^b - \beta_p\right)^2$$

Comme on l'a déjà dit, le MSE mesure la variabilité réelle des estimations autour de la vraie valeur, en opposition à la mesure de variabilité fournie par le modèle. Il est important de connaître le MSE d'un modèle si on veut l'appliquer sur un jeu de données réelles, car il représente la précision réelle des estimations et est donc indicateur de leur fiabilité. Par ailleurs, commme les biais seront tous approximativement nuls (voir chapitre 5), le MSE correspond alors aussi à la variabilité réelle des estimations autour de leur valeur moyenne, ce qui est la définition de la variance (non-corrigée) d'un estimateur :

$$\widehat{MSE}_p = \frac{1}{B}\sum_{b=1}^{B}\left(\hat{\beta}_p^b - \bar{\hat{\beta}}_p\right)^2$$

Dans notre contexte, le RMSE est donc un estimateur correct de l'écart type $\sigma_p$ , en opposition à la valeur moyenne des écarts-types $\hat{\sigma}_p^b$ qui sont sous-estimées par le modèle. Il faut être attentif à ces différentes manières d'estimer la variabilité aléatoire, qui peuvent diverger dans le cas des modèles non-spatiaux à cause de la sous-estimation des écarts-types fournis par le modèle.

Comme pour l'erreur de type I, on construit un indicateur de performance pour l'ensemble des $P$ paramètres en considérant la moyenne :

$$\widehat{RMSE} = \frac{1}{P}\sum_{p=1}^{P}\widehat{RMSE}_p$$

Contrairement à l'erreur de type I, le RMSE n'est pas un critère permettant de juger de la qualité *absolue* de l'estimation ; il permet simplement de comparer deux méthodes, en considérant que celle associée à un RMSE plus faible est plus performante.

### 4.4.2 Shifts entre les estimations des coefficients des modèles

Les critères de performances présentés précédemment mesurent la performance intrinsèque des différentes méthodes, que l'on pourra alors comparer sur la base de ces critères. On tient cependant à comparer nos modèles à un autre niveau, en se demandant s'ils fournissent, pour un jeu de donnée fixé, des estimations $\hat{\beta}$ similaires ou différentes. C'est toute la problématique de

l'étude des $shifts$, c'est à dire des différences entre les estimations fournies par différents modèles.

Une bonne compréhension de ces shifts est capital car, dans le cadre de l'étude de données réelles, contrairement aux approches par simulation, on ne peut pas mettre en oeuvre de critères de performances intrinsèques ; on dispose uniquement des estimations pour différents modèles, et donc des shifts.

Notons bien que dans une approche par simulation, il n'y a pas de lien direct entre les critères de performances intrinsèques et les shifts : deux méthodes différentes peuvent avoir des critères de performances optimales, en termes d'erreur de type I et de RMSE, bien que, sur chacun des jeux de données, les estimations fournies par les deux modèles différent.

Dans cette étude, on s'intéresse aux shifts entre méthodes spatiales et non-spatiales, qui ont déjà été illustrés figure 3.1, ainsi qu'aux shifts entre méthodes spatiales. De nombreuses polémiques existent dans la littérature d'écologie statistique sur l'existence et les causes de ces shifts, ce qui est à l'origine de beaucoup de confusion chez les chercheurs en écologie. Cette problématique est abordée en détails, à la lumière de quelques résultats, dans la discussion chapitre 6.

### 4.4.3 L'autocorrélation spatiale dans les résidus des modèles

On cherche enfin à comparer les modèles sur la question de la présence d'autocorrélation spatiale dans les résidus, qui pourrait expliquer les différences de performances entre les modèles. On veut savoir si une partie de l'autocorrélation spatiale présente dans les données va se retrouver dans les résidus, violant ainsi l'hypothèse classique d'indépendance du modèle linéaire.

On va, dans un premier temps, essayer d'établir un lien entre le type de méthode utilisé (non-spatial, spatial indépendant et spatial structuré) et la présence ou l'absence d'autocorrélation dans les résidus, en se demandant conjointement si les méthodes non-spatiales conservent de l'autocorrélation dans les résidus et si les méthodes spatiales parviennent à la supprimer. On va alors, sur cette base, chercher à faire un lien entre les performances intrinsèques des modèles et la présence d'autocorrélation dans les résidus, en investiguant si les méthodes qui conservent de l'autocorrélation dans les résidus sont nécessairement moins performantes que les méthodes parvenant à la supprimer.

Ces questions font, elles aussi, polémiques dans la littérature. Dans cette étude, nous allons essayer d'apporter notre vision à ce débat, à la lumière de nos résultats, et d'exhiber le comportement des modèles de la famille des GLM et des GAM.

Concrètement, on va appliquer l'indice de Moran ($I$ de Moran), présenté chapitre 2, pour quantifier le degré d'autocorrélation dans la série des résidus des modèles, sur deux scénarios spatiaux stationnaires représentatifs de l'ensemble de nos simulations, à savoir les scénarios GRID.4 et GNB.4 (voir tableau 4.1). Les corrélogrammes de Moran sont tracés pour l'ensemble des méthodes, en utilisant une matrice de voisinage définie à partir d'une distance limite égale à 1.5km. Afin de de mettre en oeuvre le test de significativité de la présence d'autocorrélation spatiale, on a vu qu'il est nécessaire que les observations, ici les résidus, soient distribuées normalement. Or, les résidus classiques de Pearson ou de la déviance sont connus, dans un contexte non-gaussien, pour s'éloigner notablement de l'hypothèse de normalité. On va donc utiliser un type de résidus plus fiables, les résidus quantiles randomisés, que l'on présente et que l'on applique à notre cas d'étude annexe C.

TABLE 4.1 – Synthèse des scénarios de simulations de base

| Scénario | Covariables | Terme d'erreur | Non-stationnarité spatiale |
|---|---|---|---|
| GRID.1 | Quatre covariables indépendantes et spatialement structurées avec différentes valeurs de portée et de force | Pas de terme d'erreur | Non |
| GRID.2 | Comme GRID.1 | Portée faible, intensité forte | Non |
| GRID.3 | Comme GRID.1 | Portée intermédiaire, intensité forte | Non |
| GRID.4 | Comme GRID.1 | Portée élevée, intensité forte | Non |
| GRID.5 | Comme GRID.1 | Portée faible, intensité faible | Non |
| GRID.6 | Comme GRID.1 | Portée intermédiaire, intensité faible | Non |
| GRID.7 | Comme GRID.1 | Portée élevée, intensité faible | Non |
| GRID.8 | Deux covariables indépendantes et spatialement structurées avec différentes valeurs de portée, et deux covariables indépendantes incorporant de la non-stationnarité spatiale | Portée élevée, intensité forte | Non-stationnarité dans les covariables |
| GRID.9 | Comme GRID.1 | Portée non-stationnaire, intensité forte | Non-stationnarité dans le paramètre de portée spatiale du terme d'erreur, avec des valeurs qui varient dans l'espace |
| GRID.10 | Comme GRID.8 | Comme GRID.9 | Non-stationnarité dans les covariables comme GRID.8 et non-stationnarité dans le terme d'erreur comme GRID.9 |
| GNB.1 | La vraie variable *Gestion* et trois covariables indépendantes et spatialement structurées avec différentes valeurs de portée et de force | Pas de terme d'erreur | Non |
| GNB.2 | Comme GNB.1 | Portée faible, intensité forte | Non |
| GNB.3 | Comme GNB.1 | Portée intermédiaire, intensité forte | Non |
| GNB.4 | Comme GNB.1 | Portée élevée, intensité forte | Non |
| GNB.5 | Comme GNB.1 | Portée non-stationnnaire inter-massif, intensité forte | Différentes valeurs selon les massifs pour le paramètre de portée spatiale dans le terme d'erreur |
| GNB.6 | Comme GNB.1 | Portée non-stationnaire intra-massif, intensité forte | Différentes valeurs de portée spatiale emboîtées intra-massif |

# Chapitre 5

# Résultats

On présente dans cette partie les résultats des simulations décrites dans le chapitre 4, en commençant par les simulations de base sur grille et sur scénarios GNB. Les résultats des simulations additionnelles mises en oeuvre, ainsi que quelques résultats concernant les shifts entre les estimations des coefficients des différents modèles, sont présentés dans un deuxième temps. On conclut enfin ce chapitre par la présentation des résultats de l'étude de l'autocorrélation spatiale dans les résidus des modèles.

Dans cette partie, seuls les résultats bruts sont présentés ; leur interprétation est laissée au chapitre 6.

## 5.1 Résultats des simulations de base

Présentons tout d'abord les résultats des 16 scénarios de base sur grille (GRID.1-10) et sur placettes GNB (GNB.1-6). Dans le présent rapport sont illustrés les résultats complets de la comparaison de modèles pour les scénarios spatiaux GRID.4, figure 5.1, et GNB.4, figure 5.2, qui sont représentatifs de l'ensemble des scénarios sur grille et sur placettes GNB.

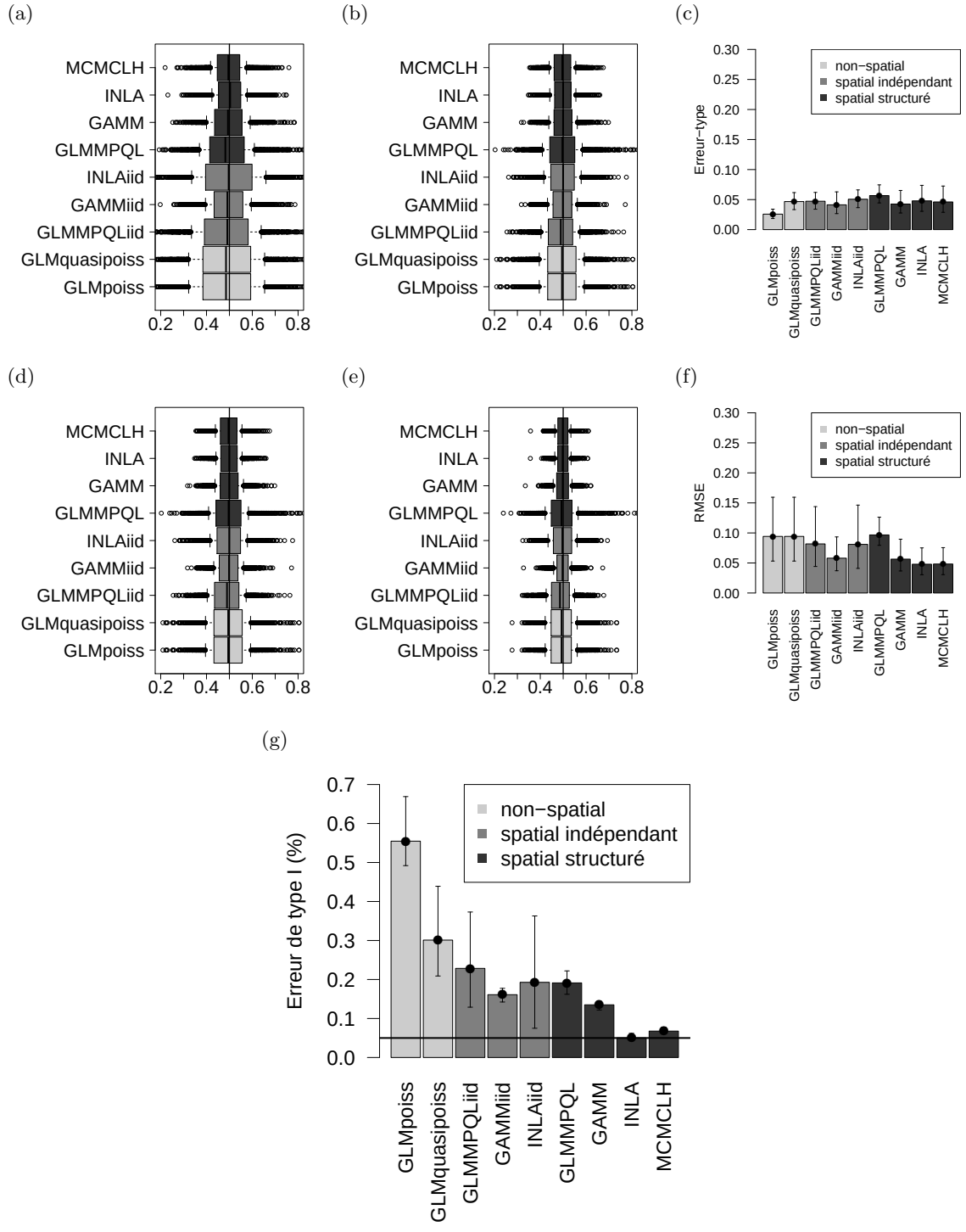
Les résultats complets de la comparaison de modèles, pour un scénario donné, comprennent tout d'abord les boîtes à moustaches, pour l'ensemble des modèles, des estimations des coefficients de régression, dont les vraies valeurs sont représentées à l'aide de lignes verticales. Les résultats comprennent aussi des représentations de la moyenne, pour l'ensemble des paramètres, des erreurs de type I et des RMSE associés aux modèles, ainsi que de la moyenne des erreurs-types estimées. Leurs valeurs minimum et maximum sont aussi reportées sur les graphes.

Les résultats donnés dans la suite, ainsi que l'interprétation des résultats qui va suivre chapitre 6, concerne tout les résultats obtenus sur l'ensemble desscénarios.

La prermier point que l'on observe, à la vue de l'ensemble des résultats, est leur homogénéité globale à la fois entre les différentes scénarios sur grille et sur placettes GNB ainsi qu'entre les deux critères de performances, l'erreur de type I et le RMSE ; un modèle performant sur un scénario a tendance a être performant sur les autres scénarios et, de manière analogue, un modèle performant en terme d'erreur de type I a tendance à être performant en terme de RMSE, et réciproquement.

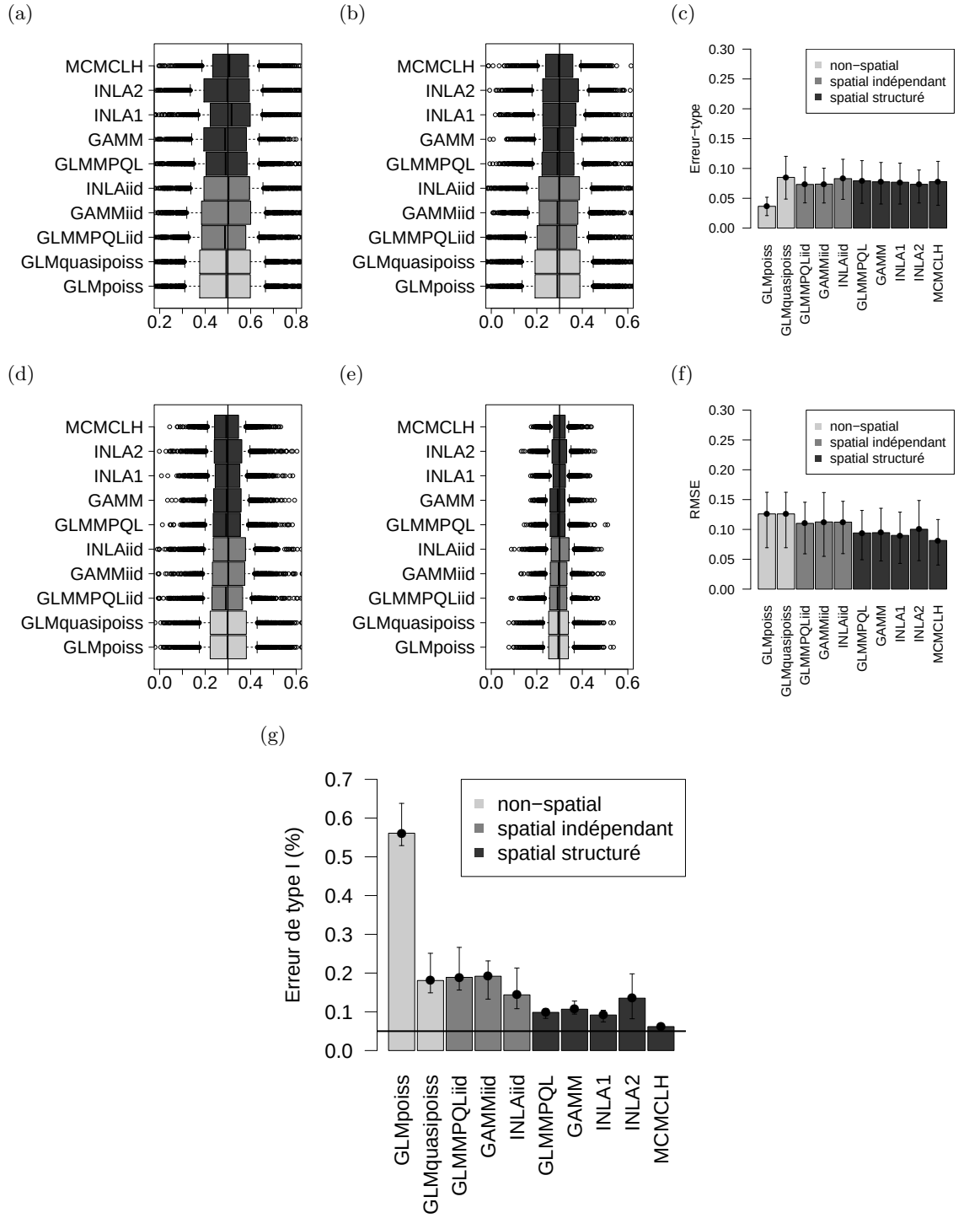
On observe ensuite, et c'est l'un des principaux résultats de l'étude, un gradient croissant de performances globales entre les méthodes non-spatiales, les méthodes spatiales indépendantes

FIGURE 5.1 – Résultats complets des comparaisons de modèles pour le scénario GRID.4




et les méthodes spatialement structurées, sur l'ensemble des scénarios et des deux critères de performances. Avant d'entrer plus en détails dans la présentation de ce gradient de performance, il faut noter en premier lieu que les méthodes testées, qu'elles soient spatiales ou non-spatiales, ne présentent aucun biais dans l'estimation des effets fixes, que ce soit dans les scénarios sta-

FIGURE 5.2 – Résultats complets des comparaisons de modèles pour le scénario GNB.4




tionnaires ou non-stationnaires. Cela peut être vue notamment sur les boîtes à moustaches des estimations sur grille figure 5.1 ou sur scénario GNB 5.2, graphes (a), (b), (d) et (e) ; on observe que les médianes, qui se confondent dans ce cas avec les moyennes car les distribution sont approximativement normales, sont environ égales aux vraies valeurs.

Détaillons à présent le gradient global de performances observé entre les trois types de modèles :

1) En ce qui concerne les modèles non-spatiaux, bien que les estimations ne sont pas biaisées, le GLM de Poisson (GLM) présente de très faibles performances, et les performances les plus mauvaises de l'ensemble des méthodes implémentées, et ce même quand le degré d'autocorrélation spatiale résiduelle est réduit. En effet, les erreurs de type I sont uniformément très élevées sur l'ensemble des scénarios, à des valeurs situées entre 50% et 60%. Elles sont associées à des estimations des erreurs-types très faibles relativement aux autres modèles, de l'ordre de 0.025 à 0.03. Enfin, le GLM de Poisson fournit dans tout les scénarios spatiaux des RMSE maximales, indiquant que les estimations fournies sont les plus imprécises d'entre toutes.

2) Le modèle non-spatial de Quasi-Poisson (GLMquasipoiss) ainsi que les modèles spatiaux indépendants (GLMMPQLiid, GAMMiid, INLAiid) présentent des performances intermédiaires, avec des erreurs de type I comprises en moyenne entre 10% et 30% sur grille, et entre 10% et 20% sur scénarios GNB, associées à des erreurs-types estimées bien plus élevées que le GLM de Poisson (de l'ordre de 0.05 sur grille et de 0.08 sur scénarios GNB). Concernant la précision des estimations, ils présentent des RMSE intermédiaires, en moyenne de 15% inférieures aux RMSE du GLM de Poisson. Alors que c'est GAMMiid qui fournit les performances les meilleurs sur grille parmi les modèles spatiaux indépendants, c'est l'implémentation bayésienne INLAiid qui est la plus performante sur scénarios GNB.

3) Les modèles spatialement structurés sont ceux qui présentent globalement les meilleurs performances sur l'ensemble des scénarios, avec des erreurs de type I et des RMSE globalement plus faibles que les modèles spatiaux indépendants, et des erreurs-type estimées de l'ordre de grandeur de celles des méthodes spatiales indépendantes. Cependant, des différences de deux types apparaissent entre méthodes spatialement structurées :

- Les modèles spatiaux se différencient tout d'abord, en termes de performances, selon qu'ils aient été estimés de manière fréquentiste ou bayésienne : les implémentations fréquentistes GLMMPQL et GAMM, basées sur l'approximation PQL, sont globalement moins performantes que les implémentations bayésiennes INLA et MCMCLH. En effet, les méthodes bayésiennes spatialement structurées présentent des performances optimales : ce sont les seules qui parviennent, à la fois sur grille et sur scénarios GNB, à atteindre des erreurs de type I correctes proches de 5%, avec des RMSE minimales. Quant aux méthodes spatiales fréquentistes, leurs erreurs de type I restent au-dessus de 5%, en moyenne entre 10% et 20%, bien que leurs erreurs-types sont du même ordre de grandeur que celles des méthodes spatiales bayésiennes. Enfin, les RMSE associés atteignent quasiment le niveau minimale des RMSE des modèles spatiaux bayésiens, si l'on met à part le modèle GLMMPQL sur grille, qui présente un RMSE encore plus élevé que le modèle spatial indépendant le plus performant INLAiid. Précisons encore que les modèles fréquentistes, à savoir GLMMPQLiid, GAMMiid, GLMMPQL et GAMM, souffrent de problèmes de convergence sur 10% à 30% des scénarios spatiaux sur grille, ce qui n'est pas le cas sur les scénarios GNB.

- Les performances des modèles spatiaux se différencient ensuite entre modèles spatiaux basés sur les distances et modèles spatiaux basés sur les voisinages, si l'on considère uniquement ces modèles dans leur implémentation bayésienne, et cela uniquement sur scénarios GNB. En effet, le modèle MCMCLH basé sur les distances, et le modèle INLA basé sur les voisinages, présentent

des performances identiques et optimales sur grille, comme on l'a déjà décrit. Sur scénarios GNB, alors que MCMCLH reste optimal sur l'ensemble des scénarios en terme d'inférence des effets fixes, les deux implémentations différentes INLA1 et INLA2 du modèle basé sur les voisinages (définies à partir d'une distance limite plus grande pour INLA2 que pour INLA1) présentent des performances différentes, avec des erreurs de type I proches de 10% pour INLA1 et de 15% pour INLA2, et des RMSE plus faibles (de l'ordre de 10% inférieures) pour INLA1. Les performances de INLA2 sont alors plus comparables à celles de la méthode spatiale indépendante INLAiid qu'à celles de INLA1.

Si l'on considère à présent en détails les performances des modèles sur les scénarios non-stationnaires, des tendances similaires se dégagent ; les méthodes performantes dans les cas stationnaires le sont aussi globalement dans les cas non-stationnaires, que ce soit sur grille ou sur scénarios GNB. Notons simplement que les méthodes spatiales indépendantes ont tendance à être moins performantes dans les cas non-stationnnaires sur grille. En effet, les méthodes GLMMPQL et INLAiid ont des erreurs de type I particulièrement élevées (de l'ordre de 30% à 40%) sur les scénarios GRID.8 et GRID.10, ce qui n'est pas le cas de GAMMiid, qui présente sur ces scénarios des erreurs de type I de l'ordre de 15% et des RMSE quasi-minimales.

Notons enfin que la méthode MCMCLH est la seule méthode spatiale basée sur les distances qui estime correctement les paramètres spatiaux de variance $\sigma^2$ et de portée $\phi$, sur l'ensemble des scénarios stationnaires sur grilles et sur scénarios GNB.

Pour finir, on observe que toutes les méthodes implémentées, qu'elles soient spatiales ou non-spatiales, fournissent des erreurs de type I correctes proches de 5% et des RMSE minimales sur les scénarios non-spatiaux de référence GRID.1 et GNB.1. Les estimations fournies sont en fait identiques entre toutes les méthodes.

## 5.2 Résultats des simulations additionnelles

### 5.2.1 Robustesse de MCMCLH à des phénomènes de sur-dispersion et de sous-dispersion

La figure 5.3 illustre les erreurs de type I associées aux modèles testés, représentatifs des différents types de modèles, dans le cas d'une augmentation progressive du degré de sur-dispersion et de sous-dispersion.

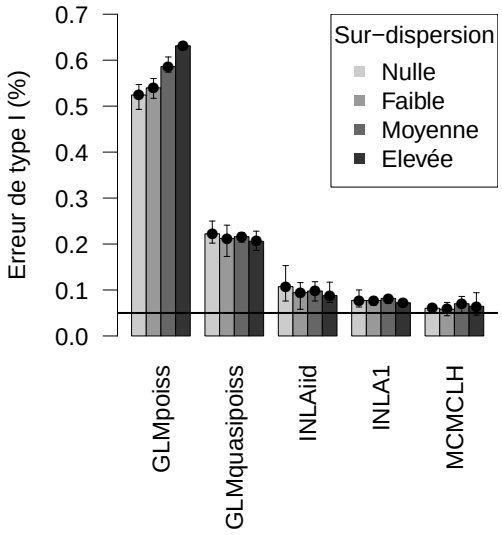
On observe globalement peu de différences en terme d'erreur de type I, pour l'ensemble des modèles, lorsqu'on fait évoluer le degré de sur-dispersion ou de sous-dispersion, mis à part le GLM de Poisson dans le cas sur-dispersé, qui voit son erreur de type I, déjà très élevée en absence de sur-dispersion (environ 52%), encore grimper jusqu'au taux de 63% dans le cas du degré de sur-dispersion maximal. Autant le modèle de Quasi-Poisson que les modèles spatiaux indépendants et spatiaux structurés s'adaptent bien en terme d'erreur de type I.

En particulier, la méthode MCMCLH, qui est déjà optimale en terme d'erreur de type I en cas d'absence de sur-dispersion ou de sous-dispersion, reste optimale lorsque la dispersion augmente, avec des erreurs de type I qui restent stable au taux nominal de 5%.

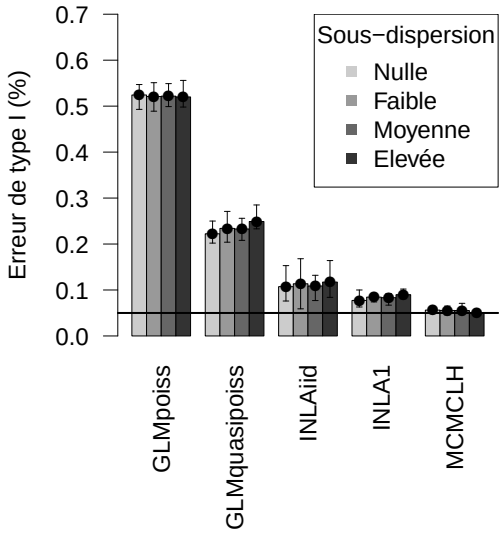
Si l'on se concentre précisément sur la méthode MCMCLH, dans le contexte de données sur-dispersées par exemple, on observe néanmoins que plusieurs problèmes d'inférence apparaissent :

FIGURE 5.3 – Erreurs de type I associées aux méthodes testées dans le cas d'une augmentation progressive du degré de sur-dispersion et de sous-dispersion

(a) Données sur-dispersées




(b) Données sous-dispersées




1) Le RMSE augmente (et ce pour toutes les méthodes) avec le degré de sur-dispersion, passant du simple au double dans le cas de sur-dispersion maximale. Cette variation de RMSE est associée à des estimations eux aussi plus importantes des écarts-types des estimateurs.

2) La constante du modèle est sous-estimée, et ce biais augmente avec le degré de sur-dispersion ; dans le cas de sur-dispersion maximale, la constante est estimée à environ la moitié de sa vraie valeur.

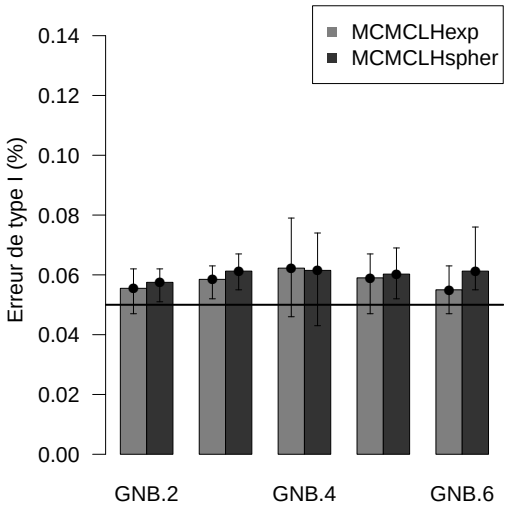
3) Les paramètres spatiaux de variance $\sigma^2$ et de portée spatiale $\phi$ sont eux aussi estimés avec un biais. En effet, les paramètres respectifs $\sigma^2$ et $\phi$ sont respectivement sur-estimés et sous-estimés, et ce de plus en plus à mesure que le degré de sur-dispersion augmente ; dans le cas de sur-dispersion maximale, $\sigma^2$ est estimé à presque le double de sa valeur et $\phi$ à moins de la moitié de la sienne.

Les cas sous-dispersés présentent le même type de phénomènes, mais de façon inverse : à mesure que le degré de sous-dispersion augmente, le RMSE baisse, la constante est sur-estimée et les paramètres $\sigma^2$ et $\phi$ sont respectivement sous-estimés et sur-estimés.

### 5.2.2 Robustesse de MCMCLH à une mauvaise spécification de la fonction de corrélation spatiale

La figure 5.4 représente les erreurs de type I, évaluant la performance globale de l'inférence des effets fixes, associées aux deux méthodes MCMCLH construites dans le but de tester la robustesse du modèle MCMCLH à une mauvaise spécification de la fonction de corrélation spatiale, à savoir le modèle MCMCLHexp dont la fonction est spécifiée en adéquation avec les données, et le modèle MCMCLHspher qui est spécifié de manière incorrecte.

FIGURE 5.4 – Erreurs de type I associées aux modèles MCMCLHexp et MCMCLHspher




On observe que les deux modèles ont globalement des erreurs de type I comparables sur l'ensemble des scénarios GNB ; les valeurs sont stables au niveau du taux nominal de 5%, même si le degré d'autocorrélation augmente (GNB.2-4) et même si les scénarios présentent des composantes de non-stationnarité spatiale (GNB.5-6). Les RMSE, qui ne sont pas illustrés, présentent des niveaux identiques de performance.

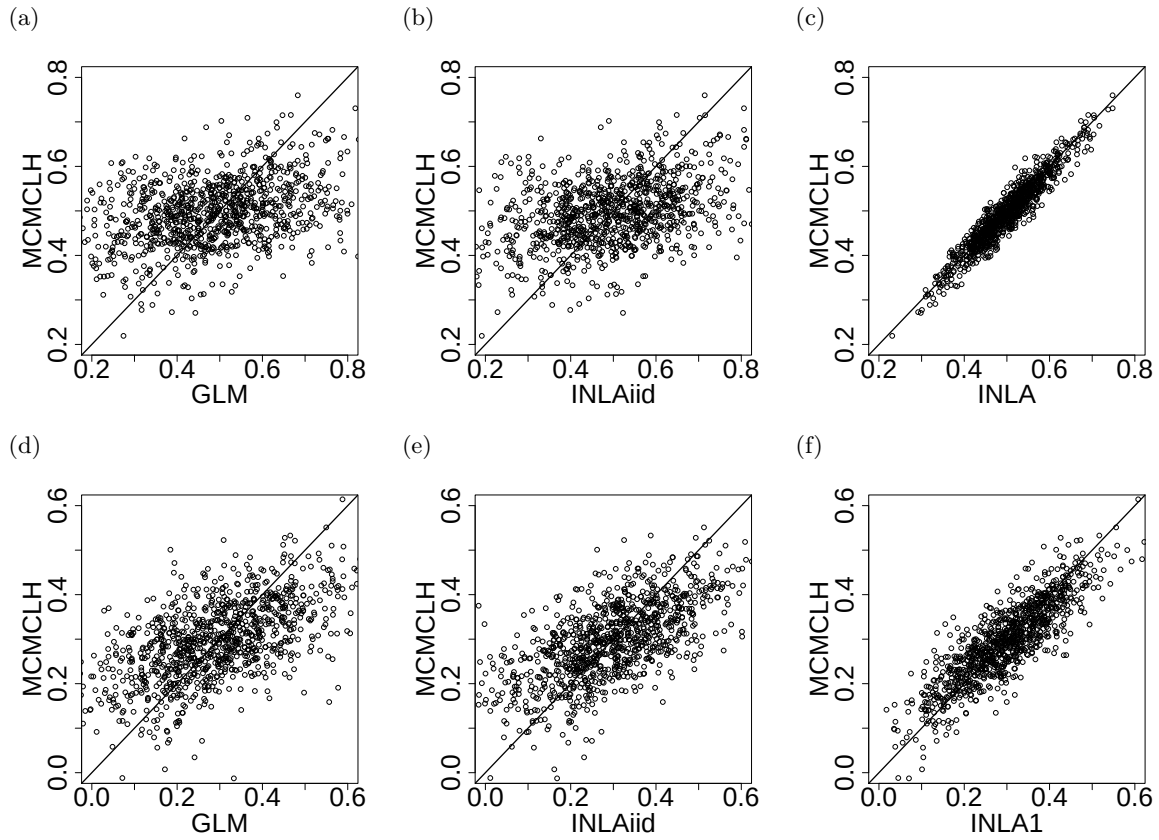
Si l'on regarde plus en détails, on s'aperçoit en fait que les deux modèles fournissent approximativement les mêmes estimations, avec les mêmes écarts-types associés, ce qui aboutit logiquement à des erreurs de type I et des RMSE équivalents.

## 5.3 Résultats de l'étude des shifts entre les modèles

On fournit à présent quelques résultats concernant l'étude des shifts entre les modèles, c'est à dire des différences entre les estimations des coefficients fournies par les modèles. La figure 5.5 présente les résultats des shifts pour un jeu de données des scénarios GRID.4, sur les graphes (a), (b) et (c), et GNB.4, sur les graphes (d), (e) et (f), en s'intéressant à trois types de shifts entre méthodes spatiales représentatives, à savoir :
- Les shifts entre méthodes spatialement structurées et méthodes non-spatiales, étudiés avec les modèles MCMCLH et GLM
- Les shifts entre méthodes spatialement structurées et méthodes spatiales indépendantes, étudiés avec les modèles MCMCLH et INLAiid
- Les shifts entre méthodes spatialement structurées elles-mêmes, étudiés avec les modèles MCMCLH et INLA sur grille, et MCMCLH et INLA1 sur scénarios GNB.

On observe tout d'abord, sur les graphes (a) et (d), la présence de shifts importants entre la méthode spatialement structurée et la méthode non-spatiale. D'autres graphes, tels que ceux de la figure 3.1, montrent que ces shifts entre méthodes spatiales et non-spatiales ont tendance à augmenter au fur et à mesure que l'autocorrélation spatiale résiduelle augmente. On observe

FIGURE 5.5 – Illustration de trois types de shifts entre modèles sur les scénarios GRID.4 et GNB.4




ensuite, sur les graphes (b) et (e), des shifts de degré à peine plus faible entre la méthode spatialement structurée et la méthode spatiale indépendante. Sur les graphes (d) et (f), on obseve enfin un degré faible de shifts entre méthodes spatiales : il n'y a quasiment aucun shift entre MCMCLH et INLA sur grille, et à peine plus sur le scénario GNB, mais de degré significativement moindre que les shifts entre méthodes spatiales et non-spatiales.

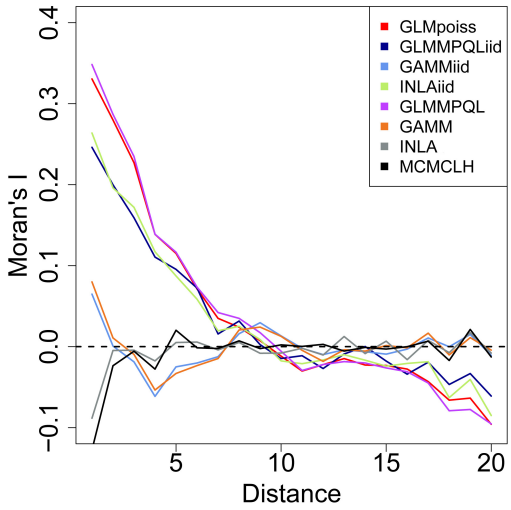
Evoquons ensuite deux résultats additionels importants, qui ne sont pas représentés graphiquement dans le présent rapport. Le premier est que lorsque le degré de RMSE est minimal ou quasi-minimal, par exemple sur grille avec les méthodes GAMMiid, GAMM, MCMCLH et INLA, ou bien sur scénario GNB avec les modèles GLMMPQL, GAMM, INLA1 et MCMCLH, les méthodes ne présentent quasiment aucun shifts les unes par rapport aux autres. Le deuxième résultat additionnel concerne les deux implémentations pratiques sur scénarios GNB, INLA1 et INLA2, du modèle basé sur les voisinages ; il existe un shift considérable entre les deux méthodes, qui reste cependant significativement moins grand que le shift entre méthodes spatiales et non-spatiales.

## 5.4 Résultats de l'étude de l'autocorrélation spatiale des résidus des modèles

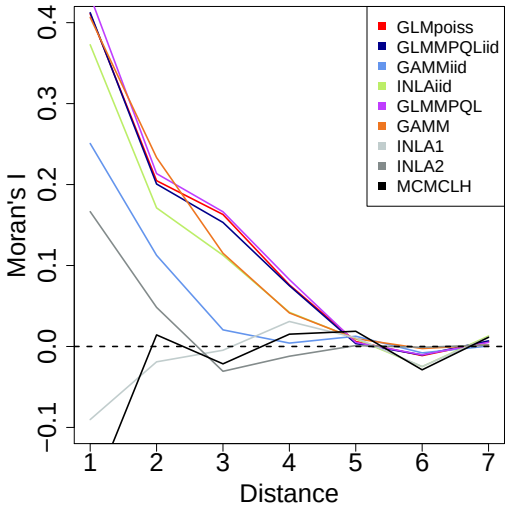
Présentons à présent les résultats liés à l'étude de l'autocorrélation spatiale des résidus, qui sont illustrés graphiquement figure 5.6 pour l'ensemble des modèles, à l'aide des corrélogrammes de Moran sur un des jeux de données des scénarios GRID.4 et GNB.4. Les résultats exhibés ci-dessous, par l'examen visuel des corrélogrammes, sont toujours en accord avec les conclusions du test de significativité de la présence d'autocorrélation spatiale présenté chapitre 2.

FIGURE 5.6 – Corrélogrammes de Moran appliqués à un jeu de données des scénarios GRID.4 et GNB.4

(a) Scénario GRID.4




(b) Scénario GNB.4




La première question à laquelle il s'agit de répondre (voir chapitre 4) concerne la présence ou l'absence effective d'autocorrélation spatiale dans les résidus des modèles testés.

Ainsi qu'on le voit figure 5.6, les modèles non-spatiaux et spatiaux indépendants ne sont pas capables d'enlever l'autocorrélation spatiale des résidus, à l'exception du modèle spatial indépendant de la famille des GAM, à savoir GAMMiid, et cela uniquement sur grille. Notons qu'il y parvient, avec d'ailleurs le modèle spatialement structuré GAMM, en conservant néanmoins un faible taux d'autocorrélation avec une portée faible environ égale à 3.

Concernant les modèles spatialement structurés, ils ne présentent pas tous le même comportement. En effet, sur grille tout d'abord, alors que les modèles spatiaux bayésiens MCMCLH et INLA parviennent à supprimer l'autocorrélation des résidus, ce n'est pas nécessairement le cas pour les modèles spatiaux fréquentistes, basés sur l'approximation PQL ; alors que le modèle GAMM y parvient, le modèle GLMMPQL échoue. Sur le scénario GNB, alors que le modèle bayésien MCMCLH, basé sur les distances, parvient à nouveau à fournir des résidus non autocorrélés, ce n'est plus du tout le cas pour les modèles de la famille des GAM (GAMMiid et GAMM), ni entièrement le cas pour le modèle basé sur les voisinages. En effet, seul le modèle INLA1, généré à partir de voisinages étroits, parvient à supprimer l'autocorrélation dans les résidus,

contrairement à INLA2, généré à partir de voisinages larges, qui conserve un degré d'autocorrélation significatif dans les résidus, bien que plus faible que celui des méthodes non-spatiales. Quant au modèle fréquentiste GLMMPQL, il ne parvient toujours pas à fournir des résidus non autocorrélés sur le scénario GNB.

Le deuxième question concerne le lien entre la présence d'autocorrélation dans les résidus des modèles et leurs performances intrinsèques en termes d'erreur de type I et de RMSE. En observant consciencieusement les corrélogrammes sur l'ensemble des scénarios, on note les deux points suivants :

1) Il n'y a pas de lien direct, de manière générale, entre l'absence d'autocorrélation dans les résidus et une erreur de type I correcte. Bien que cette relation s'observe par exemple sur grille avec INLA et MCMCLH qui présentent à la fois des erreurs de type I optimales (égales à 5%) et une absence de résidus autocorrélés, cela n'est pas vérifié par exemple pour les modèles de la famille des GAM (GAMMiid et GAMM) sur grille, qui présentent des erreurs de type I bien au-delà de la valeur nominale de 5% en l'absence d'autocorrélation spatiale significative dans les résidus. C'est aussi le cas pour la méthode INLA1 sur le scénario GNB, où une absence d'autocorrélation spatiale est associée à une erreur de type I élevée proche de 10%.

2) Un lien marginal existe entre un niveau faible pour le RMSE et l'absence d'autocorrélation dans les résidus. En effet, sur grille par exemple, les quatre méthodes qui parviennent à enlever l'autocorrélation spatiale des résidus (GAMMiid, GAMM, INLA et MCMCLH) ont toutes des RMSE quasi-minimales. Et, pour les autres méthodes qui conservent de manière significative de l'autocorrélation dans les résidus, les valeurs de RMSE sont bien plus élevées. Il existe cependant, sur scénario GNB notamment, des méthodes spatiales telles que GLMMPQL et GAMM, qui présentent un niveau de RMSE quasi-minimal tout en conservant un niveau d'autocorrélation significatif dans les résidus.

# Chapitre 6

# Interprétation des résultats et discussion

Cette dernière partie est dédiée à l'interprétation, en lien avec la littérature d'écologie statistique, des résultats présentés chapitre 5, ainsi qu'à la discussion de notions plus générales concernant les approches par simulation et l'utilisation des méthodes spatiales.

## 6.1 Interprétation des résultats

Les résultats obtenus, présentés chapitre 5, sont globalement en accord avec les conclusions issues d'études analogues (Dormann et al., 2007; Beale et al., 2007, 2010), notamment en ce qui concerne la faillite en pratique des modèles non-spatiaux lorsqu'ils sont ajustés sur des données spatialement autocorrélées. Notre étude montre en effet, de manière analogue aux travaux de Beale et al. (2010) qui se situent dans un contexte de données spatiales gaussiennes, que la régression non-spatiale de Poisson fournit des estimations peu fiables des effets fixes, et cela même dans le cas d'un faible degré d'autocorrélation résiduelle. Les erreurs de type I très élevées, observées sur tout les scénarios, sont principalement dûes à une sous-estimation des écarts-types des estimateurs, alors que, en opposition, la variabilité réelle des estimations, mesurée à l'aide RMSE, est maximale. Ce phénomène de sous-estimation de la variabilité des estimateurs apparaît dans d'autres contextes tels que celui des données longitudinales ; les phénomènes d'autocorrélation spatiale et de mesures repétées s'inscrivent tout les deux dans le problème plus large de pseudoréplication introduit par Hurlbert (1984), qui décrit comment les modèles classiques traitent des observations corrélées comme indépendantes et se trompent dans l'estimation des variabilités. Les problèmes d'inférence dont souffre le modèle de Poisson sont vraisemblablement causés par la présence d'autocorrélation spatiale dans les résidus, que le modèle n'a pa été capable de supprimer.

Certains auteurs, tels que Bini et al. (2009) ou (Hawkins, 2012), semblent ne pas avoir conscience de ce phénomène de divergence de variabilités - entre la variabilité estimée par le modèle à travers l'écart-type, et la variabilité réelle estimée à l'aide de l'approche par simulation par le RMSE ; ils considèrent que le fait que les estimateurs ne soient pas biaisés, et qu'ils présentent en plus des écarts-types estimés de faibles valeurs, suffit à rendre les estimations justes et précises. Or, l'approche par simulation montre clairement que les coefficients sont estimés en réalité avec une forte erreur aléatoire, ce qui les rend peu fiables.

Concernant le modèle de Quasi-Poisson et les modèles spatiaux indépendants, ils améliorent tout les deux les performances de l'inférence des effets fixes, par la prise en compte d'un phéno-

mène de sur-dispersion des données de comptage, lié aux composantes spatiales résiduelles des données. Ces modèles permettent, en effet, de modéliser des sources de variabilité non-observées et indépendantes, soit à travers un paramètre de sur-dispersion pour le modèle de Quasi-Poisson, soit à travers des effets aléatoires i.i.d. pour les modèles spatiaux indépendants, en faisant ainsi augmenter les écarts-types des estimateurs, ce qui fait logiquement baisser l'erreur de type I. Les estimations elles-mêmes sont aussi plus précises mais l'étude des shifts, qui sont faibles entre modèles spatiaux indépendants et modèles non-spatiaux, montre que le gain de performance réalisé par les modèles spatiaux indépendants est principalement dû à l'augmentation des estimations des écarts-types. Ni les erreurs de type I ni les RMSE ne sont pourtant optimales, bien que les écarts-types soient estimés avec des valeurs de l'ordre de celles des modèles spatiaux bayésiens les plus performants. En effet, l'amélioration des performances ne peut être complète, car les modèles spatiaux indépendants permettent uniquement la prise en compte de sources de variabilités indépendantes, alors que les dépendances spatiales constituent des sources de variabilités corrélées.

Les modèles spatialement structurés sont les seuls qui sont a priori pertinents pour modéliser l'autocorrélation spatiale, car ils sont en mesure de modéliser des sources de variabilités additionnelles corrélées. On observe en effet, sur l'ensemble des simulations, que les modèles spatiaux structurés sont globalement les plus performants, en termes d'inférence globale et de précision réelle des estimations, bien que des différences apparaissent selon le type de procédure numérique utilisée pour l'inférence et la nature de la modélisation spatiale sous-jacente.

On observe tout d'abord, en effet, des différences significatives de performances entre les méthodes spatiales bayésiennes, basés sur des algorithmes MCMC, qui réalisent des estimations globalement optimales, et les méthodes spatiales fréquentistes, basés sur l'approximation PQL, dont les performances sont médiocres et à peine meilleures que les méthodes spatiales indépendantes, et qui peinent par ailleurs à converger sur certains scénarios spatiaux élementaires. Ces mauvaises performances sont vraisemblablement liées au fait qu'ils ne parviennent pas à enlever l'autocorrélation dans les résidus, sauf marginalement les modèles de la famille des GAM sur grille. Ces résultats indiquent que l'approximation PQL n'est pas viable dans le contexte de l'estimation de modèles spatiaux structurés ajustés sur des données de comptage. Cela n'est pas si étonnant, car l'approximation PQL est connue pour fournir des problèmes d'inférence, en particulier lorsque le nombre d'effets aléatoires à estimer est élevé et que l'on se situe dans un cadre non-gaussien (McCulloch, 1997), comme c'est précisément le cas dans notre étude. Quant aux modèles spatiaux bayésiens, ce sont les seuls qui produisent globalement des estimations satisfaisantes ; l'amélioration, par rapport aux modèles spatiaux indépendants, est principalement dûe à l'augmentation de la précision réelle des estimations des coefficients de régression, et non à une meilleure estimation des écarts-types, dont les valeurs sont d'ailleurs approximativement du même ordre que celles des méthodes spatiales indépendantes.

Les modèles spatialement structurés se différencient ensuite, en termes de performance, selon la nature de la modélisation spatiale sous-jacente et le contexte spatial dans lequel on les applique. En effet, si on les considère dans leur implémentation bayésienne, le modèle spatial MCMCLH basé sur les distances et le modèle spatial INLA basé sur les voisinages fournissent des performances différentes dans le cas des scénarios GNB. Alors que le modèle spatial basé sur les distances présente des performances optimales à la fois sur grille et sur scénarios GNB, en enlevant correctement l'autocorrélation spatiale des résidus dans tout les cas, le modèle basé sur les voisinages se heurte au problème délicat de la spécification des relations de voisinages dans le contexte d'observations irrégulièrement espacées dans les scénarios GNB. La manière

dont on spécifie en pratique les relations de voisinages relève de choix a priori de distances limites arbitraires. Or, on observe des différences significatives de performances des méthodes selon le voisinage utilisé, affectant les erreurs de type I et la précision des estimations, ainsi que la capacité des modèles à fournir des résidus non-autocorrélés. Il est donc important de veiller à ne pas créer des relations de voisinages artificielles, ou d'omettre des relations de voisinages réelles. Malheureusement, on ne peut jamais être sûr en pratique de modéliser correctement les voisinages. Les méthodes basés sur les distances semblent ainsi plus souples d'utilisation dans de tels contextes spatiaux, car ils ne nécessitent pas de choix a priori, qui sont difficiles à justifier.

Les résultats indiquent, par ailleurs, que les méthodes bayésiennes spatialement structurées s'adaptent particulièrement bien en présence de composantes spatiales non-stationnaires, qu'elles soient dans les covariables ou le terme d'erreur. La méthode MCMCLH, basée sur les distances, repose sur une modélisation des dépendances spatiales à l'aide d'une fonction de corrélation stationnaire. Bien qu'une telle fonction ne permette pas a priori de modéliser des caractéristiques non-stationnaires des dépendances spatiales, le modèle MCMCLH est robuste à tout les types de non-stationnarité spatiale testés, à la fois sur grille et sur scénarios GNB. La mise en oeuvre de méthodes non-stationnaires complexes, telles que la méthode GWR (Geographically Weighted Regression) (Fotheringham et al., 2002) qui fournit différentes estimations des coefficients selon la région de l'espace, ne semble donc pas nécessaire en pratique.

Les résultats montrent aussi que les modèles de la famille des GAM, qui permettent d'introduire de manière non-linéaire des coordonnées spatiales comme covariables supplémentaires, n'améliorent pas notablement la performance des estimations, sauf de manière marginale sur les scénarios spatiaux simples sur grille, où les coordonnées spatiales ont un sens, et sur les scénarios non-stationnaires sur grille présentant des tendances linéaires continues, qui peuvent facilement être captées par ces covariables spatiales additionnelles. Cela n'est plus le cas sur les scénarios GNB où les coordonnées spatiales sont trop hétérogènes.

Concernant les résultats des simulations additionnelles, ils mettent tout d'abord en évidence la robustesse, en terme d'estimation des effets fixes, de la méthode spatiale bayésienne MCMCLH à des phénomènes de sur-dispersion et de sous-dispersion, qui invalident l'hypothèse d'adéquation au modèle de Poisson. Bien que les effets fixes soient estimés de manière satisfaisante, les estimations des paramètres spatiaux ne sont pas fiables, car biaisées : dans le contexte de données sur-dispersées, le paramètre de variance spatiale est sur-estimé, car les effets aléatoires doivent modéliser, en plus des dépendances spatiales, la sur-dispersion. Le paramètre de portée spatiale est, quant à lui, sous-estimé. Ces résultats sont en accord avec les résultats obtenus par Gschlöß l and Czado (2005) dans le cadre de données spatiales de comptage sur-dispersées. Afin d'obtenir une estimation correcte des paramètres spatiaux, il serait judicieux d'ajouter au modèle, en plus des effets aléatoires spatiaux corrélés, des effets aléatoires i.i.d. au niveau observation, afin de modéliser de manière séparée la sur-dispersion non-spatiale de l'autocorrélation spatiale résiduelle. L'implémentation de la méthode MCMCLH à l'aide du package geoRglm (Christensen and Ribeiro Jr, 2002), ne le permet cependant pas pour le moment. Les simulations additionnelles attestent aussi la robustesse de MCMCLH à une mauvaise spécification de sa fonction de corrélation spatiale. D'autres simulations, incluant d'autres fonctions de corrélations à tester, pourraient être mises en oeuvre afin de corroborer ce résultat.

Interprétons à présent, comparativement aux résultats présents dans la littérature, les résultats obtenus en terme de shift. La littérature d'écologie statistique est pleine de confusions concernant les causes des shifts ; les interprétations erronées à leur sujet remettent injustement

en cause la fiabilité des méthodes spatiales.

Plus précisément, concernant tout d'abord les shifts entre méthodes spatiales et méthodes non-spatiales, Bini et al. (2009) ont analysé un grand nombre de jeux de données réelles et ont observé de nombreux phénomènes de shifts dans les estimations fournies par les différentes méthodes spatiales et non-spatiales, sans parvenir pour autant à expliciter la cause de ces shifts. Nos résultats, de même que ceux de Beale et al. (2007, 2010), montrent pourtant qu'il existe une cause très simple permettant d'expliquer la présence de ces shifts. Elle réside dans la divergence de variabilité, présentée précédemment entre les méthodes spatiales et non-spatiales, et qui peut uniquement être décelée à l'aide d'une approche par simulation. En effet, comme les méthodes non-spatiales fournissent des estimations moins précisés que les méthodes spatiales, en terme de RMSE, des shifts importants apparaissent naturellement entre les deux types de modèles, qui ne font que traduire la différence entre un modèle correct et un modèle incorrect. Bini et al. (2009), quant à eux, dans leur approche basée sur l'étude de données réelles, considèrent de manière erronée que les estimations des méthodes spatiales ne sont pas fiables, car elles s'éloignent anormalement de celles des méthodes non-spatiales, jugées a priori correctes.

Bini et al. (2009) observent aussi sur certains jeux de données des shifts significatifs entre méthodes spatiales elles-mêmes, sans parvenir non plus à en trouver de causes autres que des différences de modélisation spatiale sous-jacente. Il conclut qu'il faut alors se méfier des méthodes spatiales et proscrire leur utilisation, car tant que l'on a pas découvert l'origine de ces différences, un doute est jeté sur la fiabilité de leurs estimations. Nous ne sommes pas d'accord avec cette manière de décrédibiliser les méthodes spatiales. En effet, obtenir des estimations un peu différentes à partir de méthodes qui modélisant de manière intrinsèquement différentes les dépendances spatiales nous semble tout à faire naturel. C'est ce que l'on observe aussi sur les simulations sur placettes GNB, notamment entre le modèle bayésien basé sur les distances et le modèle bayésien basé sur les voisinages. Ce qu'il faut simplement comprendre, c'est qu'il y a, en pratique, des choix de modélisation spatiale a priori meilleurs que d'autres selon les contextes spatiaux. Sur les scénarios GNB, il n'y a pas a priori de justification à restreindre l'information spatiale par la création de voisinages ; le modèle basé sur les distances paraît a priori plus adapté, et c'est en effet celui qui est a posteriori le plus performant. Mais, s'il semble a priori pertinent de considérer des relations de voisinages dans certains contextes spatiaux et de considérer des corrélations spatiales stables sur les voisinages, il faut implémenter une méthode performante basée sur les voisinages et s'y tenir. En somme, il s'agit de trouver la modélisation spatiale la plus appropriée au contexte spatial dans lequel on se trouve, sans chercher à comparer les estimations obtenues avec celles de méthodes spatiales moins pertinentes.

Interprétons à présent les résultats obtenus dans l'étude de l'autocorrélation spatiale des résidus des modèles, en lien avec les performances intrinsèques des modèles. Certains auteurs, tels que Zhang et al. (2005) et Hawkins et al. (2007), établissent un lien direct et réciproque entre la présence d'autocorrélation dans les résidus des modèles et leurs performances : un modèle ne présentant pas d'autocorrélation spatiale dans les résidus est fiable, et un modèle présentant de l'autocorrélation spatiale dans les résidus est nécessairement mauvais. Il leur suffirrait alors, pour juger de la fiabilité d'une méthode de régression, de vérifier qu'elle parvienne bien à fournir des résidus non-autocorrélés.

Nos résultats corroborent ceux de Beale et al. (2010) sur cette question, qui remet en question cette manière de voir les choses. Dans son étude, Beale et al. (2010) mettent en oeuvre des méthodes spatiales peu performantes, telles que la *Wavelet Revised Method* (WRM) (Carl and Kuhn, 2008), qui parviennent néanmoins à fournir des résidus non-autocorrélés. Réciproquement,

ils implémentent des méthodes spatiales qui conservent un haut degré d'autocorrélation dans les résidus tout en fournissant des performances optimales en terme d'estimation. Nos résultats vont dans le sens de (Beale et al., 2010), et indiquent qu'il ne faut pas établir trop facilement un lien entre autocorrélation dans les résidus et performances. En effet, les méthodes GAM sur grille fournissent des résidus non autocoréllés, bien qu'elles exhibent de faibles performances en terme d'erreur de type I notamment. Cependant, lorsque les modèles sont optimaux, ils parviennent tous dans notre étude, contrairement aux modèles de Beale et al. (2010), à enlever l'autocorrélation dans les résidus ; dans le cas de données de comptage spatialement autocorrélés, il nous semble légitime d'associer la présence d'autocorrélation dans les résidus à des performances réduites. Dans l'autre sens, il nous semble que l'on ne puisse pas conclure à la fiabilité d'un modèle en observant une absence d'autocorrélation dans les résidus. On obseve seulement un lien marginal entre la précision réelle des estimations et l'autocorrélation des résiduss : en absence d'autocorrélation, la précision réelle des estimations est optimale, et donc les estimations en elles-mêmes peuvent être considérées comme fiables. Mais d'autres études sont nécessaires pour corroborer ces résultats.

Notons, pour finir, que les méthodes spatiales s'adaptent bien en terme d'inférence au cas particulier de données qui ne sont pas spatialement autocorrélées. Cela confirme un principe de base de la modélisation spatiale, éconcé par Cressie (1993), qui veut que les modèles spatiaux sont rigoureusement des généralisations des modèles non-spatiaux.

## 6.2 Discussion supplémentaire

Plusieurs voix se sont élevées dans la littérature d'écologie statistique pour dénoncer une mise en oeuvre systématique et peu réfléchie des méthodes spatiales. Ils invitent les chercheurs à ne pas se précipiter sur la modélisation spatiale explicite de l'autocorrélation spatiale résiduelle, mais plutôt à tenter d'en expliciter l'origine en recherchant, à l'aide d'une investigation d'ordre écologique, les variables influentes spatialement autocorrélées qui ont été omises dans la modélisation. Le but, disent-ils, est de comprendre les mécanismes spatiaux sous-jacents, et non modéliser l'autocorrélation spatiale sans réfléchir au préalable. Théoriquement parlant, il est clair qu'introduire la ou les variables influentes spatialement structurées dans la régression statistique, et supprimer ainsi l'autocorrélation spatiale résiduelle, permet de résoudre les problèmes d'inférence en enrichissant par là nos connaissances sur les mécanismes spatiaux sous-jacents. Cependant, cela n'est pas toujours possible en pratique et on ne peut jamais être sûr d'avoir introduit toutes les variables influentes dans la modélisation. Pour traiter les effets néfastes de l'autocorrélation spatiale sur l'inférence, on n'a alors pas d'autre choix que d'appliquer une approche compensatoire, en modélisant la variabilité spatiale non-expliquée à l'aide de méthodes spatiales adaptées, et cela exactement de la même manière que l'on utilise, dans un cadre non-spatial, une loi binomiale négative par exemple pour se prémunir d'un éventuel phénomène de sur-dispersion de données de comptage.

Par ailleurs, plusieurs auteurs dont Bini et al. (2009) et Hawkins (2012) se méfient des conclusions issues des approches par simulations, dont ils ne jugent pas légitime l'extrapolation à l'étude de jeux de données réelles. En effet, ils critiquent, et ce dans une juste mesure, les approches par simulation pour leur caractère trop simpliste et irréaliste. Mais, au lieu d'essayer de les améliorer, ils se contentent de dire qu'il faut privilégier les études comparatives basées sur des jeux de données réelles. Il est vrai que les approches par simulation, de même que les approches basées sur l'étude de jeux de données réelles, présentent des avantages et des inconvénients. Les approches par simulation permettent de mettre en oeuvre des critères de

performances intrinsèques, tels que l'erreur de type I et le RMSE, sur un nombre illimité de scénarios spatiaux et de réplicats, ce que ne permettent pas les approches sur données réelles, dont le seul avantage indéniable est d'être précisément réaliste. Des études par simulations bien menées, mettant en oeuvre des caractéristiques réalistes des jeux de données d'écologie, peuvent alors légitimement prétendre pouvoir extrapoler leurs résultats à l'étude de données réelles. En somme, ces deux approches devraient simplement être considérées comme complémentaires, ce qui n'est pas encore le cas dans la littérature d'écologie statistique.

# Chapitre 7

# Conclusion

Nous sommes parvenus, dans cette étude, à généraliser les conclusions de Beale et al. (2010) dans un contexte non-gaussien de donneés de comptage. Nous avons, en particulier, mis en évidence la faillite de l'estimation des effets fixes des modèles non-spatiaux et spatiaux indépendants, ainsi que des différences de comportement entre modèles spatiaux explicites, selon le type de modélisation spatiale sous-jacente ainsi que le type d'estimation numérique.

Ainsi, on a implémenté, en plus des modèles spatiaux et spatiaux indépendants classiques, différents modèles spatialement structurés dans la famille des modèles linéaires généralisés mixtes et des modèles additifs généralisés mixtes, qui sont estimés soit de manière fréquentiste à l'aide de l'approximation PQL, soit de manière bayésienne à l'aide d'algorithmes MCMC adaptatifs. Par ailleurs, les modèles spatiaux implémentés modélisent les dépendances spatiales ,soit à l'aide d'une notion de distance, soit par la spécification de relations de voisinages.

Les performances des modèles ont été évaluées et comparées par la mise en oeuvre d'une approche par simulation, composée de scénarios spatiaux réalistes basés sur des observations irrégulièrement réparties dans l'espace. Des propriétés typiques de données de comptage spatialement autocorrélées, telles que la non-stationnarité spatiale et les phénomènes de sur-dispersion et de sous-dispersion, violant les conditions d'application classiques des modèles implémentés, ont également été simulées.

A la lumière des résultats obtenus, les implémentations bayésiennes des modèles spatialement structurés, réalisées à l'aide d'algorithmes MCMC adaptatifs, ont fournit les meilleurs résultats, alors que l'approximation PQL s'est révélée être peu fiable dans le contexte de la modélisation spatiale explicite. On privilégie ensuite l'utilisation des modèles basés sur les distances dans des contextes spatiaux généraux, car ils sont plus souples d'utilisation que les modèles basés sur les voisinages. Enfin, les modèles de la famille des GAM n'ont présenté que peu d'améliorations par rapport aux GLM.

Le modèle MCMCLH, basé sur les distances et estimé à l'aide de méthodes MCMC avec mises à jours de Langevin-Hastings, a présenté des performances optimales en terme d'inférence des effets fixes sur l'ensemble des scénarios sur grille et sur placettes GNB. De plus, ce modèle est robuste à des caractéristiques de non-stationnarité et de dispersion extra-spatiale, très fréquentes en écologie dans le contexte de données de comptage spatialement autocorrélées.

Nous sommes enfin parvenus à dissiper quelques confusions présentes dans la littérature d'écologie statistique, qui participent à mettre en doute, de manière injustifiée, la fiabilité des

méthodes spatiales de manière générale. Le domaine de la modélisation spatiale explicite nécessite en effet des études approfondies ; il serait notamment intéressant de voir converger les résultats des études basées sur des simulations et celles basées sur des jeux de données réelles, avant que les méthodes spatiales ne deviennent des outils dont on puisse se servir en toute sûreté.

# Annexes

# Annexe A

# Démonstration du caractère biaisé de l'estimateur de la variance dans un cas spatial simple

Soit $(X_1, \ldots, X_n)$ une suite de variables aléatoires i.i.d. d'espérance $\mu$ et de variance $\sigma^2$. Considérons l'estimateur corrigé de la variance, donnée par l'expression :

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$$

Où $\bar{X}$ est la moyenne des observations $X_1, \ldots, X_n$.

On peut facilement démontrer que $S^2$ est un estimateur sans biais, c'est à dire que $\mathbb{E}(S^2) = \sigma^2$.

Plaçons-nous à présent dans un cas simple de dépendances spatiales unidimensionnelles, issues d'un processus autorégressif d'ordre 1. Soit $(X_1, \ldots, X_n)$ un vecteur de variables aléatoires spatialement autocorrélées de la manière suivante :

$$cov(X_i, X_j) = \sigma^2 \phi^{|i-j|}$$

Où $\sigma^2$ est la variance des observations, supposée connue, et $\phi$ un paramètre spatial compris entre 0 et 1 permettant d'assurer la décroissance des corrélations avec l'augmentation des proximités spatiales.

Montrons, dans ce cas, que la variance est sous-estimée par l'estimateur $S^2$, c'est à dire que :

$$\mathbb{E}(S^2) < \sigma^2$$

On a :

$$\begin{aligned}
(n-1)S^2 &= \sum_{i=1}^{n} (X_i - \bar{X})^2 \\
&= \sum_{i=1}^{n} (X_i - \mu + \mu - \bar{X})^2 \\
&= \sum_{i=1}^{n} (X_i - \mu)^2 + 2(\mu - \bar{X}) \cdot \sum_{i=1}^{n} (X_i - \mu) + n(\mu - \bar{X})^2
\end{aligned}$$

$$= \sum_{i=1}^{n}(X_i - \mu)^2 - n(\mu - \bar{X})^2$$

En passant à l'espérance, on obtient :

$$\begin{aligned}(n-1)\mathbb{E}(S^2) &= \sum_{i=1}^{n}\mathbb{E}((X_i-\mu)^2) - n\mathbb{E}((\mu-\bar{X})^2) \\ &= \sum_{i=1}^{n}\sigma^2 - nVar(\bar{X}) \\ &= n\sigma^2 - nVar(\bar{X})\end{aligned}$$

Dans le contexte de données indépendantes, on aurait l'expression suivante pour la variance de la moyenne :

$$Var(\bar{X}) = \frac{1}{n^2}\sum_{i=1}^{n}Var(X_i) = \frac{\sigma^2}{n}$$

Cette expression n'est plus vraie dans le cas de données spatialement autocorrélées, car il faut tenir compte des corrélations spatiales de la manière suivante :

$$\begin{aligned}Var(\bar{X}) &= \frac{1}{n^2}Var\left(\sum_{i=1}^{n}X_i\right) \\ &= \frac{1}{n^2}cov\left(\sum_{i=1}^{n}X_i, \sum_{j=1}^{n}X_j\right) \\ &= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\sigma^2\phi^{|i-j|}\end{aligned}$$

On obtient alors :

$$(n-1)\mathbb{E}(S^2) = \sigma^2\left(n - \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}\phi^{|i-j|}\right)$$

Or, un calcul de série permet de montrer que :

$$\sum_{i=1}^{n}\sum_{j=1}^{n}\phi^{|i-j|} = n + \frac{2\phi}{1-\phi}(n-1) - 2\left(\frac{\phi}{1-\phi}\right)^2(1-\phi^{n-1})$$

D'où, on obtient :

$$(n-1)\mathbb{E}(S^2) = \sigma^2\left[n - 1 - \frac{2\phi}{1-\phi}\frac{n-1}{n} + 2\left(\frac{\phi}{1-\phi}\right)^2\frac{1-\phi^{n-1}}{n}\right]$$

Et donc :

$$\begin{aligned}\mathbb{E}(S^2) - \sigma^2 &= 2\left(\frac{\phi}{1-\phi}\right)^2\frac{1-\phi^{n-1}}{n(n-1)} - \frac{2\phi}{n(1-\phi)} \\ &= \frac{2\phi}{n(1-\phi)}\left[\frac{\phi}{1-\phi}\frac{1-\phi^{n-1}}{n-1} - 1\right]\end{aligned}$$

$$= \frac{2\phi}{n(1-\phi)} \left[ \frac{1}{n-1} \sum_{i=1}^{n-1} \phi^i - 1 \right]$$

Or, puisque $\phi$ est compris entre 0 et 1, on obtient :

$$\frac{1}{n-1} \sum_{i=1}^{n-1} \phi^i < \frac{n-1}{n-1} = 1$$

Ce qui fournit le résultat attendu :

$$\mathbb{E}(S^2) - \sigma^2 < 0$$

# Annexe B

# Code permettant d'implémenter les méthodes spatialement structurées dans le logiciel R

GLMMPQL

```
library(MASS)
attach(my.data)
GLMMPQL=try(glmmPQL(fixed=Y1~X1+X2+X3+X4,random=~1|obs,family=poisson(link="log"),
data=my.data,control=list(lmeControl(returnObject=TRUE,opt="optim"))))
if (is(GLMMPQL,"try-error")) { GLMMPQL.coef[[1]][b,]=rep(NA,2*P+HP) }
else { struct.cor=corSpatial(form=~x+y,nugget=T,type="exponential")
struct.cor=Initialize(struct.cor,as(my.data,"data.frame")[,c("x","y")])
GLMMPQL=try(update(GLMMPQL,correlation=struct.cor))
if (is(GLMMPQL,"try-error")) { GLMMPQL.coef[[1]][b,]=rep(NA,2*P+HP) }
else { GLMMPQL.coef[[1]][b,]=c(summary(GLMMPQL)$coef$fixed,
sqrt(diag(summary(GLMMPQL)$varFix)),as.numeric(VarCorr(GLMMPQL)[1,2]),NA) } }
```

GAMM

```
library(mgcv)
GAMM=try(gamm(Y1~X1+X2+X3+X4+te(x,y,bs=c("tp","tp"))+s(obs,bs="re"),
correlation=struct.cor,family=poisson(link="log"),data=my.data,method="ML",
control=list(lmeControl(returnObject=TRUE,opt="optim"))))
if (is(GAMM,"try-error")) { GAMM.coef[[1]][b,]=rep(NA,2*P+HP) }
else { GAMM.coef[[1]][b,]=c(coef(GAMM$gam)[1:P],
sqrt(diag(vcov(GAMM$gam))[1:P]),NA,NA) }
```

INLA

```
library(INLA)
INLA1=try(inla(Y1~X1+X2+X3+X4+f(obs,model="besagproper",graph.file="grid.dat",
hyper=list(prec=list(param=c(1,0.01),initial=1))),family="poisson",data=my.data,
verbose=F,control.predictor=list(compute=T),control.inla=list(h=1e-4)))
if (is(INLA1,"try-error") | length(INLA1$summary.fixed)==0)
{ INLA1.coef[[1]][b,]=rep(NA,2*P+HP) }
else { INLA1.coef[[1]][b,]=c(INLA1$summary.fixed[,1],INLA1$summary.fixed[,2],
```

```
INLA1$summary.hyperpar[1,1],INLA1$summary.hyperpar[2,1])
for (j in 1:P) { INLA1.quant[[1]][[b]][,j]=c(NA,INLA1$summary.fixed[j,3],
NA,NA,INLA1$summary.fixed[j,5],NA) } }
```

MCMCLH

```
library(geoRglm)
my.S.scale=0.0003
my.phi.scale=1
my.geodata=vector("list")
my.geodata$coords=cbind(x,y)
my.geodata$data=my.data$Y1
my.geodata$cov.model="exponential"
my.geodata$units.m=rep(1,n)
my.model=list(trend.d=~X1+X2+X3+X4,trend.l=~X1+X2+X3+X4,cov.model="exponential")
my.prior=prior.glm.control(beta.prior="normal",beta=rep(0,P),
beta.var.std=diag(1,P),sigmasq.prior="uniform",phi.prior="uniform",
phi.discrete=seq(0,10,l=500)),my.mcmcinput=mcmc.control(S.scale=my.S.scale,
thin=10,phi.scale=my.phi.scale,burn.in=30000,n.iter=70000,phi.start=0)
MCMCLH=try(pois.krige.bayes(my.geodata,prior=my.prior,mcmc.input=my.mcmcinput,
model=my.model))
if (is(MCMCLH,"try-error")) { MCMCLH.coef[[1]][b,]=rep(NA,2*P+HP) }
else { MCMCLH.coef[[1]][b,]=c(MCMCLH$posterior$beta$mean,
sqrt(diag(MCMCLH$posterior$beta$var)),sqrt(MCMCLH$posterior$sigmasq$mean),
MCMCLH$posterior$phi$mean)
for (j in 1:P) { MCMCLH.quant[[1]][[b]][,j]=quantile(MCMCLH$posterior$beta$sample[j,],
c(0.01/2,0.05/2,0.1/2,1-0.01/2,1-0.05/2,1-0.1/2)) } }
```

# Annexe C

# Les résidus quantiles randomisés

Avant de présenter les résidus quantiles randomisés, rappelons les expressions des résidus classiques de Pearson et de la déviance dans le contexte d'un GLM de Poisson, en considérant $\mathbf{y} = (y_1, \ldots, y_n)$ le vecteur des données de comptage, et $\hat{\mathbf{y}} = (\hat{y}_1, \ldots, \hat{y}_n)$ le vecteur des prédictions.

Les résidus de Pearson $\hat{r}_i^p$ et les résidus de la déviance $\hat{r}_i^d$ sont, dans ce cas, définis $\forall i \in \{1, \ldots, n\}$ par :

$$\begin{cases} \hat{r}_i^p = \frac{y_i - \hat{y}_i}{\hat{y}_i} \\ \hat{r}_i^d = signe\,(y_i - \hat{y}_i)\sqrt{2y_i \ln\left(\frac{y_i}{\hat{y}_i}\right)} \end{cases}$$

Ces deux types de résidus sont connus pour s'éloigner notablement de l'hypothèse de normalité dans un contexte non-gaussien (Dunn and Smyth, 1996), et ce indépendamment d'une éventuelle structuration spatiale des données, malgré les résultats théoriques de normalité asymptotique. Afin de vérifier cela, on ajuste un GLM de Poisson avec fonction de lien canonique log sur des données de comptage obtenues à partir de la simulation de 1000 observations normalement distribuées, en fixant à 1 le coefficient de régression. La figure C.1 illustre les qq-plots des résidus de Pearson et des résidus de la déviance, qui montrent clairement que les résidus ne sont pas distribués selon une loi gaussienne.

Afin d'obtenir la propriété de normalité souhaitée, Dunn and Smyth (1996) construisent un autre type de résidus, appelés les résidus quantiles randomisés $\hat{r}_i^q$ (randomized quantile residuals), définis dans le cadre du GLM de Poisson $\forall i \in \{1, \ldots, n\}$ par :

$$\hat{r}_i^q = \Phi^{-1}(u_i)$$

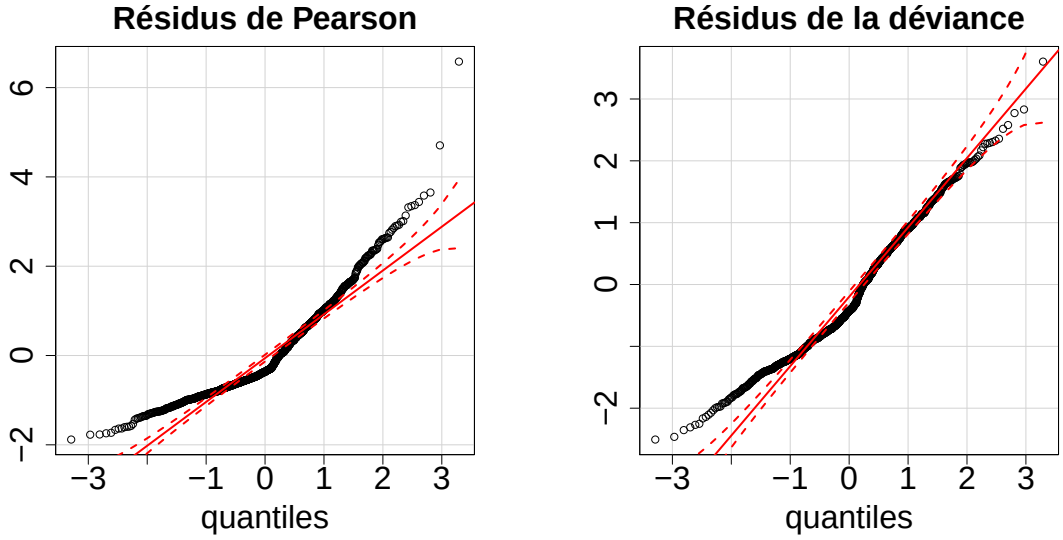
Avec :
$\Phi$ : la fonction de répartition de la loi normale centrée-réduite
$u_i$ : la variable aléatoire de loi uniforme sur l'intervalle semi-ouvert $]a_i, b_i]$

Les bornes $a_i$ et $b_i$ sont définies, en notant $F(y, \mu)$ la valeur de la fonction de répartition au point $y$ de la loi de Poisson de paramètre $\mu$, par $a_i = \lim_{y \nearrow y_i} F(y, \hat{y}_i)$ et $b_i = F(y_i, \hat{y}_i)$.

Dans le cas d'une fonction de répartition $F$ continue, dépendant du paramètre $\lambda$, la définition des résidus quantiles randomisés se simplifie de la manière suivante :

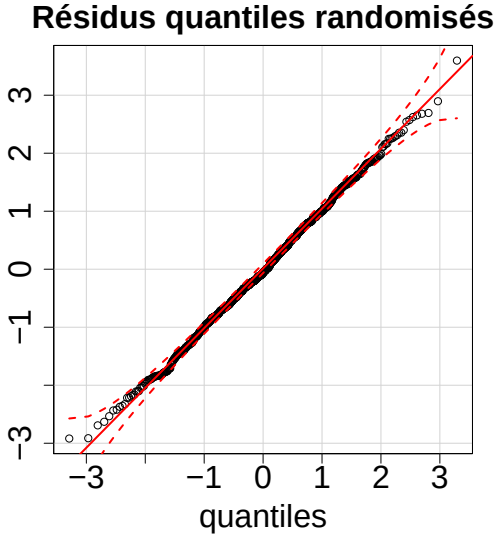
$$\hat{r}_i^q = \Phi^{-1}(F(y_i, \hat{\lambda}))$$

FIGURE C.1 – QQ-plots des résidus de Pearson et des résidus de la déviance dans le cadre d'une régression de Poisson




Dunn and Smyth (1996) montrent que, par construction, en mettant de côté la variabilité liée à l'estimation du paramètre $\lambda$, les résidus quantiles randomisés sont distribués exactement selon une loi normale. Afin d'illustrer cela, les résidus quantiles randomisés ont été calculés dans le cas simulé précédent, et leur qq-plot a été reporté figure C.2. On observe, en effet, une adéquation satisfaisante des résidus à l'hypothèse de normalité.

FIGURE C.2 – QQ-plot des résidus quantiles randomisés dans le cadre d'une régression de Poisson




On fournit enfin la fonction R permettant de calculer en pratique les résidus quantiles randomisés dans le cas d'une régression de Poisson :

```
normalized.residuals.poisson=function(Y,Y.chapeau) {
qnorm( ppois(Y-1,Y.chapeau) + dpois(Y,Y.chapeau) * runif(length(Y)) }
```

# Annexe D

# La distribution binomiale négative et la distribution de Poisson généralisée

La distribution binomiale négative

La distribution binomiale négative de paramètres $r > 0$ et $\mu > 0$, notée $NB(r, \mu)$, est une distribution discrète à support dans $\mathbb{N}$, définie par la densité de probabilité suivante :

$$f(y|r, \mu) = \frac{\Gamma(y + r)}{\Gamma(r)y!} \left(\frac{r}{\mu + r}\right)^r \left(\frac{\mu}{\mu + r}\right)^y$$

Soit $Y$ une variable aléatoire de loi $NB(r, \mu)$. L'espérance et la variance de $Y$ sont alors égales à :

$$\begin{cases} \mathbb{E}(Y|r, \mu) = \mu \\ Var(Y|r, \mu) = \mu\,(1 + \frac{\mu}{r}) \end{cases}$$

Si l'on note $\phi = 1 + \frac{\mu}{r}$ le paramètre de sur-dispersion de la loi $NB(r, \mu)$, on obtient alors une relation simple entre espérance et variance :

$$Var(Y|r, \mu) = \phi \cdot \mathbb{E}(Y|r, \mu)$$

Comme $\phi$ est supérieur à 1, la loi binomiale négative permet ainsi de générer des données sur-dispersées. Notons que, puisque le paramètre $\phi$ dépend de $\mu$, on obtient une valeur de $\phi$ différente pour chaque observation ; il s'agit d'être attentif à cette spécificité lorsqu'on simule des données sur-dispersées à l'aide d'une loi binomiale négative.

Précision enfin que la loi binomiale négative $NB(r, \mu)$ converge (en loi), quand r tend vers l'infini, vers une loi de Poisson de paramètre $\mu$. En particulier, la variance de la loi $NB(r, \mu)$, à savoir $\mu\,(1 + \frac{\mu}{r})$, tend vers la variance d'une loi de Poisson de paramètre $\mu$, à savoir $\mu$, quand r tend vers l'infini.

La distribution de Poisson généralisée

La distribution de Poisson généralisée (Consul and Jain, 1973) de paramètres $\mu > 0$ et $\lambda \in \mathbb{R}^+$, notée $GP(\mu, \lambda)$, est une distribution discrète à support dans $\mathbb{N}$, définie par la densité suivante :

$$f(y|\mu, \lambda) = \frac{\mu(\mu + \lambda y)^{y-1}}{y!} \exp^{-\mu-\lambda y}$$

Il s'agit d'une généralisation de la loi de Poisson ; lorsque $\lambda = 0$, on retrouve la densité de la loi de Poisson, à savoir :

$$f(y|\mu) = \frac{\mu^y}{y!} \exp^{-\mu}$$

Soit $Y$ une variable aléatoire de loi $GP(\mu, \lambda)$. L'espérance et la variance de $Y$ sont alors données par :

$$\begin{cases} \mathbb{E}(Y|\mu, \lambda) = \frac{\mu}{1-\lambda} \\ Var(Y|\mu, \lambda) = \frac{\mu}{(1-\lambda)^3} \end{cases}$$

Si l'on introduit le paramètre de sur-dispersion, indépendant de $\mu$, $\phi = \frac{1}{(1-\lambda)^2}$, le lien entre la variance et l'espérance s'écrit alors sous la même forme que dans le cas binomial négatif, à savoir :

$$Var(Y|\mu, \lambda) = \phi \cdot \mathbb{E}(Y|\mu, \lambda)$$

Comme $\phi = \frac{1}{(1-\lambda)^2}$ peut prendre à la fois des valeurs supérieures à 1 si $\lambda \in [0, 2]$, et inférieures à 1 si $\lambda \in [2, +\infty]$, la distribution de Poisson généralisée, contrairement à la distribution binomiale négative, permet de générer à la fois des données sur-dispersées ($\phi > 1$) et des données sous-dispersées ($\phi < 1$).

# Appendice

L'accès, via une plateforme internet, à une station de calcul basée au centre régional IRSTEA de Clermont-Ferrand a considérablement facilité la mise en oeuvre des simulations, en diminuant nettement le temps de calcul. Les simulations finales, réalisées avec 1000 jeux de données, ont néanmoins nécessité environ un mois et demi de calcul.

Durant ce stage, j'ai été en contact avec Kévin Le Rest, qui prépare une thèse en écologie statistique au centre d'études biologiques de Chizé. Son sujet d'étude s'intitule : "Modélisation des facteurs influençant la distribution et l'abondance des rapaces en France : approche méthodologique". Dans le contexte de développement d'outils méthodologiques pour le suivi de populations de rapaces dans l'espace et le temps, il étudie la modélisation de l'autocorrélation spatiale dans un cadre de données de comptage souvent sur-dispersées, en s'intéressant spécifiquement aux modèles spatiaux basés sur les voisinages.

Ce stage va vraisemblablement donner lieu à la soumission, en fin d'année, d'un article en anglais dans une revue scientifique d'écologie statistique. Environ la moitié de l'article a déjà été rédigée durant le stage.

Enfin, le présent rapport a été rédigé en LaTeX, qui est un langage déstiné à l'écriture de compositions scientifiques, auquel je me suis formé de manière parallèle durant le stage.

# Bibliographie


C.M. Beale, J.J. Lennon, and D.A. Elston. Red herrings remain in geographical ecology : a reply to hawkins et al. (2007). *Ecography*, 30 :845–847, 2007.

C.M. Beale, J.J. Lennon, and J.M. Yearsley. Regression analysis of spatial data. *Ecology Letters*, 13 :246–264, 2010.

J. Besag. Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society : Series B*, 36 :192–236, 1974.

J. Besag and P. Green. Spatial statistics and bayesian computation (with discussion). *Journal of the Royal Statistical Society : Series B*, 55 :25–37, 1993.

J. Besag, J. York, and A. Mollié. Bayesian image restoration with two applications in spatial statistics (with discussion). *Annals of the Institute of Statistical Mathematics*, 43 :1–59, 1991.

L.M. Bini, J.A.F. Diniz, and T.F.L.V. Rangel. Coefficients shifts in geographical ecology : an empirical evaluation of spatial and non-spatial regression. *Ecography*, 32 :193–204, 2009.

N.E. Breslown and D.G. Clayton. Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, 88 :9–25, 1993.

G. Carl and I. Kuhn. Analyzing spatial ecological data using linear regression and wavelet analysis. *Stochastic Environmental Research and Risk Assessment*, 22 :315–324, 2008.

O.F. Christensen and P.J. Ribeiro Jr. georglm : A package for generalised linear spatial models. *R-NEWS*, 2 :26–28, 2002.

O.F. Christensen and R. Waagepetersen. Bayesian prediction of spatial count data using generalized linear mixed models. *Biometrics*, 58 :208–286, 2002.

O.F. Christensen, J. Moller, and R. Waagpetersen. Geometric ergodicity of metropolis hastings algorithms for conditional simulation in generalized linear mixed models. *Methodology and Computing in Applied Probability*, 3 :309–327, 2001.

A.D. Cliff and J.K. Ord. *Spatial Processes : Models and Applications.* Pion, 1981.

P.C. Consul and G.C. Jain. A generalization of the poisson distribution. *Technometrics*, 15-4 : 791–799, 1973.

N.A.C. Cressie. *Statistics for spatial data.* Wiley, 1993.

M.R.T. Dale and M.-J. Fortin. Modifying the t test for assessing the correlation between two spatial processes. *Journal of Agricultural, Biological, and Environmental statistics*, 14 :188–206, 2009.

P.J. Diggle, J.A. Tawn, and R.A. Moyeed. Model based geostatistics. *Annals of Applied Statistics*, 47 :299–350, 1998.

C.F. Dormann, J.M. McPherson, and Araujo M.B. Methods to account for spatial autocorrelation in the analysis of species distribution data. *Ecography*, 30 :609–628, 2007.

P.K. Dunn and G.K. Smyth. Randomized quantile residuals. *Journal of Computational and Graphical Statistics*, 5 :236–244, 1996.

P. Dutilleul. Modifying the t test for assessing the correlation between two spatial processes. *Biometrics*, 49 :305–314, 1993.

M.-J. Fortin and M.R.T. Dale. *Spatial analysis - a guide for ecologists.* Cambridge University Press, 2005.

A.S. Fotheringham, C Brunsdon, and M.E. Charlton. *Geographically Weighted Regression : the analysis of spatially varying relationships.* Wiley, 2002.

R.C. Geary. The contiguity ratio and statistical mapping. *The Incorporated Statistician*, 5(3) : 115–145, 1954.

A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian Data Analysis.* Chapman and Hall/CRC, 2004.

A. Gemperli and P. Vounatsou. Fitting generalized linear mixed models for point-referenced spatial data. *Journal of Modern Applied Statistical Methods*, 2 :481–495, 2003.

S. Gschlöß l and C. Czado. Modelling count data with overdispersion and spatial effects. *Sonderforschungsbereich 386*, 412, 2005.

T. Hastie and R. Tibshirani. Generalized additive models. *Statistical Science*, 1(3) :297–318, 1986.

B.A. Hawkins. Eight (and a half) deadly sins of spatial analysis. *Journal of Biogeography*, 39 : 1–9, 2012.

B.A. Hawkins, J.A.F. Diniz-Filho, and L.M. Bini. Red herrings revisited : spatial autocorrelation and parameter estimation in geographical ecology. *Ecography*, 30 :375–384, 2007.

S.H. Hurlbert. Pseudoreplication and the design of ecological experiments. *Ecological Monographs*, 54 :187–211, 1984.

P. Legendre. Spatial autocorrelation - trouble or new paradigm ? *Ecology*, 74 :1659–1673, 1993.

P. Legendre and M.-J. Fortin. Spatial pattern and ecological analysis. *Vegetatio*, 80 :107–138, 1989.

J.J. Lennon. Red-shifts and red herrings in geographical ecology. *Ecography*, 23 :101–113, 2000.

C.E. McCulloch. Maximum likelihood algorithms for generalized linear mixed models. *Journal of the American Statistical Association*, 92 :162–170, 1997.

P.A.P. Moran. Notes on continuous stochastic phenomena. *Biometrika*, 37 :17–23, 1950.

A. Pettitt, I. Weir, and A. Hart. A conditional autoregressive gaussian process for irregularly spaced multivariate data with application to modelling large sets of binary data. *Statistics and Computing*, 12 :353–367, 2002.

R Development Core Team. *R : A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2011. URL `http://www.R-project.org`. ISBN 3-900051-07-0.

Haining R.P. *Spatial Data Analysis : Theory and Practice*. Cambridge University Press, 2003.

H. Rue and L. Held. *Gaussian Markov Random Fields : Theory And Applications*. Chapman and Hall/CRC, 2005.

H. Rue, S. Martino, and N. Chopin. Approximate bayesian inference for latent gaussian models using integrated nested laplace approximations (with discussion). *Journal of the Royal Statistical Society : Series B*, 71 :309–392, 2009.

Oliveau S. Autocorrélation spatiale : leçons du changement d'échelle. *L'Espace géographique*, 39 :51–64, 2010.

W.R. Tobler. A computer movie simulating urban growth in the detroit region. *Journal of Economic Geography*, 46 :234–240, 1970.

W.N. Venables and B.D. Ripley. *Modern applied statistics with S*. Springer, 2002.

S.N. Wood. *Generalized Additive Models : An Introduction with R*. Chapman and Hall/CRC, 2006.

L. J. Zhang, J. H. Grove, and L. S. Heath. Spatial residual analysis of six modeling techniques. *Ecological Modelling*, 186 :154–177, 2005.