Université de Montréal

Modélisation bayésienne des changements aux niches écologiques causés par le réchauffement climatique

par
Akpoué Blache Paul

Département de Mathématiques et de Statistique
Faculté des arts et des sciences

Thèse présentée à la Faculté des études supérieures
en vue de l'obtention du grade de Philosophiæ Doctor (Ph.D.)
en statistique

Mai, 2012

Université de Montréal
Faculté des études supérieures

Cette thèse intitulée:

Modélisation bayésienne des changements aux niches écologiques causés par le réchauffement climatique

présentée par:

Akpoué Blache Paul

a été évaluée par un jury composé des personnes suivantes:

Bilodeau Martin,	président-rapporteur
Angers Jean-François,	directeur de recherche
Bédard Mylène,	membre du jury
Choulakian Vartan,	examinateur externe
Lapointe François-Joseph,	représentant du doyen de la FES

Thèse acceptée le: . . . . . .

RÉSUMÉ

Cette thèse présente des méthodes de traitement de données de comptage en particulier et des données discrètes en général. Il s'inscrit dans le cadre d'un projet stratégique du CRNSG, nommé CC-Bio, dont l'objectif est d'évaluer l'impact des changements climatiques sur la répartition des espèces animales et végétales. Après une brève introduction aux notions de biogéographie et aux modèles linéaires mixtes généralisés aux chapitres 1 et 2 respectivement, ma thèse s'articulera autour de trois idées majeures.

Premièrement, nous introduisons au chapitre 3 une nouvelle forme de distribution dont les composantes ont pour distributions marginales des lois de Poisson ou des lois de Skellam. Cette nouvelle spécification permet d'incorporer de l'information pertinente sur la nature des corrélations entre toutes les composantes. De plus, nous présentons certaines propriétés de ladite distribution. Contrairement à la distribution multidimensionnelle de Poisson qu'elle généralise, celle-ci permet de traiter les variables avec des corrélations positives et/ou négatives. Une simulation permet d'illustrer les méthodes d'estimation dans le cas bidimensionnel. Les résultats obtenus par les méthodes bayésiennes par les chaînes de Markov par Monte Carlo (CMMC) indiquent un biais relatif assez faible de moins de 5% pour les coefficients de régression des moyennes contrairement à ceux du terme de covariance qui semblent un peu plus volatils.

Deuxièmement, le chapitre 4 présente une extension de la régression multidimensionnelle de Poisson avec des effets aléatoires ayant une densité gamma. En effet, conscients du fait que les données d'abondance des espèces présentent une forte dispersion, ce qui rendrait fallacieux les estimateurs et écarts types obtenus, nous privilégions une approche basée sur l'intégration par Monte Carlo grâce à l'échantillonnage préférentiel. L'approche demeure la même qu'au chapitre précédent, c'est-à-dire que l'idée est de simuler des variables latentes indépendantes et de se retrouver dans le cadre d'un modèle linéaire mixte

généralisé (GLMM) conventionnel avec des effets aléatoires de densité gamma. Même si l'hypothèse d'une connaissance *a priori* des paramètres de dispersion semble trop forte, une analyse de sensibilité basée sur la qualité de l'ajustement permet de démontrer la robustesse de notre méthode.

Troisièmement, dans le dernier chapitre, nous nous intéressons à la définition et à la construction d'une mesure de concordance donc de corrélation pour les données augmentées en zéro par la modélisation de copules gaussiennes. Contrairement au τ de Kendall dont les valeurs se situent dans un intervalle dont les bornes varient selon la fréquence d'observations d'égalité entre les paires, cette mesure a pour avantage de prendre ses valeurs sur $(-1;1)$. Initialement introduite pour modéliser les corrélations entre des variables continues, son extension au cas discret implique certaines restrictions. En effet, la nouvelle mesure pourrait être interprétée comme la corrélation entre les variables aléatoires continues dont la discrétisation constitue nos observations discrètes non négatives. Deux méthodes d'estimation des modèles augmentés en zéro seront présentées dans les contextes fréquentiste et bayésien basées respectivement sur le maximum de vraisemblance et l'intégration de Gauss-Hermite. Enfin, une étude de simulation permet de montrer la robustesse et les limites de notre approche.

Mots clés : modèle bayésien, données discrètes, données multidimensionnelles, Poisson, Skellam

ABSTRACT

This thesis presents some estimation methods and algorithms to analyse count data in particular and discrete data in general. It is also part of an NSERC strategic project, named CC-Bio, which aims to assess the impact of climate change on the distribution of plant and animal species in Québec. After a brief introduction to the concepts and definitions of biogeography and those relative to the generalized linear mixed models in chapters 1 and 2 respectively, my thesis will focus on three major and new ideas.

First, we introduce in chapter 3 a new form of distribution whose components have marginal distribution Poisson or Skellam. This new specification allows to incorporate relevant information about the nature of the correlations between all the components. In addition, we present some properties of this probability distribution function. Unlike the multivariate Poisson distribution initially introduced, this generalization enables to handle both positive and negative correlations. A simulation study illustrates the estimation in the two-dimensional case. The results obtained by Bayesian methods via Monte Carlo Markov chain (MCMC) suggest a fairly low relative bias of less than 5% for the regression coefficients of the mean. However, those of the covariance term seem a bit more volatile.

Later, the chapter 4 presents an extension of the multivariate Poisson regression with random effects having a gamma density. Indeed, aware that the abundance data of species have a high dispersion, which would make misleading estimators and standard deviations, we introduce an approach based on integration by Monte Carlo sampling. The approach remains the same as in the previous chapter. Indeed, the objective is to simulate independent latent variables to transform the multivariate problem estimation in many generalized linear mixed models (GLMM) with conventional gamma random effects density. While the assumption of knowledge *a priori* dispersion parameters seems too strong and not realistic, a sensitivity analysis based on a measure of goodness of fit is used to demonstrate

the robustness of the method.

Finally, in the last chapter, we focus on the definition and construction of a measure of concordance or a correlation measure for some zeros augmented count data with Gaussian copula models. In contrast to Kendall's τ whose values lie in an interval whose bounds depend on the frequency of ties observations, this measure has the advantage of taking its values on the interval $(-1, 1)$. Originally introduced to model the correlations between continuous variables, its extension to the discrete case implies certain restrictions and its values are no longer in the entire interval $(-1, 1)$ but only on a subset. Indeed, the new measure could be interpreted as the correlation between continuous random variables before being transformed to discrete variables considered as our discrete non negative observations. Two methods of estimation based on integration via Gaussian quadrature and maximum likelihood are presented. Some simulation studies show the robustness and the limits of our approach.

Keywords: bayesian model, discrete data, multivariate data, Poisson, Skellam

TABLE DES MATIÈRES

LISTE DES TABLEAUX

LISTE DES FIGURES

A ma famille

Fleur

Dorgeles

Emmanuel

et

Nicole

dont l'amour et le support m'ont accompagné tous les jours.

REMERCIEMENTS

J'aimerais exprimer ma profonde gratitude à mon directeur de thèse M. Jean-François Angers, pour sa disponibilité, son soutien et l'apprentissage continu qu'il a su me dispenser durant ces années. Je remercie aussi M. Christian Léger pour avoir facilité mon adaptation et mon insertion au Département de mathématiques et statistique. Je tiens à remercier tous les membres de l'équipe CC-Bio pour l'expérience dont j'ai pu bénéficier tout au long du premier projet stratégique du CRNSG sur la flore et la faune au Québec. Je leur suis également reconnaissant pour le soutien financier qu'ils ont bien voulu m'accorder tout au long de mes études doctorales. Merci également aux professeurs et membres du département qui ont su m'offrir un cadre tout à fait propice aux études et aussi pour leur disponibilité. Je ne saurais terminer sans remercier les instituts tels que l'ISM, le Département de mathématiques et statistique, la Faculté des études supérieures et post-doctorales pour leur soutien financier. Enfin, je remercie le Seigneur Dieu pour son soutien sans faille et inconditionnel même dans les épreuves les plus difficiles.

INTRODUCTION

Les espèces peuvent répondre aux changements climatiques de plusieurs façons. Elles peuvent migrer pour rester dans les zones climatiques qui leur sont les plus favorables, évoluer sur place, ou tout simplement disparaître. Bien que l'évolution puisse se produire rapidement, les déplacements d'aires de répartition sont les réponses les plus courantes. Des études récentes montrent que la *phénologie* et la répartition des plantes et des animaux ont changé dans les dernières $30-40$ années dans le sens induit par le réchauffement du climat. Par ailleurs, les activités humaines telles que l'exploitation abusive des ressources fossiles, les gaz à effet de serre, la pollution, la radioactivité et tant d'autres mettent sérieusement à mal l'avenir de la biodiversité.

Dans le sud du Québec, les températures de surface moyennes ont augmenté de 1,25 degrés Celsuis au cours des 4 dernières décennies et les modèles climatiques prévoient une augmentation supplémentaire de 3 à 5 degrés Celsuis au cours de ce siècle. Dans le cadre du projet CC-Bio, nous voulons explorer les effets des changements climatiques sur la biodiversité du Québec.

Cette thèse s'inscrit dans le cadre du projet stratégique du CRNSG nommé CC-Bio dont l'objectif est d'étudier l'impact des changements climatiques sur la répartition spatiale des espèces animales et végétales. L'objectif de CC-Bio est de prédire les effets potentiels des changements climatiques sur la répartition et l'abondance d'une grande panoplie d'espèces animales et végétales du Québec. CC-Bio a pour but d'alimenter les stratégies régionales d'adaptation au changement climatique en ce qui concerne la conservation de la biodiversité. Il est le premier projet en matière de biodiversité et changements climatiques supporté par le consortium Ouranos. Le projet a obtenu une subvention du programme

stratégique du CRSNG et est administré à l'Université du Québec à Rimouski. Notre tâche peut se décrire en quatre étapes :

Étape 1 Recueillir et explorer graphiquement les différentes variables observées.

Étape 2 Développer des modèles de prévision selon les différentes espèces par l'approche bayésienne. Cette approche nous permettra d'incorporer le maximum d'information disponible afin de parfaire l'ajustement de données.

Étape 3 Définir et créer un cadre et support informatique sur Matlab pour l'analyse des données.

Étape 4 Prédire et ensuite résumer l'information sur des cartes géographiques afin de faciliter la prise de décision.

Ce travail constitue donc une contribution au projet stratégique CC-Bio en proposant une démarche scientifique nouvelle qui a pour objectif de présenter des méthodes générales de traitement de données discrètes multidimensionnelles applicables dans l'étude de la répartition des espèces.

Pour ce faire, nous présentons dans un premier temps quelques notions et définitions de la biogéographie afin de se familiariser avec la biodiversité et ses caractéristiques. Cette étape bien que naturelle, paraît essentielle pour la modélisation des avis d'experts et de biologistes en lois *a priori*. Ensuite, nous faisons un bref relevé de la littérature sur les GLMM (Generalized Linear Mixed Models) selon les approches fréquentiste et bayésienne. Après avoir introduit le sujet, cette thèse s'articule autour de trois principaux axes.

Tout d'abord, nous présentons une nouvelle forme de distribution multidimensionnelle dont les composantes ont pour distributions marginales des lois de Poisson ou des lois de Skellam. Cette nouvelle spécification permet d'incorporer de l'information pertinente sur la nature des corrélations entre toutes les composantes. Contrairement à la distribution

multidimensionnelle de Poisson, celle-ci permet de traiter les variables avec des corrélations positives et négatives. Une simulation et une application dans le cas bidimensionnel permettront d'illustrer les méthodes d'estimation.

Ensuite, nous abordons le phénomène de surdispersion dans le cas multidimensionnel. Nous présentons un modèle de Poisson multidimensionnel avec effets aléatoires afin de traduire cette particularité fréquemment observée dans les données environnementales et biologiques. Une nouvelle méthode d'estimation basée sur l'intégration par Monte Carlo y est présentée, ainsi que les résultats d'une étude de simulations. Enfin, une application en écologie mettant en relief l'impact des changements climatiques sur la répartition des espèces animales et végétales au Québec selon différents scénarios servira d'illustration.

Enfin, en s'intéressant aux copules bidimensionnelles gaussiennes, nous proposons une mesure de concordance mieux adaptée aux variables aléatoires augmentées en zéro. Cette nouvelle mesure pourrait être interprétée comme la corrélation entre les variables aléatoires continues dont la discrétisation constitue les observations discrètes non négatives. Une méthode d'estimation des modèles augmentés en zéro de Poisson sera présentée à la fois dans les contextes fréquentiste et bayésien.

Certes, nous espérons obtenir des estimateurs optimaux, mais loin de nous l'idée de rejeter de façon systématique les autres méthodes. Notre principal objectif est plutôt de montrer la flexibilité de notre approche, nous permettant du coup de gérer une multitude de problèmes, tout en réduisant la difficulté.

CHAPITRE 1

CONCEPTS, DÉFINITIONS ET REPRÉSENTATION DE DONNÉES EN BIOGÉOGRAPHIE

Ce chapitre constitue une brève introduction aux notions de base en biogéographie. Il permet non seulement de présenter les modèles de niches climatiques mais aussi les incidences majeures des changements climatiques sur la répartition ou la localisation de ces niches.

1.1 Concepts d'écosystèmes

Définition 1.1.1. *Un écosystème est un ensemble dynamique d'organismes vivants (plantes, animaux et micro-organismes) qui interagissent entre eux et avec le milieu (sol, climat, eau, lumière) dans lequel ils vivent. Les caractéristiques des écosystèmes peuvent aussi se développer dans le temps et dans l'espace selon l'influence interne (espèces invasives) ou selon l'influence humaine.*

1.1.1 Évolution des écosystèmes

Définition 1.1.2. *Une succession écologique est un processus d'évolution libre d'un milieu naturel au cours du temps. Cela consiste en une série d'étapes devant se succéder dans un ordre adéquat : différentes communautés végétales et animales, sols, etc se remplacent.*

De par leur nature changeante et dynamique, les écosystèmes évoluent de façon perpétuelle jusqu'à atteindre leur état stable. En effet, ils peuvent soit évoluer de façon lente vers un autre type d'écosystème par des successions écologiques, soit changer de façon radicale sous l'effet de perturbations sporadiques et brusques. Un exemple de changement

lent est celui des espèces végétales dont l'adaptation aux variations climatiques se fait sur des longues décennies. Notons enfin que la zone existant entre deux écosystèmes, par exemple la lisière qui est la transition entre l'écosystème forêt et l'écosystème prairie, souvent appelée écotone, possède les caractéristiques conjuguées de ces deux écosystèmes et constitue donc une zone de transition très lente.

1.1.2 Classification des espèces

Nous n'avons pas l'intention de présenter une méthodologie de classification des espèces car elle s'avèrerait trop complexe compte tenu de la diversité des relations et interactions existant entre les espèces. Cependant, nous nous limiterons aux groupes fonctionnels (classification non pas selon les similitudes mais selon le fonctionnement et l'interaction avec l'écosystème), une récente classification des espèces qui sied bien à nos objectifs. Les plus anciennes classifications écologiques concernent les classifications taxonomiques en nombre d'espèces ou taxons. Bien qu'intéressantes, ces dernières ne font qu'une description statique des communautés. Elles rendent compte du degré de diversité du système écologique étudié sans réellement prendre en compte sa dynamique. Plus récemment, de nouvelles classifications dites fonctionnelles se sont développées en écologie depuis la fin des années 1990 (Gitay et Noble, 1997 et Gitay et *al.*, 1999). L'intérêt récent d'une formalisation de nouvelles classifications réside dans une meilleure compréhension du rôle de la biodiversité dans le fonctionnement des écosystèmes et en particulier pour prédire les effets de changements de facteurs environnementaux. En effet, afin de prédire les réactions d'écosystèmes riches en espèces à des perturbations, il peut être utile de regrouper certaines espèces dans des groupes fonctionnels, dont la réponse à ces perturbations est uniforme. Ces nouvelles classifications visent à identifier des groupes fonctionnels d'espèces ou groupes d'espèces présentant un comportement similaire pour un processus éco-

logique ou une fonction donnée de l'écosystème. Les groupes fonctionnels peuvent être établis soit par l'avis d'experts soit par des méthodes statistiques telles que les arbres de classification et l'analyse factorielle.

1.1.3 Niches écologiques

Le concept de niche écologique fait référence à une notion d'espace et de localisation d'une espèce, mais aussi et surtout à sa fonction (son rôle, aussi bien dans la chaîne alimentaire que dans les autres interactions) et à la manière de la remplir. La niche rassemble la totalité des relations qu'une espèce entretient avec son habitat et les autres espèces de la communauté.

Définition 1.1.3. *Une niche réalisée est un territoire occupé par une espèce, laquelle exerce son rôle d'une façon stable et clairement établie. Cette espèce établit des interactions avec son environnement et fait partie intégrante de son écosystème.*
Une niche potentielle est une niche qui n'est pas encore réalisée par une espèce, bien que le rôle de cette dernière pourrait trouver sa place au sein de l'écosystème, si une espèce venait l'exercer.

Afin de faciliter et de spécialiser les recherches au sein d'une niche écologique, plusieurs niches partielles furent définies :

Niche trophique ensemble des dimensions de la niche écologique liées à l'alimentation (le choix des proies pour les carnivores, les modes de chasse, les besoins alimentaires, l'effort fourni et l'énergie récupérée sont autant de phénomènes associés à la niche trophique).

Niche spatiale ensemble des dimensions de la niche écologique liées à l'occupation de l'espace (taille des territoires, phénomènes de migrations, répartition spatiales).

Niche temporelle ensemble des dimensions de la niche écologique liées à la gestion du temps (activités nocturnes ou diurnes, dates de mises bas, hibernations, etc.).

Niche comportementale ensemble des dimensions de la niche écologique liées à l'éthologie (étude du comportement des diverses espèces animales) : les espèces ont des comportements qui leur permettent de tirer partie des ressources disponibles (un grand carnivore n'utilise pas les mêmes techniques de capture suivant qu'il chasse une proie plus ou moins massive).

Dans la suite de notre étude, le concept de niche écologique fait référence à son caractère purement spatial. En effet, cette notion parait plus simple à observer et à évaluer et elle ne nécessite pas un suivi continu des espèces. De plus, leur caractère spatial permet de les caractériser par leur centre et leur dispersion ou périphérie.

1.2 Les effets des changements climatiques sur les espèces animales et végétales

Certaines projections (Thuiller *et al.*, 2006) confirment les effets des changements climatiques sur la flore, la faune et les écosystèmes avec des effets plus prononcés sur les plantes à cause de leur sédentarité. Ces changements affectant à la fois la distribution et la diversité des espèces sont dus au fait que leurs adaptations peuvent différer selon le groupe fonctionnel dont elles sont issues.

Les modèles d'enveloppes bioclimatiques sont souvent utilisés pour prédire les réactions ou réponses des espèces face aux changements climatiques. En effet, il serait naturel de penser que les réactions des espèces face à ces changements varient selon leurs distributions spatiales. Il s'agit donc de comparer les changements liés à l'habitat et à l'expan-

sion des espèces par des variables géographiques et des caractéristiques des niches. Selon Schoener (1989), les niches écologiques peuvent être décrites par leur position moyenne et leur dispersion.

Thuiller (2003) indique, sous l'hypothèse d'absence de migration, que les espèces avec une dispersion plus élevée ont tendance à disparaître au profit des espèces moins volatiles. Par contre, sous l'hypothèse de migration illimitée (ouverte), les premières ont tendance à migrer vers des conditions plus favorables ; il se produit ainsi des changements de paysage. L'effet du changement, mesuré par l'indice de diversité de Simpson (1949), est plus sévère en l'absence de flux migratoires.

Enfin, nous pourrons définir les variables traduisant le réchauffement global et aussi présenter l'impact de ces changements sur la biodiversité. Ce qui pourrait donc se traduire par une possible relation existant entre ces deux dernières entités.

1.3 Modèle d'interpolation de données spatiales

Nous disposons en général de données sur des sites bien précis (monitoring). Cependant, si nous étudions un phénomène à des endroits spécifiques qui ne sont pas nécessairement répertoriés dans notre base, il faudrait alors procéder à des interpolations. L'interpolation spatiale consiste donc à prédire une fonction $Z(s_0)$ pour l'emplacement s_0 à partir des données $(Z(s_1),\ldots,Z(s_n))$ observées sur les sites $(s_1,\ldots,s_n)$.

1.3.1 Méthodes déterministes

Méthodes barycentriques

Connues aussi sous le nom de moyennes mobiles ou approximation par le noyau, ces méthodes sont assez simples et intuitives car elles prédisent la valeur en un point s_0 comme

une moyenne des valeurs observées. Ainsi,

$$\widehat{Z}(s_0) = \sum_{i=1}^{n} \omega_i Z(s_i) \quad \text{avec} \quad \sum_{i=1}^{n} \omega_i = 1 \ \text{ et } \ w_i \geq 0.$$

Les poids ω_i sont fonction de la distance euclidienne entre le site d'observation s_i et le site à prédire s_0, notée $|s_i - s_0|$, tels que les sites les plus proches aient plus d'influence dans l'interpolation. En d'autres termes, seules les observations appartenant à un certain voisinage de s_0, noté $V(s_0)$, sont prises en compte. En pratique, les poids les plus utilisés sont ceux correspondant à la méthode de l'inverse de la distance élevée à une certaine puissance d. En effet, dans ce cas l'interpolation s'écrit :

$$\widehat{Z}(s_0) = \sum_{i \in V(s_0)} \frac{|s_i - s_0|^{-d}}{\sum_{i \in V(s_0)} |s_i - s_0|^{-d}} Z(s_i) \ \ , \ \ d > 0.$$

Interpolation par partitionnement de l'espace

Cette méthode est en fait un cas particulier des méthodes barycentriques. La plus simple de ces méthodes est celle du plus proche voisin. La valeur mesurée en un site d'observation est attribuée à tous les points localisés dans le polygone de ce site. En effet, les méthodes d'interpolation par partitionnement de l'espace possèdent les propriétés d'être locales et exactes. Elles n'utilisent dans l'interpolation que les observations localisées assez près du point de prévision selon un certain critère de voisinage selon la méthode utilisée. La figure 1.1 présente un exemple.

Interpolation par les splines

Contrairement aux méthodes barycentriques, ce type d'interpolation ne s'effectue pas point par point. L'objectif est plutôt d'ajuster une surface sur tout un champ. Une spline

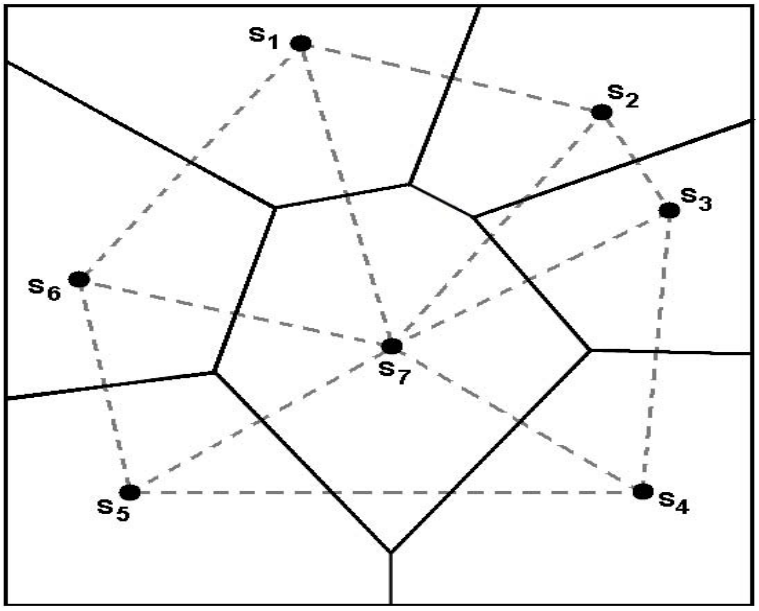

Figure 1.1 – Interpolation selon le partitionnement de l'espace

est en fait une famille de fonctions régulières de courbure minimale. Il existe deux types de splines : les splines d'interpolation contraints de passer par les points d'observation et les splines de lissage qui passent seulement à proximité de ces points (voir Hastie et Tibshirani, 1990).

1.3.2 Le krigeage

Le krigeage est une méthode stochastique d'interpolation spatiale basée sur le critère de minimisation de l'erreur moyenne quadratique, et qui dépend des propriétés de second ordre (stationnarité) du processus observé. Cressie (1993) décrit le modèle comme suit :

$$Z(s) = S(s) + \varepsilon(s),$$

où $S(.)$ désigne la structure déterministe de la fonction de localisation $Z(.)$ et $\varepsilon(.)$ est un bruit blanc dû à l'erreur de mesure.

Propriétés et caractéristiques associées au krigeage

Proposition 1.3.1. *Propriétés*

L'estimateur obtenu par krigeage possède les propriétés suivantes :

- BLUE *: linéaire, sans biais, à variance minimale.*
- *Prend en compte la* structure de dépendance spatiale des données.
- Interpolateur exact *: si un point connu est estimé, cette estimation serait égale à la valeur connue.*
- Présente un effet d'écran *: les points les plus près reçoivent les poids les plus importants. Cet effet d'écran varie selon la configuration et selon le modèle de* variogramme *utilisé pour le krigeage. Plus* l'effet de pépite *est important, moins il y a d'effet d'écran.*
- *Tient compte de la taille du champ à estimer et de la position des points entre eux.*
- *Par l'utilisation du variogramme, tient compte de la continuité du phénomène étudié (effet de pépite, anisotropie, etc).*
- Transitif *: Si la valeur krigée pour un point coïncide avec la valeur observée en ce point, alors les valeurs krigées en d'autres points ne sont pas modifiées par l'inclusion de ce nouveau point dans les krigeages. Par contre les variances de krigeage sont diminuées. De même, si un krigeage est effectué à partir d'un certain nombre de points et que ces valeurs krigées sont utilisées comme de nouvelles données, alors les krigeages subséquents ne s'en trouvent pas modifiés (sauf pour la variance de krigeage).*

1.3.3 Variables explicatives

En tenant compte de tous les types de variables, nous pourrions obtenir un meilleur ajustement des données et aussi des prévisions plus réalistes. En effet, ces variables nous permettraient de réduire le nombre de prévisions aberrantes. Par exemple, la présence des arbres pourrait être un facteur prépondérant pour la survie des oiseaux. Ainsi, en plus de prendre en compte les conditions climatiques favorables d'habitat des espèces, nous prenons en compte les conditions environnementales. Il s'agit principalement de trois types de variables :

1.3.3.1 Variables directes

Elles sont soit disponibles sur les sites d'observations, soit fournies par le consortium sur la climatologie régionale et l'adaptation aux changements climatiques (OURANOS). En effet, les variables directes font référence à celles disponibles et utilisables sans transformations. Ce sont les données climatiques telles que la température, les précipitations, etc. En effet, nous nous intéressons aux comportements extrêmes et saisonniers de ces différentes variables en considérant leurs moyennes, extrema et leurs volatilités selon les saisons.

1.3.3.2 Variables indirectes

Contrairement aux variables directes, elles n'ont pas un effet direct sur la variable dépendante. Le plus souvent, elles peuvent résumer l'information contenue dans les variables directes. Par exemple, certes avoir la température journalière peut s'avérer utile mais nous pourrions utiliser le nombre de jours de chauffe. Ces variables présentent l'avantage de résumer l'information et ainsi de réduire le nombre déjà élevé des variables explicatives.

1.3.3.3 Variables latentes

Elles ne peuvent être observées mais sont supposées avoir une influence sur les variables dépendantes. En effet, elles pourraient traduire par exemple l'effet de l'homme, la déforestation, l'urbanisation. Par exemple, la variable latente désignant l'intervention humaine peut être indirectement mesurée comme une combinaison linéaire de variables observables : la déforestation, le pourcentage de terre utilisable, la présence de sites industriels, etc.

1.4 Description des données CC-Bio

1.4.1 Méthode d'échantillonnage

La méthode d'échantillonnage consiste à effectuer des observations sur l'abondance ou la présence des espèces au sein d'un quadrat de $20km \times 20km$. Les données ainsi recueillies à l'intérieur de chacun des 761 quadrats de $400km^2$ sont ensuite agrégées pour représenter les données observées sur une période allant de 1981 à 1999. La zone d'échantillonnage concerne globalement le Québec et plus particulièrement le sud où les conditions d'échantillonnage sont plus favorables. Les données échantillonnées se trouvent dans un tableau à trois entrées :

$$\underline{\mathbf{X}} = \left(x_{tqe}\right) \quad \text{pour} \quad \begin{array}{l} t = 1981, \ldots, 1999 \\ q = 1, \ldots, 761 \\ e = 1, \ldots, E, \end{array}$$

où E désigne le nombre d'espèces observées. Ainsi, x_{tqe} représente l'abondance de l'espèce e dans le quadrat q durant l'année t. Ensuite, nous construisons le tableau binaire en

sommant toutes les observations sur la période $1981-1999$ pour obtenir :

$$\mathbf{X} = \left(\sum_{t=1981}^{1999} x_{tqe} = x_{qe} \right).$$

La figure 1.2 indique en gras sur la carte du Québec les quadrats dans lesquels les observations ont été effectuées. Il apparaît donc une forte couverture de toute la zone sud du Québec tandis que les territoires du nord ne bénéficient que d'une couverture assez irrégulière et très faible. En effet, compte tenu des conditions difficiles d'accès au nord, la méthode d'échantillonnage privilégie les quadrats du sud du Québec.

1.4.2 Changements climatiques

Le réchauffement climatique aurait pour effet de déplacer les isothermes vers le nord, c'est-à-dire que les températures du nord seront de plus en plus élevées. Cependant cette prévision est faite selon plusieurs scénarios dont les facteurs les plus importants sont l'émission des gaz à effet de serre, la déforestation, l'urbanisation, la variabilité des variables climatiques. De plus, il existe deux horizons pour nos prédictions ; la première période s'étend de 2020 à 2050 et la seconde de 2050 à 2080. Compte tenu de la multitude de scénarios existants, nous nous limiterons dans notre étude aux trois scénarios suivants :

1. **scénario optimiste : faibles changements** (GCM_{14}) Ce scénario est le plus optimiste car il ne prévoit pas de changements majeurs au niveau des principales variables climatiques. En effet, sur la période de 2010 à 2080, l'augmentation de la température annuelle moyenne ne devrait pas excéder $2C$ et celle relative aux précipitations serait de $5mm$.

2. **scénario réaliste : changements moyens** (GCM_{68}) Il représente le scénario le plus réaliste car il prévoit des changements ni trop forts ni trop faibles. Les températures

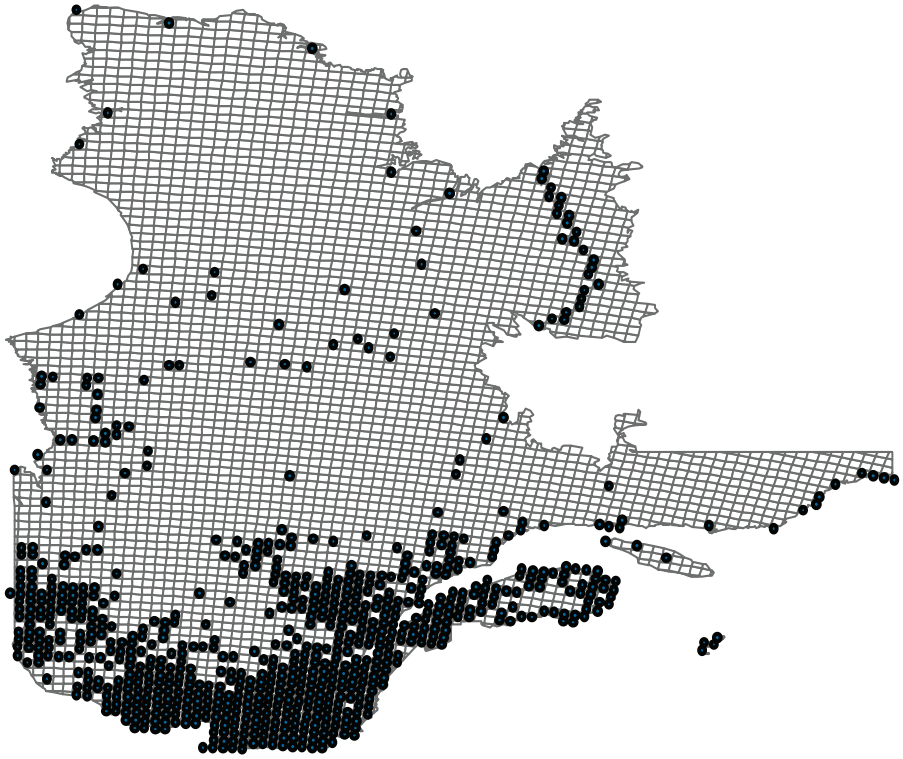

Figure 1.2 – Carte de couverture des quadrats d'observations

devraient évoluer à la hausse de $4C$ et les précipitations pourraient atteindre une augmentation de $17mm$.

3. **scénario pessimiste : forts changements** (GCM_{35}) C'est l'un des scénarios les plus extrêmes car il prédit une augmentation très importante des variables climatiques. Les températures annuelles en 2080 devraient subir une hausse de $7C$ et les précipitations annuelles augmenteraient de $29mm$.

Dans la suite, nous avons retenu seulement trois variables explicatives du climat notamment la température moyenne annuelle, la précipitation moyenne annuelle et la différence entre les températures maximale et minimale du mois le plus chaud de l'année (juillet). Les valeurs de ces variables sont obtenues par krigeage barycentrique (Le et Zidek, 2006). En effet, les valeurs sont celles du centre du quadrat obtenues grâce au krigeage de plusieurs autres valeurs provenant des stations d'observation à l'intérieur d'un même quadrat.

CHAPITRE 2

MODÈLES UNIDIMENSIONNELS EN BIOGÉOGRAPHIE

Ce chapitre a pour objectif de présenter une brève revue de littérature et de méthodologie pour le traitement des données en biogéographie. Il y existe généralement deux types de données. La première catégorie est constituée des données de présence-absence d'une espèce bien définie dont l'objectif est d'estimer la probabilité d'observer cette espèce. La seconde représente les données d'abondance qui déterminent l'abondance ou les différents foyers de concentration de l'espèce. Dans le cadre de notre étude, nous nous limiterons uniquement aux données d'abondance. Ainsi les modèles les plus fréquemment utilisés et dont nous discuterons dans ce chapitre sont les modèles de Poisson, les modèles augmentés en zéro et les modèles spatiaux.

2.1 Modèle linéaire généralisé de Poisson : Estimation Bayésienne des paramètres

Il demeure le modèle le plus utilisé compte tenu de sa simplicité.

2.1.1 Le modèle

Soit $\mathbf{Y} = (y_1, \ldots, y_n)'$ un échantillon de variables aléatoires de loi de Poisson indépendantes de paramètres respectifs $(\lambda_1, \ldots, \lambda_n)$. Le modèle s'écrit

$$\begin{aligned} Y_i &\sim Poiss(\lambda_i), \qquad i = 1, \ldots, n, \\ \log(\lambda_i) &= \mathbf{Z}_i \theta, \qquad i = 1, \ldots, n, \end{aligned}$$

où $\theta = (\theta_1, \theta_2, \ldots, \theta_p) \in \mathbb{R}^p$ représente le vecteur des paramètres à estimer et $\mathbf{Z} = (\mathbf{Z}_1, \mathbf{Z}_2, \ldots, \mathbf{Z}_n)$ désigne la matrice des variables explicatives de taille $n \times p$, et $\mathbf{Z}_i$ désigne la i^e ligne du vecteur $\mathbf{Z}$.

2.1.2 Lois *a priori* et hypothèses

La première motivation du choix de la loi *a priori* est de simplifier la forme de la densité *a posteriori* donc l'inférence. Par la suite, supposons que les p variables explicatives sont continues ; dans notre cas, ce sont les 3 variables relatives au climat, voir la section 1.3.3.1. Les lois *a priori* des paramètres s'écrivent

$$\theta \sim \mathscr{N}_p(\mu_\theta, \Sigma_\theta), \qquad \Sigma_\theta \sim \mathscr{IW}_p\left(\tilde{\Sigma}_0, m_0\right) \qquad \text{et} \qquad \mu_\theta \sim \mathscr{N}_p(\mathbf{A_0}, \mathbf{B_0}),$$

où les valeurs de m_0, $\mathbf{A_0}$, $\mathbf{B_0}$ et $\tilde{\Sigma}_0$ sont supposées connues provenant d'études antérieures ou basées sur des avis d'experts. En posant D l'ensemble des données, $\Omega_{-\theta}$ l'ensemble de tous les paramètres du modèle excepté θ, la vraisemblance du modèle peut s'écrire sous la forme suivante :

$$\begin{aligned} L(\theta|D) &= \prod_{i=1}^{n} \left\{ \frac{\lambda_i^{y_i} \mathrm{e}^{-\lambda_i}}{y_i!} \right\} \\ &\propto \left[\prod_{i=1}^{n} \left(\mathrm{e}^{\mathbf{Z}_i\theta} \right)^{y_i} \mathrm{e}^{-\mathrm{e}^{Z_i\theta}} \right] \\ &= \exp\left\{ \sum_{i=1}^{n} (\mathbf{Z}_i\theta)\, y_i \right\} \exp\left\{ -\sum_{i=1}^{n} \mathrm{e}^{Z_i\theta} \right\} \\ &= \exp\left\{ \mathbf{Y}'(\mathbf{Z}\theta) \right\} \exp\left\{ -\mathbb{I}_p' \mathrm{e}^{\mathbf{Z}\theta} \right\}, \end{aligned}$$

où $\mathbf{Y} = (y_1, \ldots, y_n)'$ désigne le vecteur des observations. L'opérateur exponentiel matriciel $e^{\mathbf{Z}\theta}$ n'est rien d'autre que l'exponentiel de chacun des termes de la matrice $\mathbf{Z}\theta$ et $\mathbb{I}_p$

désigne le vecteur colonne de dimension p constitué de termes égaux à 1.

2.1.3 Lois *a posteriori*

Lois *a posteriori* de θ

La forme de la loi *a posteriori* de θ n'est pas commune ou standard. Ainsi, il faudrait utiliser un algorithme d'acceptation rejet afin d'en tirer des échantillons. Ces algorithmes seront présentés dans le chapitre suivant. La loi *a posteriori* de θ s'écrit :

$$\begin{aligned}
P(\theta|\Omega_{-\theta},D) &\propto \pi(\theta)L(\theta|D) \\
&\propto \exp\left\{-\frac{1}{2}(\theta-\mu_\theta)^{'}\Sigma_\theta^{-1}(\theta-\mu_\theta)\right\}\exp\left\{\mathbf{Y}'(\mathbf{Z}\theta)\right\}\exp\left\{-\mathbb{I}_p^{'}e^{\mathbf{Z}\theta}\right\} \\
&= \exp\left\{-\frac{1}{2}(\theta-\mu_\theta)^{'}\Sigma_\theta^{-1}(\theta-\mu_\theta)+\mathbf{Y}'(\mathbf{Z}\theta)\right\}\exp\left\{-\mathbb{I}_p^{'}e^{\mathbf{Z}\theta}\right\} \\
&\propto \exp\left\{-\frac{1}{2}\left[\theta^{'}\Sigma_\theta^{-1}\theta-2\mu_\theta^{'}\Sigma_\theta^{-1}\theta-2\mathbf{Y}'\mathbf{Z}\theta\right]\right\}\exp\left\{-\mathbb{I}_p^{'}e^{\mathbf{Z}\theta}\right\} \\
&= \exp\left\{-\frac{1}{2}\left[\theta^{'}\Sigma_\theta^{-1}\theta-2\left(\mathbf{Y}'\mathbf{Z}+\mu_\theta^{'}\Sigma_\theta^{-1}\right)\theta\right]\right\}\exp\left\{-\mathbb{I}_p^{'}e^{\mathbf{Z}\theta}\right\} \\
&\propto \exp\left\{-\frac{1}{2}\left[\theta-\Sigma_\theta\left(\mathbf{Y}'\mathbf{Z}+\Sigma_\theta^{-1}\mu_\theta\right)\right]^{'}\Sigma_\theta^{-1}\left[\theta-\Sigma_\theta\left(\mathbf{Y}'\mathbf{Z}+\Sigma_\theta^{-1}\mu_\theta\right)\right]\right\} \\
&\quad \times\exp\left\{-\mathbb{I}_p^{'}e^{\mathbf{Z}\theta}\right\}.
\end{aligned}$$

Hyperparamètres de position

La forme conjuguée de la loi *a priori* postulée sur μ_θ permet d'obtenir une loi *a posteriori*

gaussienne. En effet, en appliquant le théorème de Bayes, nous obtenons

$$\begin{aligned}
\mathbf{P}\left(\mu_\theta|\Omega_{-\mu_\theta},D\right) &\propto \pi(\mu_\theta)\pi(\theta)L(\theta|D) \\
&\propto \pi(\mu_\theta)\pi(\theta) \\
&\propto \exp\left\{-\frac{1}{2}\left(\mu_\theta-\mathbf{A}_0\right)'\mathbf{B}_0^{-1}\left(\mu_\theta-\mathbf{A}_0\right)\right\}\exp\left\{-\frac{1}{2}\left(\theta-\mu_\theta\right)'\Sigma_\theta^{-1}\left(\theta-\mu_\theta\right)\right\} \\
&= \exp\left\{-\frac{1}{2}\left[\left(\mu_\theta-\mathbf{A}_0\right)'\mathbf{B}_0^{-1}\left(\mu_\theta-\mathbf{A}_0\right)+\left(\theta-\mu_\theta\right)'\Sigma_\theta^{-1}\left(\theta-\mu_\theta\right)\right]\right\} \\
&\propto \exp\left\{-\frac{1}{2}\left[\mu_\theta'\mathbf{B}_0^{-1}\mu_\theta-2\mathbf{A}_0'\mathbf{B}_0^{-1}\mu_\theta+\mu_\theta'\Sigma_\theta^{-1}\mu_\theta-2\theta'\Sigma_\theta^{-1}\mu_\theta\right]\right\} \\
&= \exp\left\{-\frac{1}{2}\left(\mu_\theta'\left[\mathbf{B}_0^{-1}+\Sigma_\theta^{-1}\right]\mu_\theta-2\left[\mathbf{A}_0'\mathbf{B}_0^{-1}+\theta'\Sigma_\theta^{-1}\right]\mu_\theta\right)\right\}.
\end{aligned}$$

Il suffit de remarquer attentivement que cette densité est en fait celle d'une loi normale multivariée définie telle que :

$$\mu_\theta|\Omega_{-\mu_\theta},D\sim\mathscr{N}_p\left(\left[\mathbf{B}_0^{-1}+\Sigma_\theta^{-1}\right]^{-1}\left[\mathbf{A}_0'\mathbf{B}_0^{-1}+\theta'\Sigma_\theta^{-1}\right]',\left[\mathbf{B}_0^{-1}+\Sigma_\theta^{-1}\right]^{-1}\right).$$

Hyperparamètres de dispersion

La loi *a posteriori* de Σ_θ est standard grâce au choix de la loi *a priori*.

$$\begin{aligned}
P\left(\Sigma_\theta|\Omega_{-\Sigma_\theta},D\right) &\propto \pi\left(\Sigma_\theta\right)\pi\left(\theta\right)L\left(\Theta|D\right)\\
&\propto \pi(\Sigma_\theta)\pi\left(\theta\right)\\
&\propto \sqrt{\frac{1}{|\Sigma_\theta|}}\exp\left\{-\frac{1}{2}\left(\theta-\mu_\theta\right)'\Sigma_\theta^{-1}\left(\theta-\mu_\theta\right)\right\}\frac{\left|\tilde{\Sigma}_0\right|^{m_0/2}\left|\Sigma_\theta\right|^{-(m_0+p+1)/2}}{2^{pm_0/2}\Gamma_p\left(m_0/2\right)}\\
&\quad\times\exp\left\{-\frac{1}{2}\mathrm{tr}\left(\tilde{\Sigma}_0\Sigma_\theta^{-1}\right)\right\}\\
&\propto |\Sigma_\theta|^{-(m_0+p+2)/2}\exp\left\{-\frac{1}{2}\left[\left(\theta-\mu_\theta\right)'\Sigma_\theta^{-1}\left(\theta-\mu_\theta\right)+\mathrm{tr}\left(\tilde{\Sigma}_0\Sigma_\theta^{-1}\right)\right]\right\}\\
&= |\Sigma_\theta|^{-(m_0+p+2)/2}\exp\left\{-\frac{1}{2}\mathrm{tr}\left[\left(\theta-\mu_\theta\right)'\Sigma_\theta^{-1}\left(\theta-\mu_\theta\right)+\tilde{\Sigma}_0\Sigma_\theta^{-1}\right]\right\}\\
&= |\Sigma_\theta|^{-(m_0+p+2)/2}\exp\left\{-\frac{1}{2}\mathrm{tr}\left(\left[\left(\theta-\mu_\theta\right)\left(\theta-\mu_\theta\right)'+\tilde{\Sigma}_0\right]\Sigma_\theta^{-1}\right)\right\}.
\end{aligned}$$

Enfin, nous obtenons donc :

$$\Sigma_\theta|\Omega_{-\Sigma_\theta},D\sim\mathscr{IW}_p\left(\tilde{\Sigma}_0+\left(\theta-\mu_\theta\right)\left(\theta-\mu_\theta\right)',m_0+1\right).$$

2.2 Modèles augmentés en zéro

Compte tenu du nombre élevé d'observations de modalité zéro, l'hypothèse du modèle de Poisson qui consiste à postuler l'égalité entre la moyenne et la variance n'est plus respectée. En effet, cette sur-représentation de la modalité zéro dans les observations entraîne une surdispersion. Une stratégie consiste donc à postuler un mélange de distributions de Dirac pour zéro et une loi de Poisson donnant lieu au modèle augmenté en zéro. Nous présentons donc une méthode d'estimation bayésienne basée sur l'augmentation de données de Gosh *et al.* (2006) pour les données augmentées en zéro sous forme de *power series distribution* (ZIPS) dont le cas des données augmentées de Poisson constitue un cas

particulier.

2.2.1 Définitions

Soit Y une variable aléatoire de loi $ZIPS(p,\theta)$ de densité :

$$P(Y=0) = p + (1-p)\frac{b(0)}{c(\theta)},$$
$$P(Y=k) = (1-p)\frac{b(k)\theta^k}{c(\theta)}, \;\; k = 1,2,\ldots,$$

où $c(\theta) = \sum_{k=0} b(k)\theta^k$, $0 \leq p < 1$, $b(k) > 0, \forall k$ et $\theta > 0$. De plus, si toute variable aléatoire augmentée en zéro peut s'écrire $Y = \tilde{V}(1-\tilde{B})$ où $\tilde{B}$ est une variable aléatoire de Bernoulli de paramètre p $\left(\tilde{B} \sim B(p)\right)$ indépendante de $\tilde{V}$ qui représente une variable aléatoire de distribution PS (série de puissance), alors la variable aléatoire Y ainsi définie suit une loi $ZIPS(p,\theta)$. Par ailleurs,

$$E(Y) = (1-p)E\left(\tilde{V}\right),$$
$$\text{var}(Y) = \frac{p}{1-p}\left[E(Y)\right]^2 + \delta E(Y),$$

où $\delta = \frac{\text{var}\left(\tilde{V}\right)}{E\left(\tilde{V}\right)}$ désigne le coefficient de dispersion de la variable latente $\tilde{V}$. En effet, si $\tilde{V}$ présente une surdispersion alors $\delta \geq 1$ et Y présente aussi de la surdispersion. Par contre, lorsque $\delta < 1$ alors les variables $\tilde{V}$ et Y ne présentent pas de surdispersion. La distribution de Poisson augmentée en zéro est obtenue en posant simplement $b(k) = \frac{1}{k!}$ et $c(\theta) = e^{\theta}$ pour $k = 1,2,\ldots$.

2.2.2 Modèle

Soit $(y_1, y_2, \ldots, y_n)'$ un échantillon de loi $ZIPS(p, \theta)$. La fonction de vraisemblance du modèle s'écrit :

$$\begin{aligned} L(p,\theta) &= \prod_{i=1}^{n} P(Y_i = y_i) \\ &= \left\{ \prod_{i \in \mathscr{S}} P(Y_i = 0) \right\} \left\{ \prod_{i \notin \mathscr{S}} P(Y_i = y_i) \right\} \\ &= \left[p + (1-p)\frac{b(0)}{c(\theta)} \right]^{s_0} \left\{ \prod_{i \notin \mathscr{S}} (1-p)\frac{b(y_i)\theta^{y_i}}{c(\theta)} \right\} \\ &= \left[p + (1-p)\frac{b(0)}{c(\theta)} \right]^{s_0} (1-p)^{n-s_0}\, \theta^{\sum_{i=1}^{n} y_i} \frac{c(\theta)^{s_0}}{c(\theta)^n} \prod_{i \notin \mathscr{S}} b(y_i). \end{aligned}$$

où $\mathscr{S} = \{i \in \{1, \ldots, n\} : y_i = 0\}$ et $s_0 = \text{card}(\mathscr{S})$. Par la suite, en utilisant le développement du binôme de Newton, nous obtenons

$$\begin{aligned} L(p,\theta) &\propto \left\{ \sum_{j=0}^{s_0} \binom{s_0}{j} [pc(\theta)]^j [(1-p)b(0)]^{s_0-j} \right\} (1-p)^{n-s_0} \frac{\theta^S}{c(\theta)^n} \\ &\propto \left\{ \sum_{j=0}^{s_0} \binom{s_0}{j} p^j c(\theta)^j (1-p)^{s_0-j} b(0)^{-j} \right\} b(0)^{s_0} \frac{\theta^S}{c(\theta)^n} (1-p)^{n-s_0} \\ &\propto \left\{ \sum_{j=0}^{s_0} \binom{s_0}{j} p^j (1-p)^{n-j} \left(\frac{b(0)}{c(\theta)} \right)^{n-j} \right\} \theta^S \\ &\propto \left\{ \sum_{j=0}^{s_0} w_j(\theta) p^j (1-p)^{n-j} \right\} \theta^S, \end{aligned}$$

où $w_j(\theta) = \binom{s}{j} \left(\frac{b(0)}{c(\theta)} \right)^{n-j}$ et $S = \sum_{i=1}^{n} y_i$.

Pour θ fixé, $L(p, \theta)$ peut être interprété comme un mélange de densités bêta et réciproquement pour p fixé, cette quantité peut être aussi vue comme un mélange de densités

appartenant à la famille exponentielle de loi.

2.2.3 Estimation par augmentation des données

Nous pourrions utiliser un échantillonnage de Gibbs compte tenu des densités conditionnelles des paramètres (p, θ) en choisissant des lois *a priori* conjuguées, mais lorsque des variables explicatives sont introduites dans la modélisation de p et/ou θ , la forme des densités des paramètres de régression devient alors non standard. Ainsi nous présentons à titre d'illustration une méthode d'estimation basée sur l'introduction de variables latentes $(\tilde{B}, \tilde{V})$ tout en utilisant la forme $Y = \tilde{V}(1 - \tilde{B})$.

Lois *a priori*

Les formes des lois *a priori* sur les paramètres θ et p permettent d'obtenir des lois conjuguées. Ainsi, les lois *a priori* considérées sont de la forme :

$$\pi(\theta) \propto \frac{\theta^{a_1}}{c(\theta)^{a_2}} \quad \text{et} \quad p \sim Beta(b_1, b_2).$$

De plus, pour traduire une loi *a priori* non informative sur le paramètre p, nous pourrions choisir des valeurs telles que $b_1 = b_2 = 1$ qui correspondent à une loi *a priori* uniforme. Concernant le paramètre θ, des valeurs très faibles $a_1 = a_2 = 0{,}001$ traduisent une loi *a priori* plutôt vague.

L'objectif de cette méthode est d'obtenir des échantillons provenant des lois *a posteriori* de $\left(p, \theta, \tilde{V}, \tilde{B}\right)$ plutôt que ceux provenant directement de (p, θ) qui s'avèrent plus difficiles à obtenir. Mais avant, il faut déterminer la loi de $\left(\tilde{V}, \tilde{B}\right)$ étant donné (p, θ, Y) nécessaire pour l'imputation des variables latentes. La loi conditionnelle du couple $\left(\tilde{V}, \tilde{B}\right)$ s'écrit :

$$P\left(\tilde{V}=v,\tilde{B}=0|Y=y\right)=\begin{cases}\frac{(1-p)P\left(\tilde{V}=0\right)}{p+(1-p)P\left(\tilde{V}=0\right)} & \text{si } v=y=0\\ 1 & \text{si } v=y>0\\ 0 & \text{sinon.}\end{cases} \tag{2.2.1}$$

et

$$P\left(\tilde{V}=v,\tilde{B}=1|Y=y\right)=\begin{cases}\frac{pP\left(\tilde{V}=v\right)}{p+(1-p)P\left(\tilde{V}=0\right)} & \text{si } y=0\\ 0 & \text{sinon.}\end{cases} \tag{2.2.2}$$

Ainsi, une fois que les couples $\left(\tilde{B}_i,\tilde{V}_i\right)_{i=1}^n$ sont tirés conditionnellement aux $(y_i)_{i=1}^n$, la loi *a posteriori* peut s'écrire :

$$\begin{aligned}P\left(p|\tilde{B},\tilde{V}\right) &\propto \pi(p)\prod_{i=1}^n p^{\tilde{B}_i}(1-p)^{1-\tilde{B}_i}\\ &\propto p^{b_1-1}(1-p)^{b_2-1}\,p^{\sum_{i=1}^n\tilde{B}_i}(1-p)^{\sum_{i=1}^n\left(1-\tilde{B}_i\right)}\\ &\propto p^{b_1-1+\sum_{i=1}^n\tilde{B}_i}(1-p)^{b_2-1+\sum_{i=1}^n\left(1-\tilde{B}_i\right)},\end{aligned}$$

c'est-à-dire

$$p|\tilde{B},\tilde{V}\sim Beta\left(b_1+\sum_{i=1}^n\tilde{B}_i,b_2+\sum_{i=1}^n\left(1-\tilde{B}_i\right)\right). \tag{2.2.3}$$

Concernant le paramètre θ, la forme de la loi *a posteriori* est assez complexe et peut même s'avérer non standard. Cependant, dans le cas où $\tilde{V}\sim Poiss(\theta)$, la loi *a posteriori* est :

$$\theta|\tilde{B},\tilde{V}\sim\mathscr{G}\left(a_1+\sum_{i=1}^n\tilde{V}_i,a_2+n\right). \tag{2.2.4}$$

La méthode d'estimation peut se résumer en l'algorithme ci-dessous.

Algorithme 2.2.1. *Algorithme d'estimation par augmentation des données*

1. *Initialiser les valeurs* $\left(p^{(0)}, \theta^{(0)}\right)$.
   À l'itération k,
2. *Simuler les variables latentes* $\left(\tilde{V}_i^{(k)}, \tilde{B}_i^{(k)}\right)$ *selon la loi (2.2.1) et (2.2.2) par la méthode suggérée par Gosh* et al. *(2006)*
3. *Générer* $p^{(k)}$ *selon la distribution (2.2.3)*
4. *Générer* $\theta^{(k)}$ *selon la densité (2.2.4)*
5. *Retourner à l'étape* 2 *pour estimer* $\left(p^{(k+1)}, \theta^{(k+1)}, \tilde{V}_i^{(k+1)}, \tilde{B}_i^{(k+1)}\right)$ *jusqu'à ce qu'un critère de convergence soit satisfait.*

2.3 Modèle linéaire généralisé mixte standard

Nous distinguerons deux cas selon la commplexité des effets aléatoires. En effet, nous nous interresserons dans un premier temps au modèle généralisé mixte avec un effet aléatoire unidimensionnel simple. Ensuite, nous exposerons le cas des modèles spatiaux qui introduisent une structure plus complexe des effets aléatoires.

Ce modèle a fait l'objet d'une vaste littérature et est facilement estimable grâce à de nombreux logiciels statistiques. Cette section constitue une brève présentation des différentes approches utilisées pour l'estimation des modèles généralisés linéaires mixtes. La première hypothèse consiste à postuler une distribution appartenant à une famille exponentielle pour les variables observées c'est-à-dire

$$y_i|b \sim f_{Y_i|b}(y_i|b),$$

où f peut se mettre sous la forme :

$$f_{Y_i|b}(y_i|b) = \exp\left\{\frac{y_i\gamma_i - B(\gamma_i)}{\tau^2} - c(y_i,\tau)\right\},$$

où $E(Y_i|b) = \frac{\partial B(\gamma_i)}{\partial \gamma_i}$. Le paramètre b représente le plus souvent une variable latente dont la structure de la densité est supposée connue. Par exemple, dans le cas de modèles qui prennent en compte la surdispersion en introduisant une variable latente, la forme de b la plus souvent postulée est celle d'une loi normale ou gamma. De même, dans le cas des modèles autorégressifs conditionnels, il faut préciser la matrice adjacente désignant la notion de voisinage et la distribution gaussienne conditionnelle du paramètre b.

Soient $\mathbf{Z} = (\mathbf{Z}_1,\ldots,\mathbf{Z}_n)'$ et $\mathbf{R} = (\mathbf{R}_1,\ldots,\mathbf{R}_n)'$ les matrices respectives des variables explicatives des effets fixes et aléatoires. Si m désigne la fonction de lien entre la moyenne et les variables explicatives, alors $m(\mu_i) = \gamma_i$ et le modèle s'écrit

$$E[Y_i|b] = \mu_i \; i = 1,\ldots,n,$$
$$m(\mu_i) = \mathbf{Z}_i\beta + \mathbf{R}_i b \; i = 1,\ldots,n.$$

Supposons que la densité des effets aléatoires est connue et désignée par $g(b)$. En général b possède une structure gaussienne ; cette hypothèse facilite grandement les calculs numériques de la vraisemblance donc l'inférence statistique. Cependant, il est possible de supposer une structure différente telle que la loi gamma dans le cadre des données de Poisson ; d'où le modèFle de mélange Poisson-gamma. Dans certains cas, des structures plus complexes telles que les lois conditionnelles autorégressives peuvent être utilisées afin de modéliser une dépendance spatiale ou temporelle. Les livres de Dey *et al.*, (2000) et Banergee *et al.*, (2004) constituent des références assez concises sur la dépendance spatiale modélisée par des effets aléatoires autorégressifs.

Nous présenterons une approche selon le maximum de vraisemblance et une autre selon l'approche bayésienne. Rappelons que ces exemples sont présentés dans le seul but d'illustrer cette méthode parmi tant d'autres.

2.3.1 Maximum de vraisemblance

Le principe du maximum de vraisemblance consiste à écrire ou approximer la vraisemblance pour ensuite déterminer les points maxima. Mais, en général, selon la complexité du modèle, la vraisemblance est très difficile à évaluer et elle doit être approximée par la méthode de Gauss-Hermite ou celle de Laplace. Lorsqu'elle est plutôt facile à écrire analytiquement, il suffit d'écrire les conditions d'optimalité pour ensuite résoudre les équations qui en découlent.

Maximisation par intégration

Soit ϕ les paramètres de la distribution de b et $g(b|\phi)$ la densité du vecteur b. La fonction de vraisemblance du modèle s'écrit :

$$L(\beta,\phi|y) = \int_{\mathbb{R}^n} \prod_{i=1}^{n} f_{y_i|b_i}(y_i|b_i,\beta,\phi)\, g(b_i|\phi)\, db_i.$$

Dans le cas où les effets aléatoires possèdent une structure assez simple telle que la loi gaussienne unidimensionnelle ou la loi gamma, la forme plutôt simple des effets aléatoires permet d'utiliser une approximation de la vraisemblance par la méthode numérique d'intégration de Gausse-Hermite ou celle de Laplace. Ensuite, les estimateurs de maximum de vraisemblance peuvent être déterminés par une simple procédure de maximisation classique. En pratique, des valeurs de grilles d'interpolation supérieures à 10 donnent une précision très satisfaisante.

Équations de vraisemblance

Nous décrivons les équations obtenues comme conditions d'optimalité de premier ordre. Leurs solutions sont donc des maxima potentiels ; seules les conditions de second ordre ou la convexité de la fonction de log-vraisemblance permettront de déterminer le ou les maxima. Pour la suite, posons

$$l = \log\left[\int f(y|b)g(b)db\right] = \log f(y).$$

Pour estimer les paramètres des effets fixes, il suffit de réécrire la fonction de log-vraisemblance et de dériver les conditions d'optimalité encore appelées équations de vraisemblance, c'est-à-dire

$$\frac{\partial l}{\partial \beta} = \frac{1}{f(y)}\frac{\partial}{\partial \beta}\left[\int f(y|b)g(b)db\right].$$

Sous les conditions de régularité, nous avons

$$\frac{\partial l}{\partial \beta} = \frac{\int\left[\frac{\partial}{\partial \beta}f(y|b)\right]g(b)db}{f(y)}.$$

En effet, un modèle paramétrique $\mathscr{P} = \{P_\theta\}$ est dit régulier s'il vérifie les hypothèses suivantes :

- il existe une mesure dominante μ de P_θ ($f_\theta = \frac{dP_\theta}{d\mu}$)
- $\forall\ \theta,\ \frac{\partial f(\theta,x)}{\partial \theta}$ existe μ presque partout.

- pour tout borélien A, la fonction $\theta \longmapsto P_\theta(A)$ est dérivable en θ et

$$\frac{\partial}{\partial \theta}\int_{\mathscr{A}} f(\theta,x)d\mu(x) = \int_{\mathscr{A}} \frac{\partial f(\theta,x)}{\partial \theta}d\mu(x),$$

- pour tout borélien A, la fonction $P_\theta(A)$ est deux fois dérivable en θ et

$$\frac{\partial^2}{\partial \theta^2}\int_{\mathscr{A}} f(\theta,x)d\mu(x) = \int_{\mathscr{A}} \frac{\partial^2 f(\theta,x)}{\partial \theta^2}d\mu(x).$$

En remarquant que le terme $g(b)$ est indépendant du paramètre β, nous pouvons appliquer la modification suivante :

$$\begin{aligned}\frac{\partial}{\partial \beta}f(y|b) &= \left[\frac{1}{f(y|b)}\frac{\partial f(y|b)}{\partial \beta}\right] f(y|b) \\ &= \frac{\partial \log(f(y|b))}{\partial \beta} f(y|b).\end{aligned}$$

Ensuite, nous en déduisons que :

$$\begin{aligned}\frac{\partial l}{\partial \beta} &= \int \frac{\partial \log(f(y|b))}{\partial \beta}\frac{f(y|b)g(b)}{f(y)}db \\ &= \int \frac{\partial \log(f(y|b))}{\partial \beta} g(b|y)db \\ &= \int Z'W^*\left(y-\mu\right) f_{b|y}\left(b|y\right)db \\ &= Z'E\left[W^*|y\right] - Z'E\left[W^*\mu|y\right],\end{aligned}$$

où $W^* = \frac{\partial^2 B(\gamma_i)}{\partial \gamma_i^2} m(\mu_i)$.
La condition d'optimalité de premier ordre pour β s'écrit donc :

$$\frac{\partial l}{\partial \beta} = 0 \Leftrightarrow Z'E\left[W^*|y\right] = Z'E\left[W^*\mu|y\right].$$

Quant aux effets aléatoires, une équation de vraisemblance similaire à celle des effet fixes peut être obtenue. En effet, si ϕ désigne les paramètres de la distribution de b, nous obtenons dans le cadre général :

$$\begin{aligned} \frac{\partial l}{\partial \phi} &= \int \frac{\partial \log g(b)}{\partial \phi} f_{b|y}(b|y)db \\ &= E\left[\frac{\partial \log g(b)}{\partial \phi} \mid y\right] = 0. \end{aligned}$$

Ensuite, afin de résoudre cette équation, il faudrait spécifier la distribution des effets aléatoires, c'est-à-dire la fonction $g(b)$.

2.3.2 Une approche par CMMC ou EMMC

Par la suite, nous spécifions le modèle de façon à obtenir des résultats explicites et plus simples à interpréter grâce à l'approche des chaînes de Markov par Monte Carlo (CMMC) et celle de l'espérance maximisation par Monte Carlo (EMMC). Le modèle linéaire mixte de Poisson permettant de prendre en compte le phénomène de surdispersion peut s'écrire de la forme suivante. Soient $Y_1, \ldots, Y_n$, n variables aléatoires indépendantes provenant d'une loi de Poisson telles que :

$$Y_i|b_i \sim Poiss\,(\theta_i)\,.$$

Ensuite, la forme paramétrique du modèle de régression est définie par une fonction de lien logarithmique :

$$\log(\theta_i) = Z_i\beta + b_i, \;\; i = 1, \ldots, n,$$

où $b_1, \ldots, b_n$ désignent des variables aléatoires gaussiennes unidimensionnelles non observables permettant de capter l'hétérogénéité des observations, c'est-à-dire

$$b_i \sim \mathcal{N}\left(0, \sigma^2\right) \quad i = 1, \ldots, n.$$

McCulloch (1997) a proposé un algorithme de tirage de la variable latente b_i dans un modèle linéaire généralisé en présence d'effets aléatoires. Cependant, l'efficacité de cet algorithme de Metropolis-Hastings dépend beaucoup du choix de la loi instrumentale notée h. En effet, l'auteur a montré que le choix optimal de la loi instrumentale est une normale centrée à 0 et de variance égale à celle du modèle, c'est-à-dire σ^2. Ainsi la probabilité d'acceptation de l'algorithme est égale à :

$$A_i\left(b_i, b_i^*\right) = \min\left\{1, \frac{f_{b|y}\left(b_i^*|y_i, \beta, \sigma^2\right) h_b\left(b_i\right)}{f_{b|y}\left(b_i|y_i, \beta, \sigma^2\right) h_b\left(b_i^*\right)}\right\}.$$

Pour la suite, posons

$$Q_i\left(b_i, b_i^*\right) = \frac{f_{b|y}\left(b_i^*|y_i, \beta, \sigma^2\right) h_b\left(b_i\right)}{f_{b|y}\left(b_i|y_i, \beta, \sigma^2\right) h_b\left(b_i^*\right)},$$

où b_i^* désigne un candidat tiré dans la loi instrumentale h . En choisissant $h_b = f_b$ tel que suggéré par McCulloch (1997), l'expression de Q_i se simplifie assez bien

$$Q_i\left(b_i, b_i^*\right) = \exp\left\{\left(b_i^* - b_i\right) - \left(e^{b_i^*} - e^{b_i}\right) e^{Z_i\beta}\right\}.$$

Cette étape demeure assez importante compte tenu du nombre élevé de paramètres b_i à générer. La méthode d'estimation peut se résumer en l'algorithme suivant.

Algorithme 2.3.1. *Algorithme EMMC (McCulloch, 1997)*

1. *Choisir les valeurs initiales $\beta^{(0)}$ et $\sigma^{2^{(0)}}$ pour le cas $m = 0$.*

Pour ce faire, une idée serait de maximiser la vraisemblance de chaque composante de Y qui peut se mettre sous la forme :

$$L\left(\beta,\sigma^2|y\right)=\int_{\mathbb{R}^n}\prod_{i=1}^{n}f_{y_i|b_i}\left(y_i|b_i,\beta,\sigma^2\right)f_b\left(b_i|\sigma^2\right)db_i.$$

Dans ce cas bien précis, la forme plutôt simple des effets aléatoires permet d'utiliser une approximation de la vraisemblance par la méthode numérique d'intégration de Gauss-Hermite.

2. *À l'étape $m+1$, générer N vecteurs de la forme $\widetilde{b}^{(1)},\ldots,\widetilde{b}^{(N)}$ tels que $\widetilde{b}^{(k)}=\left(\widetilde{b}_1^{(k)},\ldots,\widetilde{b}_n^{(k)}\right)$ de la loi de $f_{b|y}\left(b|y,\beta^{(m)},\sigma^{2(m)}\right)$ selon un algorithme de Métropolis-Hastings défini précédemment.*

3. *Trouver $\beta^{(m+1)}=\operatorname{argmax}\mathbb{E}\left[\log\left(f_{b|y}\left(b|y,\beta\right)\right)\right]$. Une approche Monte Carlo permet donc d'approximer cette espérance par*

$$\frac{1}{N}\sum_{k=1}^{N}\sum_{i=1}^{n}\log\left(f_{b_i|y_i}\left(\widetilde{b}_i^{(k)}|y_i,\beta\right)\right).$$

4. *Trouver ensuite $\sigma^{2(m+1)}=\operatorname{argmax}\mathbb{E}\left[\log\left(f_b\left(b|\sigma^2\right)\right)\right]$ approximée par*

$$\frac{1}{N}\sum_{k=1}^{N}\sum_{i=1}^{n}\log\left(f_{b_i}\left(\widetilde{b}_i^{(k)}|\sigma^2\right)\right).$$

Les étapes précédentes sont itérées jusqu'à la convergence de l'algorithme. Il est important de préciser que le choix des valeurs initiales augmente considérablement la vitesse de convergence de l'algorithme. En effet, il serait assez facile aussi de choisir les estimateurs obtenus par une régression de Poisson comme valeurs initiales des paramètres de la moyenne.

Exemple d'un modèle de Poisson

Tout comme McCulloch et Searle (2001), nous reprendrons l'exemple des données de Poisson corrélées tel que présenté par Diggle et *al.* (1994). Il s'agit d'une expérience à mesure répétées sur m groupes supposés corrélés. L'observation y_{ij} désigne ainsi la j^e observation effectuée sur un patient du groupe i. Le modèle s'écrit :

$$\begin{aligned} Y_{ij}|b &\sim Poiss(\mu_{ij}),\ \ i=1,\ldots,m\ , j=1,2,\ldots,n_i, \\ \log\left(\mu_{ij}\right) &= Z_{ij}\beta_j + b_i,\ \ i=1,\ldots,m\ , j=1,2,\ldots,n_i, \\ b_i &\sim \mathscr{N}\left(0,\sigma^2\right),\ \ i=1,\ldots,m. \end{aligned}$$

Il faut aussi préciser que les observations y_{ij} sont indépendantes conditionnellement au vecteur traduisant l'effet aléatoire désigné par b. La fonction de log-vraisemblance du modèle peut se mettre sous la forme suivante :

$$\begin{aligned} l &= \log\left(\prod_{i=1}^{m}\int_{-\infty}^{+\infty}\prod_{j=1}^{n_i}\frac{\mu_{ij}^{y_{ij}}e^{-\mu_{ij}}}{y_{ij}!}\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{1}{2\sigma^2}b_i^2}db_i\right) \\ &= \sum_{i=1}^{m}\log\left(\int_{-\infty}^{+\infty}\exp\left\{\sum_{j=1}^{n_i}y_{ij}b_i-\sum_{j=1}^{n_i}e^{Z_{ij}\beta_j+b_i}\right\}\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{1}{2\sigma^2}b_i^2}db_i\right) \\ &\quad +\sum_{i,j}y_{ij}Z_{ij}\beta_j-\sum_{i,j}\log(y_{ij}!). \end{aligned}$$

Le choix ou la spécification de la distribution de b est très important pour la maximisation de la fonction de log-vraisemblance. En effet, des effets aléatoires de distribution normale sont très souvent considérés car ils permettent d'approximer ces intégrales par des méthodes d'intégration numériques telles que celle de Gauss-Hermite ou encore celle de

Laplace. Cependant, cette méthode plutôt directe par le maximum de vraisemblance n'est possible que lorsque la structure des effets aléatoires est assez simple. Par contre, dans le cas d'effets aléatoires emboîtés, elle est inadéquate.

2.3.3 Le modèle binomiale négative

C'est un modèle plus général que celui de Poisson car il permet de prendre en compte la surdispersion grâce à un nouveau paramètre. Le modèle s'écrit :

$$Y_i \sim \mathscr{NB}(\theta_i, \alpha), \quad i = 1, \ldots, n,$$
$$\log(\theta_i) = Z_i\beta, \quad i = 1, \ldots, n.$$

La fonction de masse de l'observation y_i s'écrit

$$P(Y_i = y_i|\alpha, \theta_i) = \frac{\Gamma(y_i + \alpha)}{\Gamma(\alpha)y_i!}\left[\frac{\alpha}{\alpha + \theta_i}\right]^{\alpha}\left[\frac{\theta_i}{\alpha + \theta_i}\right]^{y_i}.$$

Les deux premiers moments s'écrivent

$$E(Y_i) = \theta_i \quad \text{et} \quad \text{var}(Y_i) = \theta_i + \frac{\theta_i^2}{\alpha}.$$

Ainsi spécifié, le modèle binomiale négative peut être estimé assez facilement par le maximum de vraisemblance. Par ailleurs, il faut remarquer que pour une valeur fixée de α, nous obtenons une loi de famille exponentielle qui est facilement estimable par un algorithme EM tel que proposé par Tanner et Wong (1987).

De plus, le paramètre α désigne le degré de surdispersion observé. En effet, plus la valeur de α est grande, moins le phénomène de surdispersion est prononcé. De plus, le cas $\alpha \longrightarrow +\infty$ pour θ fixé (tel que $\text{var}(Y) \to \theta$) correspond au modèle standard de Poisson

sans surdispersion. L'avantage majeur de cette formulation est l'aisance relative avec laquelle les estimateurs du maximum de vraisemblance peuvent être obtenus. La plupart des logiciels statistiques possèdent des macro-commandes qui produisent ces estimateurs.

2.4 Modèles généralisés mixtes avec des effets spatiaux

Après le modèle standard, on pourrait envisager un modèle spatial en y introduisant un effet aléatoire spatial, car l'abondance d'une espèce donnée à un site est fonction de celle de ses sites voisins. La plupart des problèmes rencontrés dans la prévision des niches écologiques peuvent être résumés comme suit :

- **la variabilité du taux d'échantillonnage** : compte tenu des caractéristiques naturelles des différents sites, il est possible que certains d'entres eux soient moins échantillonnés tandis que d'autres le seront de façon ń excessive ż. En effet, sur des sites hostiles à l'activité humaine, il serait difficile d'observer suffisamment d'espèces. Ainsi, on pourrait penser que certaines espèces sont rares alors qu'on n'a pas observé les sites adéquats ou bien parce qu'on n'a pas obtenu assez d'échantillons. De plus, les données ne sont pas toujours observées à la même échelle. Tout ceci conduit lors de la prévision à des biais d'échantillonnage et à des problèmes liés à l'alignement spatial des observations (Mugglin *et al.*, 2000 ; Gelfand *et al.*, 2002).

- **la dépendance spatiale** : très souvent ignorée ou traitée de façon insatisfaisante, ce phénomène induit des estimateurs très volatils.

- **l'incertitude** : En plus de l'approche classique, l'approche bayésienne permet non seulement d'obtenir des estimateurs des paramètres du modèle mais aussi leur distribution. Ainsi, nous possédons toute ou quasiment toute l'information sur la distribution des paramètres. Il est donc possible de quantifier l'incertitude des prévisions.

2.4.1 Le modèle

Soient $Y_1, \ldots, Y_n$, n variables aléatoires dont la distribution appartenant à une famille exponentielle c'est-à-dire

$$y_i|b \sim f_{Y_i|b}(y_i|b),$$

$$f_{Y_i|b}(y_i|b) = \exp\left\{\frac{y_i\gamma_i - B(\gamma_i)}{\tau^2} - c(y_i,\tau)\right\}, i = 1,\ldots,n.$$

Le modèle avec les effets spatiaux peut s'écrire sous la forme suivante

$$m(\gamma_i) = z_{1i}^t\theta + z_{2i}^t\mathbf{U} + e_i, i = 1,\ldots,n \ \ ,$$

où $\mathbf{e} = (e_1,\ldots,e_N)^t$ désigne les effets résiduels tels que $e_i \sim N(0,\delta_0)$. Par ailleurs $\mathbf{Z}$ et $\mathbf{e}$ sont supposés indépendants. Par la suite, posons $\mathbf{Z}_1 = (\mathbf{z}_{11},\ldots,\mathbf{z}_{1n})'$ et $\mathbf{Z}_2 = (\mathbf{z}_{21},\ldots,\mathbf{z}_{2n})'$ des matrices de tailles respectives $n \times p$ et $n \times k$. Le vecteur θ de taille $p \times 1$ représente les effets fixes et $\mathbf{U}$ de taille $k \times 1$ représente celui des effets aléatoires. Pour la suite, posons $\mathbf{V} = m(\gamma_1,\gamma_2,\ldots,\gamma_n)$.

2.4.2 Effets spatiaux

Il est généralement supposé que la corrélation spatiale entre deux sites i et j dépend uniquement de la distance les séparant. En réalité, cette forme de corrélation n'est quasiment jamais observée ; la dépendance a plutôt une structure de graphe c'est-à-dire que la distribution d'un site i est influencée ou dépend de celle de tous les autres sites.

Pour la structure AR(1) de $\mathbf{U}$, la distribution de U_i conditionnellement a $\mathbf{U}_{-i} = (U_j, j \neq i)$ dépend uniquement des variables adjacentes, notamment U_{i-1} et U_{i+1}. Besag (1974) introduisit les modèles autorégressifs conditionnels d'ordre 1 ou ARC(1). Selon l'auteur

$\mathbf{U}$ est tel que, pour $i = 1, \ldots, k$

$$f(U_i|\mathbf{U}_{-i}) = \left(\frac{\alpha_i}{2\pi\delta_1}\right)^{1/2} \exp\left\{-\frac{\alpha_i}{2\delta_1}\left(U_i - \sum_{j\neq i}^{k} \beta_{ij}\mathbf{U}_j\right)^2\right\}.$$

En supposant que $\mathbf{U}$ est une matrice symétrique et définie positive $k \times k$ définie par $U_{ij} = -\alpha_i\beta_{ij}$ avec $\beta_{ii} = 1$ alors la distribution jointe de $\mathbf{U}$ est une normale multivariée $\mathscr{N}_p\left(0, \delta_1\mathbf{B}^{-1}\right)$:

$$f(\mathbf{U}) = (2\pi\delta_1)^{-k/2}\,|\mathbf{B}|^{1/2} \exp\left\{-\frac{1}{2\delta_1}\left(\mathbf{U}'\mathbf{B}\mathbf{U}\right)\right\}.$$

Le théorème suivant nous assure l'existence de densité *a posteriori* propre pour un vaste nombre de modèles mixtes. Par exemple, lorsqu'un des paramètres possède une densité *a priori* discrète ou même une densité non informative, il se peut que la densité jointe n'existe pas. De plus en prenant $\tau = 1$, $m = \log(.)$, on vérifie aisément que toutes les conditions sont bien satisfaites pour le modèle de Poisson.

Densités *a priori*

Nous utilisons les mêmes lois *a priori* des paramètres que ceux de Sun *et al.* (1998). Ces lois facilitent les calculs car elle permettent d'obtenir des lois *a posteriori* standards. Elles s'écrivent

$$\theta \sim \mathscr{N}_p(\mu_0, \Sigma_0), \quad \delta_0 \sim \mathscr{IG}(a_0, b_0), \quad \text{et} \quad \delta_1 \sim \mathscr{IG}(a_1, b_1).$$

Contrairement à l'approche précédente, la vraisemblance complète ou jointe ne peut être utilisée. On doit plutôt se contenter de la vraisemblance conditionnelle. Notons que pour τ fixé, la fonction de vraisemblance $f(y_i|\gamma_i, \tau)$ est supposée bornée et telle que $M_i(\tau) \equiv \sup_{\gamma_i} f(y_i|\gamma_i, \tau)$.

Théorème 2.4.1. *Sun et al., (1998)*

Soit le modèle GLMM avec effets résiduels $e_i \sim N(0, \delta_0)$. *Supposons satisfaites les conditions suivantes :*

1. *Il existe a sous ensemble de* $\{1, \ldots, n\}$, *noté* $\mathscr{J}_n = (i_1, \ldots, i_n)$ *tel que*

$$\int \prod_{j \notin \mathscr{J}_n} M_j(\tau) \left\{ \prod_{j \notin \mathscr{J}_n} \int f_j\left(y_j | \gamma_j, \tau\right) m_j^{'}\left(\gamma_j\right) d\gamma_j \right\} F\left(d\tau\right) \ < \infty.$$

2. *La matrice de design* $\mathbf{Z}_1^* = (\mathbf{z}_{1,i_1}, \ldots, \mathbf{z}_{1,i_n})^t$ *est de plein rang p et* $\mathbf{Z}_2^* = (\mathbf{z}_{2,i_1}, \ldots, \mathbf{z}_{2,i_n})^t$ *est de même rang que* $\mathbf{Z}_2$.

3. *La densité de* θ *est non informative et* $\mathbf{U}$ *suit la densité :*

$$f(\mathbf{U}) \propto (2\pi\delta_1)^{r/2} |\mathbf{B}|_+^{1/2} \exp\left(-\frac{1}{2\delta_1}\mathbf{U}^{'}\mathbf{B}\mathbf{U}\right),$$

   où $|\mathbf{B}|_+$ *désigne le produit des valeurs propres strictement positives de* $\mathbf{B}$.

4. *Le rang de* $\left(\mathbf{Z}_2^{*t}\mathbf{R}_1\mathbf{Z}_2^* + \mathbf{B}\right)$ *est k où* $\mathbf{R}_1 = I_n - \mathbf{Z}_1^*\left(\mathbf{Z}_1^{*t}\mathbf{Z}_2^*\right)\mathbf{Z}_1^{*t}$.

5. *Les densités* a priori *de* (δ_0, δ_1) *satisfont la condition de moment :*

$$\mathbf{E}\left[\delta_0^{-\frac{1}{2}(n-p-k)}\delta_1^{-\frac{1}{2}(n-p)}\right] < \infty.$$

Alors la densité a posteriori *de* $(\gamma, \tau, \theta, \mathbf{U}, \delta_0, \delta_1)$ *est propre.*

Densités *a posteriori*

Nous avons tenté d'étendre les calculs de Sun *et al.*, (1998) sur la détermination des lois *a posteriori* du modèle en considérant plutôt une loi *a priori* propre sur θ de la forme

$\theta \sim N(\mu_0, \Sigma_0)$. Notons que le cas impropre peut être considérer comme cas limite lorsque nous posons $\Sigma_0^{-1} \to 0$.

En effet, on peut tirer tous les paramètres du modèle par un échantillonnage de Gibbs et un échantillonnage par tranche (*slice sampling*) ou par l'intermédiaire de l'algorithme de Metropolis-Hastings. Notons qu'ici, nous avons omis volontairement la densité *a posteriori* de τ car dans le modèle de Poisson, la valeur de τ est supposée égale à l'unité ($\tau = 1$).

Proposition 2.4.1. *Tirage des lois* a posteriori
Les lois a posteriori *des paramètres sont de la forme suivante :*

1.

$$\theta|(\gamma,\tau,\mathbf{U},\delta_0,\delta_1) \sim \mathscr{N}_p\left(\left[\Sigma_0^{-1}+\frac{1}{\delta_0}\mathbf{Z}_1'\mathbf{Z}_1\right]^{-1}\left[\Sigma_0^{-1}\mu_0'+\frac{1}{\delta_0}\mathbf{Z}_1'(\mathbf{V}-\mathbf{Z}_2\mathbf{U})\right], \left[\Sigma_0^{-1}+\frac{1}{\delta_0}\mathbf{Z}_1'\mathbf{Z}_1\right]^{-1}\right),$$

2.

$$\mathbf{U}|(\gamma,\tau,\theta,\delta_0,\delta_1) \sim \mathscr{N}_k\left(\mathbf{M}_1\mathbf{Z}_2'(\mathbf{V}-\mathbf{Z}_1\theta), \delta_0\mathbf{M}_1\right),$$

avec $\mathbf{M}_1 = \left(\mathbf{B}\delta_0\delta_1^{-1}+\mathbf{Z}_2'\mathbf{Z}_2\right)^{-1}$,

3.

$$\delta_0|(\gamma,\tau,\theta,\mathbf{U},\delta_1) \sim \mathscr{IG}\left(a_0+\frac{n}{2}, b_0+\frac{1}{2}\|\mathbf{V}-\mathbf{Z}_1\theta-\mathbf{Z}_2\mathbf{U}\|^2\right),$$

où $\|\mathbf{A}\|^2 = \mathbf{A}'\mathbf{A}$,

4.

$$\delta_1|(\gamma,\tau,\theta,\mathbf{Z},\delta_0) \sim \mathscr{IG}\left(a_1+\frac{k}{2}, b_1+\frac{1}{2}\mathbf{U}'\mathbf{B}\mathbf{U}\right).$$

5. *Étant donné* $(\tau, Z, \delta_0, \delta_1)$*, les* γ_i *sont indépendants et leurs densités conditionnelles s'écrivent*

$$P(\gamma_i|(\tau,\theta,Z,\delta_0,\delta_1)) \propto \exp\left[\frac{y_i\gamma_i - B_i(\gamma_i)}{\tau^2} - \frac{\left(h(\gamma_i) - x_{1i}^t\theta - x_{2i}^t\mathbf{U}\right)^2}{2\delta_0}\right] h'(\gamma_i) \;\; i = 1,\ldots,n.$$

Démonstration. **Loi *a posteriori* de** θ

Elle s'écrit :

$$\begin{aligned}
P(\theta|\gamma,\tau,\mathbf{U},\delta_0,\delta_1) &\propto \pi(\theta)\,\pi(m(\gamma)|\theta,\mathbf{U}) \\
&\propto \exp\left\{-\frac{1}{2}(\theta-\mu_0)'\,\Sigma_0^{-1}(\theta-\mu_0)\right\}\exp\left\{-\frac{1}{2\delta_0}(\mathbf{V}-\mathbf{Z}_1\theta-\mathbf{Z}_2\mathbf{U})'\right. \\
&\quad \left.\times(\mathbf{V}-\mathbf{Z}_1\theta-\mathbf{Z}_2\mathbf{U})\right\} \\
&\propto \exp\left\{-\frac{1}{2}\left[\theta'\Sigma_0^{-1}\theta - 2\mu_0'\Sigma_0^{-1}\theta + \frac{1}{\delta_0}\left(\theta'\mathbf{Z}_1'\mathbf{Z}_1\theta\right.\right.\right. \\
&\quad \left.\left.\left.+2\mathbf{U}'\mathbf{Z}_2'\mathbf{Z}_1\theta - 2\mathbf{V}'\mathbf{Z}_1\theta\right)\right]\right\} \\
&= \exp\left\{-\frac{1}{2}\left[\theta'\left(\Sigma_0^{-1} + \frac{1}{\delta_0}\mathbf{Z}_1'\mathbf{Z}_1\right)\theta - 2\left(\mu_0'\Sigma_0^{-1} + \frac{1}{\delta_0}\right.\right.\right. \\
&\quad \left.\left.\left.\times\left(\mathbf{V}' - \mathbf{U}'\mathbf{Z}_2'\right)\mathbf{Z}_1\right)\theta\right]\right\}.
\end{aligned}$$

En utilisant la propriété de familles conjuguées, nous en déduisons :

$$\mathrm{var}(\theta|\gamma,\mathbf{U},\delta_0,\delta_1) = \left[\Sigma_0^{-1} + \frac{1}{\delta_0}\mathbf{Z}_1'\mathbf{Z}_1\right]^{-1},$$

$$E(\theta|\gamma,\tau,\mathbf{U},\delta_0,\delta_1) = \left[\Sigma_0^{-1} + \frac{1}{\delta_0}\mathbf{Z}_1'\mathbf{Z}_1\right]^{-1}\left[\Sigma_0^{-1}\mu_0 + \frac{1}{\delta_0}\mathbf{Z}_1'(\mathbf{V}-\mathbf{Z}_2\mathbf{U})\right].$$

Loi *a posteriori* de U

En utilisant le théorème de Bayes, nous avons

$$\begin{aligned}
P(\mathbf{U}|\gamma,\theta,\delta_0,\delta_1) &\propto \pi(\mathbf{U})\,\pi(m(\gamma)|\theta,\mathbf{U}) \\
&\propto \exp\left\{-\frac{1}{2\delta_1}\mathbf{U}'\mathbf{B}\mathbf{U}\right\}\exp\left\{-\frac{1}{2\delta_0}(\mathbf{V}-\mathbf{Z}_1\theta-\mathbf{Z}_2\mathbf{U})'\right. \\
&\quad \left.\times(\mathbf{V}-\mathbf{Z}_1\theta-\mathbf{Z}_2\mathbf{U})\right\} \\
&\propto \exp\left\{-\frac{1}{2}\left[\frac{1}{\delta_1}\mathbf{U}'\mathbf{B}\mathbf{U}+\frac{1}{\delta_0}\left(-2\mathbf{V}'\mathbf{Z}_2\mathbf{U}+\mathbf{U}'\mathbf{Z}_2'\mathbf{Z}_2\mathbf{U}\right.\right.\right. \\
&\quad \left.\left.\left.+2\theta'\mathbf{Z}_1'\mathbf{Z}_2-\mathbf{Z}_2\mathbf{U}\right)\right]\right\} \\
&\propto \exp\left\{-\frac{1}{2\delta_0}\left[\frac{\delta_0}{\delta_1}\mathbf{U}'\mathbf{B}\mathbf{U}+\mathbf{U}'\mathbf{Z}_2'\mathbf{Z}_2\mathbf{U}-2\left(\mathbf{V}'-\theta'\mathbf{Z}_1'\right)\mathbf{Z}_2\mathbf{U}\right]\right\} \\
&\propto \exp\left\{-\frac{1}{2\delta_0}\left[\mathbf{U}'\left(\delta_0\mathbf{B}\delta_1^{-1}+\mathbf{Z}_2'\mathbf{Z}_2\right)\mathbf{U}-2\left(\mathbf{V}'-\theta'\mathbf{Z}_1'\right)\mathbf{Z}_2\mathbf{U}\right]\right\}.
\end{aligned}$$

En utilisant la propriété de familles conjuguées, nous en déduisons :

$$\mathrm{var}(\mathbf{U}|\gamma,\theta,\delta_0,\delta_1)=\delta_0\left[\delta_0\delta_1^{-1}\mathbf{B}+\mathbf{Z}_2'\mathbf{Z}_2\right]^{-1},$$

$$E(\mathbf{U}|\gamma,\tau,\theta,\delta_0,\delta_1)=\left[\delta_0\delta_1^{-1}\mathbf{B}+\mathbf{Z}_2'\mathbf{Z}_2\right]^{-1}\left[\mathbf{Z}_2'(\mathbf{V}-\mathbf{Z}_1\theta)\right].$$

Loi *a posteriori* de δ_0

Elle s'écrit

$$\begin{aligned}
P(\delta_0|\gamma,\theta,\mathbf{U},\delta_1) &\propto \pi(\delta_0)\,\pi(m(\gamma)|\theta,\delta_0,\mathbf{U}) \\
&\propto \frac{1}{\delta_0^{a_0+1}}\exp\left\{-\frac{b_0}{\delta_0}\right\}\frac{1}{\delta_0^{n/2}}\exp\left\{-\frac{1}{2\delta_0}\|\mathbf{V}-X_1\theta-\mathbf{Z}_2\mathbf{U}\|^2\right\} \\
&\propto \frac{1}{\delta_0^{a_0+1+n/2}}\exp\left\{-\frac{1}{\delta_0}\left[b_0+\frac{1}{2}\|\mathbf{V}-\mathbf{Z}_1\theta-\mathbf{Z}_2\mathbf{U}\|^2\right]\right\}.
\end{aligned}$$

En utilisant la propriété de familles conjuguées, nous obtenons par la suite :

$$\delta_0|\gamma,\theta,\mathbf{U},\delta_1 \sim \mathscr{IG}\left(a_0+\frac{n}{2}, b_0+\frac{1}{2}\|\mathbf{V}-\mathbf{Z}_1\theta-\mathbf{Z}_2\mathbf{U}\|^2\right).$$

Loi *a posteriori* de δ_1

Elle s'écrit

$$\begin{aligned} P(\delta_1|\gamma,\theta,\mathbf{U},\delta_0) &\propto \pi(\delta_1)\,\pi(\mathbf{U}|\delta_1) \\ &\propto \frac{1}{\delta_1^{a_1+1}}\exp\left\{-\frac{b_1}{\delta_1}\right\}\frac{1}{\delta_1^{k/2}}\exp\left\{-\frac{1}{2\delta_1}\mathbf{U}'\mathbf{B}\mathbf{U}\right\} \\ &\propto \frac{1}{\delta_1^{a_1+1+k/2}}\exp\left\{-\frac{1}{\delta_1}\left[b_1+\frac{1}{2}\mathbf{U}'\mathbf{B}\mathbf{U}\right]\right\}. \end{aligned}$$

En utilisant la propriété de familles conjuguées, nous en déduisons :

$$\delta_1|\gamma,\theta,\mathbf{U},\delta_0 \sim \mathscr{IG}\left(a_1+\frac{k}{2}, b_1+\frac{1}{2}\mathbf{U}'\mathbf{B}\mathbf{U}\right).$$

Loi *a posteriori* de γ_i

Puisque conditionnellement à $(\theta,\mathbf{U},\delta_0,\delta_1)$, les $\{\gamma_i\}_{i=1}^n$ sont indépendants, ainsi leurs lois *a posteriori* s'écrivent :

$$\begin{aligned} P(\gamma_i|\theta,\mathbf{U},\delta_0,\delta_1) &\propto \pi(\gamma_i|\theta,\mathbf{U},\delta_0,\delta_1)\,L(y_i|\gamma_i,\theta,\tau,\mathbf{U},\delta_0,\delta_1,D) \\ &\propto \pi(m(\gamma_i)|\theta,\mathbf{U},\delta_0,\delta_1)\,m'(\gamma_i)\,f(y_i|\gamma_i,\tau) \\ &\propto \exp\left[\frac{y_i\gamma_i-B_i(\gamma_i)}{\tau^2}\right]\exp\left\{-\frac{1}{2\delta_0}\left(m(\gamma_i)-\mathbf{z}_{1i}'\theta-\mathbf{z}_{2i}'\mathbf{U}\right)^2\right\}m'(\gamma_i) \\ &\propto \exp\left[\frac{y_i\gamma_i-B_i(\gamma_i)}{A_i(\tau)}-\frac{\left(h(\gamma_i)-\mathbf{z}_{1i}^t\theta-\mathbf{z}_{2i}^t\mathbf{U}\right)^2}{2\delta_0}\right]m'(\gamma_i). \end{aligned}$$

□

Remarque 2.4.1. *Pour une analyse plus robuste, nous pourrions par exemple ajouter des lois respectives normale et inverse Wishart sur* μ_0 *et* $\tilde{\Sigma}_0$. *Les densités* a posteriori *des nouveaux hyperparamètres se déduiront par l'échantillonnage de Gibbs compte tenu de la propriété de familles conjuguées.*

Conclusion

Après avoir rappelé quelques méthodes d'estimation des données de comptage dans le cas unidimensionnel, nous nous intéresserons dans les chapitres suivants à la modélisation de ces données dans le cas général multidimensionnel et en particulier dans le cas bidimensionnel. Cependant, l'extension au cas multidimensionnel des modèles précités n'est pas immédiate et s'avère plus complexe.

CHAPITRE 3

QUELQUES CONTRIBUTIONS SUR LA LOI MULTIDIMENSIONNELLE DE POISSON-SKELLAM

Nous introduisons une nouvelle forme de distribution dont les composantes ont pour distributions marginales des lois de Poisson ou des lois de Skellam. Cette nouvelle spécification permet d'incorporer de l'information pertinente sur la nature des corrélations entre toutes les composantes. De plus, nous présentons certaines propriétés de ladite distribution. Contrairement à la distribution multidimensionnelle de Poisson, celle-ci permet de traiter les variables avec des corrélations positives et négatives. Une simulation permettra d'illustrer les méthodes d'estimation. Enfin, une application sur des données de football mettra en relief les liens existant entre le nombre de points par saison et la différence de buts par des variables explicatives sélectionnées.

Introduction

L'utilisation des données discrètes multidimensionnelles demeure vaste tant en biologie, en épidémiologie qu'en sciences environnementales et dans bien d'autres domaines connexes. Cependant, jusqu'à aujourd'hui, leurs utilisations demeurent assez restreintes à cause de la forme complexe de leurs densités. Pour contourner ce problème, une des stratégies les plus courantes est l'approximation par des lois gaussiennes qui s'avèrent assez simples à manipuler. Cependant, l'approximation normale peut s'avérer trompeuse surtout dans le cas multidimensionnel où nous observons plus fréquemment de petites valeurs.
Aitchison et Ho (1989) ont introduit la loi multidimensionnelle log-normal comme étant un mélange de loi de Poisson avec une pondération bien spécifiée de la forme de densi-

tés gaussiennes multidimensionnelles. Certes elle a pour avantage de prendre en compte tant les corrélations positives que négatives, mais elle demeure inefficace dans certaines situations. En effet, selon ces auteurs, dans le cas des données bidimensionnelles, leur modèle ne peut décrire de façon fidèle les corrélations entre des variables de moyennes relativement faibles et inférieures à leurs variances respectives. La distribution multidimensionnelle de Poisson a fait l'objet d'une littérature abondante dont Mahamunulu (1967) et Johnson, Kotz et Balakrishnan (1997), pour ne citer que ceux-ci.

Les travaux de Tsionas (1999, 2001) et Karlis (2003) mettent en exergue une structure plutôt simplifiée de cette distribution dans laquelle toutes les variables possèdent la même covariance. Cependant, elle demeure assez restrictive car elle ne considère que les corrélations positives. Plus récemment, Karlis et Meligkotsidou (2005) ont proposé un cadre beaucoup plus général et flexible permettant de définir une structure libre pour chaque paire de variables. Mais, il ne traite que des corrélations positives et semble ignorer le phénomène de surdispersion.

Dans ce chapitre, nous définissons une nouvelle forme de distribution qui se veut plus flexible que le modèle multidimensionnel de Poisson. En effet, sous certaines hypothèses (si toutes les corrélations sont positives), nous obtenons la loi multidimensionnelle de Poisson comme un cas particulier. S'il existe au moins une corrélation négative, il s'agit alors de la loi multidimensionnelle de Poisson-Skellam. Cette approche nous permet donc de traiter des données multidimensionnelles discrètes de corrélations négatives et positives. De plus, dans certains cas, elle permet de tenir compte du phénomène de surdispersion très couramment observé sur ce type de données.

Tout d'abord, il s'agit de rappeler la définition et certaines propriétés de la distribution multidimensionnelle de Poisson. Cette section permettra de mettre en exergue certaines ressemblances avec la nouvelle distribution. Ensuite, nous introduirons la nouvelle distribution de Poisson-Skellam. En plus des propriétés de récurrence pour le calcul de sa den-

sité, d'autres propriétés et caractéristiques intéressantes y seront présentées. Par ailleurs, une méthode d'estimation basée sur l'augmentation des données permettra d'obtenir les différentes lois *a posteriori* nécessaires à l'inférence sur les paramètres du modèle considéré. Enfin, une simulation puis un exemple pour un modèle assez simple de dimension 2 permettront d'illustrer les aspects théoriques développés antérieurement.

3.1 Loi de Poisson multidimensionnelle

Cette partie constitue une brève présentation de la distribution multidimensionnelle de Poisson telle que présentée par Karlis et Meligkotsidou (2005).

3.1.1 Définitions

Définition 3.1.1. *Loi de Poisson multidimensionnelle*

Soit Y_r, $r = 1,\ldots,l$ *des variables aléatoires indépendantes de Poisson de paramètres respectifs* $\theta_1,\ldots,\theta_l$. *Soit* $\mathbf{A}$ *une matrice de taille* $m \times l$ *telle que* $\mathbf{A} = [\boldsymbol{\phi}_\mathbf{1},\ldots,\boldsymbol{\phi}_\mathbf{l}]$ *où* $\boldsymbol{\phi}_\mathbf{r} \in \mathbb{R}^m$ *et posons* $\mathbf{Y} = (Y_1,\ldots,Y_l)'$ *avec* $m \leq l$. *De plus, les composantes des vecteurs* $\boldsymbol{\phi}_\mathbf{j}$ *prennent leurs valeurs dans* $\{0,1\}$.

Alors $\mathbf{X} = \mathbf{AY}$ *suit une loi multidimensionnelle de Poisson. En outre,* $E(X) = \mathbf{A}\theta$ *et* $\text{var}(\mathbf{X}) = \mathbf{A}\Sigma\mathbf{A}'$ *où* $\theta = (\theta_1,\ldots,\theta_l)$, $\Sigma = diag(\theta_1,\ldots,\theta_l)$.

3.1.2 Exemple

Exemple 3.1.1. *Loi de dimension* 3, $m = 3$

Soit $Y_i \sim Poiss(\theta_i)$, $i \in \{1,2,3\}$ *et* $Y_{ij} \sim Poiss(\theta_{ij})$, $i,j \in \{1,2,3\}$ *avec* $i < j$. *Posons* $\theta = (\theta_1,\theta_2,\ldots,\theta_{23})'$. *En spécifiant la matrice* $\mathbf{A} = \begin{bmatrix} 1 & 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 & 1 \end{bmatrix}$ *et* $\mathbf{Y} = (Y_1,Y_2,Y_3,Y_{12},Y_{13},Y_{23})'$,

la transformation peut s'écrire :

$$\mathbf{X} = \mathbf{AY} \Leftrightarrow \begin{cases} X_1 = Y_1 + Y_{12} + Y_{13} \\ X_2 = Y_2 + Y_{12} + Y_{23} \\ X_3 = Y_3 + Y_{13} + Y_{23}. \end{cases}$$

La matrice de variance covariance est alors

$$\mathbf{A\Sigma A'} = \begin{bmatrix} \theta_1 + \theta_{12} + \theta_{13} & \theta_{12} & \theta_{13} \\ \theta_{12} & \theta_1 + \theta_{12} + \theta_{23} & \theta_{23} \\ \theta_{13} & \theta_{23} & \theta_3 + \theta_{13} + \theta_{23} \end{bmatrix}.$$

Malgré sa grande flexibilité lui permettant de spécifier la forme de la matrice de variance covariance (avec des termes uniquement positifs), cette distribution reste d'un point de vue numérique assez difficile à manipuler. Sa complexité analytique limite beaucoup son application. Par exemple, pour une loi multidimensionnelle m de Poisson, le calcul de la fonction de masse requiert $2^m - (m+1)$ sommations.

3.1.3 Distribution de probabilité sous forme matricielle

Considérons le vecteur $\mathbf{X}$ comme l'image de $\mathbf{Y}$ par une transformation g définie par :

$$\begin{aligned} g: \quad \mathbb{N}^l &\longrightarrow \mathbb{N}^m \\ \mathbf{Y} &\longmapsto \mathbf{X} = \mathbf{AY}. \end{aligned}$$

En utilisant le théorème de transfert sur la densité Y, facile à écrire, car constituée de variables aléatoires indépendantes de Poisson, nous aboutissons au résultat suivant :

$$\begin{aligned} P(X_1 = x_1, \ldots, X_m = x_m) &= \sum_{y \in g^{-1}(x)} \mathbf{P}(Y = y) \\ &= \sum_{y \in g^{-1}(x)} \prod_{i=1}^{l} \mathbf{P}(y_i | \theta_i) \\ &= \sum_{y \in g^{-1}(x)} \left[\prod_{i=1}^{l} \left(\frac{\theta_i^{y_i}}{y_i!} e^{-\theta_i} \right) \right], \end{aligned}$$

où $g^{-1}(x) = \left\{ \mathbf{y} \in \mathbb{N}^l | g(\mathbf{y}) = \mathbf{x} \right\}$.

Il convient de remarquer que lorsque m augmente, ce calcul peut s'avérer très difficile et lent. Cependant, Kano et Kawamura (1991) ont indiqué une méthode de calcul par récurrence un peu plus adaptée.

3.1.4 Propriétés et récurrences sur la distribution multidimensionnelle de Poisson

Kano et Kawamura (1991) ont mis en exergue plusieurs relations de récurrence sur la fonction de probabilité de la loi multidimensionnelle de Poisson. Karlis et Meligkotsidou (2005) rappellent que si $\mathbf{X} = \sum_{r=1}^{l} \phi_{\mathbf{r}} Y_r$, en posant $\mathbf{A} = [\phi_{\mathbf{1}}, \ldots, \phi_{\mathbf{l}}]$ et aussi $\theta^* = (\theta_1 P(X = \mathbf{x} - \phi_{\mathbf{1}}), \ldots, \theta_l P(X = \mathbf{x} - \phi_{\mathbf{l}}))'$, les m relations de récurrence sont données par l'écriture matricielle $\mathbf{x}P(X = \mathbf{x}) = \mathbf{A}\theta^*$. Il faudrait préciser que même dans la forme la plus simple de la matrice $\mathbf{A}$, le calcul demeure assez complexe. Pour $\mathbf{A} = [\mathbf{A_1}, \mathbf{A_2}]$, où $\mathbf{A_1}$ désigne la matrice identité d'ordre m et $\mathbf{A_2} = [\phi_{\mathbf{m+1}}, \ldots, \phi_{\mathbf{l}}]$ la matrice qui introduit les termes de covariance, chaque ligne de $\mathbf{A}$ contient au maximum m fois le chiffre 1 ; ce qui exige le calcul préalable d'au moins m probabilités. De plus, les erreurs générées à chaque

itération, aussi petites soient elles, peuvent avoir des effets néfastes sur le résultat final lorsque m devient grand. Une stratégie serait donc d'utiliser ces relations de façon parcimonieuse à l'intérieur d'un algorithme de calcul tel que celui proposé par les précédents auteurs.

Exemple 3.1.2. *Relations de récurrence en dimension 3*
Nous reprenons l'exemple 3.1.1 afin de donner les relations de récurrence pouvant être utilisées dans le calcul de la probabilité.

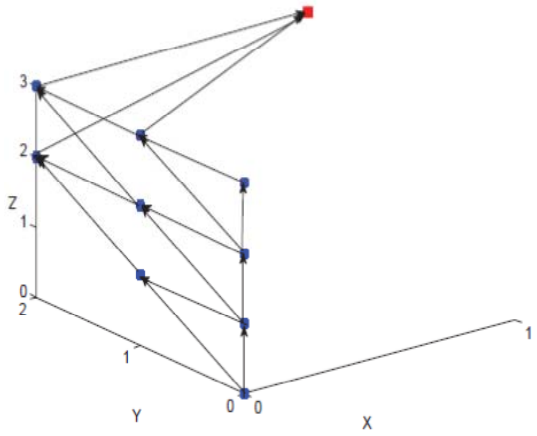

Figure 3.1 – Algorithme de calcul de la fonction de masse

La figure 3.1 présente un algorithme de calcul de la fonction de masse au point $(x_1, x_2, x_3) = (1, 2, 3)$. *Par la suite, notons* P_3 *la fonction de masse de la loi multidimensionnelle de Pois-*

son de dimension 3. *Les relations de récurrence s'écrivent donc*

$$\begin{aligned}
x_1 P_3(x_1,x_2,x_3) &= \theta_1 P_3(x_1-1,x_2,x_3)+\theta_{12} P_3(x_1-1,x_2-1,x_3) \\
&\quad +\theta_{13} P_3(x_1-1,x_2,x_3-1) \\
x_2 P_3(x_1,x_2,x_3) &= \theta_2 P_3(x_1,x_2-1,x_3)+\theta_{12} P_3(x_1-1,x_2-1,x_3) \\
&\quad +\theta_{23} P_3(x_1,x_2-1,x_3-1) \\
x_3 P_3(x_1,x_2,x_3) &= \theta_3 P_3(x_1,x_2,x_3-1)+\theta_{13} P_3(x_1-1,x_2,x_3-1) \\
&\quad +\theta_{23} P_3(x_1,x_2-1,x_3-1).
\end{aligned}$$

L'idée de l'algorithme consiste à atteindre le point de coordonnées $(1,2,3)$ *à partir du point désignant l'origine* $(0,0,0)$. *Pour cela, nous utilisons les 3 relations de récurrence. En effet, pour atteindre le point* $(1,2,3)$ *en débutant à l'origine* $(0,0,0)$, *une possibilité serait de passer par les points* $(0,0,1)$ *et* $(0,1,1)$ *pour atteindre le point* $(0,1,1)$. *Ensuite, il faut procéder de façon analogue en utilisant les relations précédentes jusqu'à aboutir au point* $(1,2,3)$.

3.2 Loi de Poisson-Skellam multidimensionnelle

Pour remédier aux problèmes de surdispersion et de corrélations négatives, Meligkotsidou (2007) propose un modèle de mélange de lois multidimensionnelles de Poisson dont le nombre de composantes peut être inconnu. Cependant, la complexité numérique du modèle et le fait que les données peuvent ne pas toujours être représentées par un mélange de lois rendent le modèle assez difficile à mettre en oeuvre.

La nouvelle approche que nous proposons a pour objectif de prendre en compte à la fois les corrélations tant négatives que positives. Cependant, elle exige un effort supplémentaire car faisant appel à l'utilisation simultanée des distributions de Poisson et de Skellam

(voir Skellam, 1946).

Définition 3.2.1. *Loi de Skellam*

Soit U et V deux variables aléatoires indépendantes de Poisson de paramètres respectifs θ_1 et θ_2, alors $Z = U - V$ suit une loi de Skellam de paramètres (θ_1, θ_2) notée $Z \sim Skellam(\theta_1, \theta_2)$. La densité de Z s'écrit alors

$$P(Z = z|\theta_1, \theta_2) = \frac{\theta_1}{\theta_2} e^{-(\theta_1+\theta_2)} \mathscr{I}_{|z|}\left(2\sqrt{\theta_1\theta_2}\right),$$

où $\mathscr{I}_{|z|}(x) = \sum_{m=0}^{\infty} \frac{(x/2)^{2m+z}}{m!\Gamma(m+z+1)}$ désigne la fonction modifiée de Bessel de première espèce.

Cette densité n'est pas très utilisée en pratique car le calcul demeure assez difficile compte tenu de la présence de la fonction de Bessel modifiée.

Un avantage de la loi de Skellam est la fermeture par addition ou par soustraction. Ainsi, toute somme ou différence finie de variables aléatoires de lois de Skellam suit une loi de Skellam.

Proposition 3.2.1. *Fermeture des lois de Skellam par addition et soustraction*

Soient Z_1 et Z_2 deux variables aléatoires indépendantes de lois de Skellam de paramètres respectifs (θ_1, θ_2) et (θ_3, θ_4). Alors leur somme et leur différence sont aussi des variables aléatoires de lois de Skellam. En effet :

$$Z_1 + Z_2 \sim Skellam(\theta_1 + \theta_3, \theta_2 + \theta_4),$$

$$Z_1 - Z_2 \sim Skellam(\theta_1 + \theta_4, \theta_2 + \theta_3).$$

Démonstration. Il existe deux couples de variables aléatoires indépendantes (Y_1, Y_2) et (Y_3, Y_4) de lois de Poisson de paramètres respectifs (θ_1, θ_2) et (θ_3, θ_4) tels que $Z_1 = Y_1 - Y_2$

et $Z_2 = Y_3 - Y_4$. Alors

$$\begin{aligned} Z_1 + Z_2 &= (Y_1 - Y_2) + (Y_3 - Y_4) \\ &= \underbrace{(Y_1 + Y_3)}_{\sim Poiss(\theta_1+\theta_3)} - \underbrace{(Y_2 + Y_4)}_{\sim Poiss(\theta_2+\theta_4)} . \end{aligned}$$

De façon similaire nous obtenons :

$$\begin{aligned} Z_1 - Z_2 &= (Y_1 - Y_2) - (Y_3 - Y_4) \\ &= \underbrace{(Y_1 + Y_4)}_{\sim Poiss(\theta_1+\theta_4)} - \underbrace{(Y_2 + Y_3)}_{\sim Poiss(\theta_2+\theta_3)} . \end{aligned}$$

Il est maintenant possible de conclure en sachant que la distribution de Skellam n'est rien d'autre qu'une différence de variables aléatoires indépendantes de lois de Poisson. □

3.2.1 Un modèle bidimensionnel hybride

Nous présentons dans cette section un modèle assez élémentaire mais qui pourrait servir de base pour les analyses futures. Le cas bidimensionnel demeure le plus simple compte tenu de la forme de la densité. Soit Y_i, $i \in \{1, 2, 12\}$ des variables aléatoires indépendantes de Poisson de paramètres respectifs θ_1, θ_2 et θ_{12}. Posons :

$$\begin{aligned} X_1 &= Y_1 + Y_{12}, \\ X_2 &= Y_2 + \delta_{21} Y_{12}, \end{aligned}$$

où δ_{21} est une constante prenant deux valeurs possibles $\{-1, 1\}$. Plus précisément, $\delta_{21} = 1$ si la corrélation entre les vecteurs X_1 et X_2 est positive et $\delta_{21} = -1$ si leur corrélation est plutôt négative.

En effet, dans le cas où les données sont positivement corrélées, la densité est celle de la loi bidimensionnelle de Poisson. Par contre, si la corrélation est négative, alors la densité obtenue est :

$$\begin{aligned} P(X_1 = x_1, X_2 = x_2) &= \sum_{y \in g^{-1}(x)} P(Y_1 = y_1, Y_2 = y_2, Y_{12} = y_{12}) \\ &= \sum_{y \in g^{-1}(x)} \prod_{i \in \{1,2,12\}} Poiss(y_i|\theta_i) \\ &= \exp(-\theta_1 - \theta_2 - \theta_{12}) \sum_{y_{12}=0 \vee -x_2}^{x_1} \frac{\theta_1^{x_1 - y_{12}}}{(x_1 - y_{12})!} \frac{\theta_2^{x_2 + y_{12}}}{(x_2 + y_{12})!} \frac{\theta_{12}^{y_{12}}}{y_{12}!} \quad , \end{aligned}$$

où $a \vee b = \max(a,b)$.

Cette section présente quelques résultats et propriétés intéressantes caractérisant la loi bidimensionnelle de Poisson-Skellam. En effet, des relations de récurrence pouvant caractériser la fonction de densité de ladite distribution y sont présentées.

Par souci de simplification, posons $P(k|\lambda) \equiv Poiss(X = k|\lambda) = \frac{\lambda^k e^{-\lambda}}{k!}$ la fonction de masse de la loi de Poisson de paramètre λ et $P(X_1 = k, X_2 = l) = P_2(k,l)$ celle de la loi bidimensionnelle négative de paramètres $(\theta_1, \theta_2, \theta_{12})$:

$$\begin{aligned} P_2(k,l) &= \exp(-\theta_1 - \theta_2 - \theta_{12}) \sum_{\delta=0 \vee -l}^{k} \frac{\theta_1^{k-\delta}}{(k-\delta)!} \frac{\theta_2^{l+\delta}}{(l+\delta)!} \frac{\theta_{12}^{\delta}}{\delta!} \\ &= \sum_{\delta=0 \vee -l}^{k} P(k-\delta|\theta_1) P(l+\delta|\theta_2) P(\delta|\theta_{12}). \end{aligned}$$

Lemme 3.2.1. *Récurrence usuelle de la loi de Poisson*

Soit X une variable aléatoire de loi de Poisson de paramètre λ notée $X \sim Poiss(\lambda)$. Alors,

$$kP(k|\lambda) = \lambda P(k-1|\lambda) \quad \forall k \geq 1.$$

Démonstration. Soit k un nombre entier naturel non nul, alors après multiplication de la fonction de masse de la loi de Poisson, nous avons successivement

$$\begin{aligned} kP(k|\lambda) &= k\frac{e^{-\lambda}\lambda^k}{k!} \\ &= \frac{e^{-\lambda}\lambda^k}{(k-1)!} \\ &= \lambda\frac{e^{-\lambda}\lambda^{k-1}}{(k-1)!} \\ &= \lambda P(k-1|\lambda). \end{aligned}$$

□

Proposition 3.2.2. *Relations de récurrence*

La fonction de masse de la loi bidimensionnelle Poisson-Skellam $P_2(k,l)$ vérifie les relations de récurrence suivantes :

$$kP_2(k,l) = \theta_1 P_2(k-1,l) + \theta_{12} P_2(k-1,l+1) \tag{3.2.1}$$

$$lP_2(k,l) = \theta_2 P_2(k,l-1) - \theta_{12} P_2(k-1,l+1), \tag{3.2.2}$$

pour tout $(k,l) \in \mathbb{N}^* \times \mathbb{Z}$ où $\mathbb{N}^*$ désigne l'ensemble des entiers naturels non nuls.

Démonstration. Voir annexe A. □

Proposition 3.2.3. *La fonction de masse de la loi bidimensionnelle Poisson-Skellam $P_2(k,l)$ vérifie les relations suivantes,*

$$kP_2(k,-k) = \theta_{12} P_2(k-1,-[k-1]) \tag{3.2.3}$$

$$lP_2(0,l) = \theta_2 P_2(0,l-1), \tag{3.2.4}$$

pour tout $(k,l) \in \mathbb{N}^* \times \mathbb{Z}$.

Démonstration. Il suffit de remarquer que :

$$\begin{aligned} P(X_1 = k, X_1 = -k) &= P(Y_1 + Y_{12} = k, Y_2 - Y_{12} = -k) \\ &= P(Y_1 = 0, Y_2 = 0, Y_{12} = k). \end{aligned}$$

Ainsi nous obtenons,

$$\begin{aligned} kP_2(k,-k) &= kP(0|\theta_1)P(0|\theta_2)P(k|\theta_{12}) \\ &= \theta_{12}P(0|\theta_1)P(0|\theta_2)P(k-1|\theta_{12}) \\ &= \theta_{12}P_2(k-1,-[k-1]). \end{aligned}$$

Quant à la seconde relation, le résultat découle directement de la définition de la loi bidimensionnelle Poisson-Skellam. En effet,

$$\begin{aligned} lP_2(0,l) &= lP(0|\theta_1)P(l|\theta_2)P(0|\theta_{12}) \\ &= \theta_2 P(0|\theta_1)P(l-1|\theta_2)P(0|\theta_{12}) \\ &= \theta_2 P_2(0,l-1). \end{aligned}$$

□

3.2.2 Modèle standard

Définition 3.2.2. *Loi de Poisson-Skellam standard*

Soient Y_i, $i = 0,\ldots,m$ des variables aléatoires indépendantes de Poisson de paramètres

respectifs $\theta_0, \ldots, \theta_m$. *Posons :*

$$\begin{aligned} X_1 &= Y_1 + \delta_1 Y_0, \\ X_2 &= Y_2 + \delta_2 Y_0, \\ &\vdots \\ X_m &= Y_m + \delta_m Y_0, \end{aligned}$$

où δ_i *est une constante prenant deux valeurs possibles* $\{-1,1\}$. *Soit* $\mathbf{A}$ *la matrice de taille* $m \times (m+1)$ *constituée de la matrice identité d'ordre m à laquelle l'on ajoute un vecteur* $\Delta = (\delta_1, \ldots, \delta_m)'$ *de* R^m. *Posons* $\theta = (\theta_1, \ldots, \theta_m, \theta_0)'$
Alors le vecteur $\mathbf{X} = (X_1, \ldots, X_m)$ *suit une loi multidimensionnelle Poisson-Skellam standard notée* $MPSS(\mathbf{A}, \theta)$.
Par la suite, notons $I^- = \{i \in \{1, \ldots, m\} : \delta_i = -1\}$ *et de façon analogue* $I^+ = \{i \in \{1, \ldots, m\} : \delta_i = 1\}$.
Alors la loi marginale de la i^e *composante du vecteur* $\mathbf{X}$ *est soit une Poisson, soit une Skellam, c'est-à-dire une différence de Poisson, selon que* δ_i *est positif ou négatif*

$$X_i \sim \begin{cases} Skellam(\theta_i, \theta_0) & si\ i \in I^-, \\ Poiss(\theta_i + \theta_0) & si\ i \in I^+. \end{cases}$$

En outre, $E(X) = \theta + \theta_0 \Delta$ *et* $\mathrm{var}(X) = \theta_0 \mathbf{I_m} + \Delta\Delta'$.

Sous forme matricielle, la définition précédente peut se résumer comme suit

$$\begin{aligned} g : \quad \mathbb{N}^{m+1} &\longrightarrow \mathbb{Z}^m \\ \mathbf{Y} = (Y_1, \ldots, Y_m, Y_0)' &\longmapsto \mathbf{X} = \mathbf{AY} \end{aligned}$$

avec $\mathbf{A} = (\mathbf{I_m}, \Delta)$.

La densité du vecteur $\mathbf{X}$ est donc

$$\begin{aligned} P(X_1 = x_1, \ldots, X_m = x_m) &= \sum_{\mathbf{y} \in g^{-1}(\mathbf{x})} P(Y = y) \\ &= \sum_{\mathbf{y} \in C[\mathbf{x}]} \left(\prod_{i=1}^{m} Poiss\left(x_i - \delta_i y_0 | \theta_i\right) \right) Poiss\left(y_0 | \theta_0\right), \end{aligned}$$

où $C[\mathbf{x}] = \{y_0 \in \mathbb{N} : \max(-x_i, i \in I^-) \leq y_0 \leq \min(x_i, i \in I^+)\}$. Ainsi, en notant par convention $x_0 = 0$ et $\delta_0 = -1$, la densité peut s'écrire comme suit :

$$P(X_1 = x_1, \ldots, X_m = x_m) = \sum_{\mathbf{y_0} \in C[\mathbf{x}]} \left(\prod_{i=0}^{m} Poiss\left(x_i - \delta_i y_0 | \theta_i\right) \right).$$

Remarque 3.2.1. *Tout d'abord, la corrélation entre deux composantes X_i et X_j est donnée par $\rho_{i,j} = \frac{\delta_i \delta_j \theta_0}{\sqrt{(\theta_i + \theta_0)(\theta_j + \theta_0)}}$. Le signe de cette corrélation dépend donc du produit $\delta_i \delta_j$. La valeur absolue de cette corrélation est identique à celle du modèle multidimensionnel de Poisson considéré par Tsiamyrtzis et Karlis (2004) lorsqu'ils proposent un algorithme efficace de calcul de densité correspondante.*

Le modèle standard possède à la fois des avantages et des inconvénients. Certes, il prend en compte le signe des corrélations entre les variables prises deux à deux mais il impose *a fortiori* une structure commune de la covariance prise en valeur absolue.

3.2.3 Généralisation

Cette section présente une généralisation de la loi multidimensionnelle de Poisson-Skellam. En effet, le modèle standard suppose une covariance identique entre toutes les composantes du vecteur aléatoire. Cette nouvelle spécification permet de spécifier un

terme de covariance différent pour chaque paire de composantes.

Définition 3.2.3. *Loi de Poisson-Skellam*

Soit $\mathbf{X}$ *un vecteur aléatoire de loi multidimensionnelle de Poisson-Skellam notée* $\mathbf{X} \sim MVPS(\mathbf{A}, \theta)$ *où* $\mathbf{A} = (\phi_1, \phi_2, \ldots, \phi_l)$ *désigne une matrice de taille* $m \times l$. *Alors il existe des variables aléatoires indépendantes de Poisson* Y_r, $r = 1, \ldots, l$ *de paramètres respectifs* $\theta_1, \ldots, \theta_l$ *tels que* $\mathbf{X} = \mathbf{AY}$ *où* $\mathbf{Y} = (Y_1, \ldots, Y_l)'$ *avec* $m \leq l$. *En outre,* $\mathbf{A} = [\mathbf{A}_1, \mathbf{A}_2]$ *où* $\mathbf{A}_1$ *est la matrice identité de taille m et la matrice* $\mathbf{A}_2$ *correspond à la matrice des signes de la corrélation. Les deux premiers moments de* $\mathbf{X}$ *s'écrivent* $E(\mathbf{X}) = \mathbf{A}\theta$ *où* $\theta = (\theta_1, \ldots, \theta_m)$ *et* $\text{var}(\mathbf{X}) = \mathbf{A}\Sigma\mathbf{A}'$ *où* $\Sigma = \text{diag}\,(\theta_1, \ldots, \theta_k)$.

Proposition 3.2.4. *Lois marginales*

La nouvelle distribution multidimensionnelle ainsi obtenue est un mélange de Poisson et de Skellam. Si nous notons :

$S^i_+ = \left\{ j > i \in \{1, \ldots, k\} : \text{cov}(X_i, X_j) > 0 \right\}$ *l'ensemble des indices dont les variables ont une corrélation positive avec* X_i *et,*

$S^i_- = \left\{ j < i \in \{1, \ldots, k\} : \text{cov}(X_i, X_j) < 0 \right\}$ *indexant celles qui sont négativement corrélées à* X_i .

Alors, $X_i = \sum_{j \in S^i_+} Y_{ij} - \sum_{j \in S^i_-} Y_{ij}$ *et la distribution marginale de chacune des composantes* X_i *peut s'écrire comme suit :*

$$X_i \sim \begin{cases} Skellam\left(\sum\limits_{j \in S^i_+} \theta_{ij}, \sum\limits_{j \in S^i_-} \theta_{ij} \right) & si\ \text{card}\left(S^i_-\right) \geq 1, \\ Poiss\left(\sum_{j=1}^{k} \theta_{ij} \right) & sinon. \end{cases}$$

La fonction jointe de probabilité s'avère assez complexe à écrire mais compte tenu de notre approche bayésienne, sa forme explicite ne nous est pas tant utile. Par ailleurs, il

n'existe pas dans la littérature de formule de récurrence lorsque les corrélations sont négatives, c'est-à-dire que les éléments de la matrice $\mathbf{A_2}$ prennent leurs valeurs dans $\{0,1,-1\}$. Cependant, nous présenterons dans la section suivante quelques résultats sur la densité de la loi multidimensionnelle Poisson-Skellam.

Proposition 3.2.5. *Généralisation des relations de récurrence*

Si $\mathbf{X} = (X_1,\ldots,X_m)$ *suit une* $MVPS(\mathbf{A},\theta)$*, alors pour tout entier* r *tel que* $1 \leq r \leq rang(\mathbf{A})$

$$P(X = x) = \sum_{i=1}^{l-r+1} \theta_i s_i P(X = x - \phi_i),$$

où $\mathbf{x^{(r)}}$ *désigne la troncature de dimension* r *du vecteur original* $\mathbf{x}$ *et*

$$s_i = \frac{\det\left(\phi_{\mathbf{i}}^{(\mathbf{r})}, \phi_{\mathbf{l-r+2}}^{(\mathbf{r})}, \phi_{\mathbf{l-r+3}}^{(\mathbf{r})}, \ldots, \phi_{\mathbf{l}}^{(\mathbf{r})},\right)}{\det\left(\mathbf{x^{(r)}}, \phi_{\mathbf{l-r+2}}^{(\mathbf{r})}, \phi_{\mathbf{l-r+3}}^{(\mathbf{r})}, \ldots, \phi_{\mathbf{l}}^{(\mathbf{r})},\right)} \quad i = 1,\ldots,l-r+1.$$

Démonstration. Voir annexe A. □

Corollaire 3.2.1. *En particulier, pour* $r = 1$*, nous avons :*

$$P(X = x) = \sum_{i=1}^{l} \frac{a_{it}}{x_t} \theta_i P(X = \mathbf{x} - \phi_{\mathbf{i}}).$$

Démonstration. Dans le cas $r = 1$, les vecteurs $\phi_{\mathbf{i}}^{(\mathbf{r})}$ sont réduits à de simples nombres réels et le résultat en découle.

□

3.3 Méthodes d'estimation

Le cas bidimensionnel demeure le plus simple compte tenu de la forme de la densité. En effet, dans le cas où les données sont positivement corrélées, la densité est celle de la

loi bidimensionnelle de Poisson. Par contre, si la corrélation est négative, alors la densité obtenue est :

$$\begin{aligned} P(X_1 = x_1, X_2 = x_2) &= \sum_{y \in g^{-1}(x)} \mathbf{P}(Y_1 = y_1, Y_2 = y_2, Y_{12} = y_{12}) \\ &= \sum_{y \in g^{-1}(x)} \prod_{i \in 1,2,12} Poiss(y_i | \theta_i) \\ &= \exp(-\theta_1 - \theta_2 - \theta_{12}) \sum_{y_{12}=0 \vee -x_2}^{x_1} \frac{\theta_1^{x_1 - y_{12}}}{(x_1 - y_{12})!} \frac{\theta_2^{x_2 + y_{12}}}{(x_2 + y_{12})!} \frac{\theta_{12}^{y_{12}}}{y_{12}!}. \end{aligned}$$

La contrainte majeure de cette forme de distribution est qu'elle postule *ipso facto* une surdispersion sur au moins une des composantes du vecteur X.

Notre objectif est d'introduire des variables explicatives afin d'expliquer les paramètres définis par θ_1, θ_2 et θ_{12} sous forme d'un modèle linéaire généralisé. Le modèle général peut donc s'écrire de la forme suivante :

$$\log\left(\theta_{ij}\right) = Z_{ij}\beta_j \;\; i = 1, \ldots, n \;, j = 1, 2, 12.$$

3.3.1 Méthode par le maximum de vraisemblance

Nous proposons l'estimation du maximum de vraisemblance par l'algorithme EM (Espérance-Maximisation) décrit par Tanner et Wong (1987). Pour la suite, posons $\Theta^{(r)} = \left(\beta_1^{(r)}, \ldots, \beta_k^{(r)}\right)$ le vecteur des paramètres obtenus après la r^e itération. Les étapes de l'algorithme se présentent comme suit :

Étape 1 **: Espérance**

Elle consiste à calculer l'espérance de la variable latente et représente l'étape la plus com-

plexe à cause de la forme de la loi considérée. Posons

$$\begin{aligned} s_{ij} &= E\left(Y_{ij}|\mathbf{Z_i}, \Theta^{(r-1)}\right) \\ &= \frac{\sum_{\mathbf{y_i} \in g^{-1}(\mathbf{x_i})} y_{ij} \prod_{j=1}^{k} Poiss(y_{ij}; \theta_{ij}^{(r-1)})}{\sum_{\mathbf{y_i} \in g^{-1}(\mathbf{x_i})} \prod_{j=1}^{k} Poiss(y_{ij}; \theta_{ij}^{(r-1)})}, \end{aligned}$$

où $\theta_{ij}^{(r-1)} = \exp\left(Z_{ij}\beta_j^{(r-1)}\right)$.

Étape 2 **: Maximisation**

Cette étape consiste à effectuer une simple régression de Poisson en utilisant les s_{ij} comme variable dépendante et les variables z_{ij} comme variables explicatives.

Il faut ensuite itérer les étapes précédentes jusqu'à la convergence de l'algorithme. Il est important de préciser que le choix des valeurs initiales augmente considérablement la vitesse de convergence de l'algorithme. En effet, les estimateurs obtenus par une régression de Poisson pourraient être considérés comme valeurs initiales des paramètres de la moyenne. Quant à ceux de la covariance, il faut d'abord penser à l'estimateur empirique de la covariance du vecteur des observations.

De simples statistiques de Wald permettent de tester si les coefficients obtenus sont significativement non nuls. Pour ce faire, les écarts types des estimateurs sont donc obtenus soit par des méthodes de ré-échantillonnage, soit par la dérivée seconde du logarithme de la vraisemblance.

3.3.2 Méthode bayésienne

L'approche bayésienne réalisée par CMMC permet d'estimer non seulement les paramètres mais aussi toute statistique issue de leur distribution *a posteriori*. Elle est basée sur l'algorithme de Metropolis-Hastings et l'échantillonnage de Gibbs.

Algorithme 3.3.1. *Algorithme de Metropolis-Hastings*

L'algorithme de Metropolis-Hastings (Metropolis et al., 1953 et Hastings, 1970) associé à la loi objectif f et la loi conditionnelle ou instrumentale q permet de produire une chaîne de Markov $(\theta^{(t)})$ convergente vers la loi stationnaire f.

- *Initialisation : Choisir une valeur initiale* $\theta^{(0)}$

- *Itération* t $(t \geq 1)$ *: étant donné* $\theta^{(t)}$,
  Générer $\eta_t \sim q\left(\eta|\theta^{(t)}\right)$.
  Prendre
  $$\theta^{(t+1)} = \begin{cases} \eta_t & \text{avec probabilité } \rho\left(\theta^{(t)}, \eta_t\right) \\ \theta^{(t)} & \text{avec probabilité } 1-\rho\left(\theta^{(t)}, \eta_t\right) \end{cases}$$
  où $\rho\left(\theta^{(t)}, \eta_t\right) = \min\left\{1, \frac{f(\eta_t)}{f\left(\theta^{(t)}\right)} \frac{q\left(\theta^{(t)}|\eta_t\right)}{q\left(\eta_t|\theta^{(t)}\right)}\right\}$

Quant à l'échantillonnage de Gibbs, il peut être interprété comme un cas particulier de l'algorithme de Metropolis-Hastings dont la probabilité d'acceptation est toujours égale à 1.

Algorithme 3.3.2. *Échantillonnage de Gibbs*

Soit $\pi(\theta_1, \ldots, \theta_p)$ *la distribution jointe de* $\theta = (\theta_1, \ldots, \theta_p)$ *avec les densités conditionnelles* $\pi_1, \ldots, \pi_p$ *où* π_j *représente la distribution de* θ_j *conditionnellement au vecteur* $\theta_{-\mathbf{j}} \equiv \left(\theta_1, \ldots, \theta_{j-1}, \theta_{j+1}, \ldots, \theta_p\right)$. *L'échantillonnage de Gibbs consiste à tirer successivement dans toutes les densités conditionnelles en ne changeant qu'une seule composante du vecteur* θ *à la fois.*

- *Initialisation : choisir une valeur initiale* $\theta^{(\mathbf{0})} = \left(\theta_1^{(0)}, \ldots, \theta_p^{(0)}\right)$

- *Itération t : étant donné* $\left(\theta_1^{(t-1)}, \ldots, \theta_p^{(t-1)}\right)$, *générer*
  $\theta_1^{(t)}$ *selon* $\pi_1\left(\theta_1 | \theta_{-\mathbf{1}}^{\mathbf{t}}\right)$,
  $\theta_2^{(t)}$ *selon* $\pi_2\left(\theta_2 | \theta_{-\mathbf{2}}^{\mathbf{t}}\right)$,
  $\vdots$
  $\theta_p^{(t)}$ *selon* $\pi_p\left(\theta_p | \theta_{-\mathbf{p}}^{\mathbf{t}}\right)$,
  où $\theta_{-\mathbf{i}}^{\mathbf{t}} = \left(\theta_1^{(t)}, \ldots, \theta_{i-1}^{(t)}, \theta_{i+1}^{(t-1)}, \ldots, \theta_{-p}^{(t-1)}\right)$.

Lois *a priori*

Elles permettent d'incorporer de l'information sur la nature ou le signe des coefficients à estimer grâce à des études similaires antérieures ou à l'avis des experts. Cependant, lorsqu'aucune information n'est disponible, il est possible d'utiliser une *loi a priori* vague qui consiste à privilégier l'indépendance des coefficients et à leur attribuer une très grande variance. Nous supposons une loi *a priori* de la forme :

$$\beta_j \sim \mathscr{N}\left(\mu_j, \mathbf{V}_j\right) \quad j = 1, 2, 12.$$

où μ_j et la matrice $\mathbf{V}_j$ sont supposés connus. Notons que $\mathbf{V}_j$ représente la précision ou la confiance des résultats obtenus dans le passé. Une loi non informative consisterait à choisir des variances assez grandes afin de ne pas contraindre les paramètres.

Loi de probabilité de $y_{i,12}$

L'objectif de la méthode d'estimation consiste à simuler les variables latentes y_1, y_2 et y_{12} pour ensuite mener l'inférence sur ces composantes. Cependant, il ne suffit que de simuler

la composante y_{12} pour en déduire les deux autres.

$$\begin{aligned} P(y_{i,12}|\mathbf{X_i},\beta,\mathbf{Z},) &= \frac{\prod_{j\in\{1,2,12\}} Poiss(y_{ij};\theta_{ij})}{\sum_{y_i\in g^{-1}(x_i)}\prod_{\in\{1,2,12\}} Poiss(y_{ij};\theta_{ij})} \\ &= \frac{\prod_{j\in\{1,2,12\}} \frac{e^{(Z_{ij}\beta_j)y_{ij}}e^{-e^{Z_{ij}\beta_j}}}{y_{ij}!}}{\sum_{y_i\in g^{-1}(x_i)}\prod_{j\in\{1,2,12\}} \frac{e^{(Z_{ij}\beta_j)y_{ij}}e^{-e^{Z_{ij}\beta_j}}}{y_{ij}!}}. \end{aligned}$$

Distribution *a posteriori* de $\beta_j, j \in \{1,2,12\}$

$$\begin{aligned} P\left(\beta_j|\mathbf{Y},\mathbf{Z}\right) &\propto \pi\left(\beta_j\right)\prod_{i=1}^{n} Poiss(y_{ij};\theta_{ij}) \\ &\propto \exp\left\{-\frac{1}{2}\left[\beta_j-\left(\mu_j+\mathbf{V_j'}\sum_{i=1}^{n}y_{ij}Z_{ij}\right)\right]'\mathbf{V_j^{-1}}\left[\beta_j-\left(\mu_j+\mathbf{V_j'}\sum_{i=1}^{n}y_{ij}Z_{ij}\right)\right]\right\} \\ &\quad\times\exp\left\{-\sum_{i=1}^{n}e^{Z_{ij}\beta_j}\right\}. \end{aligned}$$

La matrice de variance covariance de la densité instrumentale $\tilde{\Sigma}_j = c_j\Sigma_j$ où Σ_j est la matrice de covariance asymptotique du maximum de vraisemblance de β_j et c_j pour $j = 1,\ldots,k$ est un scalaire choisi de façon à calibrer l'amplitude des mouvements de l'algorithme de Metropolis-Hastings. Les valeurs de c_j devraient être ajustées de sorte qu'après la période de chauffe, le taux d'acceptation soit compris entre 0,15 et 0,5 selon Roberts, Gelman et Gilks (1997), Roberts et Rosenthal (2001). De plus, l'allure des densités conditionnelles évaluées au maximum de vraisemblance suggère une allure symétrique et même gaussienne de chacune des composantes de l'EMV conditionnellement aux autres.

Dey, Ghosh and Mallick (2000) préconisent, pour une convergence plus rapide, l'uti-

lisation des EMV de β_j comme valeurs initiales pour l'échantillonnage de Gibbs. Cette technique consisterait donc à introduire de nouvelles données non observables souvent appelées variables auxiliaires ou latentes de façon à résoudre un problème d'optimisation déterministe ou stochastique. Cependant, la forme de la distribution requérant des techniques plutôt sophistiquées (algorithme EM) pour trouver l'EMV, nous proposons d'utiliser plutôt la distribution marginale de X_1 et X_2 qui sont respectivement des lois de Poisson et de Skellam.

3.4 Simulation dans un exemple en dimension 2

Modèle avec corrélation négative

Le modèle s'écrit :

$$\log\left(\theta_{ij}\right) = \beta_j Z_{ij} \text{ avec } \beta_j = \left(\beta_j^0; \beta_j^1\right) \quad i = 1,2,\ldots,n \quad j \in \{1,2,12\}.$$

Une simulation met en évidence la performance de la méthode bayésienne. Les valeurs réelles des paramètres sont $\beta_1 = (-1{,}28; 3{,}41)$, $\beta_2 = (-2{,}16; 6{,}56)$, et $\beta_{12} = (-1{,}08; 0{,}86)$. Les variables explicatives ont été simulées selon des lois uniformes sur l'intervalle $(0;1)$ et nous avons privilégié une loi *a priori* assez vague sur les paramètres. Les résultats des simulations sont consignés dans le tableau 3.1. Les valeurs entre parenthèses représentent les écarts-types des estimateurs correspondants. Nous avons réalisé l'algorithme de Metropolis-Hastings avec marche aléatoire avec 2 500 000 itérations avec une période de chauffe de 10 000 itérations. La loi instrumentale utilisée est celle de la loi normale de paramètres de centrage et de dispersion respectifs les estimateurs du maximum de vraisemblance et leur variance asymptotique.

La figure 3.2 présente le graphique de la trace des échantillons tirés dans les lois *a*

Tableau 3.1 – Estimation des coefficients de la régression ; les écarts-types sont notés en parenthèses

coefficients	β_1^0	β_1^1	β_2^0	β_2^1	β_{12}^0	β_{12}^1
valeurs	-1,28	2,41	2,16	1,56	1,08	0,86
Maximum de vraisemblance	-1,3066 (0,0609)	2,4255 (0,1262)	2,1630 (0,0254)	1,5785 (0,0527)	1,0657 (0,0669)	0,8909 (0,1181)
Bayes	-1,2989 (0,0575)	2,4116 (0,1190)	2,1691 (0,0238)	1,5639 (0,0487)	1,0787 (0,0624)	0,8786 (0,0989)
borne inférieure à 95%	-1,4086	2,1861	2,1215	1,4709	0,9507	0,6911
borne supérieure à 95%	-1,1780	2,6591	2,2156	1,6616	1,1970	1,0780

posteriori respectives de β_1^0 et β_1^1. Le graphique suggère donc que les chaînes convergent après 500 000 itérations.

Compte tenu du fait que l'algorithme de Metropolis-Hastings produit des échantillons non indépendants, seule une observation par tranche de 50 a été retenue. En effet, la figure 3.3 montre que l'autocorrélation de l'échantillon constitué des observation par pas de 50 n'est pas significative. Une autre façon de contourner ce problème serait d'adopter un algorithme d'acceptation rejet adaptatif qui permettrait de tirer directement les échantillons selon la loi cible.

3.5 Applications

3.5.1 Données

Nous analysons les données de dix équipes de soccer de la première ligue anglaise durant les saisons de 2002-2003 à 2008-2009. Les données ont été recueillies sur le site www.soccernet.com.

Karlis et Ntzoufras (2003, 2009), McHale et Scarf (2007) ont montré que la variable aléatoire du différentiel de buts désignant la différence de buts marqués et de buts encaissés suivait une loi de Skellam. Karlis et Ntzoufras (2006) ont utilisé aussi le même principe

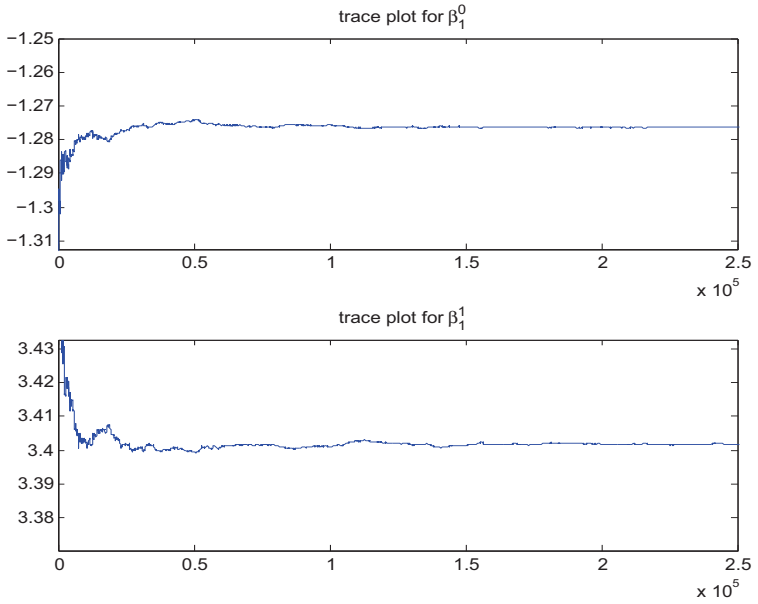

Figure 3.2 – Graphique des valeurs simulées

pour modéliser des données discrètes dans le cas de la santé dentaire. De plus, il parait légitime de penser que le nombre de points obtenus à la fin de saison pour une équipe donnée est positivement corrélée au différentiel de buts. Ainsi, les variables dépendantes retenues pour notre étude sont :

- le nombre total de points par saison (X_1). Il est important de préciser que le nombre de points est calculé uniquement selon les victoires et nuls remportés. En effet, une défaite signifie un nombre égal à zéro ; un match nul donne droit à un point et une victoire implique trois points,
- l'opposé du différentiel de buts (buts encaissés-buts marqués) noté X_2 .

Les variables explicatives retenues sont les passes décisives, le nombre de tirs non cadrés, le nombre de tirs cadrés, le nombre de cartons jaunes attribués tout au long de

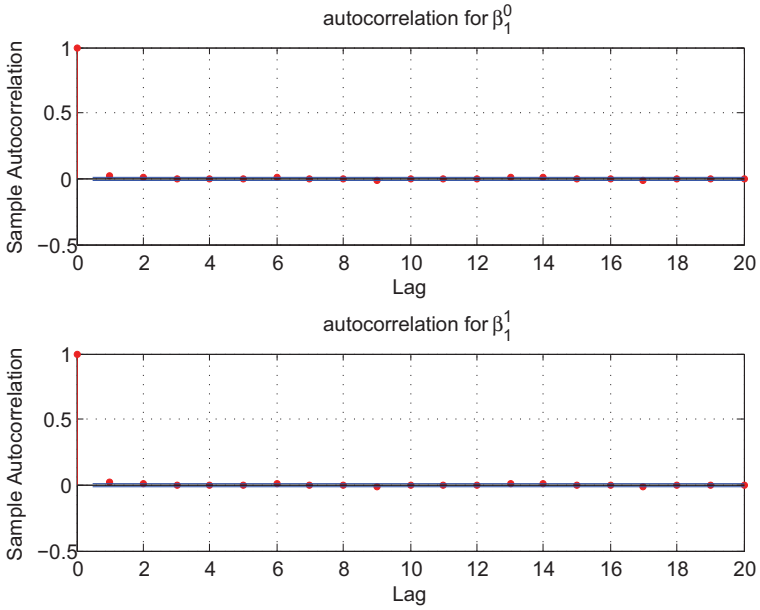

Figure 3.3 – Auto corrélation pour chaque 50^e itération

la saison. En effet, les passes indiquent la qualité et la fluidité du jeu qui demeurent des facteurs essentiels pour la victoire. Un nombre de tirs élevé traduit une forte possibilité de pouvoir marquer des buts, donc des points. Le nombre de cartons jaunes indique le niveau de discipline de l'équipe et devrait agir de façon indirecte sur la performance de l'équipe.

3.5.2 Résultats

Les figures 3.4 et 3.5 indiquent une représentation assez fidèle des observations. Cependant, il apparait une surpédiction des observations au niveau de la 60^e observation ; cela met en exergue l'équipe de Manchester United qui durant la saison 2006-2007 a eu un nombre record de tirs cadrés très largement supérieur aux années précédentes. De plus, l'une des faiblesses du modèle est qu'il ne tient pas compte de l'activité du marché des transferts et de son influence dans la ligue. En effet, l'investissement massif dans le recru-

Tableau 3.2 – Estimation des coefficients du modèle de soccer bidimensionnel ; les écarts types sont notés en parenthèses

coefficients	constante	passes	tirs cadrés	tirs non cadrés	cartons jaunes
θ_1	4,6353 (0,1516)	-0,0023 (0,0028)	-0,0011 (0,0004)	-0,0014 (0,0003)	-0,0092 (0,0020)
θ_2	1,7519 (0,2248)	-0,0080 (0,0026)	0,0001 (0,0004)	0,0013 (0,0002)	0,0242 (0,0023)
θ_{12}		0,0282 (0,0019)	0,0024 (0,0002)	0,0012 (0,0002)	0,0162 (0,0013)

tement de stars pourrait indiquer une bonne mesure de cette activité.

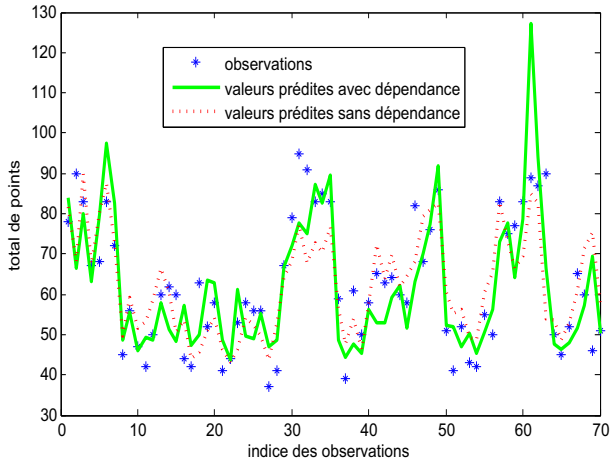

Figure 3.4 – Valeurs prédites pour le total des points X_1

Qualité de l'ajustement

Afin de comparer la qualité d'ajustement du modèle bidimensionnel, nous avons choisi deux mesures notamment l'erreur quadratique moyenne (EQM) et le critère d'information bayésien (BIC). Comme l'indique la table 3.3, le modèle avec dépendance demeure toujours le meilleur selon les deux critères précités. Cependant, lorsque nous nous intéressons à la composante X_1, l'EQM obtenue selon le modèle indépendant est nettement inférieure

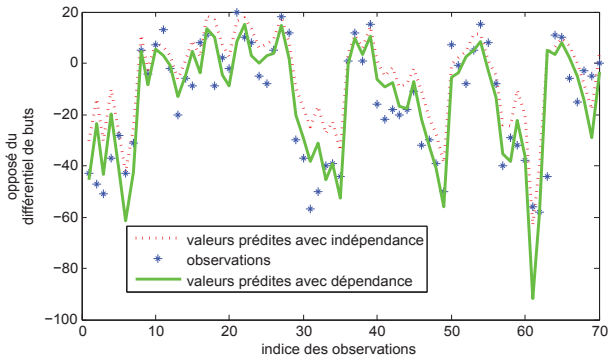

Figure 3.5 – Valeurs prédites pour l'opposé du différentiel de buts X_2

à celle obtenue dans le cas du modèle bidimensionnel. En effet, sous l'hypothèse d'indépendance, $X_1 \sim Poiss(\theta)$ tandis que dans le modèle bidimensionnel $X_1 \sim Poiss(\theta_1 + \theta_{12})$. Cela entraîne une surparamétrisation du modèle bidimensionnel donc une qualité de prédiction plus faible.

Tableau 3.3 – Qualité de l'ajustement selon les deux modèles. Les résultats sont en milliers

Modèles	critères	X_1	X_2	total
indépendance	EQM	7,17	14,5	21,67
	BIC	0,55	0,8	1,35
dépendance	EQM	8,04	9,80	17,84
	BIC	n.d.	n.d.	0,2

La désignation ńn.d.ż indique que la valeur de la statistique n'est pas disponible. En effet, dans le modèle bidimensionnel, le critère bayésien d'information (BIC) n'est pas disponible pour chaque composante.

Conclusion

Ce chapitre introduit une nouvelle forme de distribution qui permet de prendre en compte les corrélations positives et négatives et les propriétés de sa fonction de distribution. En outre, elle permet de traiter des variables discrètes pas nécessairement positives comme le montre l'exemple des données réelles de soccer. Les méthodes d'estimation proposées sont basées sur le principe d'augmentation de données à savoir l'algorithme EM et l'approche bayésienne.

Bien qu'assez simple en dimension 2, la distribution introduite peut s'avérer assez difficile à identifier, plus précisément les termes de covariances. Par ailleurs, l'algorithme EM et l'approche bayésienne avec une loi *a priori* non informative, procurent des résultats assez similaires. Cependant, une des difficultés majeures est de déterminer en pratique les constantes d'ajustement de la loi instrumentale, surtout lorsque la dimension augmente.

Enfin, les résultats obtenus par les méthodes bayésiennes par CMMC indiquent un biais relatif assez faible de moins de 0,5% pour les coefficients de régression des moyennes contrairement à ceux du terme de covariance qui semblent plus volatils.

CHAPITRE 4

MODÉLISATION MULTIDIMENSIONNELLE DE DONNÉES DE COMPTAGE PRÉSENTANT DE LA SURDISPERSION

Après avoir identifié une variante de la loi de Poisson multidimensionnelle au chapitre précédent, permettant de traiter aussi bien des corrélations positives que négatives, force est de constater que dans les données environnementales ou biologiques, il est beaucoup plus fréquent d'observer des petites valeurs ; ce qui entraîne du coup de la surdispersion dans le modèle. Le présent chapitre a pour but de modéliser les données multidimensionnelles présentant une surdispersion sur une ou plusieurs de ses composantes. En effet, l'idée est d'introduire de nouvelles variables latentes de densité gamma afin de modéliser l'hétérogénéité non observable entre les différentes observations.

Introduction

De par sa simplicité le modèle de Poisson demeure très utilisé dans la modélisation des données de comptage. Cependant, l'hypothèse d'égalité entre la variance et la moyenne de l'échantillon considéré demeure assez restrictive. En effet, lorsque vient le temps de modéliser des données caractérisées par une forte hétérogénéité, la variance est très souvent largement supérieure à la moyenne. Afin de remédier à cette situation, deux approches sont généralement privilégiées. La première consiste à introduire des effets aléatoires afin de capter la variation supplémentaire non imputable à la variable aléatoire de Poisson. Quant à la seconde, elle consiste plutôt à supposer la distribution des observations provenant d'une famille beaucoup plus vaste et qui permet de vérifier l'hypothèse de surdispersion. Cependant, dans le cadre multidimensionnel, les difficultés apparaissent assez rapidement

lorsqu'il existe de la surdispersion dans les observations. Dans toute la littérature relative à la loi multidimensionnelle de Poisson, les auteurs semblent faire fi de la présence de surdispersion. En effet, les travaux de Tsionas (1999, 2001) et Karlis (2003) mettent en exergue une structure plutôt simplifiée de la distribution multidimensionnelle de Poisson dans laquelle toutes les variables possèdent la même covariance. Ensuite, Karlis et Meligkotsidou (2005) ont présenté un cadre beaucoup plus général et flexible permettant de définir une structure libre (incorporant des variables explicatives dans les termes de covariance) pour chaque paire de variables. Mais, ils semblent tous ignorer le phénomène de surdispersion assez fréquent dans l'étude des données de comptage. Pour contourner ce problème lié à la surdispersion dans le cas unidimensionnel, une stratégie consiste à utiliser des distributions telles que la loi binomiale, la loi binomiale négative, parfois la loi géométrique et même la loi de Poisson généralisée. Cependant, jusqu'ici les effets aléatoires ne pouvaient pas être incorporés dans la distribution multidimensionnelle de Poisson. Dans ce chapitre, nous introduisons un cadre assez général permettant la modélisation des effets aléatoires avec des données multidimensionnelles de Poisson. Cette définition permet ainsi de prendre en compte le phénomène de surdispersion en introduisant une variation autre que celle de la loi de Poisson. De plus, en considérant les distributions marginales conditionnelles comme des lois de Poisson, nous pouvons aisément analyser les modèles multidimensionnels linéaires mixtes en spécifiant la matrice des effets aléatoires et la structure de corrélation. Par ailleurs, une méthode d'estimation assez proche de celle décrite au chapitre précédent, basée sur l'augmentation des données permet d'obtenir les différentes lois *a posteriori* nécessaires à l'inférence sur les paramètres du modèle considéré. Ensuite, une simulation pour un modèle assez simple de dimension 2 permettra d'illustrer les aspects théoriques développés antérieurement. Enfin, une application en écologie mettant en relief l'impact des changements climatiques sur la répartition des espèces animales notamment des oiseaux au Québec selon différents scénarios de changements climatiques servira

d'illustration.

4.1 Modélisation multidimensionnelle des données de comptage

4.1.1 Régression de Poisson dans le cas bidimensionnel du modèle log-normal

Une façon d'introduire de la corrélation entre des variables de Poisson est de leur associer des variables aléatoires gaussiennes corrélées. Mais il faut toujours se rappeler que la corrélation entre ces variables gaussiennes n'est pas identique à celle observée sur les variables de Poisson bien que ces deux mesures soient liées. Cela demeure une faiblesse de ce modèle car il ne produit pas directement une mesure de corrélation des données de comptage. En effet, la corrélation observée entre les variables de Poisson sera beaucoup plus faible que celle observée sur les variables gaussiennes qui traduisent l'hétérogénéité sous-jacente (Aitchinson et Ho, 1989). Le modèle bidimensionnel tel que spécifié par Munkin et Trivedi (1999) et Million *et al.*, (2003) s'écrit sous la forme suivante.

Soient u_1 et u_2 deux vecteurs gaussiens de moyenne nulle, de variance unité, dont la corrélation est définie par ρ ,

$$\begin{pmatrix} u_{i1} \\ u_{i2} \end{pmatrix} \sim \mathscr{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right)$$

Soient Y_{i1} et Y_{i2} des variables aléatoires de lois conditionnelles de Poisson définies par :

$$Y_{ij}|u_{ij} \sim Poiss\left(\exp\left(\mathbf{Z}_{ij}\beta_j + \sigma_j u_{ij}\right)\right) \; i = 1,\ldots,n \;\; , j = 1,2,$$

où $\mathbf{Z_1} = (\mathbf{Z}_{11},\ldots,\mathbf{Z}_{n1})'$, $\mathbf{Z}_2 = (\mathbf{Z}_{21},\ldots,\mathbf{Z}_{n2})'$ désignent les matrices de variables explicatives et $\beta = (\beta_1,\beta_2)$ représente le vecteur des paramètres à estimer. Par la suite, posons

$\lambda_{ij} = \exp\left(\mathbf{Z}_{ij}\beta_j\right)$, ainsi :

$$Y_{ij}|u_{ij} \sim Poiss\left(\lambda_{ij}\exp\left(\sigma_j u_{ij}\right)\right)\ i = 1,\ldots,n\ ,j = 1,2,$$

sont mutuellement indépendants. Le coefficient de corrélation ainsi introduit entre les variables observées Y_{i1} et Y_{i2} s'écrit :

$$\operatorname{corr}\left(Y_{i1},Y_{i2}\right) \quad = \quad \frac{\operatorname{cov}\left(Y_{i1},Y_{i2}\right)}{\sqrt{\operatorname{var}\left(Y_{i1}\right)}\sqrt{\operatorname{var}\left(Y_{i2}\right)}}. \tag{4.1.1}$$

En utilisant les expressions de variance et covariance conditionnelles aux variables latentes u_{i1} et u_{i2}, la corrélation devient donc :

$$\operatorname{corr}\left(Y_{i1},Y_{i2}\right) \quad = \quad \frac{E\left[\operatorname{cov}\left(Y_{i1},Y_{i2}|u_{i1},u_{i2}\right)\right]+\operatorname{cov}\left[E\left(Y_{i1},Y_{i2}|u_{i1},u_{i2}\right)\right]}{\prod_{j=1}^{2}\sqrt{E\left[\operatorname{var}\left(Y_{ij}|u_{ij}\right)\right]+\operatorname{var}\left[E\left(Y_{ij}|u_{ij}\right)\right]}}. \tag{4.1.2}$$

Afin de simplifier cette dernière équation, il serait avantageux d'examiner séparément les termes tant au numérateur qu'au dénominateur. La variance inconditionnelle s'écrit donc :

$$\operatorname{var}\left(Y_{ij}\right) = \lambda_{ij}\exp\left(\frac{\sigma_j^2}{2}\right)\left\{1+\lambda_{ij}\exp\left(\frac{\sigma_j^2}{2}\right)\left[\exp\left(\sigma_j^2\right)-1\right]\right\}. \tag{4.1.3}$$

Quant au numérateur, il se décompose en deux termes dont le plus complexe semble être la covariance des espérances conditionnelles des variables observées selon les variables

latentes correspondantes.

$$\begin{aligned}\operatorname{cov}\left[E\left(Y_{i1}|u_{i1}\right),E\left(Y_{i2}|u_{i2}\right)\right] &= \operatorname{cov}\left[\lambda_{i1}\exp\left(\sigma_1 u_{i1}\right),\lambda_{i2}\exp\left(\sigma_2 u_{i2}\right)\right]\\ &= \lambda_{i1}\lambda_{i2}\operatorname{cov}\left[\exp\left(\sigma_1 u_{i1}\right),\exp\left(\sigma_2 u_{i2}\right)\right]\\ &= \lambda_{i1}\lambda_{i2}\left\{E\left[\exp\left(\sigma_1 u_{i1}\right)\exp\left(\sigma_2 u_{i2}\right)\right]\right.\\ &\qquad \left.-E\left[\exp\left(\sigma_1 u_{i1}\right)\right]E\left[\exp\left(\sigma_2 u_{i2}\right)\right]\right\}.\end{aligned}$$

Après simplification, nous obtenons

$$\begin{aligned}\operatorname{cov}\left[E\left(Y_{i1}|u_{ij}\right),E\left(Y_{i2}|u_{ij}\right)\right] &= \lambda_{i1}\lambda_{i2}\left\{\exp\left(\frac{\sigma_1^2+2\rho\sigma_1\sigma_2+\sigma_2^2}{2}\right)\right.\\ &\qquad \left.-\exp\left(\frac{\sigma_1^2+\sigma_2^2}{2}\right)\right\}\\ &= \lambda_{i1}\lambda_{i2}\\ &\qquad \times\exp\left(\frac{\sigma_1^2+\sigma_2^2}{2}\right)\left\{\exp\left(\rho\sigma_1\sigma_2\right)-1\right\}. \quad (4.1.4)\end{aligned}$$

En outre, en sachant que :

$$E\left[\operatorname{cov}\left(Y_{i1},Y_{i2}|u_{i1},u_{i2}\right)\right]=0, \quad (4.1.5)$$

nous obtenons, après substitution des équations (4.1.3), (4.1.4) et (4.1.5) dans l'équation de la corrélation inconditionnelle (4.1.2) :

$$\operatorname{corr}\left(Y_{i1},Y_{i2}\right)=\frac{\sqrt{\lambda_{i1}\lambda_{i2}}\exp\left(\frac{\sigma_1^2+\sigma_2^2}{2}\right)\left\{\exp\left(\rho\sigma_1\sigma_2\right)-1\right\}}{\prod_{j=1}^{2}\sqrt{1+\lambda_{ij}\exp\left(\sigma_j^2/2\right)\left\{\exp\left(\sigma_j^2/2\right)-1\right\}}}.$$

Le coefficient de corrélation ainsi défini grâce aux variables non observables gaussiennes peut être soit positif, soit négatif selon le signe de $\left(\exp\left(\rho\sigma_1\sigma_2\right)-1\right)$. De plus, il introduit

de façon systématique de la surdispersion dans le modèle.

4.1.2 Régression de Poisson multidimensionnelle

Cette partie constitue une revue de méthodes d'estimation du modèle multidimensionnel de Poisson avec une corrélation dépendante de variables explicatives. En effet, elle résume l'approche de Karlis et Meligkotsidou (2005) qui ont proposé des méthodes d'estimations basées sur le maximum de vraisemblance et sur l'approche bayésienne par CMMC. Les méthodes d'estimation respectives sont présentées à la section 3.3 du chapitre précédent.

4.2 Modèle multidimensionnel de Poisson avec effets aléatoires

La plus grande difficulté du modèle multidimensionnel de Poisson avec des effets aléatoires, incorporés pour y introduire une surdispersion, est le nombre grandissant de variables latentes à considérer. En effet, en plus des composantes inobservables du vecteur des observations, il existe aussi les vecteurs aléatoires traduisant l'effet de surdispersion.

4.2.1 Modèle

Soient $\mathbf{Z}_1 = (\mathbf{Z}_{11}, \ldots, \mathbf{Z}_{n1})$ et $\mathbf{Z}_2 = (\mathbf{Z}_{12}, \ldots, \mathbf{Z}_{n2})$ deux vecteurs dépendants de variables observées. Le vecteur $\mathbf{Z} = (\mathbf{Z}_1, \mathbf{Z}_2)$ de n variables aléatoires observées dependantes peut s'écrire en fonction de variables aléatoires latentes indépendantes de Poisson. En effet,

$$\begin{aligned} X_1 &= Y_1 + Y_{12}, \\ X_2 &= Y_2 + Y_{12}, \end{aligned}$$

où $Y_j \sim Poiss\left(\theta_j\right),\ \ j = 1,2,12$ sont mutuellement indépendants. Le modèle de régression peut s'écrire de la forme suivante :

$$Y_{ij}|b_{ij} \sim Poiss\left(\theta_{ij}\right),\ \ i = 1,\ldots,n\ , j = 1,2,12,$$

$$\log\left(\theta_{ij}\right) = \mathbf{Z}_{ij}\beta_j + b_{ij},\ \ i = 1,\ldots,n\ , j = 1,2,12,$$

$$e^{b_{ij}} \sim \Gamma\left(\alpha_j,\alpha_j\right),\ \ i = 1,\ldots,n\ , j = 1,2,12.$$

En outre le paramètre α_j traduit la précision ou l'inverse de la variance des variables non observables b_{ij}. Il peut être estimé par la méthode des moments ou considéré comme connu, la valeur provenant d'une étude antérieure. De plus, il peut être considéré comme un paramètre de nuisance car le but ultime de l'inférence est de prédire les valeurs des variables observables X_{ij}. Cette définition des effets aléatoires garantit une moyenne égale à l'unité. Par ailleurs l'effet multiplicatif ainsi introduit peut prendre la forme additive suite à l'introduction d'une fonction de lien logarithmique.

Théorème 4.2.1. *Loi marginale de Y_{ij}*

Sous les hypothèses du modèle spécifié, la loi marginale de Y_{ij} est une loi binomiale négative, notée

$$Y_{ij} \sim \mathscr{NB}\left(\frac{e^{Z_{ij}\beta_j}}{\alpha_j + e^{Z_{ij}\beta_j}}, \alpha_j\right).$$

Démonstration. Par souci de simplification, posons $\eta_{ij} = e^{b_{ij}}$

$$\begin{aligned}\int_{-\infty}^{+\infty} f\left(y_{ij},b_{ij}\right) db_{ij} &= \int_0^{+\infty} f_\eta\left(y_{ij}|\eta_{ij}\right) f_\eta\left(\eta_{ij}\right) d\eta_{ij} \\ &= \int_0^{+\infty} \frac{e^{Z_{ij}\beta_j}}{y_{ij}!}\frac{\alpha_j^{\alpha_j}}{\Gamma\left(\alpha_j\right)}\eta_{ij}^{y_{ij}+\alpha_j-1}e^{-\eta_{ij}\left[\alpha_j+e^{Z_{ij}\beta_j}\right]}d\eta_{ij}.\end{aligned}$$

En remarquant que l'intégrale peut se simplifier comme une constante de normalisation

d'une loi gamma de paramètres $y_{ij}+\alpha_j$ et $\alpha_j+e^{Z_{ij}\beta_j}$, alors nous avons par la suite :

$$\int_0^{+\infty} f\left(y_{ij},\eta_{ij}\right)d\eta_{ij}=\frac{e^{Z_{ij}\beta_j}}{y_{ij}!}\frac{\alpha_j^{\alpha_j}}{\Gamma\left(\alpha_j\right)}\frac{\Gamma\left(y_{ij}+\alpha_j\right)}{\left(\alpha_j e^{Z_{ij}\beta_j}\right)^{\alpha_j+y_{ij}}}.$$

En réarrangeant les termes, nous obtenons finalement

$$\int_0^{+\infty} f\left(y_{ij},\eta_{ij}\right)d\eta_{ij}=\frac{\Gamma\left(y_{ij}+\alpha_j\right)}{\Gamma\left(\alpha_j\right)y_{ij}!}\left\{\frac{\alpha_j}{\left(\alpha_j+e^{Z_{ij}\beta_j}\right)}\right\}^{\alpha_j}\left\{\frac{e^{Z_{ij}\beta_j}}{\alpha_j+e^{Z_{ij}\beta_j}}\right\}^{y_{ij}}.$$

Ainsi, il suffit de remarquer que l'équation précédente n'est rien d'autre que la densité d'une variable aléatoire de loi binomiale négative de paramètres respectifs $p_{ij}=\frac{e^{Z_{ij}\beta_j}}{\alpha_j+e^{Z_{ij}\beta_j}}$ et $r_j=\alpha_j$. □

4.2.2 Une approche par Monte Carlo

Supposons que les paramètres de dispersion sont connus. Cette approche a pour but d'essayer de produire une prédiction des données latentes. Par ailleurs, il faut d'abord déterminer les lois *a posteriori* des différents paramètres pour ensuite générer les observations latentes ou non observables.

Lois *a priori* de β_j

La forme de la loi *a priori* permet de définir une loi certes propre mais elle a aussi l'avantage particulier d'obtenir une loi *a posteriori* conjuguée. Cependant, la méthode demeure toujours valide quelque soit la forme de la loi *a priori* postulée.

$$\pi\left(\beta_j\right)\propto\frac{\left(e^{\overline{\mathbf{Z}}_j\beta_j}\right)^{\overline{y}_j}}{\left(\alpha_j+e^{\overline{\mathbf{Z}}_j\beta_j}\right)^{\alpha_j+\overline{y}_j}}.$$

La forme de cette loi *a priori* permet d'incorporer de l'information contenue sur la moyenne

des observations.

Lois *a posterori*

Tout d'abord, il faut indiquer que la forme de cette loi n'est pas usuelle. Ainsi, afin d'en tirer des échantillons, un algorithme d'échantillonnage d'importance sera proposé.

Théorème 4.2.2. *Lois a posteriori*

Sous les hypothèses des lois a priori précédemment définies, les lois a posteriori conditionnelles des paramètres b_{ij} et β_j sont :

$$e^{b_{ij}} \sim \Gamma\left(y_{ij}+\alpha_j, e^{\mathbf{Z}_{ij}\beta_j}+\alpha_j\right),$$

$$P\left(\beta_j|\mathbf{Y},\mathbf{Z},\alpha_j\right) \propto \pi\left(\beta_j\right)\prod_{i=1}^{n}\left\{\frac{\left(e^{\mathbf{Z}_{ij}\beta_j}\right)^{y_{ij}}}{\left(\alpha_j+e^{\mathbf{Z}_{ij}\beta_j}\right)^{y_{ij}+\alpha_j}}\right\}.$$

Démonstration. En appliquant de façon directe le théorème de Bayes, nous avons

Lois conditionnelles des paramètres β_j

$$\begin{aligned}
P\left(\beta_j|\mathbf{Y},\mathbf{Z},\alpha_j\right) &\propto \pi\left(\beta_j\right)\exp\left\{\sum_{i=1}^{n} y_{ij}\mathbf{Z}_{ij}\beta_j\right\}\prod_{i=1}^{n}\left\{\frac{1}{y_{ij}!}\frac{\alpha_j^{\alpha_j}}{\Gamma\left(\alpha_j\right)}\int_0^{+\infty} b_{ij}^{y_{ij}+\alpha_j}e^{-\left[\alpha_j+e^{\mathbf{Z}_{ij}\beta_j}\right]b_{ij}}db_{ij}\right\}\\
&\propto \pi\left(\beta_j\right)\exp\left\{\sum_{i=1}^{n} y_{ij}\mathbf{Z}_{ij}\beta_j\right\}\prod_{i=1}^{n}\left\{\frac{\Gamma\left(y_{ij}+\alpha_j\right)}{\left(\alpha_j+e^{\mathbf{Z}_{ij}\beta_j}\right)^{y_{ij}+\alpha_j}}\frac{1}{y_{ij}!}\frac{\alpha_j^{\alpha_j}}{\Gamma\left(\alpha_j\right)}\right\}\\
&\propto \pi\left(\beta_j\right)\exp\left\{\sum_{i=1}^{n} y_{ij}\mathbf{Z}_{ij}\beta_j\right\}\prod_{i=1}^{n}\left\{\frac{\left(y_{ij}+\alpha_j-1\right)!}{y_{ij}!\left(\alpha_j-1\right)!}\frac{\alpha_j^{\alpha_j}}{\left(\alpha_j+e^{\mathbf{Z}_{ij}\beta_j}\right)^{y_{ij}+\alpha_j}}\right\}\\
&\propto \pi\left(\beta_j\right)\prod_{i=1}^{n}\left\{\frac{\left(e^{\mathbf{Z}_{ij}\beta_j}\right)^{y_{ij}}}{\left(\alpha_j+e^{\mathbf{Z}_{ij}\beta_j}\right)^{y_{ij}+\alpha_j}}\right\}.
\end{aligned}$$

Lois conditionnelles des paramètres $\eta_{ij} = e^{b_{ij}}$

$$\begin{aligned} P\left(\eta_j|\mathbf{Y},\mathbf{Z},\beta_j,\alpha_j\right) &\propto \pi\left(\eta_j\right)\prod_{i=1}^{n}\left\{\eta_{ij}^{y_{ij}}\exp\left(-\eta_{ij}e^{\mathbf{Z}_{ij}\beta_j}\right)\right\} \\ &\propto \prod_{i=1}^{n}\left\{\pi\left(\eta_{ij}\right)\eta_{ij}^{y_{ij}}\exp\left(-\eta_{ij}e^{\mathbf{Z}_{ij}\beta_j}\right)\right\} \\ &\propto \prod_{i=1}^{n}\left\{\eta_{ij}^{\alpha_j-1}e^{-\alpha_j\eta_{ij}}\eta_{ij}^{y_{ij}}\exp\left(-\eta_{ij}e^{\mathbf{Z}_{ij}\beta_j}\right)\right\} \\ &\propto \prod_{i=1}^{n}\left\{\eta_{ij}^{y_{ij}+\alpha_j-1}\exp\left(-\eta_{ij}\left(\alpha_j+e^{\mathbf{Z}_{ij}\beta_j}\right)\right)\right\}. \end{aligned}$$

Enfin, sous l'hypothèse d'indépendance des composantes de b_j, nous obtenons alors,

$$\eta_{ij} \sim \Gamma\left(y_{ij}+\alpha_j, e^{\mathbf{Z}_{ij}\beta_j}+\alpha_j\right).$$

□

Prédiction des variables latentes

Théorème 4.2.3. *Prédiction*

Sous les lois a priori spécifiées et sous la fonction de perte quadratique, la meilleure prédiction selon la loi prédictive de Y_{lj} est

$$\widehat{y_{lj}} = \frac{\left(y_{lj}+\alpha_j\right)\int\prod_{i=1}^{n}\frac{\left(e^{\mathbf{Z}_{ij}\beta_j}\right)^{y_{ij}^*}}{\left(\alpha_j+e^{\mathbf{Z}_{ij}\beta_j}\right)^{y_{ij}^*+\alpha_j}}\pi\left(\beta_j\right)d\beta_j}{\int\prod_{i=1}^{n}\frac{\left(e^{\mathbf{Z}_{ij}\beta_j}\right)^{y_{ij}}}{\left(\alpha_j+e^{\mathbf{Z}_{ij}\beta_j}\right)^{y_{ij}+\alpha_j}}\pi\left(\beta_j\right)d\beta_j},$$

où $\quad y_{ij}^* = \begin{cases} y_{ij} & si\ i \neq l, \\ y_{lj}+1 & si\ i = l. \end{cases}$

Démonstration. En utilisant les propriétés sur les espérances conditionnelles, nous obtenons

$$\begin{aligned}
\widehat{y_{lj}} &= E\left\{\eta_{lj}e^{\mathbf{Z}_{lj}\beta_j}|Y\right\} \\
&= E^{\beta}\left\{E^{\eta}\left\{\eta_{lj}|\beta_j\right\}e^{\mathbf{Z}_{lj}\beta_j}|\mathbf{Y}\right\} \\
&= E^{\beta}\left\{\frac{\left(y_{lj}+\alpha_j\right)e^{\mathbf{Z}_{lj}\beta_j}}{\alpha_j+e^{\mathbf{Z}_{lj}\beta_j}}|Y\right\} \\
&\propto \int\frac{\left(y_{lj}+\alpha_j\right)e^{\mathbf{Z}_{lj}\beta_j}}{\alpha_j+e^{\mathbf{Z}_{lj}\beta_j}}\prod_{i=1}^{n}\frac{\alpha_j^{\alpha_j}\left(e^{\mathbf{Z}_{ij}\beta_j}\right)^{y_{ij}}}{\left(\alpha_j+e^{\mathbf{Z}_{ij}\beta_j}\right)^{y_{ij}+\alpha_j}}\pi\left(\beta_j\right)d\beta_j \\
&= \frac{\left(y_{lj}+\alpha_j\right)\int\prod_{i=1}^{n}\frac{\left(e^{\mathbf{Z}_{ij}\beta_j}\right)^{y^*_{ij}}}{\left(\alpha_j+e^{\mathbf{Z}_{ij}\beta_j}\right)^{y^*_{ij}+\alpha_j}}\pi\left(\beta_j\right)d\beta_j}{\int\prod_{i=1}^{n}\frac{\left(e^{\mathbf{Z}_{ij}\beta_j}\right)^{y_{ij}}}{\left(\alpha_j+e^{\mathbf{Z}_{ij}\beta_j}\right)^{y_{ij}+\alpha_j}}\pi\left(\beta_j\right)d\beta_j}.
\end{aligned}$$

où $\quad y^*_{ij}=\begin{cases} y_{ij} & \text{si } i\neq l, \\ y_{lj}+1 & \text{si } i=l. \end{cases}$ □

4.2.3 Estimation : méthode et algorithme

4.2.3.1 Définitions

La méthode d'estimation proposée repose sur les calculs des estimateurs de Bayes qui ne sont que les moyennes *a posteriori* des paramètres notés β_j. Pour ce faire, l'échantillonnage pondéré sera utilisé à cause de ses nombreux avantages dont le plus important est la quasi absence de contrainte dans le choix de loi instrumentale.

Définition 4.2.1. *Échantillonnage pondéré*

L'échantillonnage pondéré, stratégique ou encore préférentiel, désigne toute méthode

proposant d'approcher l'intégrale $\int_{\mathscr{X}} h(x)f(x)dx$ *à partir d'un échantillon noté* $x_1, x_2, \ldots, x_N$ *généré suivant une loi g, dite instrumentale, par l'approximation :*

$$\int_{\mathscr{X}} h(x)f(x)dx \approx \frac{1}{N}\sum_{j=1}^{N} \frac{f(x_j)}{g(x_j)} h\left(x_j\right).$$

où $\mathscr{X}$ *désigne le support de la variables aléatoire X de densité* f.

Cette méthode est très pratique car elle permet une latitude quasi-totale dans le choix de la fonction d'importance g. En pratique, le choix de g se fait parmi des lois connues ou faciles à simuler.

Exemple d'échantillonnage préférentiel

Nous désirons calculer les deux premiers moments *a posteriori* des coefficients des effets fixes pour ensuite déduire la matrice de variance correspondante

$$E\left(h\left(\theta|X\right)\right) = \int_{\Theta} h\left(\theta\right)\pi(\theta|X)d\theta.$$

Par la suite, en posant $w(\theta|X) = \frac{\pi(\theta)L(\theta)}{g(\theta)}$, nous obtenons :

$$\begin{aligned} E\left(h\left(\theta|X\right)\right) &= \frac{\int_{\Theta} h(\theta)w(\theta|X)g(\theta)d\theta}{\int_{\Theta} w(\theta|X)g(\theta)d\theta} \\ &= \frac{E^g\left\{h(\theta)w(\theta|X)\right\}}{E^g\left\{w(\theta|X)\right\}} \\ &\approx \frac{\sum_{i=1}^{N} h(\theta^{(i)})w(\theta^{(i)}|X)}{\sum_{i=1}^{N} w(\theta^{(i)}|X)}. \end{aligned}$$

En choisissant successivement $h(\theta) = \theta$ et $h(\theta) = \theta^2$, les deux premiers moments *a posteriori* du paramètre θ sont obtenus de façon respective.

D'autres méthodes d'approximation d'intégrales telles que la méthode de Riemann

peuvent aussi être utilisées. Le seul inconvénient est que cette dernière méthode devient moins efficace lorsque la dimension de l'espace d'intégration augmente ; plus précisément lorsque la dimension est supérieure à 4 selon Yakowitz et *al* (1978).

4.2.3.2 Méthodologie

Algorithme 4.2.1. *Méthode d'estimation*
L'algorithme peut se décrire en quatre grandes étapes.

Étape 1 ***: Initialiser les paramètres***

$$\beta_j^{(0)} = \log(\bar{\mathbf{Z}}_j) \quad j = 1,2,12$$

Étape 2 ***: Déterminer les variables latentes*** y_{ij} ***et*** b_{ij}

2.a. Générer les composantes b_{ij} *selon leurs distributions a posteriori*

2.b. Générer y_{12} *selon la loi* $P(y_{12}|\mathbf{X},\beta,\mathbf{Z},\mathbf{b})$

2.c. En déduire les observations $y_1 = x_1 - y_{12}$ *et* $y_2 = x_2 - y_{12}$.

Étape 3 ***: Mener l'inférence via l'échantillonnage d'importance***

Calculer les deux premiers moments des paramètres par intégration selon le modèle d'approximation normale afin de trouver la loi d'importance. Une approche de Monte Carlo via l'échantillonnage d'importance sera utilisée.

Étape 4 ***: Augmenter la précision des résultats***

4.a. Augmenter la taille de l'échantillon de Monte Carlo nécessaire au calcul des intégrales ou des moments des différents paramètres.

4.b. Répéter les étapes (1), (2) *et* (3) *jusqu'à obtenir des résultats stables.*

4.3 Simulation

Supposons que les variables explicatives suivent une loi uniforme sur l'intervalle $(0;1)$. Soit $z_{ij} \sim \mathscr{U}(0;1)$ et $\beta_1 = [1{,}28;1{,}41]$, $\beta_2 = [1{,}16;3{,}56]$ et $\beta_{12} = [1{,}08;2{,}86]$. Soit n la taille de l'échantillon fixée à 200. Les valeurs de $\alpha_1^{-1} = 4$, $\alpha_2^{-1} = 3$ et $\alpha_{12}^{-1} = 2$ désignent les variance respectives des effets aléatoires $\eta_{ij} = e^{b_{ij}}$. Le modèle de régression pour chacune des composantes s'écrit

$$\log(\theta_{i1}) = 1{,}28 + 1{,}41 z_{i1} + b_{i1},$$

$$\log(\theta_{i2}) = 1{,}16 + 3{,}56 z_{i2} + b_{i2},$$

$$\log(\theta_{i,12}) = 1{,}08 + 2{,}86 z_{i,12} + b_{i,12}.$$

Les résultats de la table 4.1 indiquent que les estimateurs de Bayes obtenus sont légèrement biaisés mais demeurent satisfaisants car ils présentent un biais relatif entre 5% et 30% tel qu'indiqué par McCulloch(1997). Cependant il faut rappeler que la méthode d'intégration demande un effort supplémentaire de calibration afin d'obtenir des valeurs numériques stables.

Comme l'indiquent les figures 4.1 et 4.2, la qualité de la prédiction est assez remarquable. Un autre fait notable est la tendance à prédire le plus souvent des valeurs légère-

Tableau 4.1 – Résultats basés sur les simulations

coefficients	β_1^0	β_1^1	β_2^0	β_2^1	β_{12}^0	β_{12}^1
valeurs	1,28	1,41	1,16	3,56	1,08	2,86
estimés	1,070	1,009	0,750	3,513	1,090	2,622
	(0,1461)	(0,5127)	(0,1124)	(0,0398)	(0,0281)	(0,1052)

ment inférieures à celles observées. Par ailleurs, la qualité de la prédiction pour les valeurs relativement faibles s'avère excellente. Ce qui est très avantageux à cause de la fréquence très élevée d'observations de très faibles valeurs.

Nous avons réalisé ensuite un ré-échantillonnage afin d'étudier la distribution des coefficients avec 1000 réplications et une taille d'échantillon Monte Carlo de 5000, avec les

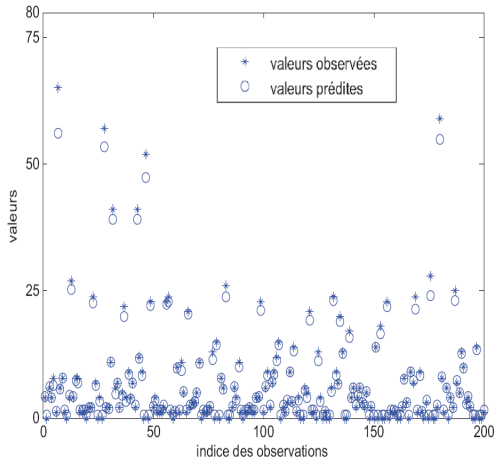

Figure 4.1 – Graphique des valeurs prédites et observées de X_1

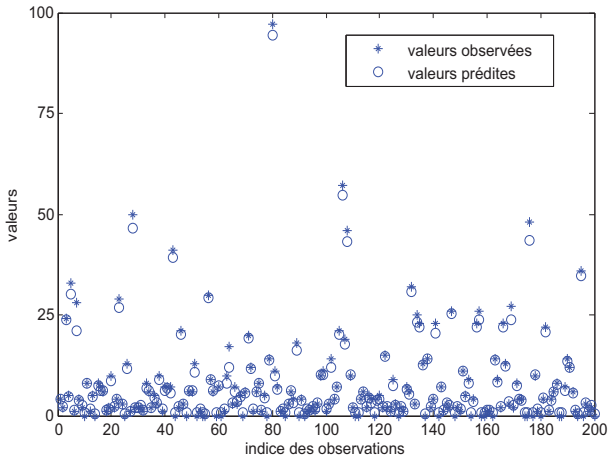

Figure 4.2 – Graphique des valeurs prédites et observées de X_2

mêmes coefficients définis précédemment. La figure 4.3 permet de visualiser la volatilité des estimateurs obtenus.

4.4 Étude de sensibilité

La modélisation du phénomène de surdispersion introduite dans ce chapitre repose principalement sur deux hypothèses assez fortes.

- La structure ou la forme des effets aléatoires est supposée connue.
- Les paramètres de ladite distribution sont aussi supposés connus.

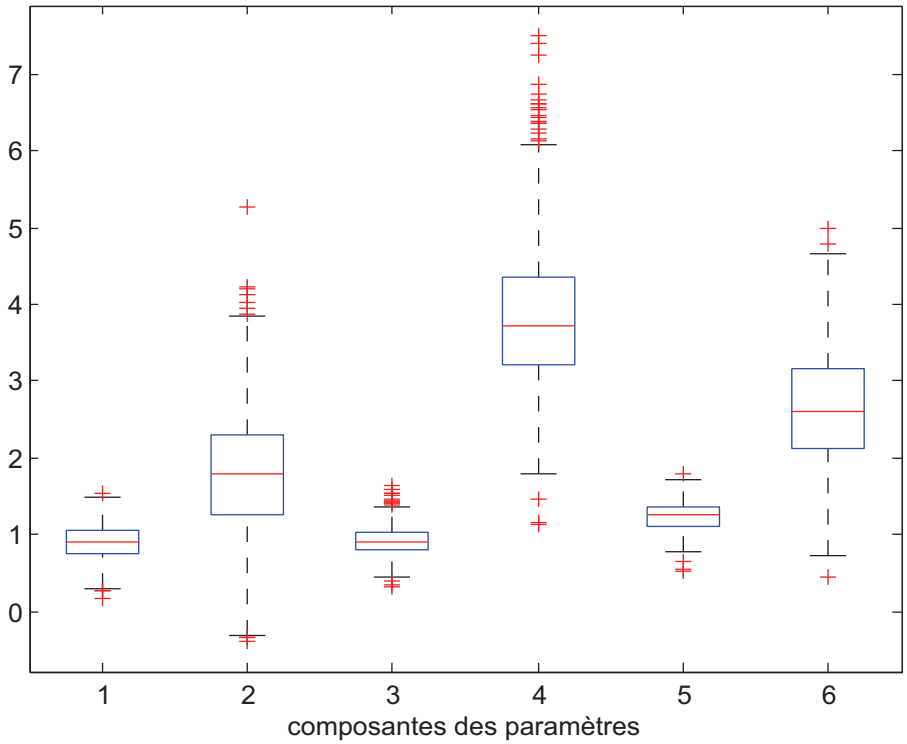

Figure 4.3 – Boîte à moustache de la distribution *a posteriori* des coefficients

L'une des hypothèses les plus contraignantes du modèle est la connaissance de la forme des effets aléatoires traduisant la surdispersion. En effet, elle reste tout de même indispensable à l'inférence compte tenu du nombre élevé de variables latentes à construire et aussi à cause du caractère de famille conjuguée qu'elle possède. Cependant, une étude de sensibilité est menée afin d'étudier l'impact d'une mauvaise spécification du paramètre de dispersion des effets aléatoires sur la qualité des estimations.

Le critère privilégié est l'erreur quadratique moyenne car il faut préciser que l'objectif ultime de la modélisation est la prédiction des variables observées suite à un ou plusieurs changements des variables explicatives considérées. C'est une mesure de la qualité de la prédiction. De plus, en géographie spatiale, l'écart quadratique moyen qui désigne la racine carrée de l'erreur quadratique moyenne constitue un excellent outil de comparaison de la qualité de prédiction d'un modèle. Si $EQM(y_j)$ désigne l'erreur quadratique moyenne de la composante j du vecteur Y, alors elle peut être approximée par :

$$EQM(y_j) \approx \frac{1}{n}\sum_{i=1}^{n}\left(y_{ij}-\hat{y}_{ij}\right)^2.$$

En outre, la qualité globale de prédiction sur toutes les composantes de Y notée EQM_T est simplement égale à la somme de l'EQM de chacune de ses composantes c'est-à-dire

$$EQM_T = \frac{1}{3n}\sum_{j\in\{1,2,12\}}\sum_{i=1}^{n}\left(y_{ij}-\hat{y}_{ij}\right)^2.$$

Ainsi, nous pourrions étudier à la fois l'effet marginal (sur chacune des composantes de Y) et l'effet total d'une ou de plusieurs erreurs de spécification des paramètres α_j.

Supposons que les vraies valeurs des paramètres de dispersion des effets aléatoires sont :

$$\alpha_1^{-1} \equiv \sigma_1^2 = 4,$$

$$\alpha_2^{-1} \equiv \sigma_2^2 = 3,$$

$$\alpha_{12}^{-1} \equiv \sigma_{12}^2 = 2.$$

Nous supposons que le biais relatif observé sur les paramètres de dispersion σ_1, σ_2 et σ_{12} peut prendre les différentes valeurs allant de -90%, -1%, 0, 50% à 200%.

La figure 4.4 indique que seule une mauvaise spécification du paramètre de dispersion du terme de covariance α_{12} entraîne un impact considérable sur la qualité de l'ajustement du modèle. Par ailleurs, même un biais relatif de 200% ou de -90% n'a pas d'impact significatif sur la qualité de prédiction du modèle.

4.5 Application

La présente section présente des applications sur 761 données d'abondance de deux espèces d'oiseaux. La première espèce est la moucherolle (*Empidonax minimus*) qui est très commune dans le sud du Québec. La seconde espèce est la paruline à joues grises (*Vermivora ruficapilla*). Ces deux espèces sont très proches tant au niveau de leurs tailles et de leurs régimes alimentaires que de leurs conditions de vie. Les variables explicatives sélectionnées sont la température moyenne annuelle (TAN), la précipitation moyenne annuelle (PAN) et l'étendue de la température du mois le plus chaud c'est-à-dire juillet (TAM).

Tableau 4.2 – Résultats basés sur les données CC-Bio

coefficients	Constante	TAN	PAN	TAM
θ_1	5,9643 (0,0272)	0,9160 (0,0022)	0,002 $< 10^{-4}$	-0,537 (0,0036)
θ_2	-6,6969 (0,3948)	0,2622 (0,0163)	0,0071 (0,0003)	0,2541 (0,0290)
θ_{12}	-34,5086 (10,4516)	1,5597 (0,3554)	0,0112 (0,0032)	1,1280 (0,7586)

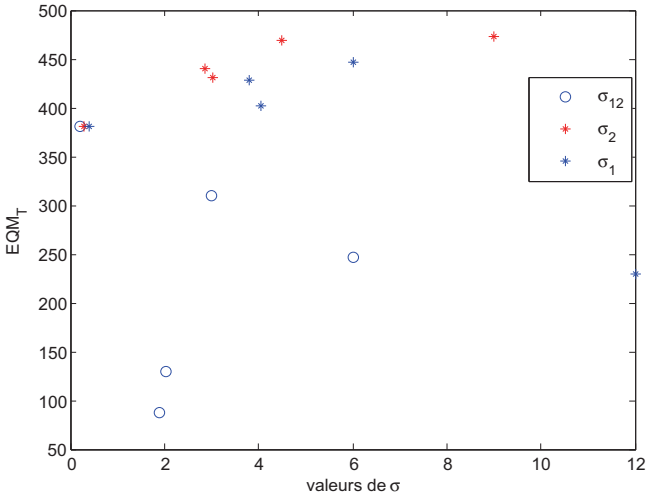

Figure 4.4 – Analyse de sensibilité

Le tableau 4.2 indique que toutes les variables expliquent de façon significative au seuil de 5% l'abondance des deux espèces excepté la variable TAM dans le terme de covariance θ_{12}. De façon globale, nous observons une légère hausse des abondances des deux espèces (environ 2%) selon les prévisions à l'an 2050 en considérant le scénario réaliste (changements moyens). Cependant, la perte ou la décroissance moyenne des espèces est plus élevée que leur accroissement. En effet, dans la région de Charlevoix et de Québec, au vu des figures 4.5, 4.6 et 4.7, il apparait une décroissance moyenne de l'abondance de la deuxième espèce et une décroissance plus prononcée de la population de la première espèce.

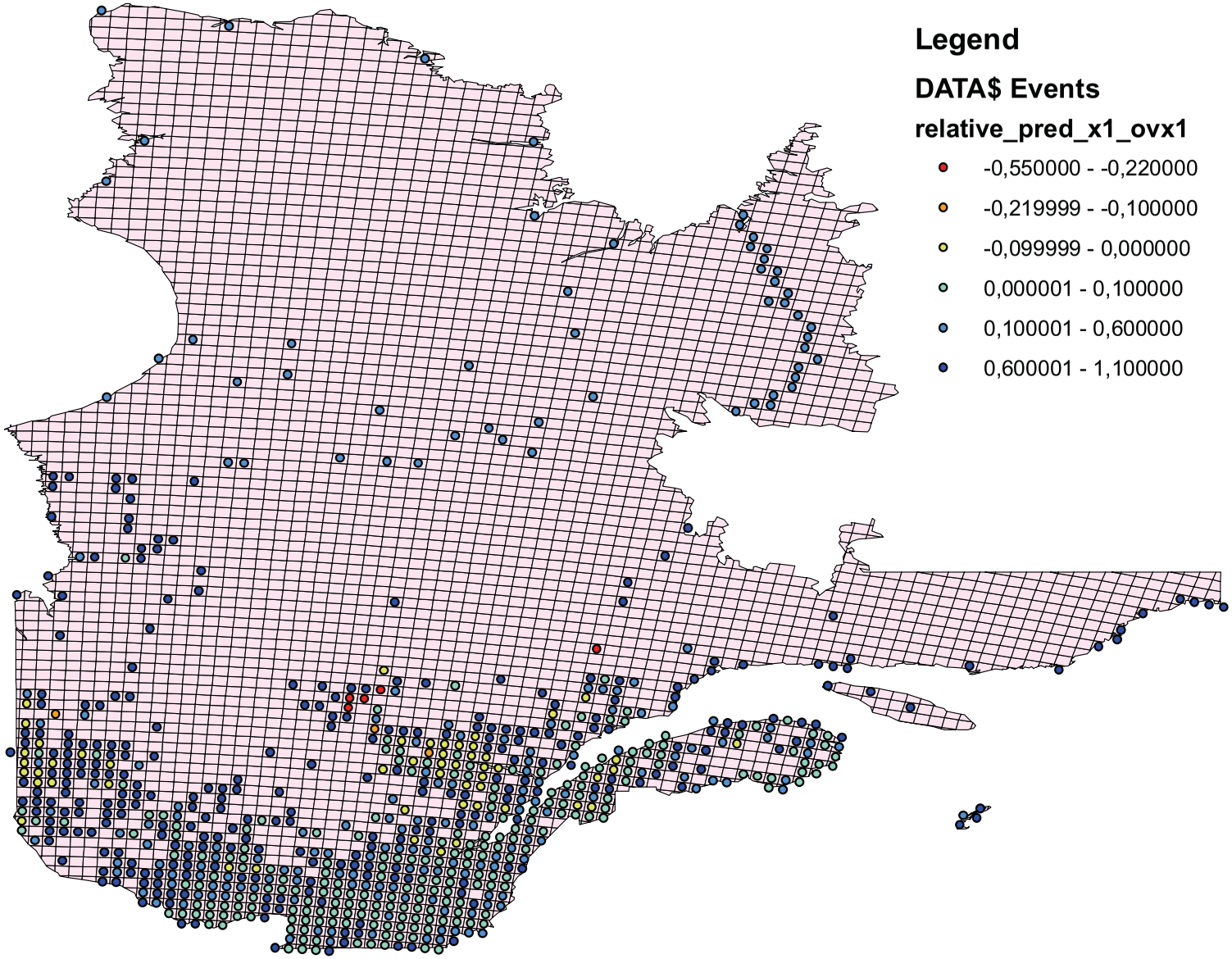

Figure 4.5 – Carte d'abondance de l'espèce 1

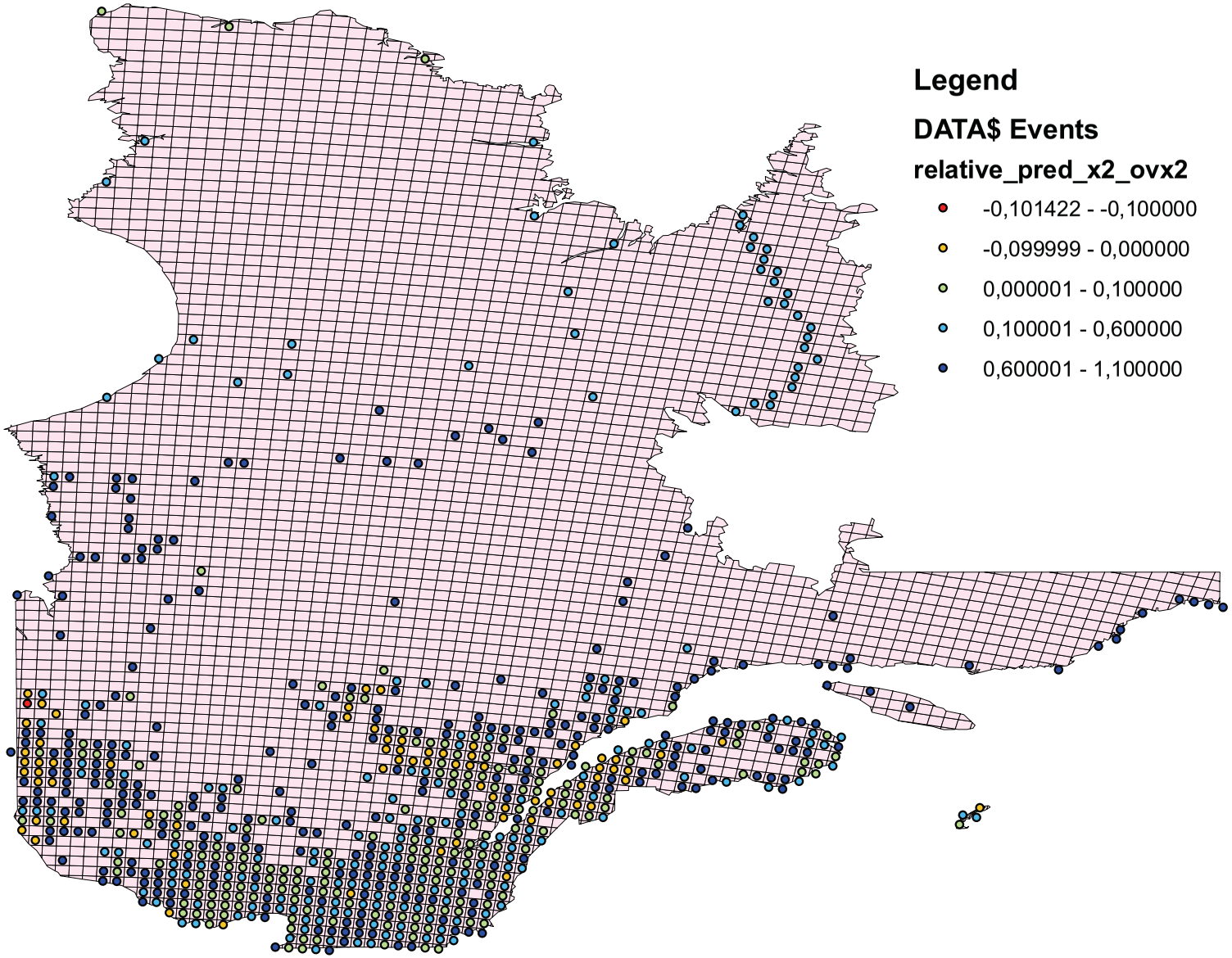

Figure 4.6 – Carte d'abondance de l'espèce 2

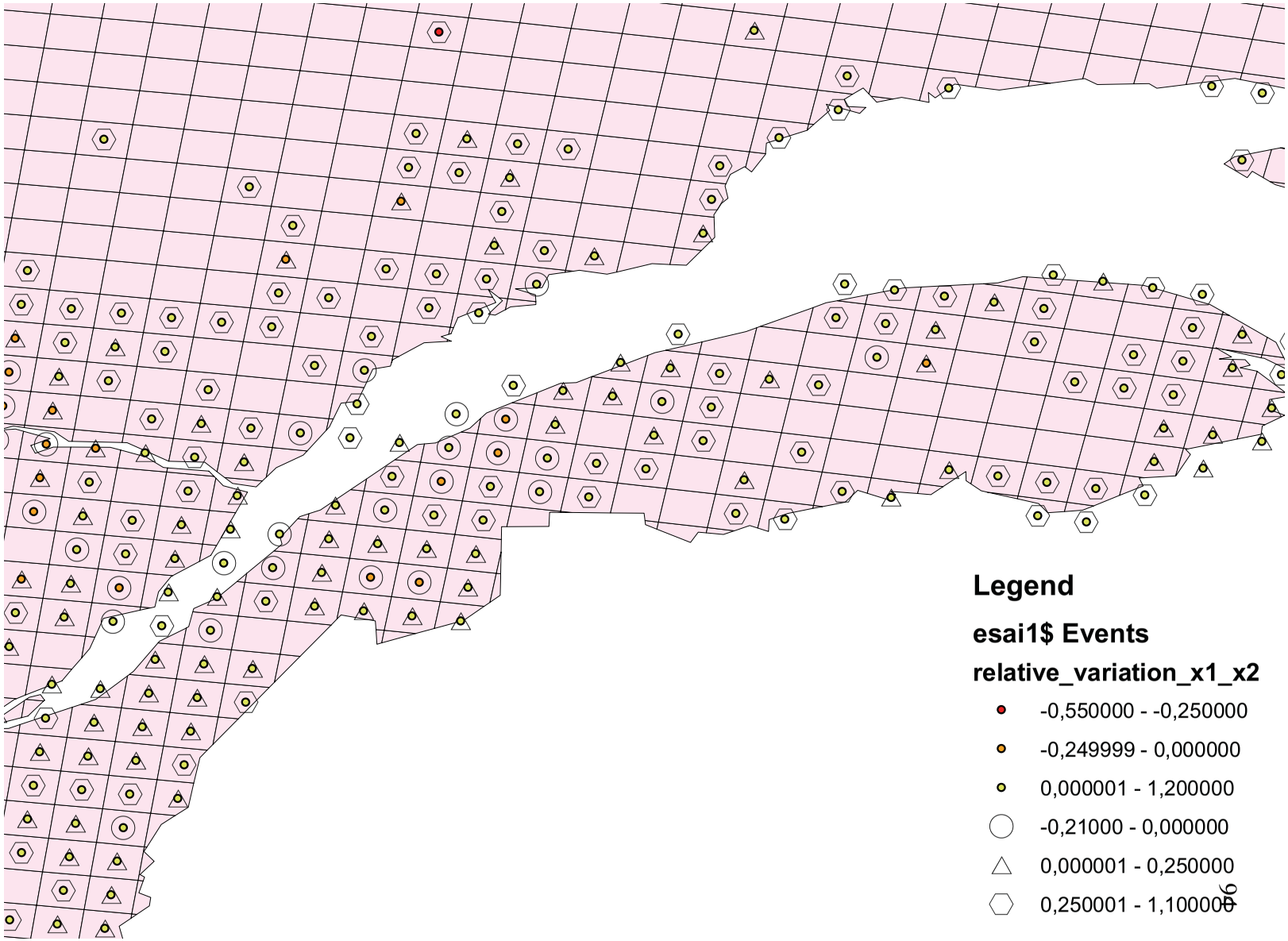

Figure 4.7 – Zoom de la carte sur les régions de Gaspésie et du Bas Saint Laurent

Conclusion

Les résultats de simulation indiquent une bonne performance de notre modèle incorporant des effets aléatoires. Cependant, l'efficacité de cette méthode dépend énormément de la qualité de l'estimation des paramètres de surdispersion. Compte tenu de la disponibilité de données (obtenues) qui présentaient une corrélation positive et une forte dispersion, nous avons appliqué ladite modélisation multidimensionnelle de Poisson avec des effets aléatoires. Au vu des résultats, il apparaît que même si le nombre de sites présentant une très faible augmentation de la population des espèces considérées semblerait être en nette augmentation en 2050, il est prépondérant de signaler que lorsque sur un site donné, la population diminue, cela se fait dans une proportion très grande. Cependant, ce modèle est loin d'être une panacée pour modéliser toutes les interactions possibles entre les espèces car il est basé sur des hypothèses assez restrictives notamment sur leurs distributions sous-jacentes et sur la structure de leurs corrélations.

CHAPITRE 5

MODÈLES BIDIMENSIONNELS AUGMENTÉS EN ZÉRO : UNE APPROCHE PAR LES COPULES

Introduit pour modéliser les corrélations entre des variables continues, l'extension du coefficient de corrélation au cas discret implique certaines restrictions. En effet, le coefficient de corrélation prend ses valeurs dans un intervalle plus restreint que $(-1;1)$, rendant du coup son interprétation plus délicate. En s'intéressant aux copules bidimensionnelles gaussiennes, nous proposons une mesure plus uniforme et par conséquent facilement interprétable. En outre, la nouvelle mesure pourrait être interprétée comme la corrélation entre les variables aléatoires continues dont la discrétisation constitue les observations discrètes non négatives. Une méthode d'estimation des modèles augmentés en zéro de Poisson sera présentée à la fois dans les contextes fréquentiste et bayésien.

Introduction

L'utilisation des données discrètes demeure assez restreinte à cause de la forme complexe de leurs densités. Pour contourner ce problème, une des stratégies les plus courantes est l'approximation par des lois gaussiennes qui s'avèrent assez simples à manipuler. Les modèles les plus récents sont le modèle log-normal d'Aitchinson et Ho (1989) et la loi multidimensionnelle de Poisson. Cependant, dans la réalité, nous observons des données avec une sur-représentation de zéros. Ainsi, une des stratégies consiste à opter pour une distribution binomiale négative ou encore utiliser un modèle de mélange de loi de Dirac concentrée en zéro et une loi discrète telle celles de Poisson, géométrique ou binomiale négative. Ainsi, la littérature sur les modèles augmentés en zéro demeure en pleine expansion

dans le domaine unidimensionnel. Lambert (1992) a proposé une méthode de régression pour les données de Poisson augmentées en zéros afin de contrôler le nombre de pièces défectueuses lors de la production en série. Gosh *et al.* (2006) ont étendu le modèle de Lambert à une plus grande classe de distributions (série de puissance). De plus, ils utilisent une approche par les chaînes de Markov et Monte Carlo (CMMC) qui s'avère posséder de meilleures propriétés en échantillon de faible ou moyenne taille que les estimateurs obtenus par la méthode classique du maximum de vraisemblance. Pour une connaissance plus complète des modèles unidimensionnels augmentés en zéro, l'ouvrage de Ridout *et al.*, (1998) constitue une bonne synthèse de ces modèles.

Lorsque vient le temps de s'intéresser à la modélisation multidimensionnelle, la littérature est très réduite. Li *et al.* (1999) ont montré que l'estimateur de maximum de vraisemblance est plus efficace que celui obtenu par la méthode des moments lorsque les paramètres à estimer sont éloignés des bords de l'espace des paramètres. Cependant, ils précisent les conditions sous lesquelles chacun de ses estimateurs respectifs s'avère meilleur. Par contre, leur approche n'incorpore pas les variables explicatives.

L'objectif de ce chapitre est de présenter une nouvelle méthode permettant de déterminer la corrélation ou le τ de Kendall pour les données bidimensionnelles augmentées en zéro. En effet, dans le cadre du projet CC-Bio, nous avons été confrontés à analyser des données augmentées en zéro, plus précisément de Poisson. Nous y privilégions une approche par les copules gaussiennes compte tenu de l'interprétation assez simple du coefficient de corrélation et son domaine de définition. Deux méthodes d'estimation sont présentées ; la première est basée sur le principe du maximum de vraisemblance en deux étapes introduit par Joe (1997). Quant à la seconde, elle privilégie une approche bayésienne en calculant les paramètres *a posteriori* grâce à la méthode proposée par Naylor et Smith (1982), Smith *et al.* (1987), Dellaportas et Wright (1991) basée sur l'intégration de Gauss-Hermite.

5.1 Copules et mesures de concordance

5.1.1 Copules bidimensionnelles

Définition 5.1.1. *Soient X_1 et X_2 deux variables aléatoires de fonctions de répartitions respectives F_1 et F_2. Alors il existe une fonction nommée copule définie sur $[0,1]^2$ telle que :*

$$H(x_1,x_2) = P(X_1 \leq x_1, X_2 \leq x_2) = C(F_1(x_1), F_2(x_2)).$$

Si les lois marginales sont continues, alors la copule est unique. Sinon, elle est uniquement déterminée sur l'espace $Ran(F_1) * Ran(F_2)$ où $Ran(F)$ désigne le support de la variable aléatoire F.

Copules gaussiennes bidimensionnelles

Nous nous limiterons au cas de la copule gaussienne compte tenu de la simplicité de sa forme mais aussi à cause du support du paramètre de corrélation $(-1;1)$ permettant de décrire tous les degrés de corrélation. Cependant, une extension aux autres copules pourrait être obtenue en appliquant la même idée présentée dans ce chapitre.

Définition 5.1.2. *Soit*

$$\begin{aligned} C_\rho : \quad [0;1]^2 \quad &\longrightarrow [0;1] \\ u = (u_1,u_2) \quad &\longmapsto C_\rho(u) = P(U_1 \leq u_1, U_2 \leq u_2), \end{aligned}$$

où $U_j \sim Unif(0,1)$. La copule gaussienne s'écrit donc

$$C_\rho(u_1,u_2) = \Phi_{X,Y,\rho}\left(\Phi^{-1}(u_1), \Phi^{-1}(u_2)\right),$$

où $\Phi_{X,Y,\rho}$ désigne la fonction de répartition de la loi gaussienne bidimensionnelle stan-

dard de corrélation ρ et Φ celle de la densité gaussienne unidimensionnelle standard.

La fonction de densité de la copule gaussienne s'obtient juste après dérivation de C_ρ

$$c_\rho(u_1,u_2) = \frac{\varphi_{X,Y,\rho}(\Phi^{-1}(u_1),\Phi^{-1}(u_2))}{\varphi(\Phi^{-1}(u_1))\varphi(\Phi^{-1}(u_2))},$$

où $\varphi_{X,Y,\rho}$ désigne la fonction de densité gaussienne bidimensionnelle standard de corrélation ρ et φ celle de la densité gaussienne unidimensionnelle standard. Nous avons donc

$$\varphi_{X,Y,\rho}(x,y) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2(1-\rho^2)}\left[x^2+y^2-2\rho xy\right]\right).$$

5.1.2 Mesures de concordance

Coefficient de corrélation ρ de Pearson

Cette mesure détermine la force de la corrélation linéaire entre deux variables X et Y. Il peut s'interpréter comme une covariance standardisée. Une valeur $\rho = 0$ indique l'absence de relation ou de dépendance linéaire entre les variables X et Y. Pour une distribution bidimensionnelle normale, Kruskal (1958) a montré que :

$$\rho = \frac{\mathrm{cov}(X,Y)}{\sqrt{\mathrm{var}(X)}\sqrt{\mathrm{var}(Y)}}.$$

τ de Kendall

Bien qu'il existe une vaste littérature sur les mesures d'association des données continues, définir celles des données discrètes ou mixtes demeurent assez complexes. L'une des mesures les plus utilisées est le τ de Kendall qui mesure la force de la dépendance entre deux variables continues. En effet, le τ de Kendall est une mesure de concordance. Si nous

notons :

$$P(concordance) = P((X_1 - X_2)(Y_1 - Y_2) > 0),$$

$$P(discordance) = P((X_1 - X_2)(Y_1 - Y_2) < 0),$$

alors il peut être défini par

$$\begin{aligned} \tau(X,Y) &= P(concordance) - P(discordance) \\ &= 2P(concordance) - 1 \\ &= 4\int_0^1 \int_0^1 C(u,v)dC(u,v) - 1. \end{aligned}$$

Genest et Favre (2007) mettent en exergue plusieurs relations existant entre les différentes mesures de concordance. Ainsi, il existe une relation entre le τ de Kendall et le ρ de Pearson donnée par :

$$\tau = \frac{2}{\pi}\arcsin(\rho).$$

Bien qu'appropriée pour les données continues, l'utilisation de cette mesure dans l'analyse de données discrètes nécessite quelques restrictions. En effet, elle omet les ensembles sur lesquels les variables aléatoires peuvent être égales, c'est-à-dire de la forme $\{X_1 = X_2\}$ et $\{Y_1 = Y_2\}$. Du coup, le τ de Kendall ainsi définit ne prend plus ses valeurs dans l'intervalle $(-1;1)$ mais plutôt sur un intervalle beaucoup plus réduit selon que les évènements d'égalité sont plus fréquents. Une autre solution proposée par Denuit et Lambert (2005) consiste à rendre continues les données discrètes à l'aide de variables aléatoires uniformes indépendantes.

Gamma de Kruskal

Souvent appelé γ de Kruskal ou de Goodman, c'est une mesure symétrique prenant ses valeurs dans l'intervalle $(-1;1)$. Elle est basée sur la différence relative entre le nombre de paires concordantes noté $P(concordance)$ et celui des paires non concordantes $P(discordance)$. Ainsi le γ de Kruskal se définit par :

$$\gamma = \frac{P(concordance) - P(discordance)}{P(concordance) + P(discordance)}.$$

Il est interprété comme le surplus de paires concordantes exprimé sous forme de pourcentage de toutes les paires excluant les égalités. Sous l'hypothèse d'indépendance entre X et Y, il est nul. Cependant cette condition n'est que suffisante car il pourrait arriver de trouver une valeur nulle de γ lorsqu'il y a autant de paires concordantes que non concordantes. De plus, elle corrige le τ de Kendall en permettant d'atteindre les bornes -1 et 1 dans les cas respectifs de dépendance parfaite négative et positive. Il convient aussi de citer la correction proposée par Mestfioui et Tajar (2005) qui consiste à standardiser le τ de Kendall afin qu'il prenne ses valeurs dans l'intervalle $(-1;1)$.

5.2 Données modifiées en zéro

Nous nous limitons aux données de Poisson modifiées en zéro et plus particulièrement aux données augmentées en zéro compte tenu de la nature des observations dans le projet CC-Bio.

5.2.1 Données de Poisson modifiées en zéro (PMZ)

Il existe deux classes de modèles, notamment les modèles traduisant un excès de zéros et ceux décrivant un déficit de zéros.

Définition 5.2.1. *Soit Y une variable aléatoire discrète de Poisson modifiée en zéro notée*

$Y \sim PMZ(\lambda, p)$. Alors sa fonction de masse s'écrit :

$$P(Y=0) = p + (1-p)\,e^{-\lambda},$$
$$P(Y=k) = (1-p)\,\frac{e^{-\lambda}\lambda^{k}}{k!},\ k = 1,2,\ldots,.$$

où p est un nombre réel tel que $P(Y=k) \geq 0\ , k \in \mathbb{N}$.

Données de Poisson diminuées en zéro

Il s'agit du cas mettant en relief une sous-représentation de la modalité zéro. En effet si $(1-e^{\lambda})^{-1} < p < 0$, alors il existe moins de zéros qu'il ne devrait y avoir pour une distribution de Poisson. Ce genre de modèles est assez rare dans la littérature. Labrecque-Synnott (2010), Dietz et Böhming (2000) proposent des méthodes d'estimation basées sur l'algorithme Espérance-Maximisation (EM). Quant à Angers et Biswas (2003), ils optent pour une optique bayésienne pour l'estimation des paramètres d'un modèle de Poisson généralisé.

Données de Poisson augmentées en zéro

Par contre lorsque $0 < p < 1$, la modalité zéro est sur-représentée. C'est le cas le plus rencontré dans la pratique notamment en écologie et sciences environnementales. De plus, cette version permet la modélisation des données surdispersées.

Notre objectif est d'estimer le coefficient de corrélation sous-jacent entre les fonctions de répartition des variables rendues continues. En effet, si nous ne faisions que rendre les variables continues tel que préconisé par Denuit et Lambert (2005), nous estimerions certes un coefficient de corrélation mais il ne pourrait pas être utilisé pour effectuer une prévision. En outre, notre étude portant sur le dénombrement des espèces animales et/ou végétales, nous aimerions prédire les abondances de ces espèces face aux changements climatiques.

Un autre avantage de la méthode proposée est qu'elle permet une interprétation tout aussi simple du nouveau coefficient de corrélation obtenu. De ce fait, la valeur du τ de Kendall qui en découle tient compte des ensembles d'égalité.

5.3 Principe de continuité

Denuit et Lambert (2005) ont montré que l'opération qui consiste à rendre continue une paire de variables aléatoires discrètes conserve le τ de Kendall.

Théorème 5.3.1 (Denuit et Lambert (2005))**.** *Soit X une variable aléatoire discrète prenant ses valeurs dans $\mathscr{X}$, un sous ensemble de $\mathbb{N}$. Supposons que sa fonction de masse s'écrit :*

$$f_x = P(X = x), \quad x \in \mathscr{X}$$

Soit $X^* = X + U - 1$ où U est une variable aléatoire continue uniforme sur $(0,1)$ c'est-à-dire $U \sim \mathscr{U}(0,1)$ de fonction de densité l_U. Alors si F et L_U désignent les fonctions de répartition respectives de X et U, alors les fonctions de masse et de répartition respectives de la variable X^* s'écrivent :

$$f^*(s) = l_U(s - [s])\, f_{[s+1]}$$

$$\begin{aligned} F^*(s) &= \sum_{x<s} P(X = x) + L_U(s - [s])\, P(X = [s+1]) \\ &= F([s]) + L_U(s - [s])\, f_{[s+1]}. \end{aligned}$$

En outre, pour tout $n \in \mathbb{N}$, nous avons

$$F(n) = F^*(n),$$

$$f_x = \int_{n-1}^{n} f^*(\eta)\, d\eta.$$

Par la suite, notons $\mathscr{L}_u$ l'opérateur qui consiste à transformer une variable discrète positive en variable continue par l'intermédiaire de la variable continue u. Ainsi, nous pouvons écrire simplement $X^* = \mathscr{L}_u(X)$.

5.4 Estimation

La littérature sur les copules s'est très rapidement développée ces dernières années avec l'apparition d'ordinateurs de plus en plus puissants, capables de résoudre plus rapidement les problèmes numériques les plus complexes. Song (2000) a présenté une forme plus simple de la densité multidimensionnelle de la copule gaussienne. Pitt *et al.*, (2006) ont présenté une méthode permettant de traiter à la fois les variables discrètes ou continues. En effet, ils proposent un algorithme par CMMC en deux étapes. Tout d'abord, il s'agit de mettre une loi *a priori* sur la matrice de covariance pour permettre de modéliser toutes sortes de matrices de corrélation, en particulier celles qui sont assez proches de la singularité. En effet, ils étendent le précédent travail de Wong *et al.*, (2003) au cas gaussien. Ensuite, ils estiment des paramètres des lois marginales par des algorithmes de tirages de Metropolis-Hastings et de Gibbs. L'un des avantages majeurs de leur méthode est l'efficacité de l'algorithme proposé tant dans le cas des données discrètes que celui des données continues.

Les méthodes d'estimation proposées sont celles du maximum de vraisemblance en deux étapes décrite précédemment et celle basée sur l'intégration par la quadrature de Gauss-Hermite. La performance de l'approximation d'une intégrale par la méthode de Gauss-Hermite diminue au fur et à mesure que la dimension de l'espace d'intégration augmente. En effet, la matrice de covariance a plus de chance d'être singulière. De plus, la méthode d'orthogonalisation des paramètres d'inférence telle que décrite par Naylor et

Smith(1982) afin de remédier à une possible dépendance de ceux-ci s'avère moins efficace en grande dimension.

5.4.1 Maximum de vraisemblance

Données discrètes

Dans le cas des données discrètes, la densité de la copule s'écrit :

$$\begin{aligned} c\left(F_1(y_{i1}),F_1(y_{i2});\theta\right) &= C\left(F_1(y_{i1}),F_2(y_{i2}),\theta\right)-C\left(F_1(y_{i1}-1),F_2(y_{i1}),\theta\right) \\ &\quad -C\left(F_1(y_{i1}),F_2(y_{i2}-1),\theta\right)+C\left(F_1(y_{i1}-1),F_2(y_{i2}-1),\theta\right). \end{aligned}$$

Ainsi, la fonction de log-vraisemblance prend la forme :

$$\mathscr{L}\left(\beta_1,\beta_2,\theta\right)=\sum_{i=1}^{n}\sum_{j=1}^{2}\log\left(c\left(F_1(y_{i1}),F_1(y_{i2});\theta\right)\right).$$

Données continues

La méthode introduite par Joe(1997) sous le nom d'inférence marginale produit des estimateurs convergents. En effet, elle permet d'effectuer une maximisation de la fonction de vraisemblance en deux étapes.
La fonction de densité de la copule s'écrit :

$$c\left(F_1(\cdot),F_2(\cdot)\right)=\frac{\partial C\left(F_1|x_1;\beta_1,F_2|x_2;\beta_2,\theta\right)}{\partial F_1\partial F_2}f_1\left(y_1|x_1;\beta_1\right)f_2\left(y_2|x_2;\beta_2\right).$$

Ensuite, la fonction de log-vraisemblance s'écrit donc :

$$\mathscr{L}(\beta_1,\beta_2,\theta)=\sum_{i=1}^{n}\sum_{j=1}^{2}\log\left(f_j\left(y_{ij}|x_{ij};\beta_j\right)\right)+\sum_{i=1}^{n}\log\left(\frac{\partial C\left(F_1|x_{i1};\beta_1,F_2|x_{i2};\beta_2,\theta\right)}{\partial F_1\partial F_2}\right).$$

Ainsi, il faut remarquer qu'elle s'écrit comme une somme de deux composantes. La première est indépendante du paramètre de la copule et seule la deuxième composante en dépend. En effet,

$$\mathscr{L}(\beta_1,\beta_2,\theta)=\mathscr{L}_1(\beta_1,\beta_2)+\mathscr{L}_2(\beta_1,\beta_2,\theta),$$

D'où l'idée de la maximisation par parties ou en deux étapes. Nous obtenons successivement

$$\hat{\beta}_j=\operatorname{argmax}\sum_{i=1}^{n}\log\left(f_j\left(y_{ij}|x_{ij};\beta\right)\right),\quad j=1,2,$$

$$\hat{\theta}=\operatorname{argmax}\sum_{i=1}^{n}\frac{\partial C\left(F_1|x_{i1};\hat{\beta}_1,F_2|x_{i2};\hat{\beta}_2,\theta\right)}{\partial F_1\partial F_2}.$$

Ainsi, l'algorithme d'estimation par les marges peut s'écrire comme suit :

Algorithme 5.4.1. *Algorithme d'inférence marginale (Joe, 1997)*

1. *Obtenir par le maximum de vraisemblance en deux étapes les estimateurs $\hat{\beta}_1$, $\hat{\beta}_2$, $\hat{\theta}$.*
2. *Tirer un échantillon aléatoire avec remise des observations, de taille K inférieure ou égale à n.*
3. *En utilisant le nouvel échantillon obtenu, ré-estimer $\hat{\beta}_1$, $\hat{\beta}_2$, $\hat{\theta}$ et les sauvegarder.*
4. *Répéter les étapes (2) et (3) pour obtenir $\hat{\beta}_1(r)$, $\hat{\beta}_2(r)$, $\hat{\theta}(r)$ òu r désigne le r^e échantillon ou réplication utilisé de taille K.*

5. *Estimer enfin la matrice de variance* $\hat{\Omega} = \left(\hat{\beta}_1, \hat{\beta}_2, \hat{\theta}\right)'$ *des paramètres par*

$$\frac{1}{R}\sum_{r=1}^{R}\left(\hat{\Omega}(r)-\hat{\Omega}\right)\left(\hat{\Omega}(r)-\hat{\Omega}\right)',$$

où R désigne le nombre de ré-échantillonnages effectués.

En outre, sous certaines conditions de régularité, l'estimateur de maximum de vraisemblance $\hat{\Omega}$ est convergent. Cependant les propriétés des estimateurs de vraisemblance sont toutes asymptotiques et l'efficacité en présence d'échantillon de petite taille n'est pas garantie.

5.4.2 Quadrature de Gauss-Hermite

Elle représente une méthode d'intégration numérique permettant de calculer les moments *a posteriori* d'un paramètre donné.

Approximation dans le cas unidimensionnel

Soit g une fonction telle que,

$$g(t) = f(t)\left(\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{1}{2}\left(\frac{t-\mu}{\sigma}\right)^2\right\}\right)^{-1}.$$

Ainsi, nous avons :

$$\int_{-\infty}^{\infty} f(t)dt \approx \sum_{i=1}^{m} m_i g(z_i),$$

$$\text{où } \; w_i = \frac{2^{m-1}m!\sqrt{\pi}}{m^2\left[H_{m-1}(t_i)\right]^2}, \quad m_i = w_i \exp\left(t_i^2\right),$$

et $z_i = \mu + \sigma\sqrt{2}t_i$ avec t_i la i^e racine du polynôme d'Hermite $H_m(t)$.

Il faut noter que cette méthode est très efficace lorsque la fonction f peut être approxi-

mée ou s'écrire comme le produit d'une fonction polynomiale et d'une densité normale. En effet, si la densité *a posteriori* peut être approchée par le produit d'une densité de loi normale et d'un polynôme de degré maximal de $2m-3$, alors une grille de m points serait largement suffisante pour estimer les paramètres. En pratique, il est recommandé de commencer avec une petite grille (4 ou 5 points) dont la taille augmenterait progressivement jusqu'à l'obtention de résultats jugés stables.

Extension au cas multidimensionnel

Bien que la généralisation au cas multidimensionnel s'avère assez simple, son efficacité repose essentiellement sur l'hypothèse d'indépendance *a posteriori* des paramètres d'intérêts. Et comme cela ne s'avère pas toujours vrai, Naylor et Smith (1982), Smith et *al.*, (1987) ont proposé une méthode d'orthogonalisation de l'espace des paramètres.

$$\int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} f(t_1,\dots,t_k)\, dt_1 \dots dt_k \approx \sum m_{i_k}^{(k)} \dots \sum m_{i_1}^{(1)} g\left(z_{i_1}^{(1)},\dots,z_{i_k}^{(k)}\right).$$

Il s'agit de trouver des combinaisons linéaires indépendantes des paramètres d'intérêt et y mener l'inférence, pour enfin déduire par la transformation réciproque les résultats désirés. Cette approche constitue un véritable défi compte tenu du nombre élevé de variables. En effet, la dimension de l'espace d'intégration augmente considérablement la complexité de la méthode et aussi le temps nécessaire au calcul numérique.

5.4.3 Algorithme pour le maximum de vraisemblance

Nous présentons une méthode d'estimation par le maximum de vraisemblance en deux étapes. L'objectif est de générer les variables latentes permettant de rendre continues les variables de Poisson augmentées en zéro et de maximiser la fonction de vraisemblance. Par ailleurs, dans le cas gaussien, le paramètre de la copule θ est tout simplement égal au coefficient de corrélation ρ. L'algorithme se décrit comme suit

Étape 0

Elle consiste tout d'abord à initialiser les paramètres. Pour ce faire, il faut générer :

- $u_i \sim \mathscr{U}(0;1)$ pour ensuite obtenir $y_{i1}^* = \mathscr{L}_u(y_{i1})$
- $v_i \sim \mathscr{U}(0;1)$ pour ensuite obtenir $y_{i2}^* = \mathscr{L}_v(y_{i2})$

Ensuite, nous obtenons les valeurs initiales des paramètres

$$\hat{\beta}_j^{(0)} = \operatorname{argmax} \sum_{i=1}^{n} \log\left(f_j\left(y_{ij}^* | z_{ij}; \beta\right)\right),$$

$$\hat{\theta}^{(0)} = \operatorname{argmax} \sum_{i=1}^{n} \frac{\partial C\left(F_1^* | y_{i1}^*; \hat{\beta_1}^{(0)}, F_2^* | y_{i2}^*; \hat{\beta_2}^{(0)}, \theta\right)}{\partial F_1^* \partial F_2^*},$$

$$\hat{\tau}^{(0)} = \tau(Y_1, Y_2),$$

À l'étape k

Si $\hat{\tau}^{(0)} > 0$, alors $\tau^{(k)} = \tau^{(k-1)} + \varepsilon$,

Si $\hat{\tau}^{(0)} < 0$, alors $\tau^{(k)} = \tau^{(k-1)} - \varepsilon$,

où $\varepsilon = \frac{2}{n(n-1)}$ représente le pas. Ensuite, on retrouve $\hat{\rho}^{(k)}$ grâce à la relation

$$\tilde{\tau}^{(k)} = \frac{2}{\pi} \arcsin\left(\hat{\rho}^{(k)}\right).$$

L'inverse du pas ε est choisi comme étant égal à $\frac{n(n-1)}{2}$. Cela correspond à ajouter une nouvelle paire concordante ; en effet, le nombre de combinaisons de deux observations parmi n est bien égal au pas. Par contre, si la corrélation était négative, il faudrait plutôt introduire une paire discordante donc considérer un pas négatif.

Il faut ensuite générer les variables latentes (u_i, v_i) selon la densité suivante

$$(u_i, v_i) \sim \begin{cases} C\left(u, v|\tilde{\rho}^{(k)}\right) & \text{si } x_{i1} = x_{i2} = 0 \\ \mathscr{U}\left((0;1)^2\right) & \text{sinon} \end{cases}$$

Obtenir ensuite $y_{i1}^* = \mathscr{L}_u(y_{i1})$ et $y_{i2}^* = \mathscr{L}_v(y_{i2})$. Puis, les estimateurs sont obtenus en résolvant les programmes d'optimisation suivants :

$$\hat{\beta}_j^{(k+1)} = \text{argmax} \sum_{i=1}^{n} \log\left(f_j\left(y_{ij}^*|z_{ij};\beta_j\right)\right),$$
$$\hat{\rho}^{(k+1)} = \text{argmax} \sum_{i=1}^{n} \frac{\partial C\left(F_1^*|y_{i1}^*;\hat{\beta}_1^{(k)}, F_2^*|y_{i2}^*;\hat{\beta}_2^{(k)}, \rho\right)}{\partial F_1^* \partial F_2^*}.$$

Répéter l'étape k jusqu'à ce que la valeur de τ soit proche des valeurs extrêmes correspondant à -1 et 1 selon que la corrélation soit négative ou positive.

5.5 Modélisation

Cette section présente deux méthodes d'estimation des paramètres basés sur le maximum de vraisemblance et l'approche bayésienne. Une simulation permettra d'illustrer ces différentes méthodes.

Soient Y_1 et Y_2 des variables aléatoires discrètes positives dont les fonctions de répartitions respectives sont F_1 et F_2. De plus, supposons que

$$Y_{ij} \sim PMZ\left(\lambda_{ij}, p_j\right), \quad i = 1, \ldots, n \quad j = 1, 2,$$

$$\log\left(\lambda_{ij}\right) = Z_{ij}\beta_j, \quad i = 1, \ldots, n \quad j = 1, 2.$$

Lois *a priori*

Nous avons opté pour des lois *a priori* impropres ou non informatives pour les différents paramètres. Cependant, dépendamment de l'information disponible, ces hypothèses pourraient être modifiées.

$$p_j \sim \mathscr{U}(0;1) \quad \text{et} \quad \pi\left(\beta_j\right) \propto 1 \quad j = 1,2,$$

$$\rho \sim \mathscr{U}(-1;1).$$

Pour cette étude, nous avons choisi des lois *a priori* impropres pour les paramètres compte tenu du manque d'information mais aussi afin de comparer les estimateurs obtenus par le maximum de vraisemblance et ceux de Bayes. Cependant, la méthode demeure toujours valide quelque soit la loi *a priori* choisie. Par ailleurs, dans le cas présent, la probabilité d'observer des égalités plus particulièrement au point $(0;0)$ augmente avec p_1, p_2 et ρ.

5.5.1 Un exemple de simulation

Supposons que les variables explicatives suivent une loi uniforme sur l'intervalle $(0;1)$. Soit $z_{ij} \sim \mathscr{U}(0;1)$ un échantillon de $n = 200$ variables explicatives. Posons $\beta_1 = [3{,}28;2{,}41]$, $\beta_2 = [2{,}16;4{,}56]$, $p_1 = 0,75$, $p_2 = 0{,}84$ et enfin $\rho = 0{,}81$.

Les résultats de la table 5.1 indiquent que pour un échantillon assez grand de 200 observations, les paramètres de la régression semblent être sans biais. Cependant, les paramètres relatifs à l'excédent de la valeur égale à zéro notamment p_1 et p_2 présentent un biais relatif beaucoup plus grand. Par ailleurs, les estimateurs du paramètre de la copule ρ tant de l'optique bayésienne que du principe du maximum de vraisemblance sont sans biais. Mais il serait plus judicieux d'étudier les propriétés de ces estimateurs selon la va-

Tableau 5.1 – Résultats basés sur les simulations par le maximum de vraisemblance et par l'approche bayésienne

coefficients	β_1^0	β_1^1	β_2^0	β_2^1	p_1	p_2	ρ
valeurs	3,28	2,41	2,16	4,56	0,75	0,84	0,81
estimateurs du MV	3,222	2,491	2,125	4,458	0,655	0,817	0,803
	(0,068)	(0,199)	(0,115)	(0,323)	(0,056)	(0,063)	(0,026)
estimateurs de Bayes	3,289	2,408	2,189	4,334	0,681	0,799	0,831
	(0,051)	(0,089)	(0,062)	(0,077)	(0,051)	(0,053)	(0,028)

riation de p_1, p_2 et ρ, ce que nous ferons. En effet, une étude de sensibilité permettrait de vérifier l'efficacité des estimateurs selon les critères respectifs de biais relatif, de variance et d'erreur quadratique moyenne (EQM).

5.5.2 Une étude de sensibilité

Dans le cas de l'EMV, nous nous intéressons aux propriétés asymptotiques notamment celle du biais et de l'efficacité des estimateurs. En effet, il semble assez cohérent de penser que la qualité des estimateurs plus particulièrement celui de la copule gaussienne dépende de la taille de l'échantillon, de l'abondance relative de la modalité zéro dans les observations et enfin de la force ou ampleur de la liaison des variables dépendantes.

En effet, les tables 5.2 et 5.3 indiquent non seulement que la qualité des estimations augmente avec la taille de l'échantillon mais aussi que leur qualité augmente selon la force de la corrélation des variables latentes continues. Par ailleurs, puisque nous essayons de reconstruire la relation d'ordre totale existant entre les observations dont la transformation par l'opérateur $\mathscr{L}$ donne la modalité zéro, il est clair que la qualité de l'estimateur augmente lorsque le nombre de couples $(0,0)$ est élevé. En d'autres termes, l'inférence sur la corrélation sous-jacente des variables latentes sera meilleure d'autant plus que les valeurs

de p_1 et p_2 augmentent simultanément.

Tableau 5.2: Table pour $n = 100$

vrais paramètres				maximum de vraisemblance				estimateurs de Bayes			
p_1	p_2	ρ		estimateurs	biais relatif	variance	EQM	estimateurs	biais relatif	variance	EQM
0,4	0,3	0,81	β_1	3,2673	-0,0039	0,0008	0,001	3,2731	-0,0021	0,0004	0,0005
				2,4518	0,0174	0,0081	0,0099	2,4496	0,0164	0,0014	0,0029
			p_1	0,45	0,125	0,0025	0,005	0,432	0,0799	0,0019	0,003
			β_2	2,1645	0,0021	0,0026	0,0026	2,1872	0,0126	0,0006	0,0014
				4,6143	0,0119	0,0212	0,0242	4,5869	0,0059	0,001	0,0017
			p_2	0,3011	0,0038	0,0024	0,0024	0,3182	0,0608	0,0019	0,0023
			ρ	0,8155	-0,0033	0,0007	0,0007	0,8084	-0,012	0,0001	0,0001
		0,6	β_1	3,3026	0,0069	0,0007	0,0013	3,2929	0,0039	0,0005	0,0007
				2,3341	-0,0315	0,0086	0,0144	2,3552	-0,0227	0,0021	0,0051
			p_1	0,36	-0,1	0,0023	0,0039	0,3564	-0,1091	0,0023	0,0042
			β_2	2,0728	-0,0403	0,0029	0,0105	2,0681	-0,0425	0,0008	0,0093
				4,8351	0,0603	0,024	0,0997	4,818	0,0566	0,0015	0,0681
			p_2	0,2017	-0,3277	0,0019	0,0115	0,2017	-0,3276	0,0019	0,0115
			ρ	0,602	0,0026	0,0029	0,0029	0,2894	-0,5181	0,0001	0,0968
		0,4	β_1	3,2765	-0,0011	0,0007	0,0007	3,2835	0,0011	0,0005	0,0005
				2,3535	-0,0234	0,0059	0,0091	2,3552	-0,0227	0,0012	0,0042
			p_1	0,35	-0,125	0,0023	0,0048	0,3577	-0,1057	0,0022	0,004
			β_2	2,1636	0,0017	0,0026	0,0026	2,172	0,0055	0,0007	0,0009
				4,6324	0,0159	0,0229	0,0281	4,6339	0,0162	0,0001	0,0055
			p_2	0,2911	-0,0296	0,0024	0,0024	0,3086	0,0287	0,0023	0,0024
			ρ	0,4997	0,2493	0,0038	0,0137	0,5009	0,2521	0,0029	0,0131
0,75	0,84	0,81	β_1	3,3111	0,0095	0,0012	0,0022	3,3359	0,017	0,0007	0,0039
				2,3762	-0,014	0,0156	0,0167	2,3402	-0,029	0,0037	0,0086
			p_1	0,65	-0,1333	0,0023	0,0123	0,6455	-0,1393	0,0014	0,0123
			β_2	2,1933	0,0154	0,0095	0,0106	2,2527	0,0429	0,004	0,0126
				4,4449	-0,0252	0,0913	0,1045	4,232	-0,0719	0,007	0,1146
			p_2	0,8227	-0,0206	0,0015	0,0018	0,7924	-0,0567	0,0015	0,0038
			ρ	0,8546	0,0446	0,0004	0,0017	0,8617	0,0533	0,0005	0,0024
		0,6	β_1	3,1422	-0,042	0,0026	0,0216	3,1876	-0,0282	0,0012	0,0097
				2,8557	0,1849	0,0202	0,2188	2,829	0,1739	0,0024	0,178
			p_1	0,76	0,0133	0,0018	0,0019	0,7679	0,0238	0,0014	0,0017
			β_2	2,0653	-0,0438	0,0149	0,0238	2,1679	0,0036	0,0032	0,0033
				4,7455	0,0407	0,135	0,1694	4,5839	0,0052	0,006	0,0065
			p_2	0,8473	0,0087	0,0014	0,0015	0,8338	-0,0074	0,0013	0,0013
			ρ	0,6818	0,7045	0,0018	0,0812	0,7055	0,7636	0,0022	0,0955
		0,4	β_1	3,1422	-0,042	0,0026	0,0216	3,1876	-0,0282	0,0012	0,0097
				2,8557	0,1849	0,0202	0,2188	2,829	0,1739	0,0024	0,178
			p_1	0,76	0,0133	0,0018	0,0019	0,7679	0,0238	0,0014	0,0017
			β_2	2,0653	-0,0438	0,0149	0,0238	2,1679	0,0036	0,0032	0,0033
				4,7455	0,0407	0,135	0,1694	4,5839	0,0052	0,006	0,0065
			p_2	0,8473	0,0087	0,0014	0,0015	0,8338	-0,0074	0,0013	0,0013
			ρ	0,6818	0,7045	0,0018	0,0812	0,7055	0,7636	0,0022	0,0955
0,85	0,9	0,81	β_1	3,1704	-0,0334	0,0043	0,0163	3,2949	0,0045	0,0016	0,0018
				2,3289	-0,0337	0,0402	0,0467	2,2448	-0,0686	0,0046	0,0319
			p_1	0,87	0,0235	0,0011	0,0015	0,8365	-0,0158	0,0008	0,001
			β_2	2,0469	-0,0523	0,0809	0,0936	2,1874	0,0127	0,0425	0,0433
				3,6429	-0,2011	2,1432	2,9842	3,1104	-0,3179	0,9948	3,0961
			p_2	0,948	0,0533	0,0005	0,0028	0,9069	0,0077	0,0007	0,0007
			ρ	0,8146	-0,0044	0,0006	0,0007	0,8622	0,0538	0,0004	0,0023

		0,6	β_1	3,3311	0,0156	0,0025	0,0051	3,4751	0,0595	0,0015	0,0395
				2,3518	-0,0241	0,0234	0,0268	2,1225	-0,1193	0,0074	0,0901
			p_1	0,83	-0,0235	0,0014	0,0018	0,8221	-0,0329	0,0009	0,0016
			β_2	2,0825	-0,0359	0,0291	0,0351	2,3403	0,0835	0,0044	0,0368
				4,9891	0,0941	0,1925	0,3766	4,5814	0,0047	0,0054	0,0059
			p_2	0,9111	0,0123	0,0009	0,001	0,8634	-0,0407	0,001	0,0023
			ρ	0,8524	0,4197	0,0004	0,0639	0,9042	0,5059	0,0004	0,0926
		0,4	β_1	3,2795	-0,0002	0,0037	0,0037	3,3952	0,0351	0,0019	0,0151
				2,4216	0,0048	0,0381	0,0383	2,2414	-0,0699	0,0069	0,0354
			p_1	0,87	0,0235	0,0011	0,0015	0,853	0,0035	0,0009	0,0009
			β_2	1,8348	-0,1506	0,0282	0,134	1,9983	-0,0749	0,0032	0,0293
				5,4046	0,1852	0,1809	0,8943	5,2625	0,1541	0,0039	0,4974
			p_2	0,8964	-0,004	0,0011	0,0011	0,8522	-0,0531	0,0011	0,0034
			ρ	0,7903	0,9755	0,0009	0,1531	0,8413	1,1031	0,0009	0,1956

Tableau 5.3: Table pour $n = 200$

vrais paramètres				maximum de vraisemblance				estimateurs de Bayes			
p_1	p_2	ρ		estimateurs	biais relatif	variance	EQM	estimateurs	biais relatif	variance	EQM
0,4	0,3	0,81	β_1	3,2973	0,0053	0,0003	0,0006	3,3068	0,0082	0,0002	0,0009
				2,3351	-0,0311	0,004	0,0096	2,3519	-0,0241	0,0008	0,0041
			p_1	0,34	-0,15	0,0011	0,0047	0,3438	-0,1406	0,0008	0,0039
			β_2	2,1762	0,0075	0,0012	0,0014	2,1941	0,0158	0,0003	0,0014
				4,5192	-0,0089	0,0099	0,0115	4,4873	-0,0159	0,0003	0,0056
			p_2	0,2042	-0,3192	0,0009	0,0101	0,22	-0,2666	0,0004	0,0068
			ρ	0,8023	-0,0194	0,0004	0,0006	0,8087	-0,0116	0,0002	0,0003
		0,6	β_1	3,2789	-0,0004	0,0004	0,0004	3,2865	0,002	0,0003	0,0003
				2,368	-0,0174	0,0043	0,0061	2,3821	-0,0116	0,0014	0,0021
			p_1	0,485	0,2125	0,0012	0,0085	0,5138	0,2844	0,001	0,0139
			β_2	2,1409	-0,0088	0,0014	0,0018	2,1666	0,0031	0,0004	0,0005
				4,6928	0,0291	0,0106	0,0282	4,6652	0,0231	0,0006	0,0117
			p_2	0,351	0,1699	0,0012	0,0038	0,3998	0,3325	0,0008	0,0108
			ρ	0,7152	0,1912	0,0008	0,014	0,7318	0,2188	0,0005	0,0178
		0,4	β_1	3,2711	-0,0027	0,0004	0,0005	3,275	-0,0015	0,0002	0,0003
				2,425	0,0062	0,0038	0,004	2,4207	0,0044	0,0006	0,0007
			p_1	0,36	-0,1001	0,0012	0,0028	0,3581	-0,1049	0,0011	0,0029
			β_2	2,1343	-0,0119	0,0015	0,0022	2,1424	-0,0082	0,0004	0,0007
				4,6746	0,0251	0,0143	0,0274	4,6507	0,0199	0,0001	0,0083
			p_2	0,3112	0,0372	0,0012	0,0014	0,3118	0,0394	0,0012	0,0013
			ρ	0,3934	-0,0166	0,0024	0,0025	0,3919	-0,0203	0,0024	0,0024
0,75	0,84	0,81	β_1	3,2226	-0,0175	0,0007	0,004	3,2886	0,0026	0,0004	0,0005
				2,4915	0,0338	0,007	0,0137	2,4078	-0,0009	0,0014	0,0014
			p_1	0,655	-0,1267	0,0011	0,0102	0,6812	-0,0917	0,0007	0,0054
			β_2	2,1247	-0,0163	0,0058	0,0071	2,1894	0,0136	0,0015	0,0023
				4,4581	-0,0223	0,0497	0,0601	4,3345	-0,0495	0,0025	0,0533
			p_2	0,8171	-0,0272	0,0008	0,0013	0,7985	-0,0494	0,0007	0,0025
			ρ	0,8028	-0,0187	0,0004	0,0006	0,8314	0,0162	0,0004	0,0006
		0,6	β_1	3,2579	-0,0067	0,0012	0,0017	3,3164	0,0111	0,0007	0,0021
				2,5494	0,0579	0,0111	0,0306	2,4525	0,0177	0,0027	0,0045
			p_1	0,8	0,0667	0,0008	0,0033	0,8084	0,0778	0,0006	0,004
			β_2	2,2237	0,0295	0,0045	0,0085	2,3485	0,0873	0,0012	0,0367
				4,5428	-0,0038	0,0354	0,0357	4,2925	-0,0587	0,002	0,0735
			p_2	0,8037	-0,0432	0,0009	0,0022	0,8133	-0,0318	0,0006	0,0013
			ρ	0,7768	0,2938	0,0005	0,0316	0,7999	0,3322	0,0005	0,0403
		0,4	β_1	3,2393	-0,0124	0,001	0,0027	3,3094	0,009	0,0007	0,0015
				2,4053	-0,002	0,0109	0,0109	2,3194	-0,0376	0,0025	0,0107
			p_1	0,775	0,0333	0,0009	0,0015	0,7641	0,0188	0,0007	0,0009
			β_2	2,1457	-0,0066	0,0071	0,0073	2,2507	0,042	0,0018	0,01
				4,6183	0,0128	0,0609	0,0643	4,4845	-0,0166	0,0035	0,0092
			p_2	0,857	0,0203	0,0007	0,0009	0,8342	-0,0069	0,0006	0,0007
			ρ	0,7197	0,7991	0,0007	0,1029	0,7602	0,9002	0,0008	0,1305
0,85	0,9	0,81	β_1	3,3273	0,0144	0,0011	0,0033	3,3818	0,031	0,0006	0,0109
				2,3882	-0,0091	0,0121	0,0126	2,3379	-0,0299	0,0026	0,0078
			p_1	0,8	-0,0588	0,0008	0,0033	0,7989	-0,0601	0,0005	0,0031
			β_2	2,084	-0,0352	0,0081	0,0138	2,171	0,0051	0,0014	0,0016
				4,8983	0,0742	0,0497	0,1642	4,6887	0,0282	0,0015	0,0181
			p_2	0,8659	-0,0379	0,0006	0,0018	0,8445	-0,0617	0,0005	0,0036
			ρ	0,881	0,0769	0,0001	0,0041	0,8939	0,0926	0,0002	0,0059

		0,6	β_1	3,2752	-0,0015	0,0014	0,0014	3,4139	0,0408	0,0008	0,0188
				2,6098	0,0829	0,0143	0,0542	2,4059	-0,0017	0,0036	0,0037
			p_1	0,84	-0,0118	0,0007	0,0008	0,8313	-0,022	0,0004	0,0008
			β_2	1,9542	-0,0953	0,0157	0,0581	2,1116	-0,0224	0,003	0,0054
				4,9218	0,0794	0,1349	0,2658	4,8402	0,0615	0,0048	0,0833
			p_2	0,9171	0,019	0,0004	0,0007	0,8783	-0,0241	0,0005	0,0009
			ρ	0,8107	0,3502	0,0004	0,0446	0,8665	0,4431	0,0003	0,0711
		0,4	β_1	3,1713	-0,0331	0,0021	0,0139	3,3133	0,0102	0,0008	0,0019
				2,6808	0,1124	0,0177	0,091	2,555	0,0602	0,0019	0,0229
			p_1	0,865	0,0176	0,0006	0,0008	0,8436	-0,0075	0,0004	0,0004
			β_2	2,1096	-0,0233	0,0233	0,0259	2,4037	0,1128	0,008	0,0673
				4,4841	-0,0166	0,2979	0,3037	3,8712	-0,1511	0,0388	0,5133
			p_2	0,9575	0,0638	0,0002	0,0035	0,9196	0,0217	0,0003	0,0007
			ρ	0,7848	0,9619	0,0004	0,1485	0,852	1,1297	0,0004	0,2046

5.6 Limites et approches alternatives

Joe(1997) a proposé une estimation en deux étapes qui consiste à estimer les paramètres des lois marginales et ensuite estimer le paramètre de la copule conditionnellement aux estimateurs précédemment estimés. Même si cette approche demeure valide pour les copules avec des lois marginales discrètes, l'interprétation du τ de Kendall demeure inappropriée, car il omet les probabilités des ensembles où les variables aléatoires sont égales. L'estimation par le maximum de vraisemblance permet également de développer des critères pour la sélection du modèle de copule ajustant le mieux les données discrètes ; le cas des données ordinales a été couvert par Choulakian et Tibeiro(2000).

Denuit et Lambert(2005) suggèrent à leur tour de rendre continues les variables aléatoires discrètes par le retrait/ajout d'une variable uniforme (le plus souvent sur $(0,1)$ par souci de simplification). Soit τ_n, la valeur du τ de Kendall pour un échantillon de taille n défini par :

$$\tau_n = \frac{N_c - N_d}{N} \tag{5.6.1}$$

où l'on définit les termes $N_c = \sum_{1 \leq i < j \leq n} 1\left\{(X_i - X_j)(Y_i - Y_j) > 0\right\}$ et
$N_d = \sum_{1 \leq i < j \leq n} 1\left\{(X_i - X_j)(Y_i - Y_j) < 0\right\}$. Notons que N désigne le nombre total de paires. Dans le cas continu, N est égal à $\frac{n(n-1)}{2}$. Lorsque l'on s'intéresse au cas discret, Genest et Neslehova(2007) ont effectué une vaste revue des propriétés des copules dans le cas discret. Il en résulte plusieurs points principaux.

La méthode qui consiste à rendre continues les variables discrètes en leur ajoutant des variables aléatoires indépendantes uniformes a certes pour avantage de conserver la mesure de concordance notamment le τ de Kendall mais il faut se garder de l'interpréter sans tenir compte des bornes de Carley.

Dans le cas discret, les ensembles d'égalité sont de probabilité non nulle. Plusieurs auteurs ont proposé des corrections notamment Kendall(1945), Goodman et Kruskal(1954), Vandenhende et Lambert(2003) et bien d'autres. Cependant, ces nouvelles mesures obtenues doivent être interprétées avec beaucoup de prudence. En effet, il est à noter que ces nouveaux estimateurs sont biaisés.

Enfin, étant donné le fait que le processus de discrétisation est irréversible, c'est-à-dire que la structure d'ordre qui prévalait sur les données initialement continues ne peut être reconstruite, Genest et Neslehova(2007) précisent que résoudre le problème de biais relié aux ensembles d'égalité exige de tenir compte des faits suivants :

- les définitions actuelles des ensembles N_c et N_d ne prennent pas en compte les ensembles d'égalité.
- considérer chaque cas d'égalité comme étant une moitié de concordance et une moitié de discordance n'affecte pas le numérateur de τ_n.

- l'allocation aléatoire des cas d'égalité comme des cas de concordance ou des cas de discordance ne saurait représenter une meilleure alternative car cela entrainerait des résultats inconsistants ; plusieurs analyses mèneraient à différents résultats.

Pour la suite, définissons le τ_b de Kendall obtenu sur un échantillon de taille n par :

$$\tau_{b,n}(X,Y) = \frac{N_c - N_d}{\sqrt{N_x N_y}} \tag{5.6.2}$$

où $N_x = \sum_{i<j} 1(X_i \neq X_j)$ et $N_y = \sum_{i<j} 1(Y_i \neq Y_j)$.

Lorsque les lois marginales sont continues, c'est-à-dire en l'absence d'égalité (de probabilité nulle), ces deux mesures $\tau_{b,n}$ et τ_n coïncident. Par contre, lorsque l'on considère des lois marginales discrètes, leurs comportements diffèrent nettement et les mesures définies en (5.6.1) et (5.6.2) ne peuvent plus être interprétées de la même façon. Par ailleurs, l'objectif ultime de notre étude est de déterminer la relation existant entre l'abondance de certaines espèces et les variables climatiques.

Nous proposons dans ce chapitre, une correction pour les données de Poisson augmentées en zéro. En effet, l'idée est plutôt de travailler sur la copule de données continues qui subira ensuite la discrétisation. Le paramètre de concordance ainsi obtenu pourrait être interprété seulement dans certains cas que nous spécifions.

Considérons pour la suite le cadre de copule bidimensionnelle gaussienne avec comme marginales des données continues avec de très larges proportions de données dans l'intervalle $(0,1)$. Ainsi après discrétisation, l'on se retrouve avec une abondance de zéros.

Supposons que la corrélation entre les données continues est très forte. Cela implique que l'ensemble d'égalité le plus fréquent qu'on obtiendrait après la discrétisation serait le couple $(0,0)$. Il est donc possible de réduire le biais des estimateurs de la mesure de

concordance en corrigeant uniquement pour les cas d'égalité en ce point. De plus, si les proportions sont de l'ordre de 75 à 90% avec une corrélation très forte les simulations indiquent que cette correction s'avère acceptable. En effet, les seuls cas d'égalité qu'il reste à corriger ne représentent qu'au plus 10% des observations.

L'avantage d'utiliser une telle mesure est qu'elle nous permet d'obtenir des valeurs de corrélation dans un intervalle plus proche de $(-1,1)$. De plus, il faut constater que l'efficacité de cette méthode réside dans le fait que les proportions de zéro sont très élevées et une corrélation très forte.

Étant donné l'ensemble des données discrètes, notre correction consiste à introduire de façon séquentielle une paire concordante ou discordante selon le signe de la corrélation jusqu'à maximiser la vraisemblance de la copule gaussienne modélisant les données continues. Certes nous utilisons le principe de continuité de Denuit et Lambert (2005) mais seulement pour obtenir les valeurs initiales de l'algorithme. Par ailleurs, Genest et Neslehova (2007) ont clairement établi les dangers et les limites d'une approche basée sur ce principe. Notre approche consiste plutôt à effectuer une correction uniquement au point d'égalité à l'origine de façon à élargir l'ensemble des valeurs possibles du coefficient de corrélation et du τ de Kendall. Une question naturelle serait : pourquoi ce point ? Le choix de ce point est conforté par le fait que pour des lois marginales discrètes de Poisson augmentées en zéro et possédant une forte corrélation, la valeur $(0,0)$ est de loin la plus fréquente. Cette approche permet d'avoir des estimateurs avec un biais relativement faible, car à chaque introduction de paire concordante ou discordante au point $(0,0)$, l'on diminue à la fois les quantités N_x et N_y de telle sorte que les valeurs de $\tau_{b,n}(X,Y)$ et $\tau_n(X,Y)$ tendent à se rapprocher. Il convient aussi de noter que, par la même occasion, le numérateur de l'équation (5.6.2) est modifié.

En tenant compte du fait qu'il existe, même après la correction, des cas d'égalité c'est-à-dire des ensembles de la forme $\{X_i = X_j, X_j \neq 0\}$ et $\{Y_i = Y_j, Y_j \neq 0\}$, cette correction est à utiliser avec précaution. Les simulations montrent l'effet néfaste notamment l'augmentation considérable du biais lorsque les proportions de zéros ne sont pas grandes et la corrélation devient faible.

5.7 Application

Nous présentons une application sur 761 données d'abondance de deux espèces d'oiseaux. La première espèce est le tangara écarlate (*piranga olivacea*) tandis que la seconde choisie est le pic à tête rouge (*melanerpes erythrocephalus*). Les variables explicatives sélectionnées sont la température moyenne annuelle (TAN), la précipitation moyenne annuelle (PAN) et l'étendue de la température du mois le plus chaud c'est-à-dire juillet (TAM).

Tableau 5.4 – Résultats basés sur les données par le maximum de vraisemblance et par l'approche bayésienne

coefficients	λ_1			λ_2			p_1	p_2	ρ
	TAN	PAN	TAM	TAN	PAN	TAM			
estimateurs du MV	0,381	0,001	-0,085	0,736	-0,013	0,975	0,574	0,936	0,476
	(0,017)	(0,0002)	(0,019)	(0,113)	(0,0016)	(0,123)	(0,022)	(0,017)	(0,023)
estimateurs de Bayes	0,380	-0,029	-0,089	0,736	-0,015	0,601	0,589	0,935	0,405
	(0,019)	(0,015)	(0,010)	(0,030)	(0,018)	(0,014)	(0,1016)	(0,020)	(0,024)

Les résultats de la table 5.4 indiquent que tous les coefficients obtenus par le maximum de vraisemblance sont statistiquement non nuls au seuil de 5%. Par contre, les estimateurs de Bayes obtenus diffèrent uniquement pour une seule variable explicative. En effet, les résultats selon les approches basées sur le MV et sur l'estimateur de Bayes bien que si-

milaires en général, présentent une divergence seulement pour la variable explicative PAN relative à la précipitation. En effet, la variable PAN explique, au seuil de 5% de façon significative les abondances du pic à tête rouge et du tangara écarlate dans le cas fréquentiste mais devient non significative selon l'approche bayésienne.

Conclusion

Ce dernier chapitre avait pour objectif de présenter une nouvelle mesure plus adaptée du τ de Kendall pour les données augmentées en zéro. D'un point de vue numérique, l'approche classique du maximum de vraisemblance semble très attrayante compte tenu de sa simplicité et aussi de ses propriétés asymptotiques (Joe 1997). Par contre, la méthode bayésienne basée sur l'intégration de Gauss-Hermite s'avère hautement complexe lorsque le nombre de paramètres augmente mais permet d'incorporer de l'information *a priori* disponible.

CONCLUSION

Ce travail présente une méthodologie de traitement des données discrètes en général et de comptage en particulier. Il constitue une ébauche dans le traitement des données multidimensionnelles pseudo ou quasi Poisson. Il serait intéressant d'étendre le modèle multidimensionnel de Poisson aux données spatiales afin de non seulement modéliser la dépendance spatiale mais aussi structurelle des variables dépendantes. Cette thèse avait pour objectif d'introduire trois concepts majeurs. Tout d'abord, nous introduisons une nouvelle forme de distribution qui permet de prendre en compte les corrélations positives et négatives et les propriétés de sa fonction de distribution. Les méthodes d'estimation proposées sont basées sur le principe d'augmentation de données à savoir l'algorithme EM et l'approche bayésienne.

Compte tenu de la disponibilité de données (obtenues) dans le cadre du projet CC-Bio qui présentaient une corrélation positive et une forte dispersion, nous avons par la suite développé une méthode d'estimation pour la loi multidimensionnelle de Poisson avec des effets aléatoires.

Enfin, le dernier chapitre introduit une nouvelle méthodologie d'estimation des copules dont les lois marginales sont des lois discrètes augmentées en zéro. L'idée est de construire une mesure de dépendance donc de corrélation dont l'ensemble des valeurs se rapproche autant que possible de l'intervalle $(-1;1)$. En effet, cette mesure serait donc égale au τ de Kendall des variables aléatoires continues dont la discrétisation permet d'obtenir des variables aléatoire discrètes augmentées en zéro.

Cependant, les différents modèles et méthodologies présentés dans ce document sont loin d'être une panacée pour modéliser toutes les interactions possibles entre les espèces car ils sont basés sur des hypothèses spécifiques notamment sur leurs distributions sous-jacentes et sur la structure de leurs corrélations.

TRAVAUX FUTURS

Dans le futur, nous nous intéresserons à développer un algorithme pour le calcul de la densité de la loi multidimensionnelle de Poisson-Skellam. Nous pourrions nous inspirer du travail de Tsiamyrtzis et Karlis (2004) afin de développer des stratégies efficaces de calcul de la densité de la loi multidimensionnelle de Poisson-Skellam en utilisant de façon parcimonieuse les relations de récurrences déjà établies.

Ensuite, il serait intéressant de se pencher sur la modélisation spatiale multidimensionnelle. Initialement introduite par Besag (1974), la littérature sur la modélisation autorégressive conditionnelle (ARC) abonde dans l'analyse des données spatiales. En effet, sa très grande simplicité demeure un atout majeur surtout dans le domaine bayésien quand il s'agit du tirage de lois. On peut citer dans ce sens Geman et Geman (1984), Besag *et al.* (1991), Clayton et Bernardinelli (1993) sans oublier Sun *et al.*, (1998). Cependant, la plupart des travaux ne considéraient que les modèles univariés. Cette partie consistera donc à introduire des effets aléatoires spatiaux comme variables explicatives par la modélisation ARCM (autorégressif conditionnel multidimensionnel) dans le cadre de la loi de Poisson-Skellam.

Enfin, il serait possible d'étendre la méthodologie d'estimation par les copules appliquée aux variables augmentées en zéro aux variables discrètes. En effet, il s'agirait de modéliser les données bidimensionnelles binaires comme cas limite et aussi des variables mixtes, c'est-à-dire une variable discrète et une autre continue.

BIBLIOGRAPHIE

Aitchison, J. et Ho C.H. (1989). The multivariate Poisson-log normal distribution. *Biometrika*, **vol. 76**, pp. 643-653.

Angers, J.-F. et Biswas, A. (2003). A Bayesian analysis of zero-inflated generalized Poisson model. *Computational Statistics and Data Analysis*, **vol. 42**, pp. 37-46.

Anscombe, F.J. (1948). The Transformation of Poisson, Binomial and Negative-Binomial Data. *Biometrika Trust*, **vol. 35**, pp. 246-254.

Banergee, S., Carlin, B., Gelfand A. (2004). Hierarchiecal modeling and analysis for spatial data. *Monographs on statistics and applied probability*, 452 p.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems (with discussion). *Journal of the Royal Statistical Society, Series B*, **vol. 36**, pp 192-236.

Besag, J., York, J. C. et Mollié, A. (1991). Bayesian image restoration, with two applications in spatial statistics (with discussion). *Annals of the Institute of Statistical Mathematics*, **vol. 43**, pp 1-59.

Choulakian V. et De Tibeiro, J. (2000). Copules archimédiennes et tableaux de contingence à variables qualitatives ordinales. Revue de Statistique Appliquée, **vol. 48**, pp 83-96.

Clayton, D. G. et Bernardinelli, L. (1993). Bayesian methods for mapping disease risks. *Small Area Studies in Geographical and Environmental*, Epidemiology, Oxford University Press, pp. 205-220.

Cressie, Noel A.C. (1993). Statistics for Spatial Data. New York : John Wiley and Sons, Inc.

Dellaportas, P. et Wright, D. (1991). Positive embedded integration in bayesian analysis. *Statistics and Computing*, **vol. 1**, pp. 1-12.

Denuit, M. et Lambert, P. (2005). Constraints on concordance measures in bivariate discrete data. *Journal of Multivariate Analysis*, **vol. 93**, pp. 40-57.

Dey, D.K., Ghosh, S.K., et Mallick, B.K. (2000). Generalized Linear Models : A Bayesian Perspective. *Marcel-Dekker, Inc., New York*.

Dietz, E. et Böhning, D. (2000). On estimation of the Poisson parameter in zero modified Poisson models. *Computational statistics and data analysis*, **vol. 34**, pp. 441-459.

Diggle, P.J., Liang, K.Y., et Zeger, S.L. (1994). Analysis of Longitudinal Data. *Oxford University Press*, Oxford.

Gelfand, A., Agarwal, D. et Silander, J.A. (2002). Investigating tropical deforestation using two stage spatially misaligned regression models. *Journal of Agricultural, Biological and Environmental Statistics*, **vol. 7**, pp. 420-439.

Gelman, A., Gilks, W. R. et Roberts, G. O. (1997). Weak convergence and optimal scaling of random walk Metropolis algorithms. *Annals of Applied Probability*, **vol. 7**, pp. 110-120.

Geman, S. et D. Geman (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Trans. Pattern Analysis and Machine Intelligence*, **vol. 6**, pp. 721-741.

Genest, C., Favre A.-C. (2007). Everything you always wanted to know about copula modeling but were afraid to ask. *Journal of Hydrologic Engineering*, **vol. 12**, pp. 347-368.

Genest C. et Neslehova J.(2007). A primer on copulas for count data. *The Astin Bulletin*, **vol. 37**, pp. 475-515.

Gitay, H. et Noble, I.R. (1997). What are the functional types and how should we seek them ? Plant functional types : their relevance to ecosystem properties and global change. T.M. Smith, H.H. Shugart, F.I. Woodward (Eds). Cambridge, Royaume Uni, Cambridge University Press : pp. 3-19.

Gitay, H., Noble, I.R. et Connelle, J.H. (1999). Deriving functional types for rainforest trees. *Journal of vegetation science* **vol. 177**, pp. 762-765.

Ghosh, S.K., Mukhopadhyay, P. and Lu, J.C. (2006). Bayesian analysis of zero inflated regression models. *Journal of Statistical Planning and Inference*, **vol. 136**, pp. 1360-1375.

Goodman L.A. et Kruskal W.H. (1954). Measures of association for cross classifications. *Journal of the American Statistical Association*, **vol. 49**, pp. 732-764.

Hastie, T. et Tibshirani, R. (1990). Generalized Additive Models. *Chapman and Hall*, London.

Hastings, W.K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, **vol. 57**, pp. 97-109.

Joe, H.,1997. Multivariate Models and Dependence Concepts. Chapman and Hall, London.

Johnson, Kotz et Balakrishnan (1997). Discrete Multivariate Distributions. Wiley Series in Probability and Statistics : Applied Probability and Statistics.

Kano, K., et Kawamura, K. (1991). On recurrence relations for the probability function of multivariate generalized Poisson distribution. *Communications in Statistics-Theory and Methods*, **vol. 20**, pp. 165-178.

Karlis, D. (2003). An EM algorithm for multivariate Poisson distribution and related models. *Journal of Applied Statistics*, **vol. 30**, pp. 63-77.

Karlis, D. et Meligkotsidou, L. (2005). Multivariate Poisson regression with covariance structure. *Statistics and Computing*, **vol. 15**, pp. 255-265.

Karlis, D. et Ntzoufras, I. (2006). Bayesian analysis of the differences of count data. Statistics in Medicine, **vol. 25**, 1885-1905.

Karlis, D. et Ntzoufras, I. (2009). Bayesian modelling of football outcomes : Using the Skellam's distribution for the goal difference. *Journal of Management Mathematics*, **vol. 20**, pp. 133-146.

Karlis, D. et Ntzoufras, I. (2003). Analysis of sports data using bivariate Poisson models. Journal of the Royal Statistical Society : Series D (The Statistician), **vol. 52**, pp. 381-393.

Kazutomo Kawamura (1985). A note on the recurrent relations for the bivariate Poisson distribution. *Kodai Math Journal*, **vol. 8**, pp. 70-78.

Kendall, M.G. (1945). The traitment of ties in ranking problems. *Biometrika*, **vol. 33**, pp. 239-251.

Kruskal, W.H. (1958). Ordinal measures of association. *Journal of the American Statistical Association*, **vol. 53**, pp. 814-861.

Labrecque-Synnott, F. (2010). Analyse bayésienne et classification pour modèles continus modifiés à zéro. Ph.D. thesis, Université de Montréal, 104 p.

Lambert, D. (1992). Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics*, **vol. 34**, pp. 1-14.

Le, N.D. et Zidek, J.V. (2006). Statistical Analysis of Environmental Space-Time Processes. Springer Series in Statistics, 341 p.

Li, C.-S., Lu, J.-C., Park, J., Kim, K., Brinkley, P., Peterson, J. (1999). Multivariate zero inflated Poisson models and their applications. *Technometrics*, **vol. 41**, pp. 29-38.

Mahamunulu (1967). A note on regression in the multivariate Poisson distribution. *Journal of the American Statistical Association*, **vol. 62**, pp. 251-258.

Mardia, K.V.(1988). Multi-dimensional multivariate gaussian Markov random fields with application to image processing. *Journal of Multivariate Analysis*, **vol. 24**, pp. 265-284.

McCulloch, Charles E. (1997). Maximum likelihood algorithms for generalized linear mixed models. *Journal of the American Statistical Association*, **vol. 92**, pp. 162-170.

McCulloch, C.E. et Searle, S.R. (2001). Generalized , Linear, and Mixed Models, *Wiley, New York*.

McHale, I. et Scarf, P. (2007). Modelling soccer matches using bivariate discrete distributions with general dependence structure, *Statistica Neerlandica*, **vol. 61**, pp.434-447.

Meligkotsidou, L. (2007). Bayesian multivariate Poisson mixtures with an unknown number of components. *Statistics and Computing*, **vol. 17**, pp. 93-107.

Mesfioui, M. et Tajar, A. (2005). On the properties of some nonparametric concordance measures in the discrete case. *Journal of Nonparametric Statistics*, **vol. 17**, pp. 541-554.

Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H., Teller, E. (1953). Equations of State Calculations by Fast Computing Machines. *Journal of Chemical Physics*, **vol. 21**, pp. 1087-1092.

Million, A., Riphahn T. R. et Wambach A. (2003). Incentive Effects in the Demand for Health Care : A Bivariate Panel Count Data Estimation, *Journal of Applied Econometrics*, **vol. 18**, pp. 387-405.

Mugglin, A. S., Carlin, B. P., and Gelfand, A. E. (2000). Fully Model-Based Approaches for Spatially Misaligned Data. *Journal of the American statistical Association*, **vol. 95**, pp. 877-887.

Murat, K.M. et Trivedi, P.K. (1999). Simulated maximum likelihood estimation of multivariate mixed-Poisson regression models, with application. *Econometrics Journal, Royal Economic Society*, **vol. 2**, pp. 29-48.

Naylor, J.C. et Smith, A.F.M. (1982). Applications of a Method for the Efficient Computation of Posterior Distributions. *Journal of the Royal Statistical Society. Series C*, **vol. 31**, pp. 214-225.

Pitt, M., Chan, D., Kohn, R. (2006). Efficient Bayesian interface for Gaussian copula regression models. *Biometrika*, **vol. 93**, pp. 537-554.

Ridout, M., Demetrio, C.G.B., Hinde, J. (1998). Models for count data with many zeros. International Biometric.

Roberts, G.O. et Rosenthal, S. (2001). Optimal scaling for various Metropolis-Hastings algorithms. *Statistical Science*, **vol. 16**, pp. 351-367.

Schoener, T.S. (1989). Food webs from the small to the large. *Ecology*, **vol. 70**, pp. 1559-1589.

Siddhartha, Chib et Rainer Winkelmann (2001). Markov Chain Monte Carlo analysis of correlated count data. *Journal of Business and Economic Statistics*, **vol. 19**, pp. 428-435.

Simpson Edward H. (1949) Measurement of diversity. *Nature*, pp. 163-688.

Skellam, J. G. (1946). The frequency distribution of the difference between two Poisson variates belonging to different populations. *Journal of the Royal Statistical Society : Series A*, **vol. 109**, pp. 296.

Sklar, A. (1959). Fonctions de répartition à n dimensions et leurs marges. *Publications de l'Institut de Statistique de l'Université de Paris*, **vol. 8**, pp. 229-231.

Smith, A. F. M., Skene, A. M., Shaw, J.E.H., Naylor, J.C. (1987). Progress with Numerical and Graphical Methods for Practical Bayesian Statistics. *Journal of the Royal Statistical Society. Series D*, **vol. 36**, pp. 75-82.

Song, P.X. (2000). Multivariate dispersion models generated from Gaussian copula. *Scandinavian Journal of Statistics*, **vol. 27**, pp. 305-320.

Sun, D., Speckman, P.L., Tsutakawa R.K. (1998). Random Effects in Generalised linear Mixed Models. *Generalized Linear Models : a Bayesian perpective*, Eds : Dey , D.K., Ghosh, S.K., Mallick, B.K.

Tanner, M.A., et Wong, W.H. (1987). The calculation of posterior distribution by data augmentation. *Journal of the American Statistical Association*, **vol. 82**, pp. 520-540.

Thuiller, W. (2003). BIOMOD : Optimising predictions of species distributions and projecting potential future shifts under global change. *Global Change Biology*, **vol. 9**, pp. 1353-362.

Thuiller, W., Lavorel, S., Sykes, M.T., Araújo, M.B. (2006). Using niche-based modelling to assess the impact of climate change on tree functional diversity in Europe. *Diversity and Distributions*, **vol. 12**, pp. 49-60.

Tsiamyrtzis, P. et Karlis, D.(2004). Strategies for efficient computation of multivariate Poisson Probabilities. *Communications in Statistics : Simulation and Computation*, **vol. 33**, pp. 271-292.

Tsionas, E. G. (1999). Bayesian analysis of the multivariate Poisson distribution. *Communications in Statistics-Theory and Methods*, **vol. 28**, pp. 431-451.

Tsionas, E.G. (2001). Bayesian multivariate Poisson regression. *Communications in Statistics - Theory and Methods*, **vol. 30**, pp. 243-255.

Vandenhende F. et Lambert, P. (2003). Improved rank-based dependance measures for categorical data. *Statistics and Probability Letters*, **vol. 63**, pp. 157-163.

Wong, F., Carter, C., Kohn, R. (2003). Efficient estimation of covariance selection models. *Biometrika*, **vol. 90**, pp. 809-830.

Yakowitz, S., Krimmel, J.E. et Szidarovszky, F.(1978). Weighted Monte Carlo Integration. *Journal on Numerical Analysis*, **vol. 15**, pp. 1289-1300.

ANNEXE A

Démonstrations des propositions énoncées

Démonstration de la proposition 3.2.2

Il faut distinguer deux cas selon que l est positif ou non.

Preuve de la première relation

Cas 1 : Supposons que $l > 0$

La fonction de masse de la loi bidimensionnelle de Poisson-Skellam s'écrit

$$\begin{aligned} P_2(k,l) &= \sum_{\delta=0}^{k} P(k-\delta|\theta_1)\,P(l+\delta|\theta_2)\,P(\delta|\theta_{12}) \\ &= P(k|\theta_1)\,P(l|\theta_2)\,P(0|\theta_{12}) + P(k-1|\theta_1)\,P(l+1|\theta_2)\,P(1|\theta_{12}) + \dots \\ &\quad + P(1|\theta_1)\,P(l+k-1|\theta_2)\,P(k-1|\theta_{12}) + P(0|\theta_1)\,P(l+k|\theta_2)\,P(k|\theta_{12})\ . \end{aligned}$$

En multipliant l'égalité précédente par k, nous obtenons

$$\begin{aligned} kP_2(k,l) &= kP(k|\theta_1)\,P(l|\theta_2)\,P(0|\theta_{12}) + (k-1)P(k-1|\theta_1)\,P(l+1|\theta_2)\,P(1|\theta_{12}) + \dots \\ &\quad + P(0|\theta_1)\,P(l+k|\theta_2)\,P(k|\theta_{12}) + P(k-1|\theta_1)\,P(l+1|\theta_2)\,P(1|\theta_{12}) \\ &\quad + 2P(k-2|\theta_1)\,P(l+2|\theta_2)\,P(2|\theta_{12}) + (k-1)P(1|\theta_1)\,P(l+k-1|\theta_2) \\ &\quad \times P(k-1|\theta_{12}) + kP(0|\theta_1)\,P(l+k|\theta_2)\,P(k|\theta_{12})\,. \end{aligned}$$

En utilisant la formule de récurrence de la fonction de masse de la loi unidimensionnelle

de Poisson, nous obtenons de façon successive :

$$\begin{aligned}
kP_2(k,l) &= \theta_1 P(k-1|\theta_1)P(l|\theta_2)P(0|\theta_{12}) + \theta_1 P(k-2|\theta_1)P(l+1|\theta_2)P(1|\theta_{12}) + \dots \\
&\quad + \theta_1 P(0|\theta_1)P(l+k-1|\theta_2)P(k-1|\theta_{12}) + \theta_{12} P(k-1|\theta_1)P(l+1|\theta_2) \\
&\quad \times P(0|\theta_{12}) + \theta_{12} P(k-2|\theta_1)P(l+2|\theta_2)P(1|\theta_{12}) + \dots \\
&\quad + \theta_{12} P(1|\theta_1)P(l+k-1|\theta_2)P(k-2|\theta_{12}) \\
&\quad + \theta_{12} P(0|\theta_1)P(l+k|\theta_2)P(k-1|\theta_{12}).
\end{aligned}$$

Ensuite, il suffit juste de réarranger les termes précédents et d'utiliser la définition de la fonction de masse de la loi bidimensionnelle de Poisson-Skellam pour obtenir :

$$\begin{aligned}
kP_2(k,l) &= \theta_1 \sum_{\delta=0}^{k-1} P(k-1-\delta|\theta_1)P(l+\delta|\theta_2)P(\delta|\theta_{12}) \\
&\quad + \theta_{12} \sum_{\delta=0}^{k-1} P(k-1-\delta|\theta_1)P(l+1+\delta|\theta_2)P(\delta|\theta_{12}) \\
&= \theta_1 P_2(k-1,l) + \theta_{12} P_2(k-1,l+1).
\end{aligned}$$

Cas 2 : Supposons que $l < 0$

La densité s'écrit alors

$$\begin{aligned}
P_2(k,l) &= \sum_{\delta=|l|}^{k} P(k-\delta|\theta_1)P(l+\delta|\theta_2)P(\delta|\theta_{12}) \\
&= P(k+l|\theta_1)P(0|\theta_2)P(|l||\theta_{12}) + P(k+l-1|\theta_1)P(1|\theta_2)P(|l|+1|\theta_{12}) + \dots \\
&\quad + P(1|\theta_1)P(l+k-1|\theta_2)P(k-1|\theta_{12}) + P(0|\theta_1)P(l+k|\theta_2)P(k|\theta_{12}).
\end{aligned}$$

En multipliant l'égalité précédente par k, nous obtenons

$$\begin{aligned} kP_2(k,l) &= (k+l)P(k+l|\theta_1)P(0|\theta_2)P(-l|\theta_{12})+(k+l-1)P(k+l-1|\theta_1) \\ &\quad \times P(1|\theta_2)P(1-l|\theta_{12})+\ldots+P(1|\theta_1)P(l+k-1|\theta_2) \\ &\quad \times P(k-1|\theta_{12})+(-l)P(k+l|\theta_1)P(0|\theta_2)P(-l|\theta_{12}) \\ &\quad +(-l+1)P(k+l-1|\theta_1)P(1|\theta_2)P(-l+1|\theta_{12}) \\ &\quad +\ldots+(k-1)P(1|\theta_1)P(l+k-1|\theta_2)P(k-1|\theta_{12}) \\ &\quad +(k)P(0|\theta_1)P(l+k|\theta_2)P(k|\theta_{12}). \end{aligned}$$

Ensuite,

$$\begin{aligned} kP_2(k,l) &= \theta_1 P(k+l-1|\theta_1)P(0|\theta_2)P(-l|\theta_{12})+\theta_1 P(k+l-2|\theta_1)P(1|\theta_2) \\ &\quad \times P(-l+1|\theta_{12})+\ldots+\theta_1 P(0|\theta_1)P(l+k-1|\theta_2)P(k-1|\theta_{12}) \\ &\quad +\theta_{12}P(k+l|\theta_1)P(0|\theta_2)P(-l-1|\theta_{12})+\theta_{12}P(k+l-1|\theta_1)P(1|\theta_2) \\ &\quad \times P(-l|\theta_{12})+\ldots+\theta_{12}P(1|\theta_1)P(l+k-1|\theta_2)P(k-2|\theta_{12}) \\ &\quad +\theta_{12}P(0|\theta_1)P(l+k|\theta_2)P(k-1|\theta_{12}). \end{aligned}$$

Après simplification, nous obtenons

$$\begin{aligned} kP_2(k,l) &= \theta_1 \sum_{\delta=-l}^{k-1} P(k-1-\delta|\theta_1)P(l+\delta|\theta_2)P(\delta|\theta_{12}) \\ &\quad +\theta_{12}\sum_{\delta=-(l+1)}^{k-1} P(k-1-\delta|\theta_1)P(l+1+\delta|\theta_2)P(\delta|\theta_{12}) \\ &= \theta_1 P(k-1,l)+\theta_{12}P(k-1,l+1). \end{aligned}$$

Preuve de la seconde relation

Cas 1 : Supposons que $l > 0$

Après multiplication, nous obtenons

$$\begin{aligned} lP_2(k,l) &= lP(k|\theta_1)P(l|\theta_2)P(0|\theta_{12}) + (l+1)P(k-1|\theta_1)P(l+1|\theta_2)P(1|\theta_{12}) + \dots \\ &\quad +(l+k)P(0|\theta_1)P(l+k|\theta_2)P(k|\theta_{12}) - P(k-1|\theta_1)P(l+1|\theta_2)P(1|\theta_{12}) \\ &\quad -2P(k-2|\theta_1)P(l+2|\theta_2)P(2|\theta_{12}) - (k-1)P(1|\theta_1)P(l+k-1|\theta_2) \\ &\quad \times P(k-1|\theta_{12}) - kP(0|\theta_1)P(l+k|\theta_2)P(k|\theta_{12}). \end{aligned}$$

En utilisant la formule de récurrence de la fonction de masse de la loi unidimensionnelle de Poisson, nous obtenons de façon successive :

$$\begin{aligned} lP_2(k,l) &= \theta_2 P(k|\theta_1)P(l-1|\theta_2)P(0|\theta_{12}) + \theta_2 P(k-1|\theta_1)P(l|\theta_2)P(1|\theta_{12}) + \dots \\ &\quad +\theta_2 P(0|\theta_1)P(l+k-1|\theta_2)P(k|\theta_{12}) - \theta_{12}P(k-1|\theta_1)P(l+1|\theta_2) \\ &\quad \times P(0|\theta_{12}) - \theta_{12}P(k-2|\theta_1)P(l+2|\theta_2)P(1|\theta_{12}) - \dots \\ &\quad -\theta_{12}P(0|\theta_1)P(l+k|\theta_2)P(k-1|\theta_{12}), \end{aligned}$$

puis

$$\begin{aligned} lP_2(k,l) &= \theta_2 \sum_{\delta=0}^{k} P(k-\delta|\theta_1)P(l-1+\delta|\theta_2)P(\delta|\theta_{12}) \\ &\quad -\theta_{12} \sum_{\delta=0}^{k-1} P(k-1-\delta|\theta_1)P(l+1+\delta|\theta_2)P(\delta|\theta_{12}) \\ &= \theta_2 P_2(k,l-1) - \theta_{12}P_2(k-1,l+1). \end{aligned}$$

Cas 2 : Supposons que $l < 0$

Après multiplication, nous avons

$$\begin{aligned} lP_2(k,l) &= lP(k+l|\theta_1)P(0|\theta_2)P(-l|\theta_{12}) + (l-1)P(k+l-1|\theta_1)P(1|\theta_2) \\ &\quad \times P(-l+1|\theta_{12}) + \ldots + (l-(l+k-1))P(1|\theta_1)P(l+k-1|\theta_2) \\ &\quad \times P(k-1|\theta_{12}) + (l-(l+k))P(k+l|\theta_1)P(0|\theta_2)P(-l|\theta_{12}) \\ &\quad + P(k+l-1|\theta_1)P(1|\theta_2)P(-l+1|\theta_{12}) \\ &\quad + \ldots + (l+k-1)P(1|\theta_1)P(l+k-1|\theta_2)P(k-1|\theta_{12}) \\ &\quad + (l+k)P(0|\theta_1)P(l+k|\theta_2)P(k|\theta_{12}). \end{aligned}$$

Ensuite,

$$\begin{aligned} lP_2(k,l) &= -\theta_{12}P(k+l|\theta_1)P(0|\theta_2)P(-l-1|\theta_{12}) - \theta_{12}P(k+l-1|\theta_1)P(1|\theta_2) \\ &\quad \times P(-l|\theta_{12}) - \ldots - \theta_{12}P(0|\theta_1)P(l+k|\theta_2)P(k-1|\theta_{12}) \\ &\quad + \theta_2 P(k+l-1|\theta_1)P(0|\theta_2)P(-l+1|\theta_{12}) + \ldots + \theta_2 P(1|\theta_1) \\ &\quad \times P(l+k-2|\theta_2)P(k-1|\theta_{12}) + \theta_2 P(0|\theta_1)P(l+k-1|\theta_2)P(k|\theta_{12}). \end{aligned}$$

Enfin,

$$\begin{aligned} lP_2(k,l) &= -\theta_{12}\sum_{\delta=-(l+1)}^{k-1} P(k-1-\delta|\theta_1)P(l+1+\delta|\theta_2)P(\delta|\theta_{12}) \\ &\quad + \theta_2 \sum_{\delta=-(l-1)}^{k} P(k-\delta|\theta_1)P(l-1+\delta|\theta_2)P(\delta|\theta_{12}) \\ &= \theta_2 P_2(k,l-1) - \theta_{12}P_2(k-1,l+1). \end{aligned}$$

■

Démonstration de la proposition 3.2.5

La preuve est similaire à celle de Kano et Kawamura (1991) sauf qu'ici le support de la distribution change. Cependant la démarche et les résultats demeurent inchangés. Notons $f(y) = \prod_{j=1}^{l} P(y_i|\theta_i)$, la vraisemblance de l'échantillon $(y_1, \ldots, y_n)$ et I_i le vecteur dont la i^e composante est égale à l'unité et toutes les autres sont nulles. Soit $y \in g^{-1}(x)$, alors $y - I_i \in g^{-1}(x - \phi_i)$; d'où l'inégalité

$$\sum_{y \in g^{-1}(x)} f(y - I_i) \leq \sum_{z \in g^{-1}(x-\phi_i)} f(z).$$

Par ailleurs, si $z \in g^{-1}(x - \phi_i)$ alors $z + \phi_i \in g^{-1}(x)$ et donc $y \in g^{-1}(x)$. Ainsi,

$$\sum_{z \in g^{-1}(x-\phi_i)} f(z) \leq \sum_{y \in g^{-1}(x)} f(y - I_i).$$

Par conséquent,

$$\begin{aligned} P(X = x - \phi_i) &= \sum_{y \in g^{-1}(x)} f(y - I_i) \\ &= \sum_{y \in g^{-1}(x)} \left(\prod_{j \neq i}^{l} P(y_j|\theta_j) \right) P(y_i - 1|\theta_i) \\ &= \sum_{y \in g^{-1}(x)} \left(\prod_{j \neq i}^{l} P(y_j|\theta_j) \right) \frac{y_i}{\theta_i} P(y_i|\theta_i) \\ &= \frac{1}{\theta_i} \sum_{y \in g^{-1}(x)} y_i f(y) \quad i = 1, \ldots, l \quad . \end{aligned}$$

En multipliant par $\theta_i s_i$, nous obtenons en faisant varier l'indice $i = 1, \ldots, l$

$$\theta_i s_i P(X = x - \phi_i) = \sum_{y \in g^{-1}(x)} y_i s_i f(y).$$

En sommant les l équations précédentes,

$$\sum_{i=1}^{l} \theta_i s_i P(X = x - \phi_i) = \sum_{y \in g^{-1}(x)} \left(\sum_{i=1}^{l} y_i s_i \right) f(y).$$

Par la suite, s'il existe un vecteur $s = (s_1, \ldots, s_l)$ tel que $\sum_{i=1}^{l} y_i s_i = 1$

$$\sum_{i=1}^{l} \theta_i P(X = x - \phi_i) = \sum_{y \in g^{-1}(x)} f(y) \equiv P(X = x).$$

La suite de la démonstration est exactement la même que celle de Kano et Kawamura (1991) car elle consiste à résoudre un système d'équations linéaires par la méthode de Cramer afin de déterminer le vecteur s.

Pour tout r tel que $1 \leq r \leq \text{rg}(A)$, il existe un mineur A_r tel que $\det(A_r) \neq 0$, donc il nous suffit d'utiliser r équations à l inconnues. Ainsi le système d'équations suivant est obtenu :

$$\begin{cases} \alpha_1 a_{11} + \alpha_1 a_{21} + \cdots + \alpha_l a_{l1} = x_1, \\ \alpha_1 a_{12} + \alpha_2 a_{22} + \cdots + \alpha_l a_{l2} = x_2, \\ \qquad\qquad \vdots \\ \alpha_1 a_{r1} + \alpha_2 a_{2r} + \cdots + \alpha_l a_{lr} = x_r. \end{cases}$$

Soient les lignes de A_r allant de $1, \ldots, r$ et les colonnes allant de $l-r+1, l-r+2, \ldots, l$. Cela équivaut à considérer le nouveau système :

$$\sum_{i=1}^{l-r} \alpha_i \phi_i^{(r)} + \sum_{i=l-r+1}^{l} \alpha_i \phi_i^{(r)} = x^{(r)},$$

$$\sum_{i=l-r+1}^{l} \alpha_i \phi_i^{(r)} = x^{(r)} - \sum_{i=1}^{l-r} \alpha_i \phi_i^{(r)}.$$

Posons $A_r = \left(\phi_{l-r+1}^{(r)}, \ldots, \phi_l^{(r)}\right)$ alors $\det(A_r) = \det\left(\phi_{l-r+1}^{(r)}, \ldots, \phi_l^{(r)}\right)$ et $B_{l-r} = \left(\phi_1^{(r)}, \ldots, \phi_{l-r}^{(r)}\right)$.
Ainsi

$$A^{(r)} = [B_{l-r}\,;\,A_r] = \begin{bmatrix} a_{11} & a_{21} & \ldots & a_{l-r,1} & a_{l-r+1,1} & \ldots & a_{l1} \\ a_{12} & a_{22} & \ldots & a_{l-r,2} & a_{l-r+1,2} & \ldots & a_{l2} \\ \vdots & \vdots & \ldots & \vdots & \vdots & \ldots & \vdots \\ a_{1r} & a_{2r} & \ldots & a_{l-r,r} & a_{l-r+1,r} & \ldots & a_{lr} \end{bmatrix}.$$

Puisque A_r par définition est de rang plein égal à r, alors $\det(A_r) \neq 0$. Ainsi, nous pouvons utiliser la méthode de Cramer pour déterminer les $(\alpha_{l-r+1}, \ldots, \alpha_l)$. D'où

$$\alpha_{l-r+1} = \frac{\det\left(x^{(r)} - \sum_{i=1}^{l-r} \alpha_i \phi_i^{(r)}, \phi_{l-r+2}^{(r)}, \ldots, \phi_l^{(r)}\right)}{\det\left(\phi_{l-r+1}^{(r)}, \ldots, \phi_l^{(r)}\right)}.$$

De façon générale, nous pouvons écrire :

$$\alpha_{l-r+j} = \frac{\det\left(\phi_{l-r+1}^{(r)}, \ldots, x^{(r)} - \sum_{i=1}^{l-r} \alpha_i \phi_i^{(r)}, \phi_{l-r+j+1}^{(r)}, \ldots, \phi_k^{(r)}\right)}{\det\left(\phi_{l-r+1}^{(r)}, \ldots, \phi_l^{(r)}\right)} \quad \forall j \in \{1, \ldots, r\}.$$

Dans la suite, supposons que $s_{l-r+1} = s_{l-r+2} = \ldots = s_k = 0$ alors la condition sur les α_i devient

$$\sum_{i=1}^{l} \alpha_i s_i = 1 \Rightarrow \sum_{i=1}^{l-r+1} \alpha_i s_i = 1$$

$$\sum_{i=1}^{l-r} \alpha_i s_i + \alpha_{l-r+1} s_{l-r+1} = 1$$

$$\alpha_{l-r+1} s_{k-r+1} = \frac{s_{l-r+1} \det\left(x^{(r)} \phi_{l-r+2}^{(r)}, \ldots, \phi_l^{(r)}\right) - \sum_{i=1}^{l-r} \alpha_i s_{l-r+1} \det\left(x^{(r)} \phi_{l-r+2}^{(r)}, \ldots, \phi_l^{(r)}\right)}{\det\left(\phi_{l-r+1}^{(r)}, \ldots, \phi_l^{(r)}\right)}$$

$$\left(1-\sum_{i=1}^{l-r}\alpha_i s_i\right)\det\left(\phi_{l-r+1}^{(r)},\ldots,\phi_l^{(r)}\right) = s_{l-r+1}\det\left(x^{(r)}\phi_{l-r+2}^{(r)},\ldots,\phi_l^{(r)}\right) - s_{l-r+1}\sum_{i=1}^{l-r}\alpha_i\det\left(x^{(r)}\phi_{l-r+2}^{(r)},\ldots,\phi_l^{(r)}\right)$$

$$\sum_{i=1}^{l-r}\alpha_i\left[s_i\det\left(\phi_{l-r+1}^{(r)},\ldots,\phi_l^{(r)}\right)-s_{l-r+1}\det\left(\phi_i^{(r)},\phi_{l-r+2}^{(r)},\ldots,\phi_l^{(r)}\right)\right] + \left[s_{l-r+1}\det\left(x^{(r)}\phi_{l-r+2}^{(r)},\ldots,\phi_l^{(r)}\right)-\det\left(\phi_{l-r+1}^{(r)},\ldots,\phi_l^{(r)}\right)\right]=0 \quad \forall\alpha_i.$$

Finalement, étant vraie pour tout α_i, l'équation précédente implique :

$$\begin{cases} s_i\det\left(\phi_{l-r+1}^{(r)},\ldots,\phi_l^{(r)}\right)-s_{l-r+1}\det\left(\phi_i^{(r)},\phi_{l-r+2}^{(r)},\ldots,\phi_l^{(r)}\right)=0, \\ s_{l-r+1}\det\left(x^{(r)}\phi_{l-r+2}^{(r)},\ldots,\phi_l^{(r)}\right)-\det\left(\phi_{l-r+1}^{(r)},\ldots,\phi_l^{(r)}\right)=0. \end{cases}$$

$$s_i=\frac{\det\left(\phi_i^{(r)},\phi_{k-r+2}^{(r)},\ldots,\phi_k^{(r)}\right)}{\det\left(x^{(r)},\phi_{k-r+2}^{(r)},\ldots,\phi_k^{(r)}\right)} \quad \forall i\in\{1,\ldots,k-r+1\}.$$

■

ANNEXE B : ÉCHANTILLONNAGE PONDÉRÉ

Détermination de la loi instrumentale

Cette section présente la méthode utilisée pour estimer les coefficients comme moyenne *a posteriori*, estimateur de Bayes sous la fonction de perte quadratique.

Approximation normale

Selon Anscombe (1948), la loi de Poisson peut être approximée par une loi normale grâce à une transformation stabilisatrice de la variance du type racine carrée. Le nombre réel c est choisi de façon à ajuster au mieux les données. En général, l'une des valeurs les plus utilisées est $c = 1/8$

$$\sqrt{Y+c} \sim \mathcal{N}\left(\sqrt{\mu}, \frac{1}{4}\right).$$

Cette transformation permet ainsi d'obtenir une approximation normale de variance constante et égale à $1/4$. Ainsi, il est beaucoup plus aisé de trouver une loi instrumentale.

Ensuite, après avoir transposé le problème au cas gaussien, il s'agit de déterminer les coefficients β_j de la régression en résolvant le programme suivant :

$$\begin{aligned} \beta_j &= \operatorname{argmin} \sum_{i=1}^{n} \left(\sqrt{(y_{ij}+c)} - e^{\frac{1}{2}\mathbf{Z}_{ij}\beta_j} \right)^2 \\ &= \operatorname{argmin} \sum_{i=1}^{n} \left(v_{ij} - \frac{1}{2}\mathbf{Z}_{ij}\beta_j \right)^2, \end{aligned}$$

où $v_{ij} = \sqrt{(y_{ij}+c)} - 1$; Par la suite, nous obtenons aisément l'estimateur des moindres carrés ordinaires (MCO) :

$$\beta_j^{mco} = 2\left(\mathbf{Z}_j\mathbf{Z}_j'\right)^{-1}\mathbf{Z}_j'\mathbf{V}.$$

La variance de cet estimateur est donc

$$Var\left(\beta_j^{mco}\right) = \left(\mathbf{Z}_j\mathbf{Z}_j'\right)^{-1}.$$

Inférence via l'intégration par Monte Carlo

La seconde étape consiste à utiliser une loi normale de moyenne $\beta_{\mathbf{j}}^{mco}$ et de variance $\left(Z_jZ_j'\right)^{-1}$ comme loi instrumentale afin d'obtenir la moyenne et la variance *a posteriori* des β_j notées respectivement $\tilde{\beta}_j^{(1)}$ et $\widetilde{\Sigma}_j^{(1)}$.

Une fois ces nouveaux estimateurs obtenus, à l'étape k, nous utiliserons donc une loi normale de moyenne $\tilde{\beta}_j^{(k)}$ et de variance $\widetilde{\Sigma}_j^{(k)}$ comme loi instrumentale afin d'obtenir la moyenne et la variance *a posteriori* des β_j notées respectivement $\tilde{\beta}_j^{(k+1)}$ et $\widetilde{\Sigma}_j^{(k+1)}$.

Continuer les itérations jusqu'à ce qu'un critère de convergence préalablement défini soit atteint. En général, utiliser l'erreur relative comme critère permet d'obtenir des résultats satisfaisants.