Champs aléatoires de Markov cachés pour la cartographie du risque en épidémiologie

Lamiae Azizi

UNIVERSITÉ DE GRENOBLE

THÈSE

Pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ DE GRENOBLE

Spécialité : **Mathématiques Appliquées**

Arrêté ministériel : 7 août 2006

Présentée par

« Lamiae AZIZI »

Thèse dirigée par **« Florence FORBES »** et
codirigée par **« Myriam CHARRAS-GARRIDO»**

préparée au sein du laboratoire **LJK**
dans l'école doctorale **MSTII**

Champs aléatoires de Markov cachés pour la cartographie du risque en épidémiologie

Thèse soutenue publiquement le ,
devant le jury composé de :

M. Christophe BIERNACKI
Professeur à l'Université de Lille 1, Rapporteur
M. Nicolas MOLINARI
Maître de conférence à l'Université de Montpellier 1, Rapporteur
M. Olivier FRANCOIS
Professeur à l'INP Grenoble, Examinateur
M. Jean-François MARI
Professeur à l'Université Nancy 1, Examinateur
Mme. Florence FORBES
Directeur de recherche à l'INRIA Rhône-Alpes, Directeur de thèse
Mme. Myriam CHARRAS-GARRIDO
Chargé de recherche à l'INRA Clermont-Ferrand – Theix, Co-Directeur

Résumé

La cartographie du risque en épidémiologie permet de mettre en évidence des régions homogènes en terme du risque afin de mieux comprendre l'étiologie des maladies. Nous abordons la cartographie automatique d'unités géographiques en classes de risque comme un problème de classification à l'aide de modèles de Markov cachés discrets et de modèles de mélange de Poisson. Le modèle de Markov caché proposé est une variante du modèle de Potts, où le paramètre d'interaction dépend des classes de risque.
Afin d'estimer les paramètres du modèle, nous utilisons l'algorithme EM combiné à une approche variationnelle champ-moyen. Cette approche nous permet d'appliquer l'algorithme EM dans un cadre spatial et présente une alternative efficace aux méthodes d'estimation de Monte Carlo par chaîne de Markov (MCMC).
Nous abordons également les problèmes d'initialisation, spécialement quand les taux de risque sont petits (cas des maladies animales). Nous proposons une nouvelle stratégie d'initialisation appropriée aux modèles de mélange de Poisson quand les classes sont mal séparées. Pour illustrer ces solutions proposées, nous présentons des résultats d'application sur des jeux de données épidémiologiques animales fournis par l'INRA.

Mots-clés : Classification, Champs aléatoires de Markov cachés discrets, Cartographie du risque, Mélanges de Poisson, Modèle de Potts, EM champ-moyen.

Abstract

The analysis of the geographical variations of a disease and their representation on a map is an important step in epidemiology. The goal is to identify homogeneous regions in terms of disease risk and to gain better insights into the mechanisms underlying the spread of the disease. We recast the disease mapping issue of automatically classifying geographical units into risk classes as a clustering task using a discrete hidden Markov model and Poisson class-dependent distributions. The designed hidden Markov prior is non standard and consists of a variation of the Potts model where the interaction parameter can depend on the risk classes.
The model parameters are estimated using an EM algorithm and the mean field approximation. This provides a way to face the intractability of the standard EM in this spatial context, with a computationally efficient alternative to more intensive simulation based Monte Carlo Markov Chain (MCMC) procedures.
We then focus on the issue of dealing with very low risk values and small numbers of observed cases and population sizes. We address the problem of finding good initial parameter values in this context and develop a new initialization strategy appropriate for spatial Poisson mixtures in the case of not so well separated classes as encountered in animal disease risk analysis. We illustrate the performance of the proposed methodology on some animal epidemiological datasets provided by INRA.

Key-words : Classification, Discrete hidden Markov random field, Disease mapping, Poisson mixtures, Potts model, Variational EM.

À mes parents, mes deux soeurs et mon petit frère ;

Remerciements

Je commence par remercier les membres de mon jury...

Ma sincère reconnaissance à Christophe Biernacki et Nicolas Molinari qui ont accepté d'évaluer mon manuscrit et qui l'ont jugé d'une manière constructive. Un grand merci notamment à Olivier Françis pour avoir présidé mon jury. Je remercie aussi Jean-François Mari d'avoir accepté d'être examinateur de ce travail. Je remercie ensuite Florence Forbes d'avoir dirigé cette thèse. Un grand merci à ma codirectrice de thèse Myriam Charras-Garrido pour son soutien, malgré la distance qui nous séparait, j'ai toujours apprécié nos échanges professionnels et personnels.

Au BSHM où l'aventure a commencé...

Toute aventure de thèse, et avant de se terminer, a un début, la mienne a commencé dans les couloirs et les salles de cours du Bâtiment des Sciences Humaines et Mathématiques de l'UPMF à Grenoble, avec mes anciens enseignants du Master MASSS. À toutes ces personnes, merci de m'avoir toujours fait confiance, aidé, conseillé et de m'avoir fait découvrir le monde de la recherche. J'en oublie certainement quelques uns mais je remercie en particulier Abdel abdali, Karim Benhenni, Mohammed el-methni, Mustapha Rachdi, Gérard et Catherine d'aubigny. Enfin, mes plus grands remerciements s'adressent à mes deux grands amis Rémy Drouilhet et Jean-François Coeurjolly. Merci Rémy d'avoir toujours été là, merci pour ton soutien dans les moments de doute, merci pour les discussions constructives que l'on a eues pendant ces longues années. Merci Jean-François pour tes conseils, ton amitié, ta gentillesse et ta rigeur. Les mots ne suffiront jamais pour vous remercier tous les deux mais ...Merci.

Chez bioMérieux où la deuxième étape de l'aventure s'est réalisée...

Le stage de fin détudes dans l'équipe IT chez bioMérieux était très stimulant. C'était un plaisir de pouvoir discuter avec de nombreuses personnes issues de différentes disciplines. Cet environnement multidisciplinaire m'a permis d'apprécier l'application des formalismes mathématiques à cette discipline aussi impressionnante qu'est la biologie. Je remercie tous les chercheurs, ingénieurs ou stagiaires que j'ai pu côtoyer pendant mon stage. Je remercie mon ancien maître de stage René Vachon pour son soutien infaillible pendant le stage et la thèse, son amitié et sa confiance. Un grand merci à mes excellentes amies Céline Vidal, Sophie Grosz, Magali Jaillard et Tioka Rabeony pour leur humour, gentillesse et leur soutien inconditionnel tout au long de mon travail de thèse. Merci à vous pour tous les moments qu'on a passés ensemble ces quatre dernières années. Enfin je remercie mon «grand frère» Hafid Abaibou pour ses encouragements, ses conseils lumineux et pour plein d'autres choses que je

ne pourrai pas citer ici.

À l'INRA qui a permis à cette aventure de continuer...

Ce travail de thèse n'aurait jamais eu lieu sans le soutien financier de l'INRA de Clermont-Ferrand-Theix. Je les remercie de m'avoir confié ce sujet de thèse intéressant et passionnant. Mes différents séjours dans l'unité d'épidémiologie animale m'ont permis de découvrir une remarquable et singulière atmosphère qui prévaut dans ce collectif au très bel esprit ! Cela résulte en partie de la qualité de management de Christian Ducrot et Gwenael Vrouc'h que je remercie énormèment pour leur soutien, même à distance, et leur souci du bien être de chaque membre de l'unité et spécialement les thésards. Je remercie David Abrial de m'avoir appris toutes les bases de l'épidémiologie, pour sa disponibilité et son incommensurable optimisme. Je remercie également Françoise Ondet pour sa disponibilité et sa gentillesse. Un grand Merci à Valérie Poux et Nelly Dorr pour leurs côtés maternelles et leurs messages de soutien tout au long de ces trois années. Une mention spéciale à Maud Marsot, Elsa Jourdan, laurent Crespin, Severine Bord, Sebastien Masseglia, Xavier Bailly, Jocelyn De Goer, Mathieu Gonnet, Patrick Gasqui pour leur bonne humeur, leur gentillesse et les inoubliables pauses cafés passées avec eux.

À l'INRIA où l'aventure a eu lieu et s'est terminée...

Mon travail de thèse s'est réalisé au sein de l'INRIA Grenoble. J'ai pu côtoyer pendant ces trois années plusieurs personnes qui m'ont soutenue et avec qui j'ai partagé des moments difficiles et d'autres plus agréables. Une attention très particulière à toutes nos assisstantes Barta Beneddine, Patricia Oddos, Imma Pressenguer et Marie-anne Dauphin, qui se sont toujours occupées de toutes les formalités administratives et qui m'ont soutenue du début jusqu'à la fin chacune à sa manière. Un grand merci à mon ami Vasil Khalidov pour son sourire, sa gentillesse, ses chocolats et fleurs et son soutien systématique. Je remercie également Antoine Letouzey pour sa présence, sa joie de vivre et son support infaillible surtout lors de la dernière ligne droite. Many thanks to my friend Ramya Narasimha for the touch of sparkle she brings to my life during these hard three years, for her support and for her friendship. I needed a mentor and found you Ramya. Thank you also Darren Wraith for supporting me during all these years, sharing the same office with you for such a long time was a pleasure. I will miss our long discussions about everything and nothing, our games and all your gifts and advices. I can not forget the third person sharing this office with us, Senan Doyle, the most kind Irish that I have ever meet. Thanks to both of you for everything.

Un grand merci particulier à toute la troupe des cakes au beurre : Amael Delaunoy, Pierre-Edouard Landes (le Marc Levy des thésards), Régis Perrier, Simone Gasparini (le chef du tiramisu), Miles Hansard, Gaetan Janssens, Benjamin Petit, Xavi Alameda, Michel Amat, Kiran Varanasi, Svetlana Artemova, Sergei Grudinin, Mael Bosson, Laurentiu Trifan et Diana

Stefan pour les super moments passés ensembles, pour les pauses cafés et les pains au chocolat du matin, les BBQ sur le parking de l'INRIA, les sorties ciné, les parties de laser-game, le ski...Que d'excellents moments que je ne suis pas prête à oublier. Bonne continuation à tous.
Merci aussi à toutes mes équipes d'adoption. Je remercie donc tous les permanents et non-permanents des équipes de recherche : Nano-D, IBIS, Morpheo, Perception, Steep, Artis, POP-ART, Evasion ou encore Prima.
J'ai également une pensée affectueuse à tous les services de support à la recherche qui ont permis à leur manière à l'aboutissement de ce travail. Je pense ainsi à Gérard finet, Valérie Martinez, Alain Kersaudy, Aurélia Mouton, Brigitte Duret, Eric Angles, Gaelle Riverieux et Isabelle Rey. Merci pour votre soutien, votre humour, votre enthousiasme et la bonne ambiance d'équipe lors de nos différentes courses.
Je remercie aussi mes amis Marie-Jose Martinez, Jean-Baptiste Durand, Qai Kin, Jonathan el Methni et El-hadji Deme pour leur gentillesse et pour toutes nos amusantes discussions.

Les conférences pendant l'aventure...

Je ne pourrais pas oublier de remercier les amis rencontrés lors des différentes conférences. Je remercie en particulier Jean-Noel Bacro pour sa gentillesse, son humour, ses conseils sur le jury et l'après-thèse et son parfait anglais. Un grand Merci à Cècile Hardouin de m'avoir invitée aux sessions qu'elle organisait, pour son amitié et sa bonne humeur. Un très grand merci notamment à mes amis Sophie-dabo Niang, Anne Françoise Yao, Baba Thiam et Besnik Pumo. Les rues de Rome et de Tunis n'auraient pas pu être étincellantes sans la compagnie de mes super amis Bordelais : Marie Chavent, Jérôme Saracco, Benoît Liquet et Vanessa Kuentz. Une pensée particulière à Stéphane Robin, Ali Gannoun, Aurore Lavigne, Nicolas Eckert et Liliane Bel. Je m'excuse de tous ceux que j'ai pu oublier mais merci à tous ceux que j'ai croisé un jour et qui m'ont poussée à aller au delà des difficultés.

Et finalement pour leur soutien constant...

Je termine ces remerciements par remercier les personnes qui m'ont toujours soutenue que ce soit dans mon travail ou dans la vie de tous les jours : ma petite famille. Affectueusement merci à mes excellents parents, Ahmed et Fatiha qui m'ont toujours poussée vers l'avant et qui m'ont toujours soutenue dans mes choix même les plus «tordus» d'entre eux. Malgré la distance, ils ont su m'encourager, m'apprendre à ne jamais baisser les bras et à dépasser les obstacles. Je remercie aussi mes deux soeurs Meriem et Asmae et mon petit frère Zakariae d'avoir toujours été souriants, prévenants et patients. Merci à vous cinq d'avoir toujours été là pour moi et surtout dans les moments difficiles, sans vous je n'aurai jamais pu être là.

Table des matières

CHAPITRE 1

Introduction

Sommaire

L'évolution actuelle des problèmes de santé dans le monde montre certaines faiblesses de la médecine moderne face aux risques d'émergence de maladies nouvelles ou de ré-émergence de maladies que l'on pensait avoir éradiquées. Cette évolution incite aujourd'hui certains scientifiques à parler d'une quatrième transition épidémiologique, celle de la "ré-émergence" (McMichael, 2001). La médecine et les disciplines connexes se sont essentiellement préoccupées, jusqu'à présent, de comprendre les pathologies humaines dans un contexte socio-économique où les agents étiologiques pouvaient être éradiqués grâce au progrès de la médecine. Cependant, l'évolution des maladies a incité la pensée scientifique à s'imprégner de la notion de complexité et de la nécessité d'une approche systémique (Froment, 1997). Au sein du système d'interactions hôte-pathogène s'est donc ajoutée au fil du temps, la prise en compte d'un troisième élément, l'environnement. Depuis quelques dizaines d'années, tend donc à se développer une approche pluridisciplinaire visant à prendre en compte les causes «externes» dans le développement des pathologies, telles que les facteurs physiques, chimiques, climatiques, écologiques mais aussi les comportements personnels ou culturels pouvant favoriser l'émergence ou la ré-émergence de maladies infectieuses et parasitaires. C'est dans cette optique, que ce sont développées un ensemble de thématiques de recherche sur les relations entre environnement et santé animale ou humaine (Washino and Wood, 1994; Epstein, 1999; Avruskin et al., 2004).

Un certain nombre de facteurs ou de déterminants influant sur la santé ont des caractéristiques spatiales et temporelles distinctes. Si l'on en vient aujourd'hui à étudier la spatialisation des données épidémiologiques, c'est tout d'abord parce que l'on se rend compte de l'importance de l'espace dans la structuration des phénomènes de contagion-diffusion de maladies, mais

également parce qu'il existe maintenant de nouveaux outils permettant d'intégrer cette information. Depuis plusieurs années, les outils de télédétection et les Systèmes d'Information Géographiques (SIG) sont utilisés pour de telles recherches. Ces outils permettent d'intégrer une composante spatiale à l'observation des dynamiques de maladies et d'identifier certains paramètres environnementaux.

1.1 Motivations

En épidémiologie, comprendre comment certains facteurs peuvent influencer une dynamique épidémiologique requiert de pouvoir prendre en compte la variation spatiale et temporelle de l'occurence des maladies et des infections afin de mettre en évidence des hétérogénéités liées par exemple à des facteurs de risque. Que ce soit dans le contexte humain ou animal, l'épidémiologie vise à résumer l'information concernant la variation spatiale (et/ou temporelle) du risque pour les maladies étudiées afin d'évaluer l'hétérogenéité spatiale (et/ou temporelle) et les structures sous-jacentes qui lui y sont associées. Il s'agit, par exemple, de mettre en évidence les régions à risque et d'obtenir des indices quant à l'étiologie (l'étude des causes et des facteurs) des maladies.
Dans le domaine vétérinaire, comme dans le médical, l'épidémiologie réunit différentes étapes d'une démarche globale qui vise à lutter contre les maladies. Les données pour les maladies animales d'élevage sont obtenues le plus souvent au travers d'enquêtes conduites en élevage et/ou dans les laboratoires de diagnostic. Pour les maladies humaines elles peuvent aussi être extraites des dossiers médicaux. Les données d'épidémiologie animale sont de nature complexe du fait qu'elles peuvent provenir de différentes enquêtes et de différents environnements.
Un des enjeux du XXI^e siècle est d'améliorer la collaboration entre le monde médical et le monde vétérinaire en matière d'épidémiologie et le partage d'information car la plupart des maladies émergentes préoccupantes sont liées à l'environnement et souvent à des réservoirs dans le monde animal. Parfois les pathogènes de l'homme peuvent aussi infecter les animaux d'élevage et sauvages. D'où l'importance d'étudier les maladies animales, en plus des maladies humaines, et de comprendre leur mécanismes. Cette compréhension servira de fondement à la logique des interventions faites dans l'intérêt des populations visées.
L'objectif principal de la modélisation statistique spatialisée dans ce domaine est d'estimer le risque pour chaque unité géographique de la zone étudiée. Les mesures de lutte étant différentes en épidémiologie animale et en épidémiologie humaine, cela influe sur le traitement

des données et leur représentation. Il en découle des motivations de modélisation sensiblement différentes. En épidémiologie animale, on est souvent amené à utiliser des mesures de lutte ciblées, comme l'abattage, la restriction de circulation ou encore la vaccination, dans certaines régions à risque. On a donc besoin de précisément délimiter ces régions. Au contraire, dans le cas humain, les vaccinations se font rarement en fonction des régions mais plutôt en fonction de la population à risque (par métiers : médecins ou éléveurs, par catégorie d'âge :personnes âgées ou encore enfants) et la restriction de circulation n'existe pas vraiment.
Ainsi un objectif courant de la modélisation en épidémiologie animale est d'obtenir une segmentation des niveaux de risque en un nombre fini de classes en plus d'une estimation raisonnable de ces niveaux. Une telle classification peut également être requise pour les maladies humaines mais elle est en général moins indispensable.
Le présent travail de recherche vise à élaborer une méthode de classification des niveaux de risque permettant de comprendre la dynamique spatiale de la propagation des maladies. la compréhension d'un tel processus passe par une approche à la fois épidémiologique, statistique et géographique.

1.2 Principales Contributions

Nous présentons ici les contributions de ce travail, en détaillant avant les points sur lesquels elles vont porter :

Classification en niveaux de risque. Nous assimilons dans ce travail la cartographie du risque à un problème de classification. L'objectif de la classification est de regrouper les populations qui se "ressemblent" dans une même classe et de fournir alors une vue résumée de l'ensemble des données. Cela peut permettre de regrouper les unités géographiques qui ont un niveau de risque semblable ou encore de mettre en évidence les unités à risque élevé.
L'approche considérée dans ce travail est l'approche probabiliste dans laquelle observations (nombres de cas infectés en cartographie) et classes (niveaux de risque) sont supposées être des réalisations de variables aléatoires. Le problème de la classification peut être vu comme un problème à données manquantes. En effet, les classes à associer aux unités ne sont pas observées, on peut les considérer comme manquantes. Cette approche probabiliste repose alors sur la donnée d'un modèle pour le couple des observations et des classes. Généralement, ce modèle est décomposé en un modèle régissant les classes et un modèle pour la génération des observations conditionnellement aux classes (appelé le modèle d'attache aux données). Au niveau de la modélisation, on suppose que le terme d'attache aux données se factorise sur les

unités. L'attache aux données est supposée Poissonnienne dans notre problème de cartographie.
Afin de mieux intégrer les caractéristiques individuelles et données d'interaction, nous nous orientons vers les modèles probabilistes graphiques. La souplesse de ces modèles permet de construire un modèle adapté au problème considéré en le décrivant à l'aide d'une structure capable de capter les dépendances entre les variables observées de façon interprétable par les épidémiologistes.
Nous nous intéressons plus précisément au modèle de champ de Markov caché dans lequel une distribution de probabilités paramétrique permet de prendre en compte la distribution des données individuelles. Les interactions qui peuvent refléter une mesure de similarité entre les unités géographiques sont représentées par un graphe. En effet, l'idée naturelle que nous mettons en oeuvre, dans ce travail, consiste à construire un graphe dont les noeuds représentent les unités géographiques et les arêtes des relations de voisinage entre elles. Les observations (qui sont le nombre de cas infectés) sont affectées à chaque noeud du graphe correspondant et considérées comme des réalisations de variables aléatoires. La notion de voisinage est à prendre au sens large. Elle peut représenter des relations de dépendance statistique entre variables aléatoires, être liée à la proximité géographique des sites mais éventuellement aussi à des similarités entre sites dûes à des facteurs environnementaux communs par exemple. Dans ce travail, nous nous limitons à la proximité géographique comme système de voisinage.

Variante du modèle de Potts. L'assimilation de la cartographie du risque épidémiologique à un problème de classification nous conduit à une situation avec données incomplètes, pour lesquelles une partie des données est manquante. L'objectif est, alors, de retrouver la vraie carte de risque (les données cachées) à partir d'une version bruitée ou dégradée de cette carte (données observées). Les observations sont dans notre contexte les nombres de cas enregistrés pour chaque unité géographique et la vraie carte que l'on cherche à retrouver est un ensemble d'étiquettes. Une étiquette représente l'appartenance de l'unité à une des classes de risque. Un modèle adapté pour prendre en compte les dépendances spatiales entre les étiquettes est le modèle de champ de Markov caché. Un des modèles de Markov caché le plus souvent choisi pour représenter les variables aléatoires associées aux données cachées est le modèle de Potts à K classes. La version la plus simple de ce modèle est le modèle de Potts homogène et isotrope sans champ externe qui est décrit par un seul paramètre b qui décrit la force des interactions entre les unités. Ce paramètre est lié à l'importance du lissage et par conséquent à l'homogénéité des zones de risque.
En cartographie du risque, l'ordre des classes en lien avec leur situation géographique et leur

interprétation en terme de gravité du risque, est très important. Cet ordre des classes n'est pas pris en compte par les modèles de type Potts. En effet, ces modèles ne prennent en compte que l'égalité ou la différence de variables cachées voisines. Par exemple, les classes à plus fort et plus faible risque (les deux classes extrêmes) peuvent se retrouver côte à côte, alors que l'on s'attend plutôt à une gradation des risques et à des transitions progressives entre les classes. Dans ce travail, nous proposons une variante du modèle de Potts, pour laquelle le paramètre d'interaction dépend des classes de risque afin de prendre en compte l'hypothèse épidémiologique de gradation du risque.

L'algorithme EM champ-moyen pour l'estimation des paramètres. Dans le cas des problèmes avec données cachées, l'algorithme EM est un algorithme classique d'estimation. Il est largement utilisé pour la classification basée sur les modèles de mélanges indépendants. Pour ces modèles, la mise en oeuvre de l'algorithme EM est simple du fait de l'hypothèse d'indépendance. Par contre, dans le modèle de champs de Markov cachés des approximations sont nécessaires du fait de la complexité des modèles. Nous utilisons, dans ce travail, une approximation basée sur la théorie du champ-moyen qui conduit à un algorithme aussi simple à mettre en oeuvre que pour les mélanges tout en préservant l'information spatiale.

Stratégie d'initialisation pour les mélanges de Poisson. Les algorithmes de type EM donnent des résultats fortement dépendants de l'initialisation. Selon le point de départ choisi le résultat peut être plus ou moins bon. Pour appréhender ce problème, plusieurs techniques ont été proposées dans la litérature. La stratégie la plus généralement utilisée est de lancer plusieurs fois l'algorithme d'une position aléatoire et de retenir ensuite la solution fournissant la plus grande log-vraisemblance ou log-vraisemblance complétée. D'autres techniques d'initialisation ont été mises en place pour palier à cette limite. Ces stratégies ont été proposées en général dans le cas des mélanges gaussiens avec quelques techniques proposées pour les mélanges de Poisson. Dans ce travail, nous sommes en présence de mélanges de Poisson dont les composantes sont encore plus compliquées à séparer. En effet, le fait que la moyenne d'une loi de Poisson est, en même temps, sa variance rend la tâche de l'algorithme EM plus complexe pour ce qui est de l'estimation des différentes moyennes des composantes du mélange. D'autant plus que pour les maladies rares, les taux de risque sont très petits et donc les moyennes des distributions du mélange sont potentiellement petites et d'autant plus difficile à séparer. En plus d'une initalisation raisonnable des paramètres de risque, nous avons besoin dans notre contexte d'initialiser les paramètres du modèle de Markov caché. Nous proposons dans ce travail une stratégie d'initialisation complète pour l'ensemble des paramètres du mo-

dèle proposé. Cette technique d'initialisation exploite l'idée d'utiliser l'espace où vivent les trajectoires de l'algorithme EM pour obtenir des valeurs initiales raisonnables.

En résumé, les principales contributions de cette thèse consistent en :

- assimiler la cartographie du risque en épidémiologie animale à un problème de classification de données spatialement dépendantes en niveaux de risque traité à l'aide d'un modèle de champ de Markov caché discret,
- proposer pour la partie markovienne cachée, une variante du modèle de Potts prenant en compte les hypothèses épidémiologiques liées à la distribution spatiale et graduelle du risque,
- estimer les paramètres du modèle en utilisant l'algorithme EM **champ-moyen** comme alternative aux méthodes de Monte Carlo par chaînes de Markov (MCMC) qui sont habituellement utilisées pour l'estimation des paramètres des modèles en cartographie du risque épidémiologique,
- proposer une stratégie d'initialisation adaptée aux mélanges de Poisson.

1.3 Organisation du manuscrit

Dans le Chapitre 2, nous présentons quelques notions élémentaires en épidémiologie animale et humaine et donnons un aperçu sur les mesures de lutte dans le monde médical et le monde vétérinaire (section 2.1). Nous détaillons ensuite les principes de l'épidémiologie spatiale qui est une partie de l'épidémiologie qui étudie la distribution des risques en fonction de la géographie (section 2.2). Nous détaillons ensuite les méthodes statistiques proposées pour le traitement des données d'enquête en épidémiologie (section 2.3). Nous terminons ce chapitre en précisant les raisons du choix de la méthode utilisée dans ce travail (section 2.4).

Le Chapitre 3 vise à établir un état de l'art des directions retenues dans la littérature pour la cartographie du risque en épidémiologie. Nous présentons les principes de l'analyse bayésienne hiérarchique nécessaire à la compréhension des modèles hiérarchiques proposées (section 3.1). Pour l'estimation des paramètres, nous nous focalisons sur une présentation des algorithmes d'estimation de type MCMC qui sont les plus utilisés pour l'inférence bayésienne (section 3.1.5). Nous décrivons ensuite brièvement les principales familles d'approches pour la cartographie en épidémiologie (section 3.2). Nous présentons en section 3.3 un nombre de critères utilisés en inférence bayésienne. Nous terminons le chapitre par une discussion générale récapitulant le contenu des sections principales de ce chapitre (section 3.4).

Le chapitre 4 est consacré à la description des modèles de champ de Markov cachés pour la classification. Nous introduisons les principes généraux des champs de Markov (section 4.1). Nous décrivons ensuite l'utilisation des modèles de mélange pour la classification de variables indépendantes (section 4.2.1). Puis, nous abordons la classification de variables dépendantes à l'aide des champs de Markov cachés (section 4.2.2). Nous présentons l'algorithme EM et ses approximations dans le cas de ces modèles (section 4.3). Nous décrivons au passage le critère BIC de sélection de modèles que nous avons utilisé pour le choix du nombre de classes sur nos données (section 4.4).

Dans le Chapitre 5 nous détaillons notre modèle basé sur les modèles de champ de Markov cachés. Nous commençons par rappeler les motivations de cette modélisation en cartographie du risque épidémiologique (section 5.1) que nous détaillons ensuite dans la section 5.2. Nous montrons la mise en œuvre de l'algorithme EM champ-moyen pour l'estimation des paramètres de notre modèle (section 5.3). Nous détaillons alors la stratégie d'initialisation proposée pour initialiser notre EM **champ-moyen** (section 5.3.2) et l'illustrons par des exemples (sections 5.4.2 et 5.4.3). Nous concluons ce chapitre par une discussion générale (section 5.5).

Le Chapitre 6 décrit les applications de notre méthodologie à des données simulées et à des données réelles concernant la maladie de L'Encéphalopathie Spongiforme Bovine (ESB). Cette évaluation ne vise pas seulement à établir la performance du modèle et de la stratégie d'initialisation mais aussi à illustrer certains de leurs comportements sur ce type de données. Nous faisons aussi la comparaison avec d'autres modèles et d'autres techniques d'initialisation.

Le Chapitre 7 donne une conclusion générale des perspectives intéressantes à considérer pour de futurs travaux.

CHAPITRE 2

Éléments d'épidémiologie spatiale

Sommaire

L'épidémiologie est un ensemble de disciplines scientifiques dont l'objectif est d'étudier la distribution des maladies et des indicateurs de santé (facteurs influant sur la santé) dans les populations ainsi que les influences qui déterminent cette distribution (OMS, 1968).
On peut distinguer quatre branches au sein de l'épidémiologie :

- L'épidémiologie descriptive a pour objectif d'étudier la répartition des maladies dans l'espace et le temps selon les caractéristiques des personnes. Elle décrit les phénomènes de santé, en fonction des caractéristiques des individus (âge, sexe).
- L'épidémiologie analytique ou étiologique vise à établir les causes et déterminer l'ensemble des facteurs liés aux phénomènes décrits. On procède par comparaisons, par exemple en examinant la fréquence d'une affection entre des groupes exposés ou non à certains facteurs.
- L'épidémiologie prospective s'appuie sur des connaissances déjà obtenues, elle vise à faire des projections sur l'avenir, afin de prévoir l'évolution d'un problème pathologique.

– L'épidémiologie d'intervention consiste à mettre en œuvre sur une maladie une méthode de prévention et à étudier ses conséquences (*a priori* bénéfiques) sur l'apparition et/ou le développement de la maladie.

Dans ce travail, nous nous limitons à l'épidémiologie descriptive. Ce premier chapitre vise à donner une brève introduction à l'épidémiologie en section 2.1, à présenter les notions clés de l'épidémiologie descriptive spatiale (en référence à la géographie) en section 2.2, à décrire briévement les méthodes statistiques les plus courantes de traitement de données épidémiologiques en section 2.3 et à expliquer enfin les motivations du choix de la méthode proposée dans ce travail en section 2.4.

2.1 Introduction à l'épidémiologie

Nous présentons dans cette partie les notions clés en épidémiologie et l'intérêt des mesures de lutte qui doivent être mises en place pour cette discipline.

2.1.1 Quelques définitions épidémiologiques

Nous présentons dans cette section la définition des termes qui sont couramment utilisés en épidémiologie (humaine ou animale).

Le **taux de mortalité** est le nombre de cas contaminés divisé par la population. Ce taux est le plus étudié en épidémiologie.

En épidémiologie, **le risque absolu** (RA) peut être défini comme une probabilité. C'est la probabilité de survenue d'un événement (maladie ou décès), à un instant donné ou pendant un intervalle de temps donné. C'est aussi la probabilité qu'un individu pris au hasard dans la population soit malade sur la période considérée.

On parle de **facteur de risque** pour désigner tout facteur d'exposition susceptible de modifier le risque d'une maladie, c'est à dire sa probabilité de survenue. Un facteur de risque est donc une variable potentiellement liée à une maladie et présentant un éventuel lien causal avec celle-ci.

Le risque relatif (RR) est une mesure statistique souvent utilisée en épidémiologie, mesurant le risque de survenue d'un événement entre deux groupes. Dans l'étude d'une maladie, on considère souvent deux populations : une première population de personnes exposées à un certain facteur de risque dont on désire mesurer l'influence sur la maladie et une deuxième population de cas témoins autrement dit de personnes non exposées au facteur de risque. Le risque relatif est alors le rapport de deux taux. Le premier taux est la proportion de personnes

malades dans la population exposée *i.e.* le nombre de personnes malades et exposées sur le nombre totale de personnes exposées. Le deuxième taux est la proportion de personnes malades dans la population témoin, *i.e.* le nombre de personnes malades et non exposées sur le nombre de personnes non exposées. Ainsi, un risque relatif de 2 par exemple, indique que le risque d'avoir la maladie en question est deux fois plus élevé chez les personnes exposées que chez les non exposées.

L'incidence correspond au ratio du nombre de nouveaux cas dans un laps de temps réduit sur la population totale dans une zone définie. Elle permet de tenir compte d'une hétérégonéité spatiale potentielle de la population et d'évaluer la fréquence instantanée et la vitesse d'évolution d'une maladie.

La transmission est le passage d'un agent pathogène d'un organisme à un autre. Il existe différents types de transmission :

- la transmission directe : elle se fait par contact étroit entre organismes (mufle à mufle, contact cutané, lait maternel, voie vénérienne).
- la transmission indirecte : elle se fait par l'intermédiaire d'un autre organisme, d'un objet ou d'une substance (exemple : la fièvre aphteuse par voie aérienne).
- la transmission horizontale : c'est une transmission indépendante des liens de parenté, y inclus la transmission vénérienne (exemple : tuberculose).
- la transmission verticale : d'un parent à un descendant, à l'occasion de la reproduction (exemple : babésiose chez la tique).

Maladie transmissible : il s'agit d'une maladie dont l'agent peut être transmis et retransmis à différents organismes de la même espèce ou non. Une maladie peut être transmissible mais non contagieuse si elle exige pour sa transmission l'intervention d'un vecteur (comme un moustique ou une tique, voir figure 2.1).

Maladie contagieuse : c'est une maladie transmise par contact direct ou indirect avec un organisme source de l'agent pathogène (voir figure 2.1).

2.1.2 Mesures de préventions en épidémiologie humaine et animale

Les analyses statistiques, en épidémiologie humaine et animale, peuvent aider les autorités à mettre en place des mesures de sécurité afin de contrôler les épidémies par des mesures de lutte et de prévention adaptées, selon la pathologie en cause. La prévention repose en priorité sur la lutte contre la transmission et peut être effectuée à trois niveaux :

- au niveau de la source ou du réservoir de l'agent causal,
- au niveau des modes de transmission,
- au niveau des défenses des individus et des populations exposées.

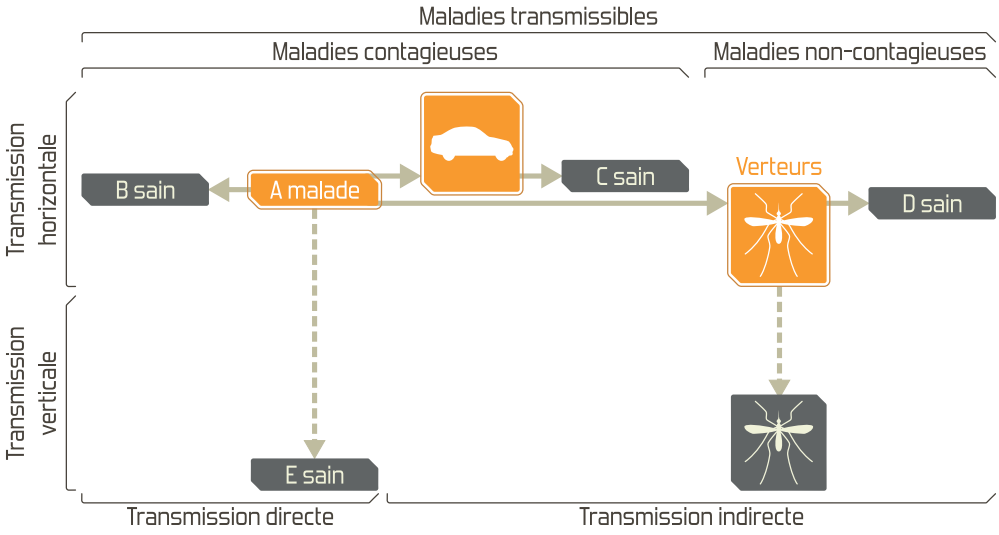


FIGURE 2.1 – Classification des maladies et mode de transmission en pratique.

En épidémiologie humaine, il existe deux stratégies complémentaires pour lutter contre la transmission :

- Les méthodes de **prévention générale** qui s'adressent à l'environnement et aux conditions de vie. Il s'agit le plus souvent de mesures d'hygiène et d'assainissement qui vont réduire les risques de transmission comme : la distribution d'eau potable contrôlée, la désinsectisation ou le contrôle d'hygiène alimentaire.
- Les mesures **préventives spécifiques** qui sont plus individuelles et ponctuelles. Elles correspondent généralement aux vaccinations de sujets plus particulièrement exposés dans une région à risque : personnel soignant (tuberculose), enfants, personnes âgées (grippe). La vaccination en général et les vaccins obligatoires des enfants font partie de ces mesures.

En épidémiologie humaine, la plupart de ces mesures sont nationales. En épidémiologie animale, il est plus classique de ne vacciner que dans certaines régions. On peut aussi restreindre la circulation des animaux, ce qui n'est pas fait pour l'humain, voire même restreindre la circulation humaine entre les fermes (pas de passage de camion de ramassage de lait par exemple). Enfin, on peut décider d'abattre les animaux si on estime que le risque est trop fort pour que toutes ces mesures de prévention ne réussissent pas à contrôler l'épidémie.

2.2 Introduction à l'épidémiologie spatiale

L'épidémiologie spatiale a pour objectif la description et l'analyse de données de santé géographiquement indexées en considérant des facteurs de risque démographiques, environnementaux, comportementaux, socio-économiques, génétiques ou infectieux. C'est une composante des analyses géographiques datant des années 1800 où les cartes de taux de maladie dans différents pays ont commencé à apparaître pour caractériser la diffusion et les causes possibles des apparitions de maladies infectieuses comme la fièvre jaune et le choléra.
Les avancées récentes dans la collecte de données et les méthodes d'analyse ont permis d'améliorer les études de maladies à l'échelle nationale ou régionale. De telles enquêtes peuvent inclure des données de facteur de risque localement appropriées comme les expositions aux sources locales de pollution environnementale et la distribution locale de la variation des facteurs socio-économiques et comportementaux. Mais elles présentent aussi de nouveaux défis puisque les données récoltées à petite échelle (commune, canton ou encore région) peuvent conduire à des évaluations de risques instables. L'analyse de la répartition spatiale d'indicateurs de santé comporte différents objectifs. D'une part, elle vise à décrire ces variations et à modéliser leur structure. D'autre part, elle a pour objectif de mettre en évidence des associations entre ces variations et celles d'exposition à des facteurs de risque environnementaux. Les variations spatiales des indicateurs de santé et des facteurs d'expositions environnementales sont étudiées en épidémiologie dans un but descriptif et afin de suggérer des hypothèses étiologiques.
La représentation cartographique du risque de maladie, la détection d'agrégats spatiaux autour d'un point source ou encore l'évaluation de l'association entre risque et exposition environnementale en fonction de facteurs de risque connus représentent les différents types d'analyse qui font intervenir des approches géographiques. Ces analyses requièrent des informations spatialisées afin d'être effectuées. Ces informations consistent en général en des nombres de cas contaminés pour une pathologie donnée, géolocalisés à partir des adresses présentes dans les bases de données existantes. Ce sont ces données là qui seront, ensuite, exploitées dans les systèmes d'informations géographiques puis dans l'analyse statistique.

2.2.1 Nature des données

En épidémiologie, deux grandes catégories de données sont définies par leur niveau de résolution.

– **Les données agrégées :** ces données sont en général mises à la disponibilité des usagers après qu'elles aient été traitées au niveau statistique. En effet, les données agrégées

sont constituées à partir d'un fichier de données brutes et sont le résultat d'une combinaison de différentes mesures. On les obtient en faisant une addition ou une moyenne des valeurs individuelles obtenues. Elles permettent d'obtenir de l'information sur des groupes qui ont des caractéristiques communes. On peut agréger par lieu géographique, par caractéristiques ou par temps. Ces données sont appelées également données de comptage.

– **Les données ponctuelles :** lorsque la précision des données est plus fine, on parle de données ponctuelles ou encore de données individuelles. Dans ce cas, un évènement peut être défini par ses coordonnées géographiques en spatial.

2.2.2 Zone et période d'étude

La zone d'étude est définie par l'étendue géographique (appelée emprise) qui concerne les observations de la maladie étudiée. Elle doit permettre aussi de disposer d'une population suffisamment large pour les données sanitaires.
La période d'étude repose habituellement sur les données les plus récentes. En fonction des pathologies étudiées et de la latence de leur survenue par rapport à l'exposition, elle peut refléter une exposition antérieure, allant de plusieurs années à plusieurs dizaines d'années. Tout comme pour la zone d'étude, afin de pallier au manque d'effectif, plusieurs années de données sont collectées.

2.2.3 L'échelle des données

Le point clé des études épidémiologiques est le choix de l'échelle de l'étude. Pour les données ponctuelles, les études sont faites au niveau de l'individu.
Quant aux données agrégées, le choix de l'unité géographique est un point crucial. Au vu de la diversité, de la disponibilité et de la qualité des données, il est indispensable de s'interroger sur le choix de l'unité spatiale de référence pour l'étude. Ce choix dépendra essentiellement de la résolution spatiale des données et de leur compatibilité. Souvent, l'unité est choisie en fonction des données de santé et des données démographiques disponibles. Elle est définie soit sur un découpage géographique de type administratif (commune, canton), soit par un découpage régulier appelé maillage (hexagone par exemple).
Par ailleurs, ce découpage peut ne pas être toujours pertinent d'un point de vue épidémiologique, et les résultats peuvent être sensibles à sa redéfinition. L'échelle de représentation et d'analyse doit donc être choisie avec précaution.

Selon l'objectif de l'étude et les données dont on dispose, les méthodes statistiques en épidémiologie spatiale au niveau de petites zones géographiques peuvent être découpées en trois grandes familles (Elliott et al., 2000) :

- Les méthodes de détection d'agrégats visent à chercher généralement les zones spatiales, appelées agrégats, dans lesquelles la densité de cas est anormalement élevée. Rappelons qu'un agrégat peut se définir comme un groupement de cas malades géographiquement proches, de taille et de concentration suffisante pour qu'il y ait peu de chance qu'il soit uniquement dû au hasard. Ces méthodes prennent en compte la densité de la population sous-jacente afin d'analyser le processus aléatoire d'apparition des événements observés (les cas des maladies). Elles s'appliquent en priorité aux données ponctuelles, mais peuvent être appliquées, par transformation, sur données agrégées.
- Les méthodes de cartographie du risque permettent d'obtenir des estimations ponctuelles des risques associés aux unités géographiques afin de pouvoir construire une représentation fiable sur le domaine étudié de la variabilité spatiale de l'indicateur de santé. Ces méthodes ont été développées pour les données agrégées.
- Les méthodes pour l'étude de corrélations géographiques ont pour objectif d'étudier, au niveau de groupes d'individus définis sur une base géographique, la relation entre un indicateur de santé et une exposition environnementale. Ces méthodes sont applicables sur les données agrégées.

Cette classification est artificielle, et dépend de l'echelle utilisée pour l'étude. Par exemple, la cartographie du risque peut apporter des informations sur les agrégats individuels. De même, l'étude des corrélations géographiques partagent des points en commun avec la cartographie.

2.3 Méthodes de traitement de données d'enquête en épidémiologie spatiale

On décrit ici brièvement les méthodes d'analyse spatiale les plus utilisées dans la littérature. On présentera un résumé des méthodes de détection d'agrégats en section 2.3.1, de l'étude de corrélations géographiques en section 2.3.3 et de la cartographie du risque en section 2.3.2. Ces différentes analyses dépendent des objectifs de l'étude, des attentes des épidémiologistes et des données disponibles.

2.3.1 Détection d'agrégats

Différentes méthodes, basées sur la notion de densité, ont été developpées pour tester une tendance à l'aggrégation de cas d'une maladie. L'objectif est de mieux comprendre la distribution géographique des maladies et d'en étudier l'hétérogénéité spatiale. Rappelons qu'un agrégat peut se définir comme un groupement de cas malades géographiquement proches, de taille et de concentration suffisante pour qu'il y ait peu de chance qu'il soit uniquement dû au hasard. Selon les questions soulevées par les épidémiologistes, ces analyses d'agrégats peuvent être classées en trois familles :

- **Méthodes de balayage spatial :** le but est de détecter les zones pour lesquelles une incidence plus élevée de cas d'une maladie est observée. Parmi ces méthodes, on retrouve la statistique de scan spatial appelée "la méthode de Kulldorff" (Kulldorff and Nagarwalla, 1995; Kulldorff, 1996) qui reste la plus populaire malgré l'émergence d'autres méthodes de détection de clusters (Tango and Takahashi, 2005). Cette méthode est basée sur un test du rapport de vraisemblance et permet d'identifier des zones ayant une incidence anormalement élevée et qui sont les moins cohérentes avec l'hypothèse nulle de risque constant. Cette méthode est réputée être très puissante.
  Une fenêtre, de forme prédéfinie (cercles ou ellipses), de taille variable, balaye la zone d'étude. Pour chaque fenêtre, une statistique, basée sur le rapport de vraisemblance et les nombres de cas observés et attendus, est calculée. Les fonctions de vraisemblance s'écrivent selon le choix de la distibution théorique associée au nombre de cas. L'hypothèse alternative, pour chaque "position spatiale" et taille de fenêtre, est qu'il existe un risque plus élevé à l'intérieur de la fenêtre par rapport à l'extérieur. La fenêtre qui correspond au maximum de vraisemblance est le cluster le plus probable, celui qui a le moins de chance de survenir par hasard. La méthode de Kulldorff permet d'ordonner les agrégats selon leur rapport de vraisemblance et d'identifier des agrégats secondaires. Les principaux inconvénients de cette méthode sont la forme prédéfinie des fenêtres et leur grand nombre. La récente méthode de Cucala (2009) permet d'aller au delà de ces inconvénients.
- **Test de concentration :** ces méthodes s'intéressent plus à l'évaluation de l'existence d'agrégats en référence à un point spécifique. Elles représentent une vraie alternative aux méthodes de balayage spatial lorsque l'on a des informations *a priori* sur la position exacte d'un possible "cluster" (agrégat).
  Ces tests nécessitent une mesure du facteur de risque dans l'espace. Souvent, la distance au point source tient lieu d'indicateur d'exposition.
  Plusieurs tests sont disponibles (Morris and Wakefield, 2000) et les plus utilisés sont le

test de Stone du maximum de vraisemblance (Bithel and Stone, 1989) et le **test du score de risque linéaire** (Bithel et al., 1994; Bithel, 1995) qui sont utilisés pour tester une augmentation de risque en relation à un point prédéfini.

– **Les tests de clustering global (global clustering tests) :** L'objectif de ces méthodes est de tester l'existence d'une hétérogénéité globale de la distribution spatiale d'une maladie. Ces méthodes ne donnent pas la localisation des clusters mais permettent d'étudier la surdispersion, la corrélation spatiale et de détecter la tendance des cas au clustering. Il existe de nombreuses méthodes de global clustering (Kulldorff, 2006). Les tests les plus utilisés dans les études de corrélation spatiale sont les suivants :

1. **Test de Potthoff et Whittinghil :** ce test largement utilisé en épidémiologie consiste à tester l'existence d'une hétérogénéité spatiale globale en terme de surdispersion (Wakefield et al., 2000).
   Sous l'hypothèse nulle d'une distribution aléatoire des cas d'une maladie, les taux d'incidence sont les mêmes sur toute la zone étudiée et les seules variations des cas observés sont liés aux fluctuations de la loi de Poisson. Le nombre de cas observés est supposé suivre une loi de Poisson de moyenne et de variance égale au nombre de cas attendus.
   Sous l'hypothèse alternative de l'existence d'une surdispersion des cas, un certain nombre de cas apparaissent dans certaines zones plus fréquemment que ce qui était prédit sous l'hypothèse d'une distribution de Poisson. Le test de Potthoff et Whittinghill suppose que le rapport entre la variance et la moyenne est égal à $1+\gamma$, où γ est défini comme la variation extra-poissonienne. Pour évaluer la surdispersion du risque de maladie, on évalue le rapport $\gamma/SD(\gamma)$ (où SD est l'écart-type). En l'absence de surdispersion et lorsque le nombre de zones géographiques est grand, la distribution de $\gamma/SD(\gamma)$ suit approximativement une loi Normale $\mathcal{N}(0,1)$.

2. **La statistique de Moran :** cette méthode évalue l'existence d'une hétérogénéité spatiale globale en terme d'autocorrélation spatiale. La statistique de Moran est l'indice d'autocorrélation spatiale le plus utilisé. Cette statistique résume le degré de ressemblance des unités géographiques voisines par une moyenne pondérée de la ressemblance entre observations. Toutefois, le calcul de cette statistique est déséquilibré, car certaines observations sont plus représentées que d'autres. Certaines localités centrales ont en effet plus de voisins que d'autres, situées par exemple sur les limites du territoire ou dans des zones éparses (Huang et al., 2008).

3. **La statistique de Tango :** elle teste si les cas de maladie sont regroupés dans des clusters à l'intérieur de la région d'étude (Tango, 1995).

2.3.2 La cartographie du risque épidémiologique

L'objectif de la cartographie du risque est de prédire des risques relatifs (ou absolus) par unité géographique et de produire des cartes de risque des maladies. Cette représentation cartographique des indicateurs de santé permet la description de leur distribution spatiale. Elle permet aussi la mise en évidence de zones avec un risque anormalement élevé pour la suggestion des hypothèses étiologiques. La difficulté est de présenter des images fiables qui séparent les réelles variations géographiques des indicateurs de santé du bruit inhérent et qui modélisent correctement la structure de ces variations. Les cartes de risque présentent souvent le **Taux de Mortalité Standardisé** (SMR). Il est défini comme étant le rapport entre le nombre de cas observés et un nombre de cas attendus sous l'hypothèse d'une incidence de référence.

Soit S la région étudiée et N le nombre d'unités géographiques qui forment cette région. Soit y_i et n_i respectivement le nombre observé de cas et la population cible dans l'unité géographique i ($i \in S = \{1, \ldots, N\}$), E_i le nombre attendu de cas, avec :

$$E_i = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} n_i} \, n_i,$$

r_i représentera dans la suite le risque relatif (RR) et λ_i le risque absolu (RA) de l'unité i.

Soit Y_i la variable aléatoire associée au nombre de cas, la supposition commune à tous les modèles de cartographie en épidémiologie est que :

$$Y_i \sim \text{Pois}(E_i r_i) \;\; (\text{ou } Y_i \sim \text{Pois}(n_i \lambda_i)).$$

Cela traduit le fait que les fluctuations aléatoires du nombre de cas de maladie observés sont modélisées par une loi de Poisson.

Le SMR correspond à l'estimateur de maximum de vraisemblance de r_i (ou λ_i) avec :

$$\hat{r}_i = \text{SMR}_i = Y_i / E_i,$$

(ou encore $\hat{\lambda}_i = \text{SMR}_i = Y_i / n_i$) avec une variance égale à Y_i / E_i^2 (ou Y_i / n_i^2).

On peut observer à partir de cela, que la variabilité des SMR est différente selon les unités géographiques et que cette variabilité peut donner des cartes bruitées où les SMR les plus extrêmes correspondent le plus probablement aux unités les moins peuplées (Wakefield, 2007). Pour illustrer cette instabilité, nous présentons à la figure 2.2 un ensemble de cas d'Encéphalopathie Spongiforme Bovine (ESB) en France et les SMR correspondants. Nous pouvons voir par exemple que le SMR d'un hexagone au Sud-Ouest de la France est extrême alors que la population bovine dans cette zone n'est pas élevée (voir chapitre 6 pour la carte de la population bovine). Ce problème résulte en partie du fait que les risques sont considérés comme

indépendants, d'une unité géographique à l'autre, sans prendre en compte l'autocorrélation spatiale.
Pour palier à cette limite, des méthodes de lissage des SMR ont été developpées (voir chapitre 3) pour produire des estimations plus fiables en prenant en compte la dépendance spatiale qui implique que des zones proches géographiquement sont susceptibles d'avoir des risques similaires.

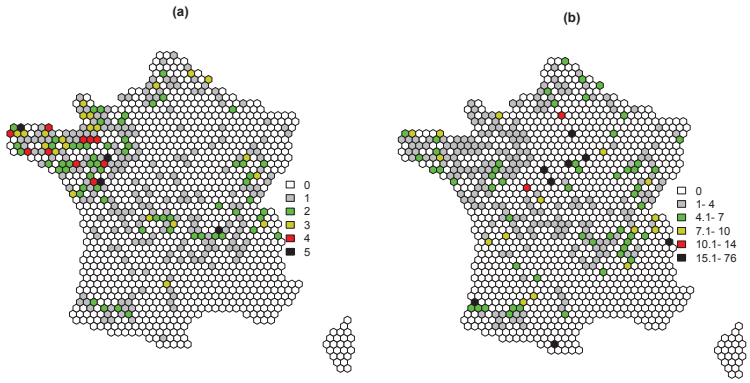


FIGURE 2.2 – Exemple de l'ESB pour les 1264 hexagones découpant la France. (a) : nombres de cas enregistrés et (b) : les SMR (Y_i/E_i) correspondants.

2.3.3 L'étude des corrélations géographiques

L'objectif de l'étude des corrélations géographiques (appelée souvent modèles de régression) est d'estimer l'association entre les variations géographiques d'un indicateur de santé et celles des variables environnementales.
L'étude d'une maladie rare ou des petites unités spatiales conduit à utiliser un modèle de régression de Poisson. Il faut souligner que les modèles utilisés ici sont similaires à ceux utilisés pour la cartographie du risque. Mais il est important de rappeler que les objectifs des deux études sont différents étant donné que le but de la représentation cartographique est la prédiction des risques relatifs par unité géographique alors que l'objectif des modèles de la régression est l'estimation de la relation entre indicateur de santé et exposition.
On rappelle que le modèle de régression de Poisson classique est rarement adapté à cause de la sur-dispersion qui n'est pas prise en compte. Il est adapté quand la variabilité intra-zone est négligeable comparée à la variabilié inter-zones (large zone d'étude et/ou maladies communes).
Il est peu réaliste de supposer l'indépendance des résidus de la régression : en général, les

nombres de cas dans les zones voisines géographiquement présentent de la dépendance spatiale résiduelle. Dans le cadre de la représentation cartographique, cette dépendance, comme mentionné auparavant, peut être exploitée dans l'estimation des risques en lissant localement entre unités voisines. Dans le cadre de la régression, la dépendance doit être prise en compte et la famille des modèles linéaires généralisés (**GLM**) spatialisés permet la prise en compte de cette dépendance.

Discussion Les méthodes présentées à la section 2.3 permettent de répondre aux différentes questions épidémiologiques que l'on peut se poser. Elles permettent, entre autres, de tester si une maladie est distribuée aléatoirement dans la région étudiée ou encore de détecter des zones à incidence élevée.
Les tests d'analyse de clusters ou encore les méthodes de cartographie permettent non seulement de mettre en évidence des contrastes entre les valeurs des indicateurs de santé mais aussi de guider la recherche des facteurs de risque environnementaux pour formuler des hypothèses étiologiques. Toutefois, elles ne peuvent être considérées que comme des méthodes de dépistage derrière lesquelles des études plus ciblées doivent être mises en oeuvre pour confirmer les hypothèses qu'elles permettent de dégager.
Le schéma général des études épidemiologiques est de générer, dans un premier temps, des hypothèses quant à l'effet de certains facteurs de risque sur la santé à partir d'une représentation adéquate des données, et dans un deuxième temps de confirmer (ou pas) ces hypothèses. Toutes ces méthodes statistiques qui permettent d'estimer les risques liés à la survenue d'événements rares, soit pour obtenir une représentation cartographique, la plus informative possible, soit pour quantifier les liens entre un indicateur sanitaire et des covariables environnementales, répondent aux objectifs des études épidémiologiques et donc ont toute leur place dans les activités de veille sanitaire.

2.4 Choix de la méthode

Les méthodes présentées dans la section 2.3 sont complémentaires et peuvent être utilisées pour décrire les mécanismes des maladies (humaines ou animales) contagieuses ou non. Les études de corrélations géographiques sont plus appropriées aux études écologiques qui n'ont pas pour but l'étude des risques au niveau individuel mais l'étude des effets de groupe expliquant une partie de la variation entre les unités géographiques de l'incidence de la pathologie étudiée.
Les méthodes de cartographie du risque et les méthodes de détection d'agrégats sont plus

adaptées au cadre épidémiologique. En effet, ces méthodes sont plus adpatées pour la détection des zones à incidence élevée ou pas.
La présente étude vise à l'élaboration d'une méthode descriptive pour données spatialement groupées. Afin de préparer à terme le développement d'une méthode spatio-temporelle du risque de maladies animales rares non contagieuses, nous nous concentrerons dans ce travail sur l'élaboration au préalable d'une méthode spatiale en vue de son extension au contexte spatio-temporel.
Nous avons choisi de travailler sur les maladies non contagieuses. En effet, il est évident que construire une méthode de cartographie pour les maladies contagieuses demanderait des études de très longue durée et nécessiterait une modélisation encore plus complexe que celle pour les maladies non contagieuses, du fait d'une dépendance plus complexe entre les cas observés.
Notre objectif est de pouvoir obtenir en plus d'une estimation raisonnable du risque en chaque unité géographique, une classification de ces risques en un petit nombre de classes. En épidémiologie animale, on a essentiellement besoin de la classification pour déterminer clairement des zones où appliquer des mesures de lutte, ou des zones à cibler plus particulièrement (zones très atteintes ou très peu atteintes) pour de futures études en vue de déterminer des facteurs pouvant influer sur les différences de risque observées.
Dans ce travail, nous écartons les méthodes utilisées dans les études des corrélations géographiques qui ont plus leur place dans les études écologiques qu'épidémiologiques. Quand aux méthodes de détection d'agrégats, elles pourraient éventuellement répondre à nos objectifs d'estimation et de classification. Toutefois, leur objectif principal est la recherche des zones à risque fort et non pas spécialement des classes avec une certaine gradation du risque. Étant donné notre intérêt pour la classification des risques, en priorité, nous préférons nous orienter vers les méthodes de cartographie du risque et nous développons une méthode inspirée des méthodes existantes dans la littérature.

CHAPITRE 3

Modèles pour la cartographie du risque épidémiologique

Sommaire

L'objectif des différents modèles de cartographie développés est de lisser les différences de précision des estimations initiales, les SMR, en partageant l'information qu'apportent les différentes unités géographiques. Rappelons que les SMR sont différentes selon les unités géographiques et que cette variabilité peut donner des cartes bruitées. Il s'agit donc de produire des estimations de risque plus fiables en considérant la dépendance spatiale entre sites. Le lissage prend généralement en compte les informations apportées par d'autres unités géographiques pour obtenir une estimation plus stable dans chaque unité géographique. Plusieurs

modèles de type hiérarchique bayésien ont été proposés pour ce faire. Ils ont été largement utilisés en cartographie du fait des possibilités qu'ils offrent pour tenir compte des phénomènes complexes qui régissent la structure du risque.
Dans l'approche classique, les observations de chacune des unités géographiques sont considérées comme des réalisations d'une variable aléatoire ayant une distribution de Poisson car on considère des maladies non contagieuses ou non transmissibles. La moyenne de cette distribution, correspondant au risque relatif (RR) (respectivement au risque absolu (RA)), est considérée comme fixe et inconnue. Dans l'approche bayésienne, on suppose que ce paramètre est lui même une variable aléatoire. Cette distribution est appelée distribution *a priori*. L'estimation du RR est alors le résultat de la combinaison de l'information supposée *a priori* et de l'information apportée par les observations.
Ce chapitre est consacré à une introduction aux principes de l'analyse bayésienne en section 3.1, à une brève présentation des différents modèles hiérarchiques bayésiens usuellement utilisés en épidémiologie en section 3.2 et aux critères de sélection de modèles en section 3.3.

3.1 Éléments d'analyse bayésienne hiérarchique

Le noyau de l'inférence bayésienne (voir par exemple Robert (1992) pour une référence en français) est de mettre sur un pied d'égalité observations et paramètres. Les manipulations conditionnelles (formule de Bayes) permettent d'interchanger leurs positions respectives. Dans une approche bayésienne, la distinction entre les paramètres et les variables aléatoires se fait uniquement en fonction de leur place dans la structure hiérarchique. L'innovation principale par rapport à l'analyse statistique standard est de proposer en plus une loi de probabilité sur les paramètres. La notion de vraisemblance reste donc une notion centrale en statistique bayésienne également. Elle est considérée comme une fonction décrivant les paramètres d'une loi en fonction des valeurs observées. Fondamentalement, il s'agit de remonter des effets (les observations) aux causes (les valeurs de paramètres) dans une démarche d'inversion.

3.1.1 Principe de vraisemblance et loi des observations

L'idée classique est que toute l'information sur les paramètres $\boldsymbol{\theta}$ tirée des observations est contenue dans la vraisemblance. Dans une approche dite hiérarchique (voir section 3.1.2), la vraisemblance se situe donc au premier niveau de la hiérarchie (potentiellement à plusieurs

niveaux) qui lie les données aux paramètres. Bien souvent on suppose en plus que les observations, étant donnés les paramètres, sont indépendantes. La vraisemblance des paramètres sécrit alors :

$$L(\mathbf{y}|\boldsymbol{\theta}) = \prod_{i \in S} f(y_i|\boldsymbol{\theta}) \ ,$$

où f une densité de probabilité indiquant la contribution individuelle de chaque observation sur les paramètres. Dans un cadre bayésien, cela revient à considérer que les observations sont indépendantes conditionnellement à $\boldsymbol{\theta}$. Bien sûr, dans la majorité des applications, a fortiori en spatial, les données ne sont pas indépendantes et cette hypothèse d'indépendance conditionnelle est une hypothèse commode en pratique qui permet de repousser la modélisation de la dépendance à un niveau différent de celui de la vraisemblance. Par exemple, dans le cas de données de comptage en épidémiologie, l'observation y_i dont on dispose pour l'unité géographique i peut être vue comme indépendante des autres unités si l'on a des informations sur les paramètres du modèle. Cela signifie que la dépendance n'existerait qu'inconditionnellement (la dépendance est dûe peut-être aux effets non observables ou inconnus). Cette hypothèse est souvent considérée comme vraie en épidémiologie, où l'on suppose qu'une partie de l'hétérogenéité spatiale est expliquée par l'exclusion des variables non observées ou voire même inconnues du modèle. Dans tous les cas, l'hypothèse de l'indépendance inconditionnelle ne peut être valide que si on prend en compte dans le modèle, et à n'importe quel niveau, la corrélation spatiale. L'inclusion de la corrélation spatiale à un niveau hiérarchique différent de celui de la vraisemblance est une idée très exploitée dans les modèles bayésiens en épidémiologie. En effet, la dépendance spatiale apparaît plutôt dans les paramètres de la vraisemblance, pour lesquels on spécifie une distribution *a priori* (voir section 3.1.3).

La pseudo-vraisemblance. L'approximation par pseudo-vraisemblance a été proposée comme une alternative au calcul de la vraisemblance en présence de corrélations. Elle bénéficie d'un nombre important de variantes (Lindsay, 1988; Thibshirani and Hastie, 1987; Kauermann and Opsomer, 2003; Nott and Rydén, 1999; Varin et al., 2005). Cette approximation est donnée, dans sa forme générale, par :

$$L_p(\mathbf{y}|\boldsymbol{\theta}) = \prod_{i \in S} f(y_i|y_j, j \in \mathcal{N}_i, \boldsymbol{\theta}),$$

où $\mathcal{N}_i$ est l'ensemble des voisins du site i susceptibles d'être corrélés avec i.

3.1.2 Les modèles hiérarchiques

Les modèles hiérarchiques rencontrent un succès croissant dans les domaines d'applications tels que l'écologie, la génétique, l'épidémiologie ou encore les sciences de l'environnement. Ils permettent de traduire relativement simplement les hypothèses sur les mécanismes, les spécificités des processus d'observations, les problèmes de censure ou des connaissances *a priori* sur les paramètres. Ils ont révolutionné le domaine des statistiques spatiales et spatio-temporelles par les possibilités qu'ils offraient de prendre en compte des phénomènes de plus en plus complexes. Ils répondent aussi à un besoin nouveau du fait de l'accessibilité croissante par les Systèmes d'Information Géographique (SIG) à des données spatiales de plus en plus nombreuses et multi-sources. L'utilisation des modèles hiérarchiques se répand du fait d'une montée en puissance des approches bayésiennes ces dernières années : les algorithmes de simulation Monte Carlo (Markov Chain Monte Carlo (MCMC) et filtres particulaires) permettent d'exploiter efficacement les propriétés de ces modèles. Comme les modèles statistiques élémentaires, une structure hiérarchique est composée de grandeurs observables $\mathbf{Y}$ et de paramètres $\boldsymbol{\theta}$. Toutefois, ils ont la spécificité d'être basés sur des variables aléatoires $\mathbf{Z}$ non observables dites latentes. Ces variables définissent généralement un mécanisme aléatoire non observable. La démarche d'une modélisation hiérarchique consiste à décomposer un phénomène aléatoire complexe en processus aléatoires plus simple qui peuvent être modélisés avec des structures probabilistes standards.
Ainsi, la distribution jointe des observations et des variables aléatoires latentes s'écrit sous la forme :

$$P(\mathbf{Y}, \mathbf{Z}|\boldsymbol{\theta}) = P(\mathbf{Y}|\mathbf{Z}, \boldsymbol{\theta})P(\mathbf{Z}|\boldsymbol{\theta}),$$

avec :

- Le modèle $P(\mathbf{Y}|\mathbf{Z}, \boldsymbol{\theta})$, dit *modèle des observations*, décrit l'occurrence des données conditionnellement aux variables latentes et à certains paramètres.
- Le modèle $P(\mathbf{Z}|\boldsymbol{\theta})$, dit *modèle du processus interne*, décrit l'occurence des variables latentes conditionnellement aux paramètres.

L'idée principale de la construction des modèles hiérarchiques repose alors sur l'emboîtement de ces différents modèles élémentaires par conditionnements successifs. Chaque source d'incertitude est décrite dans une couche différente du modèle hiérarchique. Au premier niveau de la hiérarchie se situe le modèle des observations $P(\mathbf{Y}|\mathbf{Z}, \boldsymbol{\theta})$, qui permet de modéliser la variabilité naturelle des observations et tenir compte des erreurs de mesure. Au second niveau le fonctionnement du processus interne, est décrit, par l'intermédiaire des variables latentes conditionnellement à d'autres paramètres. La variabilité du processus non observé ainsi que les erreurs de modélisation sont prises en compte à ce niveau.

Dans une structure hiérarchique, les variables latentes permettent de fomaliser et de décrire le phénomène aléatoire étudié et font le pont entre les paramètres du modèle et les variables observables. Elles jouent un double rôle : celui des paramètres quand elles conditionnent la naissance d'observations et celui d'un résultat potentiellement observable quand elles sont générées dans les couches internes de la structure d'un modèle.
Dans les modèles hiérarchiques, le raisonnement à base de conditionnement probabiliste s'appuie commodément sur la modélisation graphique qui se propose de représenter les connaissances sous la forme de graphes orientés acycliques (DAG pour Direct Acyclic Graph selon Spiegelhalter et al. (1996)). Ils sont constitués de noeuds désignant soit une variable observable, soit une variable latente soit un paramètre (voir figure 3.1). Les relations de dépendance sont représentées par des flèches orientées qui partent des grandeurs conditionnantes et pointent vers les grandeurs conditionnées. Les noeuds initiaux qui n'ont pas de parents sont les paramètres : ils ne dépendent d'aucune autre variable aléatoire. Les noeuds terminaux sans enfants, sont les variables observables.

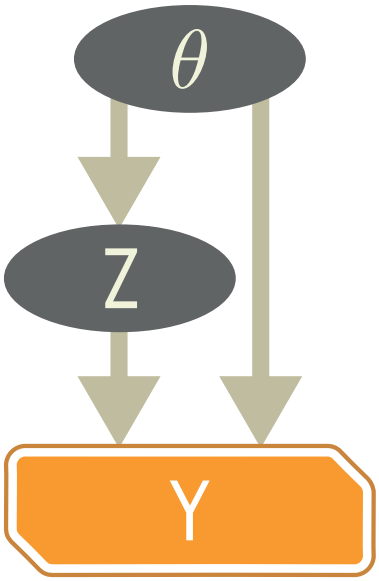


FIGURE 3.1 – Forme générale du DAG pour les modèles hiérarchiques représentant la relation entre les variables observées, latentes et les paramètres : Les grandeurs aléatoires sont représentées par des cercles, les grandeurs observées par des rectangles.

3.1.3 Des informations *a priori* aux lois *a priori*

Dans un contexte bayésien, les paramètres sont considérés comme des variables aléatoires au même titre que les variables observées. La construction d'un modèle hiérarchique nécessite, en plus de la spécification du modèle des observations et du modèle du processus interne, d'assigner une loi *a priori* sur les paramètres inconnus du modèle. Cette distribution de probabilité peut être établie sur la base d'autres données similaires ou refléter l'avis d'experts. L'association du modèle d'occurence des observations, du modèle du processus interne et de la loi *a priori*, appelé modèle d'expertise, permet d'intégrer dans une même structure nos connaissances sur toutes les grandeurs (observables ou non) du phénomène étudié.
La distribution jointe des données observées, des variables latentes et des paramètres s'écrit sous la forme :

$$P(\mathbf{Y}, \mathbf{Z}, \boldsymbol{\theta}) = P(\mathbf{Y}|\mathbf{Z}, \boldsymbol{\theta})P(\mathbf{Z}|\boldsymbol{\theta})P(\boldsymbol{\theta})$$

Dans la pratique on a recours à des lois usuelles (lois normales, lois gamma, etc) ou à des lois dites conjuguées. L'information *a priori* peut être utilisée pour déterminer les paramètres de la loi *a priori*, appelés hyperparamètres.
En l'absence d'information (ou de connaissance d'expert) *a priori* on utilisera le plus souvent la notion de loi *a priori* non informative. On peut citer par exemple les lois de Bernardo-Berger (Berger and Bernardo, 1992), qui minimisent l'information apportée par la loi *a priori* face à l'information apportée par les données.

3.1.4 La distribution *a posteriori* et l'inférence *a posteriori*

La loi *a priori* assignée aux paramètres du modèle et la vraisemblance des paramètres apportent deux types d'information complémentaires à propos du phénomène étudié. La vraisemblance fournit des informations sur les paramètres *via* les données, alors que la loi *a priori* apporte de l'information *via* des connaissances ou des suppositions. Lorsque la taille de l'échantillon est grande, la vraisemblance contribuera plus à l'estimation des paramètres (par exemple les paramètres du risque pour la cartographie). En revanche, lorsque l'on dispose de peu de données, la loi *a priori* aura plus de poids dans l'analyse. Le produit de la vraisemblance et de la loi *a prioi* est appelé la distribution *a posteriori* et s'écrit selon la formule de Bayes :

$$P(\boldsymbol{\theta}|\mathbf{y}) = \frac{P(\mathbf{y}|\boldsymbol{\theta})P(\boldsymbol{\theta})}{P(\mathbf{y})}.$$

Cette distribution décrit le comportement des paramètres une fois que les données sont observées et que les connaissances *a priori* sont intégrées dans le modèle. Le choix de l'*a priori* reste donc crucial étant donné que cela peut affecter significativement la distribution *a posteriori*. Quand un simple modèle de vraisemblance est utilisé, le maximum de vraisemblance fournit un estimateur classique des paramètres. En revanche, lorsqu'un modèle hiérarchique bayésien est spécifié, une estimation de ce type n'est plus possible pour les paramètres $\boldsymbol{\theta}$. Cela est dû au fait que les paramètres ne sont plus supposés fixes mais considérés comme des variables aléatoires provenant d'une certaine distribution. Etant donné les grandeurs observables, les paramètres d'intérêt sont décrits par la distribution *a posteriori* qui doit être calculée et interprétée. Il est possible de considérer la moyenne $E(\boldsymbol{\theta}|\mathbf{y}) = \int \boldsymbol{\theta} P(\boldsymbol{\theta}|\mathbf{y}) d\boldsymbol{\theta}$ ou le mode $\underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, P(\boldsymbol{\theta}|\mathbf{y})$ de la distribution *a posteriori* pour obtenir une estimation ponctuelle des paramètres d'intérêt. Comme le maximum de vraisemblance est le mode de la vraisemblance, le *maximum a posteriori* est le mode de la distribution *a posteriori*. Pour des distributions symétriques et unimodales, le mode et la moyenne coincident.
Pour des distributions *a posteriori* simples, il est possible d'expliciter la forme exacte de la densité et donc de la moyenne ou du mode *a posteriori*. Néanmoins, pour la plupart des modèles réalistes en cartographie du risque, il n'est pas possible d'obtenir une forme analytique pour la distribution *a posteriori* et donc il n'est pas toujours facile d'obtenir une estimation simple des paramètres. Quand l'expression de la distribution *a posteriori* est trop difficile à déterminer, des algorithmes de simulation permettent de l'approcher.
Par la suite plusieurs techniques, dont l'objectif est de simuler un échantillon de la loi *a posteriori* pour estimer les paramètres du modèle bayésien, vont être présentées.

3.1.5 Méthodes de Monte Carlo par chaîne de Markov (MCMC)

Souvent en cartographie du risque, les modèles réalistes sont construits comme un empilement de deux niveaux ou plus de la hiérarchie et la complexité résultante de la distribution *a posteriori* des paramètres nécessite l'utilisation d'algorithmes de simulation.
Les méthodes MCMC sont apparues dans le domaine de la physique et de l'analyse d'image pour prendre en compte les problèmes de grande échelle dans l'estimation. Dans les dernières décennies, ces méthodes ont été adaptées pour des classes de problèmes plus générales comme présentées par Gilks et al. (1993) et Gilks et al. (1996).
Les méthodes MCMC sont un ensemble de méthodes qui simulent une chaîne de Markov dont la loi stationnaire est la loi souhaitée $P(\boldsymbol{\theta}|\mathbf{y})$, leur construction ne pose pas de grandes difficultés, contrairement à leur mise en pratique.

Les méthodes MCMC permettent de calculer la constante $P(\mathbf{y}) = \int P(\mathbf{y}|\boldsymbol{\theta})P(\boldsymbol{\theta})d\boldsymbol{\theta}$ lorsqu'elle n'est pas calculable explicitement, au moins dans un délai raisonnable. En effet, elles permettent de simuler un échantillon de loi $P(\boldsymbol{\theta}|\mathbf{y})$, en évitant le calcul de $P(\mathbf{y})$. Ces techniques reposent sur la construction d'une chaîne de Markov $\{\boldsymbol{\theta}_t; t \in \mathbb{N}\}$ qui possède un noyau de transition ergodique tel que la loi stationnaire soit la loi souhaitée, $P(\boldsymbol{\theta}|\mathbf{y})$. Il existe principalement deux algorithmes permettant de construire une telle chaîne, l'algorithme de Metropolis-Hastings et l'echantillonnage de Gibbs, qui est un cas particulier du premier. Lorsque l'inférence consiste à comparer des modèles, les probabilités *a posteriori* de ces derniers sont évaluées grâce aux méthodes MCMC à saut réversible.

3.1.5.1 L'algorithme de Metropolis-Hastings (MH)

Cet algorithme peut être utilisé lorsque la loi à simuler $P(\boldsymbol{\theta}|\mathbf{y})$ est connu à sa constante de normalisation près qui est inconnue et difficile à calculer. L'algorithme de Metropolis-Hastings (Metropolis et al., 1953; Hastings, 1970) propose à chaque itération un nouvel état qui est ensuite accepté ou rejeté. L'algorithme MH permet de mettre à jour les paramètres dont la loi conditionnelle est difficile à simuler.
Soit $q(.|\boldsymbol{\theta})$ une distribution qui permet de proposer un état pour la chaîne de Markov à partir de l'état actuel, $\boldsymbol{\theta}$. L'état candidat généré, $\boldsymbol{\theta}^{'}$, sera le nouvel état du processus avec la probabilité :

$$\alpha(\boldsymbol{\theta}, \boldsymbol{\theta}^{'}) = \min\{1, \frac{P(\boldsymbol{\theta}^{'}|\mathbf{y})q(\boldsymbol{\theta}^{'}|\boldsymbol{\theta})}{P(\boldsymbol{\theta}|\mathbf{y})q(\boldsymbol{\theta}|\boldsymbol{\theta}^{'})}\} \tag{3.1}$$

sinon la chaîne restera dans son état actuel. Dans l'équation ci-dessus, comme nous considérons le rapport $P(\boldsymbol{\theta}^{'}|\mathbf{y})/P(\boldsymbol{\theta}|\mathbf{y})$ la constante de normalisation problématique n'apparaît pas.
Pour compléter le noyau de transition de la chaîne de Markov, il faut donner la probabilité pour que le processus reste dans le même état $\boldsymbol{\theta}$ soit :

$$1 - \int q(\boldsymbol{\theta}^{'}|\boldsymbol{\theta})\alpha(\boldsymbol{\theta}, \boldsymbol{\theta}^{'})d\boldsymbol{\theta}^{'}$$

Le noyau de transition $\alpha(\boldsymbol{\theta}, \boldsymbol{\theta}^{'})q(\boldsymbol{\theta}^{'}|\boldsymbol{\theta})$ est bien réversible (par construction) et, lorsque la propriété d'ergodicité est vérifiée, admet $P(\boldsymbol{\theta}|\mathbf{y})$ pour loi stationnaire.
L'algorithme 1 permet de simuler une chaîne de Markov dont la loi stationnaire est la loi *a posteriori* $P(\boldsymbol{\theta}|\mathbf{y})$.
Le comportement de la chaîne se comprend à partir de l'équation 3.1 : à l'équilibre, l'exploration de l'espace des paramètres se concentre sur les régions pour lesquelles la probabilité

Algorithme 1 Algorithme de Metropolis-Hastings

{Initialiser la chaîne de Markov}
$\boldsymbol{\theta}_0$
répéter
{Proposer un nouvel état}
$\boldsymbol{\theta}^{'} \sim q(.|\boldsymbol{\theta}_t)$
{Calculer la probabilité d'accepter ce nouvel état}
$\alpha(\boldsymbol{\theta}, \boldsymbol{\theta}^{'}) = min\left(1, \frac{P(\boldsymbol{\theta}^{'}|\mathbf{y})q(\boldsymbol{\theta}|\boldsymbol{\theta}^{'})}{P(\boldsymbol{\theta}|\mathbf{y})q(\boldsymbol{\theta}^{'}|\boldsymbol{\theta})}\right)$
{Accepter ou rejeter la mise à jour avec la probabilité α}
$u \sim \mathcal{U}(0,1)$
si $u \leq \alpha(\boldsymbol{\theta}_t, \boldsymbol{\theta}^{'})$ **alors**
$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}^{'}$
sinon
$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t$
fin si
{Passer à l'itération suivante}
$t = t + 1$
jusqu'à l'atteinte du régime stationnaire et l'obtention d'un échantillon de la loi *a posteriori*

(ou densité) *a posteriori* est forte. En effet, les sauts vers des régions de plus haute probabilité *a posteriori* sont presque toujours acceptés, les petits (resp. grands) sauts vers des densités un peu (resp. beaucoup) plus faibles sont parfois (resp. rarement) acceptés.

3.1.5.2 Echantillonneur de Gibbs

Cette algorithme peut être utilisé lorsque le paramètre $\boldsymbol{\theta}$ considéré est multidimensionnel. Dans ce cas, les lois conditionnelles univariées, *i.e.* les lois dans lesquelles tous les paramètres sont fixés sauf un, sont bien souvent plus facile à simuler que la loi jointe de tous les paramètres. L'échantillonneur de Gibbs consiste donc à générer d valeurs successivement plutôt qu'un seul vecteur de paramètre de dimension d en 1 itération. Il s'agit en fait d'un cas particulier de l'algorithme MH dans lequel la distribution $q(.|\boldsymbol{\theta})$ est égale à la loi conditionnelle de $\boldsymbol{\theta}$. Du fait de cette utilisation toutes les mises à jour des paramètres concernés sont acceptés car, dans ce cas, la probabilité de passer de l'état θ à l'état $\boldsymbol{\theta}^{'}$, $\alpha(\boldsymbol{\theta}, \boldsymbol{\theta}^{'})$, vaut 1.

Algorithme 2 Echantillonneur de Gibbs pour $\boldsymbol{\theta} = \{\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(d)}\}$:

{Initialiser la chaîne de Markov}
$\boldsymbol{\theta}_0$
répéter
Pour $i = 1, \ldots, d$, faire
$\boldsymbol{\theta}_{t+1}^{(i)} \sim P(.|\boldsymbol{\theta}_t^{(1)}, \ldots, \boldsymbol{\theta}_t^{(i-1)}, \boldsymbol{\theta}_t^{(i+1)}, \ldots, \boldsymbol{\theta}_t^{(d)}, y)$
{Passer à l'itération suivante}
$t = t + 1$
jusqu'à l'atteinte du régime stationnaire et l'obtention d'un échantillon de la loi *a posteriori*

Du Metropolis-Hasting à l'échantillonneur de Gibbs. Chacun des deux algorithmes présente des avantages et inconvénients. L'échantillonneur de Gibbs fournit une unique nouvelle valeur pour $\boldsymbol{\theta}$ à chaque itération, mais se base sur une évaluation d'une distribution conditionnelle. De l'autre côté, l'étape MH ne demande pas une évaluation de la distribution conditionnelle mais ne garantit pas l'acceptation d'une nouvelle valeur. Si les distributions conditionnelles sont difficiles à obtenir ou si leur calcul est assez coûteux, alors l'algorithme MH peut être utilisé. En résumé, l'echantillonneur de Gibbs permet d'avoir une convergence rapide de la chaîne si le calcul des probabilités conditionnelles n'est pas très long à chaque itération. L'algorithme MH sera en général plus rapide à chaque itération mais ne garantit pas nécessairement l'exploration de l'espace des paramètres.
Pour les modèles hiérarchiques où les distributions conditionnelles sont faciles à obtenir et à simuler, alors l'echantillonneur de Gibbs est favorisé. Par contre, dans des problèmes plus complexes, on a plus recours à l'algorithme de Metropolis-Hastings.

3.1.5.3 Méthodes MCMC à saut réversible (RJMCMC)

Les méthodes à saut réversible proposées dans Green (1995) étendent les possibilités des techniques MCMC à la sélection de modèles (dont le nombre de paramètres peut varier). Le but des méthodes RJMCMC est d'estimer les probabilités *a posteriori* des modèles considérés pour identifier celui (ou ceux) qui explique(nt) le mieux les données observées et pondérer les estimations des paramètres.
Chaque modèle $\mathcal{M}$ comprend un jeu de paramètres $\boldsymbol{\theta}_{\mathcal{M}}$ dont la dimension $n_{\mathcal{M}}$ peut changer d'un modèle à l'autre. Si $P(\mathcal{M})$ est la probabilité *a priori* du modèle $\mathcal{M}$, la formule de Bayes donne la distribution *a posteriori* :

$$P(\boldsymbol{\theta}_{\mathcal{M}}, \mathcal{M}|\mathbf{y}) \propto P_{\mathcal{M}}(\mathbf{y}|\boldsymbol{\theta}_{\mathcal{M}}) P_{\mathcal{M}}(\boldsymbol{\theta}_{\mathcal{M}}) P(\mathcal{M})$$

avec $P_{\mathcal{M}}(\mathbf{y}|\boldsymbol{\theta}_{\mathcal{M}})$ la vraisemblance des paramètres selon le modèle $\mathcal{M}$ et $P_{\mathcal{M}}(\theta_{\mathcal{M}})$ la distribution *a priori* des paramètres selon le même modèle. L'algorithme RJMCMC, de façon analogue à l'algorithme MH, propose une mise à jour du modèle. Si le modèle courant est $\mathcal{M}$, le modèle $\mathcal{M}^{'}$ est proposé avec la probabilité $P(\mathcal{M}^{'}|\mathcal{M})$. Il s'agit ensuite de passer de $\boldsymbol{\theta}_{\mathcal{M}}$ à $\boldsymbol{\theta}_{\mathcal{M}'}$ sans oublier de prendre en compte les changements de dimension. Dans ce but, un vecteur u de taille d est généré à partir d'une distribution $q_{\mathcal{M},\mathcal{M}'}$ indépendante de $\boldsymbol{\theta}_{\mathcal{M}}$. Les paramètres du nouveau modèle se calculent à partir du difféomorphisme (fonction différentiable bijective dont la réciproque est également différentiable) $g_{\mathcal{M},\mathcal{M}'}$:

$$\left(\boldsymbol{\theta}_{\mathcal{M}}^{'}, u^{'}\right) = g_{\mathcal{M},\mathcal{M}'}\left(\boldsymbol{\theta}_{\mathcal{M}}, u\right),$$

de telle sorte que les dimensions des espaces de départ et d'arrivée de $g_{\mathcal{M},\mathcal{M}'}$ correspondent. On a donc,

$$n_{\mathcal{M}}^{'} + d^{'} = n_{\mathcal{M}} + d.$$

Le passage de $(\boldsymbol{\theta}_{\mathcal{M}}, \mathcal{M})$ à $\left(\boldsymbol{\theta}_{\mathcal{M}'}, \mathcal{M}^{'}\right)$ est accepté avec la probabilité :

$$\alpha(\mathcal{M}, \mathcal{M}^{'}) = \min\left(1, \frac{P(\boldsymbol{\theta}_{\mathcal{M}'}, \mathcal{M}^{'}|\mathbf{y}) q_{\mathcal{M}',\mathcal{M}}(u^{'}) P(\mathcal{M}^{'}|\mathcal{M})}{P(\boldsymbol{\theta}_{\mathcal{M}}, \mathcal{M}|\mathbf{y}) q_{\mathcal{M},\mathcal{M}'}(u) P(\mathcal{M}|\mathcal{M}')} \left|\frac{\partial_{g_{\mathcal{M},\mathcal{M}'}}(\boldsymbol{\theta}_M, u)}{\partial(\boldsymbol{\theta}_{\mathcal{M}}, u)}\right|\right)$$

Les probabilités *a posteriori* des modèles se calculent à partir des fréquences de visite de chaque modèle. Ces probabilités permettent de tenir compte de l'incertitude liée au choix du modèle lors de l'estimation des paramètres.

Implémentation et convergence des algorithmes de type MCMC. La mise en pratique des méthodes MCMC décrites dans cette section présente des difficultés qui peuvent compliquer l'interprétation des résultats. En particulier, le temps d'atteinte du régime stationnaire est inconnu, il convient d'attendre un certain nombre d'itérations (*burn-in*) avant d'échantillonner la chaîne de Markov.

La stratégie d'exploration de l'espace des paramètres, *i*.e. le choix de $q(.|\boldsymbol{\theta})$ et des conditions initiales, influencent le temps de convergence. La conception d'un noyau de transition adapté au problème posé est donc essentielle à la qualité des estimations *a posteriori*. Par ailleurs, il est difficile d'obtenir un échantillon de réalisations indépendantes de la loi stationnaire. En effet, les valeurs issues de la simulation d'une chaîne de Markov sont généralement corrélées. Il convient alors d'observer un certain intervalle de temps entre deux échantillonnages de la chaîne. Une des difficultés présentée par les algorithmes MH et Gibbs est qu'il n'existe pas de critère absolu qui indique que l'on a atteint le régime stationnaire de la chaîne de Markov.

Plusieurs critères sont suggérés par Gelman et al. (2003). Une des techniques les plus couramment utilisées consiste à exécuter l'algorithme plusieurs fois de façon indépendante, en changeant les conditions initiales à chaque fois. On compare ensuite les échantillons obtenus par les différentes instances de l'algorithme. Si le régime stationnaire est atteint, alors ces différents échantillons sont de même loi. On peut par exemple comparer la variance propre à chaque échantillon avec la variance entre les échantillons, qui doivent être du même ordre si le régime stationnaire est atteint.

3.2 Les modèles hiérarchiques bayésiens en épidémiologie

Dans cette section nous présentons les modèles hiérarchiques en cartographie qui ont été développés pour lisser les différences de précision des estimations brutes des SMR. Rappelons avant de présenter ces modèles, qu'en cas de maladie non contagieuse rare, le modèle le plus approprié pour les données de comptage est le modèle de Poisson. On suppose que le nombre de cas y_i est indépendamment distribué selon une loi de Poisson de moyenne μ_i :

$$y_i \sim Pois(\mu_i) \tag{3.2}$$

avec $\mu_i = E_i r_i$ (ou $\mu_i = n_i \lambda_i$) pour le risque absolu (ou relatif). La moyenne de la loi de Poisson est décomposée en deux parties. Une partie représente l'effet de la population soit à travers le nombre de cas attendu E_i soit à travers la taille de la population n_i. La seconde partie représente le risque relatif r_i ou le risque absolu λ_i. Dans cette section, nous présentons les modèles existants dans la littérature pour le risque relatif r_i.

L'objectif principal en cartographie est la modélisation du risque relatif . L'approche commune est de considérer une fonction de lien logarithmique à un modèle linéaire :

$$\log(r_i) = \eta_i. \tag{3.3}$$

Cette forme de modèle a été largement utilisée dans différentes applications pour l'analyse des données de comptage dans des petites zones géographiques (Stevenson et al., 2005; Waller and Gotway, 2004, Chapitre 9).
Tous les modèles hiérarchiques présentés dans cette section sont formulés par le modèle hiérarchique générique suivant :

$$\begin{aligned} \forall i = 1, \ldots, N \; Y_i &\sim Pois(E_i r_i) && (3.4)\\ r_i &\sim P(.|\boldsymbol{\theta}) && (3.5)\\ \boldsymbol{\theta} &\sim P_{\boldsymbol{\theta}}(.). \end{aligned}$$

où $P(.|\boldsymbol{\theta})$ est une distribution *a priori* appropriée pour le risque r_i et les paramètres $\boldsymbol{\theta}$ sont les hyperparamètres du second niveau hiérarchique avec une distribution $P_{\boldsymbol{\theta}}(.)$. La différence entre ces modèles réside dans le choix de la distribution *a priori* assignée au risque r_i.

3.2.1 Modèles non spatiaux

Dans cette section nous présentons les modèles non spatiaux qui ont été proposés pour lisser les estimations brutes des SMR. On reproche à ces modèles le fait que la dépendance spatiale entre les unités géographiques n'est pas prise en compte. Néanmoins, un ensemble de modèles ont été développés et sont utiles dans certains cas où l'on soupçonne une faible corrélation.

3.2.1.1 Les modèles à effets aléatoires non corrélés

La flexibilité de l'approche décrite par l'équation 3.3 a donné naissance à une large gamme de modèles de régression incorporant des covariables à effets fixes ou encore aléatoires. Le problème de ces modèles de régression est qu'ils ne capturent pas complètement la variabilité présente dans les données de comptage.
La surdispersion ou encore la corrélation spatiale présentes dans ces données et qui sont dues à des facteurs de confusion (des facteurs d'exposition secondaires), ne peuvent pas être capturées par la simple introduction de covariables dans le modèle mais nécessitent d'inclure en plus un terme ou plusieurs termes capables de capturer de tels effets. Initialement, la surdispersion ou l'extra-variabilité peuvent être prises en compte dans ce modèle de deux manières :

- Considérer une loi *a priori* pour le risque relatif (le modèle Poisson-gamma).
- Étendre le prédicteur linéaire (ou non linéaire) pour inclure un effet aléatoire supplémentaire (le modèle log-normal).

Ces deux manières de procéder ont donné naissance à la famille des modèles à effets aléatoires non corrélés.

Le modèle Poisson-gamma Pour le second niveau de l'expression (3.4), Clayton and Kaldor (1987) proposent une variété de distributions *a priori* pour r_i. Dans le cas le plus simple, ils suggérent de lui assigner une loi gamma avec un paramètre d'échelle a et un paramètre de forme b. Cela signifie que la loi gamma est de moyenne a/b et de variance a/b^2. Les paramètres a et b sont inconnus et peuvent être soit supposés fixés soit munis d'une distribution *a priori*.

– si a et b sont fixés à partir des données, on se retrouve dans le cas des méthodes bayésiennes empiriques. Pour ce modèle, la distribution *a priori* est appelée l'*a priori conjugué* car la distribution *a posteriori* est de la même famille que celle de l'*a priori*. La distribution *a posteriori* des paramètres r_i est alors :

$$r_i|y_i, E_i, a, b \sim Gamma(y_i + a, E_i + b).$$

La moyenne et la variance *a posteriori* sont donc respectivement $(y_i + a)/(E_i + b)$ et $(y_i + a)/(E_i + b)^2$. On pourrait remarquer que l'estimation du RR de l'unité i est une combinaison pondérée du SMR de l'unité i et de l'estimation *a priori* :

$$\frac{y_i + a}{E_i + b} = SMR_i \times w_i + \frac{a}{b} \times (1 - w_i),$$

où $w_i = E_i/(E_i + b)$ peut être vu comme le poids associé au SMR de l'unité i. Cette approche a donc pour effet d'atténuer les contrastes initiaux liés aux différences de précision des estimations. En effet pour les unités avec une population importante, l'estimation sera dominée par les données et sera proche du SMR et pour les unités avec des effectifs faibles, le poids associé au SMR sera plus petit et le lissage sera plus important, ce qui fera que les estimations seront moins variables que les SMR.

– si a et b sont supposés variables, une loi *a priori* est assignée aussi aux paramètres a et b. Une paramétrisation linéaire pour les paramètres de la loi gamma peut être envisagée. Par exemple, le modèle suivant peut être considéré :

$$r_i|y_i, E_i, a, b \sim Gamma(a, b_i),$$

avec $b_i = a/\mu_i$. Pour ce modèle, la moyenne et la variance *a priori* sont égales respectivement à μ_i et μ_i^2/a.
Ces hyperparamètres (a et b) suivent eux mêmes une distribution *a priori* selon h_α et h_β. On peut noter que les estimations des r_i seront différentes selon les valeurs prises par a et b comme discuté dans Lawson and Williams (2001). Cette approche implique un lien direct entre la moyenne et la variance qui peut être vue comme un inconvénient. En général, une paramétrisation lognormale est préférée à celle là.

Le modèle Poisson-lognormal Le modèle Poisson-gamma est algébriquement pratique du fait que la distribution *a posteriori* est aussi une loi gamma. Toutefois, elle peut être restrictive car l'ajustement sur des covariables est difficile et il n'y a aucune possibilité de tenir compte de l'autocorrélation spatiale entre les risques des zones voisines. Clayton and Kaldor (1987)

proposent une distribution *a priori* log-normale (la transformation $\log$ permet de passer sur $\mathbb{R}$) pour les risques relatifs. Ce modèle est considéré comme beaucoup plus flexible que le modèle Poisson-gamma et permet facilement l'introduction de covariables comme souligné dans Ghosh et al. (1998) ou encore la prise en compte d'une structure spatiale entre les risques. La forme générale de ce modèle proposé par Clayton and Kaldor (1987) est :

$$\log r_i = \varepsilon_0 + \mathcal{V}_i,$$

où ε_0 est un terme constant qui représente l'effet moyen commun à toutes les unités géographiques et $\mathcal{V}_i$ sont des effets aléatoires gaussiens indépendants et identiquement distribués, $\mathcal{V}_i \sim \mathrm{N}(0, \sigma_v^2)$. Le choix des distributions *a priori* de ε_0 et σ_v^2 est aussi nécessaire et est différent selon les données traitées. Arora and Lahiri (1997) proposent d'assigner des *a priori* vagues, par exemple, pour ces hyperparamètres.

3.2.1.2 Modèles de mélanges

Une autre famille de modèles alternative à ceux présentés en section 3.2.1.1 est la famille des modèles de mélange (voir section 4.2.1 pour plus de détails). Le but principal de ces modèles n'est pas l'estimation mais plutôt la détection des zones à risque élevé ou pas. Le risque est considéré dans ces modèles comme une combinaison de plusieurs niveaux de risque non observés. Un premier exemple a été proposé par Schlattman and Boehning (1993), bien que ce soit dans un contexte bayésien empirique. Dans leur approche, la distribution gouvernant les données observées est un mélange de distributions de Poisson (au lieu d'une simple Poisson) :

$$f(y_i|\pi, E_i, \mathbf{r}) = \sum_{k=1}^{K} \pi_k Pois(y_i|E_i r_k),$$

avec $\mathbf{r} = \{r_1, \ldots, r_K\}$, $\pi = \{\pi_1, \ldots, \pi_K\}$ les probabilités d'appartenance à la composante k du mélange et r_k le risque pour la classe k, pour $k \in \{1, \ldots, K\}$, quand K est fixé et $\sum_k \pi_k = 1$. Les paramètres $\boldsymbol{\pi}$ et $\mathbf{r}$ sont inconnus. Le choix des distributions *a priori* est différent selon les applications considérées. Par exemple, Schlattman and Boehning (1993) proposent la distribution de Dirichlet pour les probabilités d'appartenance où : $\boldsymbol{\pi} \sim Dir(\boldsymbol{\alpha})$ avec $\boldsymbol{\alpha} = \{\alpha_k, k = 1, \ldots, K\}$ et une distribution gamma pour les α_k. Dans ce cas, les distributions *a priori* pour les r_k peuvent être des distributions gamma aussi avec des paramètres pour-lesquels des distributions (hyper-priors) doivent être spécifiées. Pour le nombre de composantes K, une distribution uniforme est souvent utilisée $K \sim \mathrm{Unif}(1, K_{max})$ comme dans Green (1995), Fernandez and Green (2002), Green and Richardson (2002), où K_{max} est fixé

par l'utilisateur. Ces modèles ne prennent pas en compte la dépendance spatiale entre les unités géographiques mais il est par contre clair, que les modèles de mélange peuvent être généralisés à des situations plus complexes. La première idée est de les étendre à des modèles avec classes latentes dépendantes.

3.2.1.3 Modèle de comptage à inflation de zéro

Les modèles de comptage à inflation de zéro sont souvent utilisés pour modéliser des données de comptage sur-dispersées et/ou comportant une grande proportion de zéro. c'est le cas notamment pour les maladies rares. Il faut noter que pour ce genre de données, plusieurs régions auront zéro cas et peu de régions auront un petit nombre de cas enregistré, ce qui peut engendrer une distribution sur-dispersée et conduire à avoir une distribution marginale multimodale. Ce problème n'est pas forcément bien modélisé par une distribution surdispersée comme la loi binomiale négative.
Les modèles à inflation de zéro ont été introduits en premier par Lambert (1992) pour prendre en compte cette surdispersion engendrée par la présence d'un excés de zéros dans les données. Il y a une large littérature sur les modèles de mélange pour données surdispersées (Boehning et al., 1999; Agarwal et al., 2002; Ghosh et al., 2006, parmi d'autres).
Ils sont un cas particulier des modèles de mélanges (voir section 4.2.1) à deux composantes. Lorsque un mélange de Poisson est considéré, le cas simple est celui à deux composantes où les zéros appartiennent à la composante dont la probabilité est : $(1-\pi)+\pi\exp(-\mu)$, avec μ la moyenne de la loi de Poisson (avec $1-\pi$ et π sont appelés les poids du mélange) et les valeurs non nulles appartiennent à la composante dont la probabilité est : $\pi\exp(-\mu).\frac{\mu^{\mathbf{y}}}{\mathbf{y}!}$.
En particulier, pour nos données observées y_i et les cas attendus E_i, le modèle est défini comme suivant :

$$y_i|E_i, r_i \sim (1-\pi)Pois(0) + \pi Pois(E_i r_i)$$

Pour mieux comprendre ces deux composantes, on peut considérer que c'est un problème où une variable latente traite les zéros soit comme des zéros structurés (faux zéros) donc $z=0$, soit comme des vrais zéros et donc $z=1$. Dans ces modèles, z est non observée et doit être estimée. La vraisemblance des données complètes (Marin and Robert, 2007) utilisée pour estimer les paramètres est alors :

$$L(\mathbf{y}|\mathbf{z}) = \prod_{i\in S} \pi_{z_i} Pois(y_i; E_i r_{z_i}).$$

Lambert (1992) ont implémenté ce modèle dans un contexte de régression en utilisant l'algorithme EM (voir section 4.2.1.3) pour l'estimation des paramètres. Ces modèles sont adaptés

aux problèmes des excés de zéros dans les données de maladies rares mais l'absence d'une composante modélisant la corrélation spatiale fait que ces modèles sont moins performants que les modèles qui prennent en compte cette corrélation : une partie de la variabilité reste inexplicable par cette modélisation.

3.2.2 Les modèles spatiaux hétérogènes

Les modèles présentés dans la section 3.2.1 aident à améliorer les estimations du risque obtenues par les SMR, mais ne prennent pas en compte la corrélation spatiale entre les unités géographiques et présentent plusieurs limites. Par exemple, il n'y a pas de généralisation simple et adaptée pour la distribution gamma avec des paramètres corrélés spatialement. Wolpert and Ickstadt (1998) ont présenté un exemple de modèles de champ de gamma corrélés, mais il a été démontré par Best et al. (2005) que ces modèles ont une faible performance dans les études d'evaluation qu'ils ont menées.
Ces limitations ont poussé à se tourner vers les modèles gaussiens qui permettent une incorporation plus simple d'une quelconque structure de corrélation. En effet, quand les effets aléatoires sont supposés être corrélés, il est plus simple de spécifier une forme lognormale pour n'importe quel extra-variabilité présente. Une simple extension est de considérer des composantes additives qui décrivent les différentes variabilités dont on soupçonne la présence dans les données. Les modèles présentés dans cette section sont en effet une adaptation du modèle Poisson-lognormal présenté en section 3.2.1.1.

3.2.2.1 Modèle BYM (Besag, York et Mollié)

Il est naturel dans la cartographie du risque de supposer que le risque dans les unités géographiques voisines est similaire. Donc la prise en compte d'une autocorrélation spatiale dans les modèles est justifiable. Un des modèles les plus utilisés en épidémiologie est le modèle hiérarchique bayésien de Besag, York et Mollié (Besag et al., 1991). Ce modèle est largement utilisé car il permet la prise en compte de la surdispersion en introduisant un terme d'hétérogénéité non structurée, l'autocorrélation spatiale et l'ajustement des covariables. Ce modèle se présente dans sa forme générale comme suit :

$$\log(r_i) = \beta_0 + \mathcal{U}_i + \mathcal{V}_i,$$

avec $\mathcal{U}_i$ et $\mathcal{V}_i$ des effets aléatoires décrivant respectivement l'hétérogénéité structurée (corrélation spatiale) et l'hétérogénéité non structurée. Ces effets aléatoires sont considérés comme

des variables latentes capturant les effets des facteurs de risque inconnus ou non mesurés structurés ou non spatialement. La composante d'hétérogénéité non structurée est supposée suivre une loi normale définie par :

$$\mathcal{V}_i \sim \mathcal{N}(0, \sigma_v^2),$$

où σ_v^2 contrôle la variabilité des RR, dans sa composante non spatiale.
La composante spatiale suppose que les unités géographiques spatiales proches ont tendance à avoir des RR similaires. Le modèle gaussien auto-régressif conditionnel (le Conditionnal Autoregressif (CAR)) Kunsch (1987) intrinsèque, permet de prendre en compte cette hypothèse avec :

$$(\mathcal{U}_i|\mathcal{U}_j = u_j, j \neq i) \sim \mathcal{N}\left(\frac{\sum_{j\neq i} w_{ij}u_j}{\sum_{j\neq i} w_{ij}}, \frac{\sigma_u^2}{\sum_{j\neq i} w_{ij}}\right),$$

où les poids w_{ij} décrivent la proximité des unités i et j et σ_u^2 contrôle la variabilité conditionnelle des RR, dans sa composante spatiale.
Le critère de proximité géographique le plus utilisé en épidémiologie est celui d'adjacence. Les unités i et j sont voisines si elles partagent une frontière commune :

$$w_{ij} = \begin{cases} 1 & \text{si les unités i et j sont voisines} \\ 0 & \text{sinon.} \end{cases}$$

D'autres options pour w_{ij} apparaissent dans la littérature comme proposée par Best et al. (1999). Les variances σ_u^2 et σ_v^2 modulent les niveaux d'hétérogénéité locale et globale respectivement. On peut remarquer que plus σ_v^2 est petit, plus les effets aléatoires ont tendance à être similaires entre toutes les unités géographiques. De même plus σ_u^2 est petit, plus les effets aléatoires ont tendance à être similaires entre unités géographiques voisines.
Le modèle CAR intrinsèque a l'avantage d'être facilement estimable. En revanche, c'est un modèle impropre : sa moyenne est non définie et sa variance est infinie. Pour que le modèle soit identifiable, la contrainte $\sum_i \mathcal{U}_i = 0$ doit être imposée. Besag (1974), Besag et al. (1991) et Cressie (1993)(pages 407-408, 410-423) donnent des détails sur les structures des modèles autorégressifs conditionnels.
Les distributions *a priori* de β_0, σ_v^2 et σ_u^2 doivent être spécifiées. Le choix des distributions *a priori* des paramètres de variance est délicat (Wakefield, 2007). En pratique, les distributions conjuguées gamma inversé sont populaires et Ghosh et al. (1999) et Sun et al. (1999) ont présenté les restrictions concernant les distributions hyperpriors pour ces paramètres afin d'assurer l'identifiabilité. L'inférence de ce modèle se fait *via* les algorithmes de type MCMC présentés en section 3.1.5.

Récemment, plusieurs auteurs (Bernadinelli et al., 1995; Heisterkamp et al., 2000; Knorr-Held and Rasser, 2000; Sun et al., 2000; Waller et al., 1997; Xia and Carlin, 1998) ont étendu ce modèle au contexte spatio-temporel en permettant à la structure spatiale de varier dans le temps.
Le modèle BYM présenté ici est le plus utilisé en épidémiologie humaine et animale. Ce modèle a été étendu par Clayton and Bernadinelli (1992) et appelé le modèle de convolution par Mollie (1999). Le modèle CAR a tendance à produire un degré élevé de lissage et ne permet pas de détecter les discontinuités dans la structure spatiale du risque. D'autres modèles alternatifs ont été proposés pour remédier à cette limite (voir section 3.2.3.3). Il existe un large choix d'approches différentes moins utilisées, certes, mais qui peuvent répondre mieux aux objectifs attendus pour certaines données.

3.2.2.2 Modèles à covariance spatiale

Une éventuelle alternative au BYM est un modèle où les termes aléatoires corrélés et non corrélés sont tous les deux modélisés par un seul terme. Cela peut être réalisé en spécifiant une distribution *a priori* à deux paramètres représentant ces deux effets. La matrice de covariance d'une distribution *a priori* multivariée normale (MVN) peut être paramétriquement modélisée avec de tels termes (Diggle et al., 1998; Wikle, 2002). Cette approche est apparentée à l'approche du "Krigeage universel" (Wackernagel, 2003; Cressie, 1993). Pour les données de comptage, la moyenne de la loi de Poisson dans ce cas là serait :

$$E_i \exp\{\beta + S_i\},$$

avec $\mathbf{S} \sim MVN(0, \Gamma)$ où $\mathbf{S} = \{S_1, \ldots, S_N\}$ est le vecteur d'effets aléatoires et Γ est une matrice de covariance spatiale (Kelsall and Wakefield, 2002). En pratique, cette matrice de covariance spatiale consiste en des fonctions paramètriques qui définissent la covariance comme fonction des localisations relatives pour n'importe quelle paire d'observations. Cressie (1993), Waller and Gotway (2004) ont présenté des introductions à ces fonctions de covariance. Ces modèles sont plutôt utilisés en géostatistique et ne représente qu'une petite fraction de la littérature concernant la cartographie en épidémiologie.
Best et al. (2005); Henderson et al. (2002) rapportent que ces modèles surlissent en général les estimations de risques extrêmes et conduisent à des mauvaises inférences comparées aux modèles où un CAR est spécifié pour l'*a priori*. Ils sont aussi très coûteux en terme de calcul.

3.2.3 Les modèles spatiaux alternatifs

Nous avons présenté à la section 3.2.2 les modèles les plus utilisés dans les applications en cartographie du risque. Toutefois, il existe un ensemble de modèles alternatifs qui sont moins communs mais qui doivent être pris en considération dans certaines applications. La première famille de modèles utilise l'approximation de la vraisemblance par la pseudo-vraisemblance proposée par Besag (1975). La deuxième famille repose sur le relâchement de la supposition paramètrique inhérente à la spécification de la distribution *a priori* pour les modèles de convolution et pour la covariance.

3.2.3.1 Les modèles auto-logistiques

Besag (1975) a proposé de modéliser les variables distribuées continûment dans l'espace en conditionnant par les voisins. La vraisemblance sera donc remplacée par un produit de distributions conditionnelles en chaque site i. La *pseudo-vraisemblance* est simplement de la forme :

$$L_p(\mathbf{y}|\mathbf{r}) = \prod_{i=1}^{n} P(y_i|y_j, j \in \mathcal{N}_i; \mathbf{r}), \tag{3.6}$$

avec $\mathcal{N}_i$ l'ensemble des voisins pour le site i. Cette pseudo-vraisemblance peut prendre différentes formes selon la spécification paramétrique donnée à la distribution conditionnelle $P(y_i|y_j, j \in \mathcal{N}_i; \mathbf{r})$.
Le modèle dérivant de cette famille et le plus utilisé en épidémiologie est celui où les données y_i ne prennent que les valeurs 1 ou 0 (présence ou absence de l'épidémie), dans ce cas :

$$P(y_i|y_j, j \in \mathcal{N}_i; \mathbf{r}) = r_i^{y_i}(1 - r_i)^{1-y_i},$$

Besag (1974) montre que la forme la plus naturelle de r_i est : $r_i = \exp(\alpha_i + \delta S_{\mathcal{N}_i})/1 + \exp(\alpha_i + \delta S_{\mathcal{N}_i})$ où $S_{\mathcal{N}_i} = \sum_{j \in \mathcal{N}_i} y_j$ est la somme des voisins du site i et δ représente le paramètre de la dépendance spatiale.
La pseudo-vraisemblance est donnée alors par :

$$L_p(y_i|\mathbf{r}) = \prod_{i=1}^{n} \left[\frac{[\exp(\alpha_i + \delta S_{\mathcal{N}_i})]^{y_i}}{1 + \exp(\alpha_i + \delta S_{\mathcal{N}_i})} \right].$$

Dans les différentes applications en cartographie du risque ou encore en agriculture ou la modélisation des distributions d'espèces, il est important de pouvoir introduire des covariables dans le modèle.
Gumpertz et al. (1997), Wu and Huffer (1997), Hoeting et al. (2000) ont proposé d'introduire

des covariables dans ce modèle *via* les constantes α_i, donc $\mathbf{r}$ peut s'exprimer sous la forme suivante :

$$r_i = \frac{\exp(\delta S_{\mathcal{N}_i} + \mathbf{x_i^t}\boldsymbol{\gamma})}{1 + \exp(\delta S_{N_i} + \mathbf{x_i^t}\boldsymbol{\gamma})},$$

où $\mathbf{x_i^t}\boldsymbol{\gamma}$ est le prédicteur linéaire avec $\mathbf{x_i^t}$ le vecteur transposé des valeurs des covariables pour chaque unité i et $\boldsymbol{\gamma}$ le vecteur des paramètres. Ce modèle peut aussi être étendu pour inclure des effets aléatoires. Des termes d'effets aléatoires ν_i (au niveau de l'unité) non corrélés, par exemple, peuvent être introduits comme :

$$r_i = \frac{\exp(\delta S_{\mathcal{N}_i} + \mathbf{x_i^t}\boldsymbol{\gamma} + \nu_i)}{1 + \exp(\delta S_{\mathcal{N}_i} + \mathbf{x_i^t}\boldsymbol{\gamma} + \nu_i)}.$$

Dans un contexte bayésien, le terme $\delta S_{\mathcal{N}_i} + \mathbf{x_i^t}\boldsymbol{\gamma} + \nu_i$ est traité de la même manière que les autres modèles spatiaux où les covariables et les termes d'effets aléatoires peuvent être présents.

Comme précédemment, les distributions *a priori* des paramètres $\boldsymbol{\gamma}$, δ et le terme d'effets aléatoires ν doivent être spécifiées au second niveau de la hiérarchie. Encore différents choix de distributions *a priori* pour les hyperpriors peuvent être proposés en fonction des données traitées (Caragea and Kaiser, 2006; Gumpertz et al., 1997). Il faut noter que la corrélation spatiale est introduite au premier niveau de la hiérarchie contrairement aux modèles présentés en section 3.2.2.

Il existe bien sûr différents auto-modèles basés sur les différentes distributions de Poisson, binomial ou encore de la famille exponentielle. En général, c'est la pseudo-vraisemblance qui est utilisée pour l'estimation des paramètres vu que la constante de normalisation n'est généralement pas analytique et plusieurs contraintes sur les paramètres doivent être imposées pour assurer l'identifiabilité. Parmi tous ces modèles, le plus utilisé est le modèle auto-logistique, en raison notamment de la simplicité de son implémentation et de son interprétation.

La pseudo-vraisemblance est connue pour être une approximation raisonnable à la vraisemblance lorsque la corrélation spatiale n'est pas très forte. Une extension de ce modèle au spatio-temporel est possible. Besag and Tantrum (2003) ont proposé d'utiliser le modèle auto-logistique pour un ensemble de données spatio-temporelles.

3.2.3.2 Modèles à base de splines

les modèles à base de splines est une approche semi-paramétrique qui représente une alternative aux modèles paramétriques pour la modélisation de la composante spatiale structurée du risque.

L'idée de cette approche est de considérer un paramètre de lissage pour représenter la structure de la moyenne d'un processus. En plus de l'hypothèse usuelle pour les données observées ($y_i \sim Pois(\mu_i), \forall i \in \{1, \ldots, N\}$), on définit un point de référence $s_i = (s_{i1}, s_{i2})$ qui peut être, par exemple, le centre de la région.
On suppose, pour cette famille de modèles, que :

$$\log \mu_i = Spl(s_i)$$

où $Spl(.)$ est la fonction de lissage. Un large choix est possible pour cette fonction. On présente ici les modèles splines qui sont attractifs dans différentes applications et qui ont un lien assez étroit avec les processus gaussiens (French and Wand, 2004).
La moyenne est définie, dans ce cas, par :

$$\log(\mu_i) = \alpha_0 + \sum_{j=1}^{2} \alpha_j s_{ij} + \sum_{j=1}^{\eta_k} \psi_j C\left(||s_i - \kappa_j||\right),$$

où $\kappa_j, j \in \{1, \ldots, \eta_\kappa\}$ est un ensemble de noeuds (des points fixés dans l'espace), ψ_j est un effet aléatoire gaussien et C est la fonction de covariance qui est définie dans ces modèles par :

$$C(d) = (1 + |d|)e^{-|d|}.$$

La distribution *a priori* jointe des effets aléatoires donnée par French and Wand (2004) est :

$$\psi \sim \mathcal{N}(0, \tau \mathbf{w}^{-1}),$$

où $\mathbf{w}$ est une matrice carrée qui peut être définie comme :

$$\mathbf{w} = [C\{||\kappa_i - \kappa_j||/\rho\}]_{1 \leq i,j \leq \eta_\kappa},$$

où ρ est considéré comme le paramètre de lissage. La valeur de ρ est fixée dans French and Wand (2004), mais il serait plus approprié d'estimer ce paramètre vu qu'il contrôle le degré de lissage. Des modèles à base de splines alternatifs ont été proposés par Zhang et al. (2006), Macnab (2007) pour la modélisation spatio-temporelle en cartographie du risque.

3.2.3.3 Modèles de mélange spatiaux

Dans cette section nous présentons une autre famille des modèles spatiaux à classes latentes. Le risque est considéré ici comme une combinaison de niveaux de risque. Ces niveaux sont des variables latentes et on n'a aucune connaissance *a priori* de leurs distributions. Ce

type de modèles a comme objectif d'estimer le risque pour un regroupement (classe) d'unités géographiques contrairement aux modèles précédents où le résultat obtenu est l'estimation du risque pour chaque unité géographique.
Les modèles présentés ici sont des extensions des modèles de mélange, présentés en 3.2.1.2, qui prennent en compte la dépendance spatiale entre les unités géographiques. Différentes approches ont été proposées pour inclure cette corrélation. Une approche proposée par Fernandez and Green (2002), est de supposer que la probabilité d'appartenance de l'unités i à la classe k, π_{ik} a une structure de dépendance spatiale au lieu de considérer qu'elle est donnée par une loi de Dirichlet (voir section 3.2.1.2). Pour cela, ils proposent une variété de modèles. Un choix possible pour prendre en compte la dépendance spatiale à travers les probabilités π_{ik} serait, le modèle auto-logistique normal (voir, par exemple McCullagh and Nelder (1989), chapitre 5) qui est spécifié dans ce cas de la manière suivante :

$$
\begin{aligned}
f(y_i|\pi, e_i, \mathbf{r}) &= \sum_{k=1}^{K} \pi_{ik} Pois(y_i|e_i r_k) \\
\pi_{ik} &= \eta_{ik}(\phi) / \sum_{l=1}^{K} \eta_{il}(\phi) \\
\text{où} \quad \eta_{il}(\phi) &= \exp(x_{ik}/\phi)
\end{aligned}
$$

où $\{x_{ik}, \forall k = 1, \ldots, K, i = 1, \ldots, N\}$ est un champ aléatoire spatialement corrélé et ϕ est le paramètre de corrélation spatiale. L'*a priori* donné au champ $\{x_{ik}\}$ est un **CAR** afin d'assurer l'identifiabilité. L'inférence de ce modèle se fait *via* des algorithmes MCMC. Un ensemble de distributions hyperpriors est proposé par Fernandez and Green (2002) pour les différents paramètres du modèle.
Les connections entre les modèles de mélange et les modèles de partition est simple à mettre en évidence. Par exemple, le modèle proposé par Green and Richardson (2002) suppose qu'il existe un petit nombre de niveaux de risque $\{r_j, j = 1, \ldots, K\}$ et les unités géographiques sont affectées à un de ces niveaux de risque par l'intermédiaire d'une variable d'allocation $\{z_i, i = 1, \ldots, N\}$. Différemment des autres approches, les variables d'allocation sont supposées issues d'un modèle de Potts (voir section 4.1.3). Ce modèle permet d'obtenir une estimation de risque pour les groupes (classes) de risque. Ce modèle est aussi hiérarchique où un niveau d'hiérarchie supplémentaire est ajouté à la forme générique donnée par la formule (3.4). Souvent, les auteurs considérent ces méthodes comme concurrentes des modèles de lissage en cartographie comme le modèle de convolution de Besag présenté en section 3.2.2.1 (voir les discussions dans Knorr-Held and Rasser (2000), Ferreira et al. (2002), Fernandez and Green (2002)). Elles ont été comparées sur des simulations par Best et al. (2005).

3.3 Critères de sélection de modèles

La variété des modèles présentés en (section 3.2) confronte l'utilisateur au problème de choix du "meilleur" modèle (celui qui est le plus en adéquation avec les données). Dans un grand nombre de situations, les connaissances *a priori* sur les données ne permettent pas de déterminer un unique modèle dans lequel se placer pour réaliser l'inférence. Depuis la fin des années 70, les méthodes pour la sélection de modèles à partir des données ont été developpées. Nous présenterons dans cette partie un certain nombre de ces critères qui permettent de comparer différents modèles entre eux et qui sont fréquemment utilisés afin de vérifier si le modèle hiérarchique inferé est en bonne adéquation avec les données (voir Green and Richardson (2002) et Durand (2009) pour l'application des différents critères). Nous présentons ces algorithmes avec une forme générale des paramètres $\boldsymbol{\theta}$.

3.3.1 Divergence de Kullback-Leibler

La divergence de Kullback-Leibler (Kullback and Leibler, 1951) mesure l'écart entre deux distributions de probabilité. Soient $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ deux vecteurs de paramètres quelconques. La divergence de Kullback-Leibler entre les distributions de probabilités $p_{\boldsymbol{\theta}_1} = P(.|\boldsymbol{\theta}_1)$ et $p_{\boldsymbol{\theta}_2} = P(.|\boldsymbol{\theta}_2)$ est alors :

$$d(p_{\boldsymbol{\theta}_1}, p_{\boldsymbol{\theta}_2}) = \int P(\mathbf{y}|\boldsymbol{\theta}_1)[\log\left(P(\mathbf{y}|\boldsymbol{\theta}_1)\right) - \log\left(P(\mathbf{y}|\boldsymbol{\theta}_2)\right)]d\mathbf{y}$$

Si l'on suppose qu'il existe un "vrai modèle" (inconnu) de densité $p_{\boldsymbol{\theta}^{(t)}}$ duquel les données sont issues, alors $d(p_{\boldsymbol{\theta}^{(t)}}, p_{\tilde{\boldsymbol{\theta}}})$ mesure la distance entre la vraie distribution et la distribution du modèle ayant pour paramètres le vecteur $\tilde{\boldsymbol{\theta}}$. L'objectif est de minimiser cette distance et donc la valeur $\boldsymbol{\theta}_0 = \text{argmin}_{\tilde{\boldsymbol{\theta}}}(d(p_{\boldsymbol{\theta}^{(t)}}, p_{\tilde{\boldsymbol{\theta}}}))$ est la valeur la plus "correcte" pour le vecteur des paramètres $\boldsymbol{\theta}$. Le problème c'est qu'on ne connait pas $p_{\boldsymbol{\theta}^{(t)}}$, ce qui signifie qu'on ne peut pas calculer $\boldsymbol{\theta}_0$. Il est donc nécessaire d'utiliser une approximation pour $d(p_{\boldsymbol{\theta}^{(t)}}, p_{\tilde{\boldsymbol{\theta}}})$.

3.3.2 La déviance

La déviance sert en général à comparer la précision prédictive de plusieurs modèles alternatifs. Elle est définie par :

$$D(\mathbf{y}, \boldsymbol{\theta}) = -2 \log P(\mathbf{y}|\boldsymbol{\theta}).$$

La déviance est une notion importante de par sa connexion avec le divergence de Kullback-Leibler. Si on moyenne par la "vraie distribution" $P(\mathbf{y}|\boldsymbol{\theta}^{(t)})$, on remarque que la déviance est proportionnelle, à une constante indépendante de $\boldsymbol{\theta}$ près, à la divergence de Kullback-Leibler :

$$\mathbb{E}_{\mathbf{y}|\boldsymbol{\theta}}[D(\mathbf{y},\boldsymbol{\theta})] = 2d(p_{\boldsymbol{\theta}^{(t)}}, p_{\boldsymbol{\theta}}) - 2\int P(\mathbf{y}|\boldsymbol{\theta}^{(t)}) \log\left(P(\mathbf{y}|\boldsymbol{\theta}^{(t)})\right) d\mathbf{y}. \tag{3.7}$$

La déviance permet de mesurer l'erreur commise, en terme de divergence de Kullback-leibler, lorsque le modèle $P(\mathbf{y}|\boldsymbol{\theta})$ est supposé pour les données $\mathbf{y}$. On peut remarquer que la déviance dépend des données $\mathbf{y}$ et des paramètres $\boldsymbol{\theta}$. Il serait plus intéressant, pour comparer différents modèles, d'avoir une mesure qui ne dépend que des données $\mathbf{y}$. Il est naturel d'un point de vue bayésien, de calculer la moyenne *a posteriori* de la déviance :

$$D_{avg}(\mathbf{y}) = \mathbb{E}_{\boldsymbol{\theta}|\mathbf{y}}[D(\mathbf{y},\boldsymbol{\theta})].$$

3.3.3 Critère d'information de la déviance (DIC)

Afin de comparer différents modèles et de choisir celui qui convient le mieux au vu des données, Spiegelhalter et al. (2002) ont proposé le DIC (Deviance Information Criterion) qui est une mesure d'adéquation, pénalisée par une mesure de complexité d'un modèle :

$$DIC = \hat{D}_{avg}(\mathbf{y}) + p_D \tag{3.8}$$

où

- $\hat{D}_{avg}(\mathbf{y})$ est l'estimation de $D_{avg}(\mathbf{y})$, la moyenne *a posteriori* de la déviance (donnée par la formule (3.7)).
- p_D est le nombre de paramètres effectifs qui est calculé comme la différence entre la déviance moyenne et la déviance de la moyenne du paramètre :

$$p_D = D_{avg}(\mathbf{y}) - D_{\hat{\boldsymbol{\theta}}}(\mathbf{y}),$$

où $\hat{\boldsymbol{\theta}}$ représente un estimateur $\boldsymbol{\theta}$, qui peut être sa moyenne *a posteriori*.

Seuls les écarts entre les DIC obtenus par différents modèles ont du sens, puisque le DIC n'est pas normalisé dans l'absolu (Carlin et al., 2006). En pratique, il est assez difficile de décider quand un tel écart est significatif ou non. Dans le cadre d'un algorithme MCMC, il convient généralement de relancer plusieurs fois la chaîne de Markov (de façon indépendante), de calculer le DIC pour chaque exécution. On peut ainsi se faire une idée de la variance du DIC pour chaque modèle que l'on souhaite comparer et décider qu'une différence de DIC est significative si la variance observée pour les modèles testés ne suffit pas à l'expliquer.

3.4 Discussion

Les méthodes présentées brièvement dans ce chapitre ont pour objectif de fournir des représentations cartographiques des risques qui soient les plus informatives possibles. L'intérêt du lissage est de permettre de mieux apprécier la structure spatiale sous-jacente en lissant le bruit causé par l'instabilité des SMR dans les unités à petit nombre de cas. L'enjeu de ces méthodes est de lisser les risques relatifs pour éliminer le bruit lié aux petits effectifs et en même temps, de ne pas trop lisser les risques relatifs pour pouvoir mettre en évidence leur structure spatiale.

Nous avons essayé dans ce chapitre de résumer l'ensemble des modèles les plus utilisés en épidémiologie. Toutes ces méthodes sont formulées dans un cadre bayésien. La différence entre ces modèles réside dans le niveau de la hiérarchie où l'on met l'*a priori* sur le risque. L'objectif de remonter plus loin dans le niveau de la hiérarchie est de permettre au modèle d'être plus flexible. Ces modèles ont été développés pour permettre des éventuels discontinuités et des changements abrupts dans la distribution spatiale des risques. Tous ces modèles sont confrontés en plus au problème du choix des distributions hyperprior qui reste délicat et qui peut mener à des résultats très différents. L'estimation des paramètres de ces modèles est faite en général *via* les algorithmes MCMC qui nécessitent un nombre élevé d'itérations afin de garantir la convergence indispensable à toute estimation. Ces méthodes MCMC peuvent être très coûteuses en temps. De plus, le nombre de fois où il est nécessaire d'utiliser ces algorithmes peut être très important dans l'estimation de ces modèles comme il est indispensable de faire les analyses de sensibilité aux différents paramètres des modèles (les distributions *a priori*, par exemple).

Tous ces modèles ont été proposés pour les maladies rares et non-contagieuses humaines. Dans le cadre de l'épidémiologie animale, le modèle qui a été largement utilisé est le modèle BYM. Comme mentionné auparavant, cela est dû à sa flexibilité quant à l'introduction des covariables ou encore de la corrélation spatiale. Ce modèle produit des estimations du risque pour chaque unité géographique i et si l'on est intéressé par une classification de ces estimations du risque, une méthode de classification standard est utilisée *a posteriori* pour l'obtenir. Étant donnés les enjeux en épidémiologie animale et notre intérêt pour la classification dans ce contexte, nous nous orientons vers les modèles à classes latentes (modèles de mélange) et nous choisirons dans ce travail de développer un modèle à base de champs de Markov cachés discrets. L'objectif est de mettre en place une méthodologie différente des autres approches présentées ici où la classification et l'estimation du risque font partie intégrante de la procédure.

CHAPITRE 4

Champs aléatoires de Markov cachés

Sommaire

Les enjeux principaux de la cartographie du risque en épidémiologie animale sont, comme précisé dans le chapitre 3, l'obtention d'une classification en niveaux de risque et une estimation raisonnable de ces niveaux pour chaque classe (ou groupe) de risque à l'aide d'un modèle statistique facile à interpréter et à mettre en oeuvre par les épidémiologistes mais qui prenne en compte toute la complexité des phénomènes étudiés.
Nous abordons dans ce travail la cartographie du risque comme un problème de classification. L'objectif principal de la classification est de grouper des unités en les associant à des classes selon un ou plusieurs critères.
Les modèles de champs de Markov permettent de prendre en compte la proximité spatiale et de pondérer la portée de l'influence de ces dépendances spatiales. Ces modèles permettent de décrire le phénomène étudié par ses caractéristiques locales, plutôt que globales.

Le but de ce chapitre est de donner une vue générale des champs de Markov cachés. Après leur présentation générale à la section 4.1, nous montrons la mise en œuvre de ces modèles pour la classification à la section 4.2. Enfin, nous décrivons des procédures d'estimation des paramètres d'un champ de Markov à la section 4.3 et le critère BIC pour la sélection de modèles à la section 4.4.

4.1 Champ de Markov et distribution de Gibbs

Les modèles markoviens sont des modèles largement utilisés dans de nombreux domaines comme la physique statistique, l'apprentissage machine, le traitement du signal et l'analyse d'image (Besag, 1974; Geman and Geman, 1984). Ces modèles permettent une modélisation explicite des dépendances entre les individus considérés *via* l'utilisation d'une structure de voisinage ou d'un graphe d'interactions. Considérons un ensemble S de sites spatialement organisés et indicés par $\{1, \ldots, N\}$ sur lequel un système de voisinage est défini. Puisque nous nous intéressons à des applications en épidémiologie, les éléments de S seront intérprétés comme des unités géographiques. La première question à se poser avant de définir un champ de Markov est alors, à partir d'un ensemble d'individus N, comment définir le système de voisinage à leur associer.
Nous définissons en section 4.1.1 les notions relatives au système de voisinage. Nous présentons ensuite, les notions élémentaires liées au champ de Markov cachés en section 4.1.2. À la section 4.1.3, nous donnons les exemples des champs de Markov cachés les plus utilisés et enfin nous présentons en section 4.1.4 la simulation de ces champs.

4.1.1 Système de voisinage

La définition d'un champ de Markov repose sur celle d'un système de voisinage. Ce système de voisinage peut être vu comme un graphe G reliant les individus $i \in S$ (considérés comme des *sommets*), par des branches ou *arêtes* : si j est voisin de i (c'est à dire si $j \in \mathcal{N}_i$, avec $\mathcal{N}_i$ l'ensemble des voisins de i), alors une arête relie i à j. En général, le système de voisinage est supposé symétrique :

$$j \in \mathcal{N}_i \Leftrightarrow i \in \mathcal{N}_j.$$

Cela signifie que les arêtes du graphe G sont non-orientées. Pour simplifier la notation nous noterons, parfois, $i \sim j$ si j est voisin de i. Il est également possible de définir un champ

de Markov à partir d'un système de voisinage non symétrique (dans ce cas, le graphe G sera orienté). Mais cette asymétrie entraîne des complications supplémentaires sur le modèle, même pour des graphes très simples. Nous n'utilisons dans ce travail que des systèmes de voisinage symétriques. La présence (respectivement l'absence) d'une arête entre deux noeuds de G représente une interaction (ou non) entre les éléments de S correspondants. Une relation de voisinage statue quelle(s) dépendance(s) existe(nt) entre les éléments du graphe. La définition d'un système de voisinage est une des premières questions que l'on se pose quand on est on présence d'un ensemble d'individus et que l'on veut construire un champ de Markov. Deux systèmes de voisinage souvent utilisés lorsque S est une grille régulière sont illustrés sur la figure 4.1 qui sont de premier ordre (figure 4.1 (a)) et de second ordre (figure 4.1 (b)). Pour le voisinage de premier ordre, chaque individu est relié à ses deux voisins horizontaux et à ses deux voisins verticaux. En ce qui concerne le voisinage de second ordre, chaque individu est relié à ses quatres voisins horizontaux et à ses quatres voisins diagonaux. Quant au système de voisinage illustré sur la figure 4.1 (c), il est propre à notre transformation en grille régulière d'hexagones de la structure du territoire Français (voir section 6.1). Il existe deux grandes

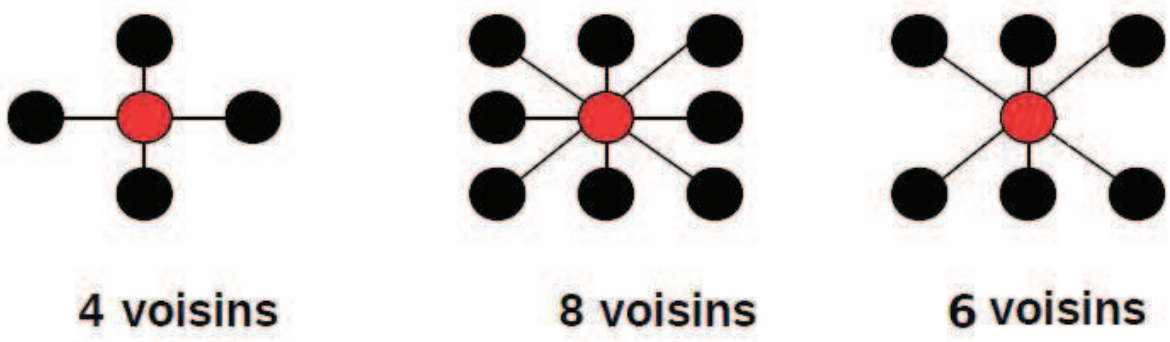


FIGURE 4.1 – Différentes relations de voisinage : (a) Relation de 4-connexité (voisinage de premier ordre), (b) Relation de 8-connexité (voisinage de second ordre), (c) : Système des 6-plus proches voisins.

familles de graphes :

– *Graphes de proximité* : ce sont des graphes construits à partir d'une métrique : deux sites sont reliés l'un à l'autre s'ils sont "suffisamment proches". Cette proximité peut être fonction de la disposition spatiale des individus, ou d'une certaine similarité entre ces individus. Il existe de nombreuses manières de construire un graphe de proximité à partir d'une métrique. Nous citons ci-aprés les graphes de cette famille les plus utilisés :

  1. **Graphe des k-voisins réciproques.** Pour chaque individu, on cherche ses k plus proches voisins (au sens d'une distance à définir) et on ne retient que les voisins ré-

ciproques afin d'avoir un graphe symétrique. C'est en particulier le type de graphe utilisé dans le cas des grilles régulières. Le voisinage de premier ordre illustré sur la figure 4.1 est un graphe des 4-plus proches voisins et et celui de second ordre est un graphe des 8-plus proches voisins.

2. **Triangulation de Delaunay.** Une telle tesselation est obtenue comme duale du diagramme de Voronoi. Une cellule de Voronoi associée à un site i de S est composée de l'ensemble des points qui sont plus proches de i que de tout autre point de S. Deux points i et j créent alors une arête dans le graphe de Delaunay si et seulement si les régions de Voronoi associées à i et j sont adjacentes.

3. **Graphe de voisinage relatif.** La construction d'un tel graphe est fondée sur la notion de voisins "relativement proches" définie par Lankford (1969). Deux sites i et j de S sont dits "relativement proches" s'il n'existe pas de point plus proche à la fois de i et j, c'est à dire si :

$$d(i,j) \leq \min_{s \in S \setminus \{i,j\}} \max\{d(i,s), d(j,s)\}$$

où d est une distance (euclidienne en général) et $S\setminus\{i,j\}$ est l'ensemble des sites de S privé des sites i et j. Le graphe de voisinage relatif (Toussaint, 1980) est alors obtenu en reliant les points "relativement proches".

– *Graphes aléatoires* : ces graphes furent initialement introduits par Erdos and Rényi (1959) et sont construits à partir d'une loi de probabilité. Le modèle de graphe aléatoire le plus étudié est celui noté G_{np} composé de n sommets, chaque paire de sommets étant reliée par une arête avec une probabilité p.

En épidémiologie, le système de voisinage le plus utilisé est le système le plus simple reliant les unités géographiques contigues. C'est un graphe de proximité. Un système de voisinage donné définit un ensemble $\mathcal{C}$ de parties de S appelées *cliques*. Une clique est définie comme étant soit un singleton, soit un ensemble de sites deux à deux voisins. On appelle *ordre* d'une clique le nombre de ses éléments. La figure 4.2 représente les cliques associées au voisinage de premier ordre qui sont des cliques d'ordre 1 ou 2 et celles associées au voisinage de second ordre où des cliques d'ordre 3 et 4 interviennent en plus.

4.1.2 Définition d'un champ de Markov

Considérons un système de variables aléatoires $\mathbf{Z} = \{Z_1, \ldots, Z_N\}$ définies sur S. Soit $\mathcal{N}_i$ l'ensemble des sites voisins du site i. De manière générale, notons $\mathbf{Z}_A$ l'ensemble des

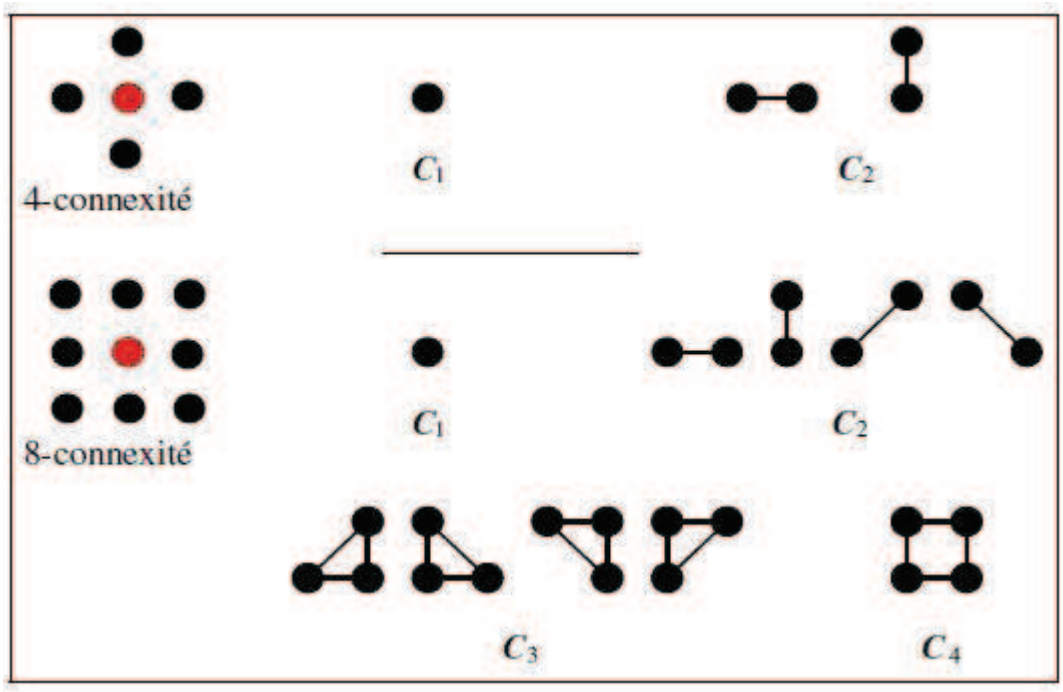


FIGURE 4.2 – Les cliques associées à deux systèmes de voisinage en dimension 2.

variables $\{Z_i, i \in A\}$ pour un sous-ensemble A quelconque de S.

Definition 1 *Un champ aléatoire* $\mathbf{Z}$ *est un champ de Markov sur* S *si la distribution jointe* $P(\mathbf{z})$ *vérifie les deux propriétés suivantes :*

$$\forall \mathbf{z}, \forall i, \quad P(Z_i = z_i | \mathbf{Z}_{S\setminus\{i\}} = \mathbf{z}_{S\setminus\{i\}}) = P(Z_i = z_i | \mathbf{Z}_{\mathcal{N}_i} = \mathbf{z}_{\mathcal{N}_i}) \tag{4.1}$$

$$\forall \mathbf{z}, \quad P(\mathbf{Z} = \mathbf{z}) > 0. \tag{4.2}$$

La propriété markovienne donnée par l'équation (4.1) traduit le fait que l'influence des autres sites de S sur un site i se réduit à l'influence de son voisinage $\mathcal{N}_i$. Il s'agit d'une extension de la notion de chaîne de Markov. Les caractéristiques locales définies par la formule (4.1) sont en général faciles à calculer. Dès lors que la condition de positivité donnée par l'équation (4.2) est vérifiée, la distribution jointe $P(\mathbf{z})$ est définie de manière unique à partir des caractéristiques locales. Cette condition (4.2) de positivité signifie que toute configuration doit avoir une probabilité non nulle d'occurence. Elle est suffisante pour assurer la consistance de $P(\mathbf{z})$ mais non nécessaire et peut donc être relâchée (Lauritzen et al., 1990).
La formulation (4.1), ne permet pas en général d'avoir une expression simple de la distribu-

tion jointe. On préfère caractériser le champ de Markov par une distribution jointe, ce qui est rendu possible grâce au théorème d'Hammersley-Clifford (Besag, 1974) :

Definition 2 *Soit G un graphe de voisinage et C un ensemble de cliques associées à ce graphe. Un champ aléatoire $\mathbf{Z}$ est régi par une distribution de Gibbs si la distribution de probabilité jointe est de la forme :*

$$P_G(\mathbf{z}) = W^{-1} \exp\left(-H(\mathbf{z})\right) \tag{4.3}$$

où la fonction énergie H, définie à une constante près, se décompose en une somme de fonctions potentiels V_c associées aux cliques $c \in \mathcal{C}$,

$$H(\mathbf{z}) = \sum_{c \in \mathcal{C}} V_c(\mathbf{z}_c) \tag{4.4}$$

et $W = \sum_{\mathbf{z}} \exp\left(-H(\mathbf{z})\right)$ est la constante de normalisation, appelée aussi fonction de partition.

La constante de normalisation W se calcule différemment dans les cas discret ou continu :

$$\text{cas discret :} \quad W = \sum_{z'} \exp\left(-H(\mathbf{z}')\right)$$

$$\text{cas continu :} \quad W = \int_{z'} \exp\left(-H(\mathbf{z}')\right) d\mathbf{z}'$$

Théorème 1 *(**Hammersley-Clifford**) Le champ aléatoire $\mathbf{Z}$ est un champ de Markov si et seulement si sa loi jointe $P_G(\mathbf{z})$ est une distribution de Gibbs.*

À l'origine, les distributions de Gibbs étaient utilisées en physique statistique pour modéliser le comportement de systèmes moléculaires en interaction, d'où les termes "potentiel" ou encore "énergie". Elles représentent les systèmes les plus désordonnés au sens où, parmi les distributions de probabilité P telles que l'espérance de l'énergie $\mathbb{E}_p[H(\mathbf{Z})]$ est fixée, la distribution de l'équation (4.3) est celle qui maximise l'entropie :

$$\xi(P) = -\sum_{\mathbf{z}} P(\mathbf{z}) \log P(\mathbf{z}).$$

Notons qu'un champ de Markov peut être discret ou continu. Dans ce travail, nous nous intéressons au cas où le champ $\mathbf{Z}$ est discret. La notation P_G sera reprise par la suite pour

désigner toute distribution de probabilité vérifiant la définition 2. L'équivalence champ de Markov/distribution de Gibbs permet d'avoir une formule explicite de la distribution jointe et donc une description globale du modèle.
Grâce aux fonctions de potentiels, il est possible de spécifier les dépendances spatiales locales du champ de Markov. Selon les modèles, les potentiels sont plus ou moins complexes, ce qui influe sur la richesse des caractérisations locales mais aussi sur la facilité à manipuler le modèle. Les probabilités conditionnelles (4.1) se retrouvent directement à partir de (4.3) :

$$\forall \mathbf{z}, \quad P_G(z_i|\mathbf{Z}_{\mathcal{N}_i} = \mathbf{z}_{\mathcal{N}_i}) = \frac{\exp\left(-\sum_{i\in c} V_c(\mathbf{z}_c)\right)}{\sum_{z_i'} \exp\left(-\sum_{i\in c} V_c(\mathbf{z}_c')\right)}, \tag{4.5}$$

avec $\mathbf{z}_c^{'} = \{z_i^{'}\} \cup \{z_j, j \in c, j \neq i\}$. Si ces probabilités sont facilement calculables de manière analytique, ce n'est pas le cas de la distribution de Gibbs (4.3), même si l'on dispose d'une formule explicite. En effet, son calcul requiert l'évaluation de la fonction de partition W. Lorsque la taille de S est trop importante, le nombre de configurations possibles pour $\mathbf{z}$ explose, ce qui rend le calcul de W impossible de manière directe.

4.1.3 Exemples de champs de Markov

Nous décrivons ici quelques exemples de structures possibles pour un champ de Markov. Il est important de noter qu'à travers un modèle de champ de Markov on cherche à décrire les caractéristiques locales du champ, plutôt que globales. Il s'agit de traduire une connaissance ou des contraintes *a priori*. En épidémiologie, le champ représente la vraie carte du risque que l'on cherche à restaurer.
Les modèles présentés ci-dessous correspondent ainsi à différents choix de caractéristiques locales.

4.1.3.1 Modèle d'Ising

Le modèle markovien le plus simple est le modèle d'Ising (Ising, 1925), issu de la mécanique statistique. Il correspond au cas où les variables Z_i ne peuvent prendre que deux valeurs, $+1$ ou -1. L'énergie du champ $\mathbf{Z}$ est donnée par :

$$H(\mathbf{z}) = h \sum_i z_i - J \sum_{i\sim j} z_i z_j$$

En mécanique statistique, cette énergie modélise l'interaction ferromagnétique entre *s*pins voisins, les *s*pins pouvant être orientés vers le haut ($z_i = +1$) ou vers le bas ($z_i = -1$). Le paramètre J témoigne du caractère ferromagnétique ($J > 0$) ou anti-ferromagnétique ($J < 0$) du modèle. Ce modèle est utilisé également en sciences économiques pour modéliser des systèmes de coopération ou non entre individus (Jean-Pierre and Mina, 2005).

4.1.3.2 Modèle de Potts et extensions

Dans certaines applications, la variable Z_i représente une étiquette, parmi K possibles, attribuée au site i. Les K valeurs possibles de Z_i sont appelées les classes (ou étiquettes). Largement développés en analyse d'image, ces modèles sont en général définis à partir de leur fonction d'énergie H donnée par l'équation (4.4) qui peut être décomposée en une somme sur des cliques de différentes tailles :

$$H(\mathbf{z}) = \sum_{i \in S} V_i(z_i) + \sum_{\substack{i,j \\ voisins}} V_{ij}(z_i, z_j) + \ldots + \sum_{\substack{i_1,\ldots,i_q \\ voisins}} V_{i_1,\ldots,i_q}(z_{i_1}, \ldots, z_{i_q}) \tag{4.6}$$

Les potentiels sur les cliques d'ordre 1 et 2 sont considérés, dans la plupart des cas, comme suffisants pour modéliser les dépendances spatiales. On fixe le plus souvent à zéro les potentiels sur les cliques d'ordre supérieur à 2, et l'énergie donnée par l'équation (4.6) est alors réduite à l'énergie suivante :

$$H(\mathbf{z}) = \sum_{i \in S} V_i(z_i) + \sum_{i \sim j} V_{i,j}(z_i, z_j) \tag{4.7}$$

Potentiels sur les singletons. Les potentiels sur les singletons (les cliques d'ordre 1) $V_i(z_i)$ permettent de modéliser la probabilité d'occurence de la classe z_i au site i considérée individuellement. En mécanique statistique, ces potentiels représentent l'influence du champ magnétique externe. Lorsque les potentiels $V_i(z_i)$ dépendent de i (et non seulement de z_i), on parle de champ externe non-stationnaire. Notons que sous l'hypothèse de non-stationnarité et sans paramétrisation particulière, l'estimation des potentiels sur les singletons est impossible puisqu'il faudrait estimer K potentiels V_i par site i. Néanmoins, de tels potentiels non stationnaires peuvent être intéressants pour intégrer de l'information *a priori* visant à influencer les sites individuellement (Scherrer et al., 2007).
En l'absence de connaissance particulière, l'hypothèse classique consiste à supposer que ces fonctions de potentiels sont les mêmes sur l'ensemble des sites, c'est-à-dire que $V_i(z_i)$ ne dépend du site i qu'à travers la valeur de z_i. Cette hypothèse correspond à un champ externe spatialement stationnaire et peut se traduire par :

$$V_i(z_i) = -\alpha_{z_i}. \tag{4.8}$$

Les fonctions potentiels sur les singletons sont alors caractérisées par le vecteur des poids $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)$ associés aux K classes. On adopte la notation vectorielle $\mathbf{z}_i = e_{z_i}$ où $(e_1, \ldots, e_K)$ désigne la base canonique et $\mathbf{z}_i^t$ est la transposée du vecteur $\mathbf{z}_i$, une écriture équivalente à (4.8) est :

$$V_i(z_i) = -z_i^t \alpha.$$

Si les potentiels sur les cliques de taille supérieure à 1 sont nuls, la distribution (4.3) s'écrit :

$$P_G(\mathbf{z}) = \frac{\exp\left(\sum_i \alpha_{z_i}\right)}{\sum_{z_1^t} \cdots \sum_{z_N^t} \exp\left(\sum_i \alpha_{z_i^t}\right)} = \prod_{i \in S} \frac{\exp\left(\alpha_{z_i}\right)}{\sum_{z_i^t} \exp\left(\alpha_{z_i^t}\right)}. \tag{4.9}$$

La loi jointe se décompose alors en produit de fonctions de z_i qui peuvent s'interpréter comme la probabilité d'occurence de la classe z_i au point i. Pour un site i quelconque, la probabilité de se trouver dans la classe k vaut alors :

$$\pi_k = \frac{e^{\alpha_k}}{\sum_{k'=1}^{K} e^{\alpha_{k'}}}. \tag{4.10}$$

Notons que les fonctions potentiels sont équivalentes à une constante additive près car l'ajout d'une constante ne modifie pas la distribution de Gibbs correspondante. Pour définir les potentiels sur les singletons de manière unique, on peut par exemple imposer $\sum_k \alpha_k = 0$.

Potentiels sur les paires. Les potentiels sur les paires $V_{ij}(z_i, z_j)$ permettent de modéliser la dépendance entre les classes Z_i et Z_j en des sites i et j voisins. En général, on suppose que ces fonctions de potentiels sont les mêmes sur l'ensemble des sites, ce qui peut se traduire par la notation :

$$V_{ij}(z_i, z_j) = V(z_i, z_j) = -b_{z_i, z_j}. \tag{4.11}$$

Les fonctions potentiels sur les paires sont caractérisées alors par la matrice symétrique $\mathbb{B} = (b_{kk'})_{k,k' \in \{1,\ldots,K\}}$ associée aux $K \times K$ interactions entre classes. Le terme $b_{kk'}$ peut s'interpréter comme le degré de compatibilité entre les classes k et k'. Les fonctions potentiels étant équivalentes à une constante additive près, pour définir les potentiels sur les paires de manière unique, on peut par exemple imposer $b_{11} = 0$. Une hypothèse consiste à supposer de plus que la matrice $\mathbb{B}$ s'écrit $\mathbb{B} = bI_K$ où I_K désigne la matrice identité de dimension $K \times K$. L'énergie correspondante (si les potentiels sur les singletons sont nuls) est donnée par :

$$H(\mathbf{z}) = -\sum_{i \sim j} \mathbf{z}_i{}^t \mathbb{B} \mathbf{z}_j = -b \sum_{i \sim j} 1_{z_i = z_j} = -bO(\mathbf{z})$$

où $O(\mathbf{z})$ désigne le nombre de paires homogènes (qui sont dans la même classe) pour la classification $\mathbf{z}$. Une telle distribution correspond à la distribution de Strauss avec classes interchangeables (Strauss, 1977), appelée modèle de Potts. Elle est largement utilisée en segmentation

markovienne d'image (Besag, 1986) car elle traduit de la facon la plus simple possible l'hypothèse de régularité spatiale. En effet, plus le paramètre $b > 0$ est grand, plus la probabilité que deux sites voisins i et j soient dans la même classe est élevée. À noter que lorsque $b = 0$, les classes $Z_1, \ldots, Z_N$ sont indépendantes les unes des autres. Si de plus, les potentiels sur les singletons sont nuls, toutes les classes sont équiprobables.

4.1.3.3 Auto-modèles

Les auto-modèles (Besag, 1974) sont des modèles markoviens largement utilisés dans différents domaines d'application (qui vont de l'analyse d'image à l'épidémiologie). L'énergie y est définie sur les cliques d'orde 1 et 2 uniquement. Le modèle auto-logistique qu'on a présenté en section 3.2.3.1 est un exemple de ces auto-modèles. Il existe, comme précisé auparavant, différentes variations sur les auto-modèles selon la forme donnée aux fonctions potentiels V_i et V_{ij}.

4.1.4 Simuler un champ de Markov

La loi jointe d'un champ de Markov $\mathbf{Z}$ n'étant pas calculable directement, il est fondamental de pouvoir simuler des réalisations du champ $\mathbf{Z}$. La complexité des champs de Markov conduit à utiliser des méthodes de type Monte-Carlo (Robert and Casella, 1999) pour approcher des espérances par rapport à la distribution de Gibbs car le calcul n'est pas possible directement. Il existe des méthodes adaptées au modèle de champ de Markov qui permettent de simuler de manière efficace et simple la distribution de Gibbs. Nous décrivons ici celle qui est basée sur l'échantillonneur de Gibbs.

Échantillonneur de Gibbs. Un exemple de type chaîne de Markov de Monte-Carlo (MCMC) pour la simulation d'un champ de Markov est l'échantillonneur de Gibbs. Nous avons présenté cet algorithme en section 3.1.5.2 dans sa forme générale et décrivons ici sa mise en oeuvre pour la génération d'une chaîne de Markov $\{\mathbf{z}^{(q)}\}_{q\in\mathbb{N}}$ pour laquelle les probabilités de transition sont données par les caractéristiques locales de la distribution d'intérêt $P_G(\mathbf{z})$. À chaque pas un seul site est visité de telle sorte que $\mathbf{z}^{(q-1)}$ et $\mathbf{z}^{(q)}$ ne peuvent différer que par la composante z_{i_q}. Si l'on note $\{i_1, \ldots, i_q\}$ la suite des indices visités, la probabilité de transition de

la chaîne à l'étape (q) est donnée par :

$$P(\mathbf{Z}^{(q)} = \mathbf{z}^{(q)} | \mathbf{Z}^{(q-1)} = \mathbf{z}^{(q-1)}) = \begin{cases} 0 & \text{si } \exists i \neq i_q \text{ tel que } z_i^{(q)} \neq z_i^{(q-1)}, \\ P_G(z_{i_q}^{(q)} | \mathbf{z}_{\mathcal{N}_{i_q}}^{(q-1)}) & \text{sinon.} \end{cases} \tag{4.12}$$

L'ordre de visite des sites le plus simple est l'ordre croissant des indices. Dans ce cas, la chaîne $\{\mathbf{z}^{(q)}\}_{q \in \mathbb{N}}$ est irréductible, apériodique et admet P_G comme unique mesure stationnaire, quelle que soit l'initialisation (Geman and Graffigne, 1987). Cet algorithme peut mettre longtemps à converger du fait qu'un seul site est mis à jour à chaque itération. Le calcul des probabilités de transition est simple car ne nécessite pas de recours à la fonction de partition W ce qui rend facile la mise en pratique de l'algorithme.

4.2 Modèles de champ de Markov cachés pour la classification

Les champs de Markov cachés discrets sont appropriés pour la classification. Rappelons que l'objectif de la classification est de regrouper les individus qui se "ressemblent" dans une même classe. En épidémiologie, par exemple, le but est de regrouper les unités géographiques dont les niveaux de risque sont proches dans la même classe. Les champs de Markov cachés discrets entrent dans le contexte des problèmes à données incomplètes : pour chaque individu i de l'ensemble S, deux données de types différents s'expriment, l'une observée, y_i, l'autre manquante (ou cachée), z_i. Le champ $\mathbf{z} = \{z_i, i \in S\}$ représente une classification des unités i de S en fonction des observations $\mathbf{y} = \{y_i, i \in S\}$. L'objectif est de retrouver des informations sur $\mathbf{z}$ à partir des données observées $\mathbf{y}$. En restauration d'images, $\mathbf{Z}$ représentera l'image originale dont $\mathbf{y}$ est une observation dégradée.
La forme générale d'un modèle de champ de Markov caché utilisé pour représenter les données complètes $(\mathbf{Z}, \mathbf{Y})$ est le suivant : les données cachées sont décrites par un modèle de champ de Markov discret, de distribution de probabilité P_G définie par une énergie H comme dans la définition donnée par l'équation (4.3). Chaque variable Z_i est affectée à un des K états possibles. Les observations $\mathbf{Y} = \{Y_i, i \in S\}$ sont supposées indépendantes conditionnellement à $\mathbf{Z}$, de sorte que la densité conditionnelle $f(\mathbf{y}|\mathbf{z})$ se décompose ainsi :

$$f(\mathbf{y}|\mathbf{z}) = \prod_{i \in S} f(y_i | z_i, \boldsymbol{\theta})$$

où la forme des densités conditionnelles $f(y_i | Z_i = e_k, \boldsymbol{\theta})$ est connue avec un ensemble de paramètres $\boldsymbol{\theta}$.

Un modèle de champ de Markov caché discret peut être vu comme un modèle de mélange fini où la probabilité d'appartenance aux classes est régie par des dépendances complexes entre les classes. Nous présenterons ainsi en premier les modèles de mélanges à la section 4.2.1 avant de présenter les modèles de champ de Markov cachés pour la classification à la section 4.2.2.

4.2.1 Modèle de mélange pour la classification

La description d'un problème de classification dans un contexte de données manquantes a eu pour conséquence de renforcer l'intérêt pour les modèles de mélange en terme de développement méthodologique et de domaines d'application. L'algorithme EM est devenu un outil standard d'estimation des paramètres d'un modèle de mélange. Nous présentons ici les principes du mélange à la section 4.2.1.1. Nous décrivons ensuite le principe général de l'algorithme EM à la section 4.2.1.2 et sa mise en oeuvre pour le modèle simple de mélange indépendant en section 4.2.1.3. Enfin la section 4.2.1.4 est consacrée à la présentation des régles utilisées pour la classification à partir des résultats obtenus par l'algorithme EM dans le cadre des mélanges.

4.2.1.1 Distributions de mélange

Le modèle de mélange fini de lois de probabilité consiste à supposer que les données proviennent d'une source contenant plusieurs sous-populations. On considère alors que l'échantillon des données observées provient d'un nombre de groupes inconnus *a priori* qu'il faut retrouver. Cela permet ainsi de modéliser des données complexes dont la distribution ne peut pas être l'archétype d'une loi classique, mais plutôt un mélange de distributions.
Nous nous restreindrons dans nos applications aux mélanges de lois de Poisson. Cependant, dans ce chapitre nous ferons une présentation générale des modèles de mélange.
Un modèle de mélange (McLachlan and Peel, 2000) se base sur deux hypothèses :

– La classification $\mathbf{z}$ est tirée pour chacune de ses composantes z_i de façon indépendante selon une loi multinomiale dont les paramètres sont les proportions du mélange :

$$P(\mathbf{Z}, \boldsymbol{\phi}) = \prod_{i \in S} P(Z_i, \boldsymbol{\phi}),$$

où $\boldsymbol{\phi}$ est l'ensemble des paramètres associés à la loi de $\mathbf{Z}$. Les paramètres $\boldsymbol{\phi}$ se limitent aux proportions du mélange c'est à dire aux probabilités d'appartenance à une

des classes du mélange. Donc $\boldsymbol{\phi} = (\pi_1, \ldots, \pi_K)$ avec $\pi_k \in [0,1]$, $\forall k \in \{1, \ldots, K\}$ et $\sum_{k=1}^{K} \pi_k = 1$. Une fois cette classification fixée,

– Chaque y_i est indépendamment des autres, issu d'une loi paramétrée par $\theta_{z_i} = \theta_k$ si $z_i = k$. On note $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K)$ les paramètres de classe.

Donc l'ensemble des paramètres de ce modèle de mélange est $\boldsymbol{\Psi} = (\pi_1, \ldots, \pi_K, \theta_1, \ldots, \theta_K)$
La loi marginale (ou encore la vraisemblance des paramètres $\boldsymbol{\Psi}$) pour $\mathbf{Y}$ est donnée par :

$$P(\mathbf{y}|\boldsymbol{\Psi}) = \prod_{i \in S} P(y_i|\boldsymbol{\Psi}) = \prod_{i \in S} \sum_{k=1}^{K} \pi_k f(y_i|\theta_k). \tag{4.13}$$

4.2.1.2 Prinicipe général de l'algorithme EM

L'algorithme Expectation-Maximisation (EM) (Dempster et al., 1977) est une procédure largement utilisée pour trouver le maximum de vraisemblance des paramètres pour les modèles à données manquantes.
Nous présentons ici le principe de l'algorithme EM dans un cadre général. Le but de cet algorithme itératif est de maximiser la log-vraisemblance en procédant à chaque étape (q) à la maximisation de l'espérance de la log-vraisemblance complète connaissant les observations $\mathbf{y}$ et l'estimation courante (issue de l'étape précédente) des paramètres $\boldsymbol{\Psi}^{(q)}$:

$$\begin{aligned} Q(\boldsymbol{\Psi}|\boldsymbol{\Psi}^{(q)}) &= \mathbb{E}_{\boldsymbol{\Psi}^{(q)}}[\log P(\mathbf{y}, \mathbf{Z}|\boldsymbol{\Psi})|\mathbf{Y} = \mathbf{y}] \\ &= \underbrace{\log P(\mathbf{y}|\boldsymbol{\Psi})}_{=\log L(\boldsymbol{\Psi})} + \underbrace{\mathbb{E}_{\boldsymbol{\Psi}^{(q)}}[\log P(\mathbf{Z}|\mathbf{Y} = \mathbf{y}, \boldsymbol{\Psi})|\mathbf{Y} = \mathbf{y}]}_{:=H(\boldsymbol{\Psi}, \boldsymbol{\Psi}^{(q)})} \end{aligned} \tag{4.14}$$

Une itération de l'algorithme EM se divise en deux étapes :

– **Étape Expectation (E) :** Calcul de l'espérance conditionnelle $Q(\boldsymbol{\Psi}|\boldsymbol{\Psi}^{(q)})$ qui revient au calcul des probabilités *a posteriori* $P(Z_i = k|\mathbf{Y} = \mathbf{y}, \boldsymbol{\Psi}^{(q)})$.

– **Étape Maximisation (M) :** Mise à jour de $\boldsymbol{\Psi}^{(q)}$ en $\boldsymbol{\Psi}^{(q+1)}$ en résolvant :

$$\boldsymbol{\Psi}^{(q+1)} = \underset{\boldsymbol{\Psi}}{\operatorname{argmax}}\, Q(\boldsymbol{\Psi}|\boldsymbol{\Psi}^{(q)}). \tag{4.15}$$

Ces deux itérations sont répétées jusqu'à l'atteinte d'un critère d'arrêt. Celui-ci peut consister soit en un nombre d'itérations que l'on estime suffisant pour la convergence de l'algorithme soit en un seuil sous lequel les différences entre deux étapes des paramètres du modèle, ou de la vraisemblance ne sont pas considérés significatifs. Sous des conditions suffisantes de régularité Wu (1983), l'estimateur obtenu converge vers un maximum local de la vraisemblance.
La propriété fondamentale de l'algorithme EM est que la fonction de vraisemblance $L(\boldsymbol{\Psi}^{(\mathbf{q})})$ est croissante. En effet :

$$\log L(\boldsymbol{\Psi}) = Q(\boldsymbol{\Psi}, \boldsymbol{\Psi}^{(q)}) - H(\boldsymbol{\Psi}, \boldsymbol{\Psi}^{(q)}).$$

Or, on montre que la fonction $H(\mathbf{\Psi}, \mathbf{\Psi}^{(q)})$ atteint son maximum en $\mathbf{\Psi}^{(q)}$:

$$\begin{aligned} H(\mathbf{\Psi}, \mathbf{\Psi}^{(q)}) - H(\mathbf{\Psi}^{(q)}, \mathbf{\Psi}^{(q)}) &= \mathbb{E}_{\mathbf{\Psi}^{(q)}}[\log \frac{P(\mathbf{Z}|\mathbf{Y}, \mathbf{\Psi})}{P(\mathbf{Z}|\mathbf{Y}, \mathbf{\Psi}^{(q)})}|Y] \\ &\leq \log \mathbb{E}_{\mathbf{\Psi}^{(q)}}[\log \frac{P(\mathbf{Z}|\mathbf{Y}, \mathbf{\Psi})}{P(\mathbf{Z}|\mathbf{Y}, \mathbf{\Psi}^{(q)})}|Y] \\ &= \log \int P(\mathbf{Z}|\mathbf{Y}, \mathbf{\Psi}).dZ \\ &= \log 1 = 0. \end{aligned}$$

Comme $\mathbf{\Psi}^{(q+1)}$ est choisi de façon à maximiser $Q(\mathbf{\Psi}, \mathbf{\Psi}^{(q)})$ d'où :

$$\begin{aligned} \log L(\mathbf{\Psi}^{(q+1)}) &= Q(\mathbf{\Psi}^{(q+1)}, \mathbf{\Psi}^{(q)}) - H(\mathbf{\Psi}^{(q+1)}, \mathbf{\Psi}^{(q)}) \\ &\geq Q(\mathbf{\Psi}^{(q)}, \mathbf{\Psi}^{(q)}) - H(\mathbf{\Psi}^{(q)}, \mathbf{\Psi}^{(q)}) = \log L(\mathbf{\Psi}^{(\mathbf{q})}). \end{aligned}$$

Donc augmenter la vraisemblance revient à augmenter la fonction Q. Si $\mathbf{\Psi}^{(q+1)}$ est choisi de façon à augmenter localement (donc en n'étant pas nécessairement le jeu de paramètres conduisant au maximum global), on parle d'algorithmes EM généralisés (*Generalised* EM Dempster et al. (1977)).

4.2.1.3 Algorithme EM et modèle de mélange

L'algorithme EM trouve sa plus importante contribution dans le contexte de l'estimation des paramètres d'un modèle de mélange. L'intérêt croissant pour ces modèles et la description des problèmes comme des problèmes à données manquantes ont permis d'utiliser l'algorithme EM qui est devenu un outil standard pour l'estimation des paramètres de ces modèles.
À l'itération (q), l'espérance conditionnelle Q donnée par l'équation (4.14) s'écrit comme suit :

$$Q(\mathbf{\Psi}|\mathbf{\Psi}^{(q)}) = \sum_{i \in S} \sum_{k=1}^{K} t_{ik}^{(q)} \log (\pi_k f(y_i|\theta_k)), \tag{4.16}$$

où le coefficient $t_{ik}^{(q)}$ désigne la probabilité *a posteriori* d'appartenance du site i à la classe k, $P(Z_i = k|y_i, \mathbf{\Psi}^{(q)})$.
Les deux étapes de l'algorithme EM décrites en section 4.2.1.2 deviennent alors :

– **Étape E :** Calcul des probabilités *a posteriori* :

$$\forall i \in S, \forall k \in \{1, \ldots, K\}, t_{ik}^{(q)} = \frac{\pi_k^{(q-1)} f(y_i|\theta_k^{(q-1)})}{\sum_{l=1}^{K} \pi_l^{(q-1)} f(y_i|\theta_l^{(q-1)})}, \tag{4.17}$$

– **Étape M :** Mise à jour des paramètres $\mathbf{\Psi}$:

$$\pi_k^{(q+1)} = \frac{\sum_{i \in S} t_{ik}^{(q)}}{N} \tag{4.18}$$

$$\boldsymbol{\theta}^{(q+1)} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \sum_{i \in S} \sum_{k=1}^{K} t_{ik}^{(q)} \log \left(f(y_i|\theta_k) \right). \tag{4.19}$$

L'algorithme EM allie, dans la plupart des cas, simplicité de mise en œuvre et efficacité. Néanmoins, quelques cas problèmatiques ont donné lieu à des développements complèmentaires. Parmi les variantes existantes nous évoquons l'algorithme Classification EM (CEM) permettant de prendre en compte l'aspect classification lors de l'estimation, ainsi que l'algorithme Stochastic EM (SEM) dont l'objectif est de réduire le risque de tomber dans un optimum local de vraisemblance.

Algorithme CEM. L'algorithme CEM proposé par Celeux and Govaert (1992) est un algorithme itératif qui donne simultanément les paramètres et la classification. Il maximise, par rapport à $\mathbf{z} = \{z_1, \ldots, z_N\}$ et les paramètres $\mathbf{\Psi}$, le critère de vraisemblance classifiante :

$$CL(\mathbf{z}|\mathbf{\Psi}) = \sum_{i \in S} \sum_{k=1}^{K} \mathbb{1}_{\{z_i = k\}} \log \pi_k f(y_i|\theta_k),$$

Cette procédure insère après l'étape (**E**) de l'algorithme EM standard, une étape (**C**) de classification. Les trois étapes de CEM sont alors :

– **Étape (E) :** Calcul des probabilités *a posteriori* $t_{ik}^{(q)}$ selon l'équation (4.17). Cette étape reste identique à celle du EM standard.
– **Étape (C) :** Affectation de chaque observation y_i à la classe $z_i^{(q)}$ qui maximise $t_{ik}^{(q)}$.
– **Étape (M) :** Mise à jour des paramètres $\mathbf{\Psi}$ en maximisant $CL(\mathbf{z}^{(q)}|\mathbf{\Psi})$.

Algorithme SEM. Afin de réduire le risque de tomber dans un maximum local de vraisemblance, Celeux and Diebolt (1985) proposent d'intercaler une étape stochastique **(S)** entre les étapes **(E)** et **(M)**. La forme de cet algorithme est la suivante :
– **Étape (E) :** Calcul des probabilités *a posteriori* $t_{ik}^{(q)}$ selon l'équation (4.17).
– **Étape (S) :** Les $z_i^{(q)}$ sont tirés aléatoirement dans $\{1, \ldots, K\}$ selon une loi multinomiale $\mathcal{M}(1, t_{i1}^{(q)}, \ldots, t_{ik}^{(q)})$.
– **Étape (M) :** Mise à jour des paramètres en maximisant $CL(\mathbf{z}^{(q)}|\mathbf{\Psi})$.

4.2.1.4 Classification avec l'algorithme EM

L'algorithme EM est à l'origine un algorithme d'estimation de paramètres. On peut toutefois en faire un algorithme de classification. La configuration recherchée $\mathbf{z}$ peut être obtenue à

partir des paramètres estimés. Pour cela, il faut définir une fonction qui indique le coût d'une classification $\mathbf{z}$ lorsque la vraie classification est $\mathbf{z}^*$. Nous décrivons ici deux stratégies associées à des coûts différents, définissant les estimateurs les plus utilisés dans la littérature : l'estimateur du maximum a posteriori (MAP) et l'estimateur du maximum des probabilités marginales (MPM).

La règle du MAP. La fonction de coût la plus simple et la plus utilisée est appelée coût 0-1. Elle associe un coût 0 à la bonne classification, 1 à la mauvaise :

$$c(\mathbf{z}, \mathbf{z}^*) = \mathbb{1}_{\mathbf{z} \neq \mathbf{z}^*} = 1 - \mathbb{1}_{\mathbf{z}=\mathbf{z}^*} = \begin{cases} 1 \text{ si } \exists i \text{ tel que } z_i \neq z_i^*, \\ 0 \text{ sinon.} \end{cases}$$

Le coût conditionnel est alors :

$$\bar{c}(\mathbf{z}|\mathbf{y}) = \sum_{\mathbf{z}^*} c(\mathbf{z}, \mathbf{z}^*) P(\mathbf{z}^*|\mathbf{y}) = 1 - P(\mathbf{z}|\mathbf{y}) = P(\mathbf{Z} \neq \mathbf{z}|\mathbf{y}).$$

C'est la probabilité de se tromper en choisissant la classification $\mathbf{z}$ au lieu de la vraie classification. La régle du MAP revient alors à choisir la configuration la plus probable conditionnellement aux données :

$$\mathbf{z}^{map} = \underset{\mathbf{z}}{\operatorname{argmin}}\, \bar{c}(\mathbf{z}|\mathbf{y}) = \underset{\mathbf{z}}{\operatorname{argmax}}\, P(\mathbf{z}|\mathbf{y}).$$

La règle du MPM. La règle du MAP est une règle globale qui associe le même coût $c(\mathbf{z}, \mathbf{z}^*)$ à une configuration différant en un seul site de la vraie classification qu'à une configuration quelconque. Il est plus intéressant de définir la fonction de coût comme une somme de coûts locaux :

$$c(\mathbf{z}, \mathbf{z}^*) = \sum_{i \in S} \mathbb{1}_{z_i \neq z_i^*}.$$

La règle de classification MPM consiste alors à choisir pour le site i la classe :

$$z_i^{mpm} = \underset{z_i}{\operatorname{argmax}}\, P(z_i|\mathbf{y}). \tag{4.20}$$

L'estimateur $\mathbf{z}^{mpm} = (z_1^{mpm}, \ldots, z_n^{mpm})$ correspondant porte le nom d'estimateur du maximum de probabilités marginales (MPM) .

Dans le cas d'un mélange, les règles de classement du MAP et du MPM sont équivalentes et elles s'écrivent comme suit :

$$\forall i \in S, z_i^{mpm} = z_i^{map} = \underset{z_i}{\operatorname{argmax}}\, P(z_i|y_i, \psi_{z_i}) = \underset{k}{\operatorname{argmax}}\, t_{ik}.$$

Les valeurs des t_{ik} qui sont obtenues en sortie de l'algorithme EM permettent donc de restaurer directement la configuration recherchée $\mathbf{z}$.

4.2.2 Classification des variables dépendantes par un modèle de champ de Markov caché discret

Le modèle de champ de Markov caché discret peut être interprété comme la généralisation d'un modèle de mélange fini pour lequel la classe d'une observation n'est plus indépendante (comme dans le cas des mélanges indépendants) de la classe des autres observations. On revient alors à la notion d'individus organisés spatialement et à la notion de voisinage, qui n'existent pas dans le modèle de mélange fini indépendant. Nous présentons dans cette partie la mise en pratique des champs de Markov cachés pour la classification en section 4.2.2.1 et donnons en section 4.2.2.2 les limites de l'algorithme EM standard pour ce modèle.

4.2.2.1 Modèle de champ de Markov caché

Les champs de Markov cachés permettent de prendre en compte les dépendances spatiales entre sites voisins dans la distribution de $\mathbf{Z}$. Dans ce modèle, les classes sont distribuées selon le champ de Markov $\mathbf{Z}$ défini par les équations (4.1 ou 4.3). Les observations $\mathbf{Y}$ sont alors indépendantes conditionnellement à $\mathbf{Z}$. Leur loi est paramétrée par $\boldsymbol{\theta}$:

$$P(\mathbf{Y}|\mathbf{Z}) = \prod_{i \in S} P(Y_i|Z_i, \theta_{Z_i})$$

La loi marginale d'une observation se décompose alors selon les K classes :

$$P(Y_i|\boldsymbol{\Psi}) = \sum_{k=1}^{K} P_G(Z_i = k|\boldsymbol{\phi}) P(Y_i|Z_i = k, \theta_k)$$

Nous avons donc une formule analogue à celle des modèles de mélange où les proportions π_k sont remplacées par les probabilités gibbsiennes $P_G(Z_i = k|\boldsymbol{\phi})$. La classe d'une observation n'est plus indépendante des autres observations ni du site considéré.
On note que la probabilité *a posteriori* s'exprime toujours grâce à la formule d'inversion de Bayes :

$$P(\mathbf{Z}|\mathbf{Y}, \boldsymbol{\Psi}) = \frac{P(\mathbf{Z}|\boldsymbol{\phi}) P(\mathbf{Y}|\mathbf{Z}, \boldsymbol{\theta})}{P(\mathbf{Y}|\boldsymbol{\theta})}$$

Le terme au numérateur ne dépend que des données. Donc à une constante près :

$$P(\mathbf{Z}|\mathbf{Y}, \boldsymbol{\Psi}) \propto \exp\{-H(\mathbf{Z}|\boldsymbol{\phi}) + \sum_{i \in S} \log P(Y_i|Z_i, \boldsymbol{\theta})\}$$

D'après le théorème d'Hammersley-Clifford, $\mathbf{Z}|\mathbf{Y}$ est aussi distribué selon un champ de Markov dont la fonction d'énergie comprend un terme propre au graphe, qui est la fonction d'énergie du champ $\mathbf{Z}$, et un terme dit d'attache aux données qui est : $-\sum_{i\in S}\log P(Y_i|Z_i,\boldsymbol{\theta})$. On distingue alors le champ de Markov conditionnel $\mathbf{Z}|\mathbf{Y}=\mathbf{y}$ et le champ de Markov marginal $\mathbf{Z}$. Une difficulté majeure lors de l'utilisation des champs de Markov réside dans le calcul des quantités utilisant des distributions marginales ou conditionnelles qui sont spécifiées par les paramètres.

4.2.2.2 Algorithme EM pour champ de Markov caché

Les paramètres $\boldsymbol{\Psi}=\{\boldsymbol{\theta},\boldsymbol{\phi}\}$ du modèle sont en général inconnus et doivent être estimés. Comme dans le cadre du modèle de mélange, l'idée serait d'appliquer l'algorithme EM pour l'estimation des paramètres par maximum de vraisemblance. Cependant, sous modélisation markovienne, cet algorithme ne peut être utilisé sans approximation. En effet, l'itération (q) de l'algorithme EM appliqué au champ de Markov caché $(\mathbf{Y},\mathbf{Z})$ consiste à calculer les nouveaux paramètres $\boldsymbol{\Psi}^{(q+1)}$ à partir de ceux $\boldsymbol{\Psi}^{(q)}$ de l'itération précédente de façon à maximiser l'espérance conditionnelle $Q(\boldsymbol{\Psi}|\boldsymbol{\Psi}^{(q)})$ qui s'écrit dans le cas des champs de Markov cachés comme :

$$\begin{aligned} Q(\boldsymbol{\Psi}|\boldsymbol{\Psi}^{(q)}) = &\sum_{i\in S}\sum_{z_i} P(z_i|\mathbf{y},\boldsymbol{\Psi}^{(q)})\log f(y_i|z_i,\theta_{z_i}) - \log W(\boldsymbol{\phi}) \\ &+ \sum_{c}\sum_{\mathbf{z}_c} V_c(\mathbf{z}_c|\boldsymbol{\phi})P(\mathbf{z}_c|\mathbf{y},\boldsymbol{\Psi}^{(q)}). \end{aligned} \tag{4.21}$$

Le premier terme est indépendant de $\boldsymbol{\Phi}$ tandis que les deux suivants ne dépendent pas de $\boldsymbol{\theta}$. On peut procéder alors séparément pour calculer les valeurs des paramètres qui maximisent Q :

$$\begin{aligned} \boldsymbol{\theta}^{(q+1)} &= \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, Q_{\boldsymbol{\theta}}(\boldsymbol{\theta}|\boldsymbol{\Psi}^{(q)}), \\ \boldsymbol{\phi}^{(q+1)} &= \underset{\boldsymbol{\phi}}{\operatorname{argmax}}\, Q_{\boldsymbol{\phi}}(\boldsymbol{\phi}|\boldsymbol{\Psi}^{(q)}) \end{aligned}$$

On est confronté à deux difficultés pour évaluer Q : la fonction de partition $W(\boldsymbol{\phi})$ et les probabilités conditionnelles $P_G(z_i|\mathbf{y},\boldsymbol{\Psi}^{(q)})$ et $P_G(\mathbf{z}_c|\mathbf{y},\boldsymbol{\Psi}^{(q)})$ ne peuvent être calculées exactement. Plusieurs solutions ont été proposées pour rendre les deux étapes (**E**) et (**M**) réalisables. Nous présentons dans la section suivante, quelques solutions proposées pour palier à cette limite d'utilisation de l'algorithme EM.

4.3 EM et approximation de type champ moyen pour un champ de Markov caché

Les approximations nécessaires pour appliquer l'algorithme EM au modèle de champ de Markov caché sont de nature stochastique ou déterministe. Si l'algorithme EM par essence ne demande pas le calcul des étiquettes des sites pour mettre à jour les paramètres et utilise seulement les probabilités marginale et conditionnelle du champ de Markov, les approximations nécessaires pour la mise en œuvre demandent de simuler des réalisations des variables cachées. Plusieurs solutions ont été proposées comme l'algorithme Monte-Carlo EM (MCEM) (Wei and Tanner, 1990), l'algorithme EM gibbsien (Chalmond, 1989), l'algorithme de gradient stochastique (Younes, 1988), la procédure Iterative Conditionnel Estimation (ICE) (Pieczynski, 1994) ou encore des généralisations de l'algorithme MCEM proposées par Qian and Titterington (1991). Nous nous limitons à présenter ici l'algorithme *Neighborhood Restoration* EM (NREM) proposé par Celeux et al. (2003) qui est fondé sur une approximation de type champ moyen inspiré de l'algorithme de Zhang (Zhang, 1992b). Nous détaillons dans cette partie le principe du champ moyen en section 4.3.1 et la justification de l'approche en champ moyen en section 4.3.2. Enfin, nous présentons la mise en pratique de l'algorithme (NREM) en section 4.4.

4.3.1 Principe du champ moyen

L'approximation en champ moyen est à l'origine une méthode d'approximation pour le calcul de la moyenne d'un champ de Markov. Elle a été étendue pour l'estimation de la distribution du champ (Zhang, 1992a). Plus précisement, la méthode permet d'approcher un système markovien, avec des interactions complexes, par un système de variables indépendantes plus facile à manipuler.
Elle trouve sa base en mécanique statistique (Chandler, 1987) où elle a par exemple été utilisée pour l'étude de phénomène de transition de phase dans des matériaux ferro-magnétiques. Elle a été largement employée dans diverses applications telles que la vision par ordinateur (Li, 2001), la recherche de solution de problèmes issus de la théorie de graphes ou encore des réseaux de neurones.
L'idée principale du champ moyen est de négliger les fluctuations des sites voisins $\mathcal{N}_i$ quand on considère un site i, et de les fixer à leur valeur moyenne. Le système qui résulte d'une telle approximation se comporte comme un système de variables indépendantes. Les distributions deviennent dans ce cas factorisables et les calculs possibles.

On fixe Z_j pour tous les $j \in \mathcal{N}_i$ à leur valeur moyenne $m_j = \mathbb{E}_G[Z_j]$. Notons alors que les variables Z_i ne seront plus à valeurs dans $\{0,1\}^K$ mais plutôt dans $[0,1]^K$ (Cette généralisation ne pose pas de problème Wu and Doerschuk (1995)). Une nouvelle fonction d'énergie peut être définie à partir de l'équation (4.4) pour chaque individu i :

$$H_i^{mf}(z_i) = H(\mathbf{z})|_{z_j=\mathbb{E}_G[Z_j], j\in S\setminus\{i\}}. \tag{4.22}$$

Notons que cette définition est locale et il lui correspond une nouvelle mesure de probabilité :

$$P_i^{mf}(\mathbf{z}) = \frac{\exp\left[-H_i^{mf}(z_i)\right]}{W_i^{mf}} \prod_{j\in S\setminus i} I_{Z_j=\mathbb{E}_G[Z_j](z_j)},$$

concentrée sur l'hyperplan $\{Z_j = m_j, j \in S \setminus i\}$. La constante $W_i^{mf} = \sum_{z_i} \exp\left[-H_i^{mf}(z_i)\right]$ garantit que la mesure de probabilité P_i^{mf} est bien une mesure de probabilité. Dans la décomposition de l'énergie du champ d'une mesure gibbsienne (4.4), les termes qui font intervenir z_i peuvent être isolés. Cela conduit à une décomposition de l'énergie en champ moyen à la formule (4.22) du site i en un terme $H_i^{mf,loc}(z_i)$ correspondant à l'énergie en champ moyen locale au site i et un terme $H_i^{mf,ind}(\mathbb{E}_G[\mathbf{Z}_{S\setminus i}])$ indépendant de z_i. La constante de normalisation correspondante à l'énergie locale moyenne est : $W_i^{mf,loc} = \sum_{z_i} \exp\left[-H_i^{mf,loc}(z_i)\right]$
Le principe du champ moyen consiste alors à remplacer la distribution marginale du champ $P_G(z_i)$ par :

$$\begin{aligned} P_i^{mf}(z_i) &= \exp\left[-H_i^{mf}(z_i)\right]/W_i^{mf} \\ &= \exp\left[-H_i^{mf,loc}(z_i)\right]/W_i^{mf,loc}, \end{aligned} \tag{4.23}$$

qui est aussi la probabilité Z_i conditionnellement à $\mathbf{Z}_{\mathcal{N}_i} = \mathbb{E}_G[\mathbf{Z}_{\mathcal{N}_i}]$. La distribution en champ moyen qui approche la distribution jointe gibbsienne $P_G(\mathbf{z})$ est alors donnée par :

$$P^{mf}(\mathbf{z}) = \prod_{i\in S} P_i^{mf}(z_i) \tag{4.24}$$

$$= \frac{\exp\left[-\sum_{i\in S} H_i^{mf,loc}(z_i)\right]}{W^{mf}} \tag{4.25}$$

où $W^{mf} = \prod_{i\in S} W_i^{mf,loc}$.
La différence majeure avec la pseudo-vraisemblance de Besag est que les voisins ne sont plus autorisés à fluctuer. Ainsi, fixés à des constantes (leur moyenne), les termes du produit sont bien indépendants et P^{mf} est bien une distribution de probabilité. Mais pour utiliser l'approximation définie par l'équation (4.24), il est nécessaire d'évaluer les moyennes $\mathbb{E}_G[Z_i]$, pour les sites différents de i, qui sont inconnues et l'objectif de l'approximation en champ

moyen est justement de les calculer.
Afin de de définir l'approximation en champ moyen une condition de cohérence doit être vérifiée par les valeurs approchées. Cette condition stipule que les valeurs moyennes z_i^{mf} calculées à partir de l'approximation doivent être égales aux valeurs moyennes utilisées pour définir l'approximation. Ceci s'écrit $\forall i \in S$:

$$z_i^{mf} = \mathbb{E}_i^{mf}[Z_i] = \sum_{z_i} z_i \exp\left[-H_i^{mf,loc}(z_i)\right]/W_i^{mf,loc},$$

On peut remarquer que ette expression est une fonction de $\{z_j^{mf}, j \in \mathcal{N}_i\}$ qu'on peut noter $g_i(\{z_j^{mf}, j \in \mathcal{N}_i\})$.
Si l'on répète cette démarche pour tous les sites $i \in S$ on peut obtenir l'équation :

$$\mathbf{z}^{mf} = g(\mathbf{z}^{mf}) = \begin{cases} g_1(\{z_j^{mf}, j \in \mathcal{N}_1\}), \\ \dots \\ g_N(\{z_j^{mf}, j \in \mathcal{N}_N\}). \end{cases} \tag{4.26}$$

L'approximation en champ moyen consiste alors à résoudre ce problème de point fixe. La solution $\mathbf{z}^{mf} = \{z_i^{mf}, i \in S\}$ est alors l'estimation en champ moyen de l'espérance $\mathbb{E}_G[\mathbf{Z}]$.

4.3.2 Justification de l'approche en champ moyen

L'approche en champ moyen se justifie par un principe variationnel issu de la physique statistique. Ce principe repose sur l'idée de minimisation de la fonctionnelle énergie libre $F(P)$ appelée aussi *énergie libre variationnelle*. Avec une distribution d'énergie H et une probabilité P définies sur l'espace des configurations possibles d'un système d'entropie ξ, la fonctionnelle $F(P)$ est définie par :

$$F(P) = \mathbb{E}[H(\mathbf{Z})] - \xi(P),$$

où ξ est l'entropie de la probabilité P avec $\xi(P) = -\mathbb{E}[\log P(\mathbf{Z})]$. Dans le cas des champs de Markov cachés, si P_G est la distribution gibbsienne associée à l'énergie H alors :

$$F(P) = -\log W + \mathbb{E}[\log P(\mathbf{Z})/P_G(\mathbf{Z})].$$

On peut voir le problème comme une minimisation de la quantité $\mathbb{E}[\log P(\mathbf{Z})/P_G(\mathbf{Z})]$ sur les distributions de probabilité P. On minimise la divergence de Kullback-Leibler (voir section 3.3.1) entre P et la vraie distribution de Gibbs P_G. Si $P = P_G$, l'énergie libre minimale vaut

alors : $F(P_G) = -\log W$. On montre facilement que l'approximation en champ moyen est optimale pour le critère de la divergence de Kulback-Leibler pour les systèmes à variables indépendantes (Peyrard, 2001).
En effet, les probabilités se factorisent alors selon :

$$\begin{aligned} P(\mathbf{z}) = W^{-1} \exp\left[-H(\mathbf{z})\right] &= \prod_{i \in S} P(z_i) \\ &= \prod_{i \in S} \frac{\exp\left[-H_i(z_i)\right]}{W_i} \\ &= \frac{\exp - \sum_{i \in S}[H_i(z_i)]}{\prod_{i \in S} W_i}. \end{aligned}$$

On se restreint ainsi à des systèmes de variables indépendantes. Plus précisément cela revient à considérer les systèmes de variables dont la fonction d'énergie associée est de la forme :

$$H^{fac}(\mathbf{z}) = \sum_{i \in S} z_i^t [V_i + \delta V_i], \tag{4.27}$$

où les vecteurs $V_i = \{V_i(e_1), \ldots, V_i(e_K)\}, i \in S$ sont des fonctions de potentiel sur les singletons de l'énergie H et où δV_i doit être identifié. La probabilité gibbsienne associée à cette fonction d'énergie factorisable est :

$$F(P^{fac}) = -\log W^{fac} + \mathbb{E}_{P^{fac}}[H(\mathbf{Z}) - H^{fac}(\mathbf{z})], \tag{4.28}$$

où W^{fac} est la fonction de partition du modèle défini par l'équation (4.27). Les quantités δV_i satisfaisant le principe d'énergie libre minimale peuvent être identifiées par annulation du gradient de la fonction (4.28). On retrouve alors le problème de point fixe (4.26) et la solution du problème de minimisation de l'énergie libre sur l'ensemble restreint des distributions factorisables est la distribution P^{mf}. Ce choix est donc optimal au sens de la divergence de Kullback-Leibler : pour ce critère c'est la meilleure approximation de P_G par un système de variables indépendantes. On pourra aussi montrer qu'une bonne approximation de W est obtenue en utilisant une approximation de type champ moyen (voir Peyrard (2001) pour plus de détails).
L'approche variationnelle permet de distinguer l'approximation en champ moyen parmi l'ensemble des approximations d'un champ de Markov correspondant à un système de variables indépendantes et de montrer son optimalité au sens de la divergence de Kullback-Leibler.

4.3.3 Mise en œuvre de l'algorithme EM de type champ moyen

Dans Celeux et al. (2003), on parle d'approximation de type champ moyen quand la valeur d'un site i ne dépend pas des valeurs des sites voisins mais de constantes (pas forcément

la moyenne donc). Cette idée sera utilisée pour le calcul fastidieux de la probabilité jointe $P_G(\mathbf{y}, \mathbf{z}|\mathbf{\Psi})$ dans une procédure de type EM (Celeux et al., 2003). Le principe de l'algorithme EM basé sur l'approximation de type champ moyen que Celeux et al. (2003) proposent consiste à alterner une étape de restauration du voisinage (**NR**) (**neighborhood restauration**), puis une étape (**EM**) d'estimation des paramètres de champ de Markov caché *via* l'algorithme EM pour le modèle de mélange fini indépendant correspondant au modèle simplifié courant. La forme générale de ces algorithmes de type champ moyen à l'itération (q) est donc :

– **(NR) Choix des voisins :** on crée à partir des observations $\mathbf{y}$ et de l'estimation courante des paramètres $\mathbf{\Psi}^{(q-1)}$ une nouvelle configuration du champ des voisins $\tilde{\mathbf{z}}^{(q)}$. Trois choix naturels apparaissent pour la valeur du champ voisin $\tilde{\mathbf{z}}^{(q)}$:

1. **EM champ-moyen** : $\tilde{z}_i^{(q+1)} = \mathbb{E}_{\tilde{\mathbf{z}}^{(q)}}[Z_i]$. Cette configuration du champ voisin est celle dont on a vu qu'elle était optimale au sens de la divergence de Kullback-Leibler pour les distributions factorisables ;
2. **EM champ-modal** : $\tilde{z}_i^{(q+1)} = \underset{z_i}{\operatorname{argmax}}\, P_{\tilde{\mathbf{z}}^{(q)}}(z_i|y_i, \mathbf{\Psi}^{(q)})$.
3. **EM champ-simulé** : $\tilde{z}_i^{(q)} \sim P_{\tilde{\mathbf{z}}^{(q)}}(z_i|y_i, \mathbf{\Psi}^{(q)})$.

Ainsi, pour chaque site i, l'état de ses voisins $\mathbf{z}_{\mathcal{N}_i}^{(q)}$ est fixé à $\tilde{\mathbf{z}}_{\mathcal{N}_i}^{(q)}$ et la distribution marginale peut être approchée par :

$$P_{\tilde{\mathbf{z}}^{(q)}}(\mathbf{z}|\boldsymbol{\phi}) = \prod_{i \in S} P_G(z_i|\tilde{\mathbf{z}}_{\mathcal{N}_i}^{(q)}, \boldsymbol{\phi}). \tag{4.29}$$

Notons que la mise à jour est séquentielle sur les sites de 1 à N : quand on met à jour le site i à l'itération $(q+1)$, les sites $1, \ldots, i-1$ ont déjà été mis à jour et on utilise alors $\tilde{z}_1^{(q+1)}, \ldots, \tilde{z}_{i-1}^{(q+1)}$ tandis qu'on utilise $\tilde{z}_{i+1}^{(q)}, \ldots, \tilde{z}_n^{(q)}$ pour cette mise à jour.

– **(EM) estimation :** on applique l'algorithme EM pour le modèle de champ de Markov caché défini par la formule (4.29) et la loi d'observations des données $\mathbf{y}$ conditionnellement aux classes. On part de valeurs initiales $\mathbf{\Psi}^{(q-1)}$ pour les mettre à jour en $\mathbf{\Psi}^{(q)}$. La distribution jointe gibbsienne $P_G(\mathbf{y}, \mathbf{z}|\mathbf{\Psi})$ est remplacée par :

$$P_{\tilde{z}^{(q)}}(\mathbf{y}, \mathbf{z}|\mathbf{\Psi}) = \prod_{i \in S} f(y_i|z_i, \theta_{z_i}) P_G(z_i|\tilde{\mathbf{z}}_{\mathcal{N}_i}^{(q)}, \boldsymbol{\phi}), \tag{4.30}$$

qui correspond à une vraisemblance des observations :

$$\begin{aligned} P_{\tilde{\mathbf{z}}^{(q)}}(\mathbf{y}|\mathbf{\Psi}) &= \sum_{\mathbf{z}} f(\mathbf{y}|\mathbf{z}, \boldsymbol{\theta}) P_{\tilde{\mathbf{z}}^{(q)}}(\mathbf{z}|\boldsymbol{\phi}) \\ &= \prod_{i \in S} \sum_{z_i} f(y_i|z_i, \theta_{z_i}) P_G(z_i|\tilde{\mathbf{z}}_{\mathcal{N}_i}^{(q)}, \boldsymbol{\phi}) \\ &= \prod_{i \in S} P_G(y_i|\tilde{\mathbf{z}}_{\mathcal{N}_i}^{(q)}, \mathbf{\Psi}). \end{aligned} \tag{4.31}$$

L'étape d'estimation de l'algorithme EM est décomposée en deux étapes :

– **(E)** Calcul des probabilités *a posteriori* :

$$\tilde{t}_{ik}^{(q)} = \frac{\tilde{\pi}_{ik}^{(q)} f(y_i|\theta_k^{(q-1)})}{\sum_{l=1}^{K} \tilde{\pi}_{il}^{(q)} f(y_i|\theta_l^{(q-1)})}, \tag{4.32}$$

avec : $\tilde{\pi}_{ik}^{(q)} = P_{\tilde{\mathbf{z}}^{(q)}}(z_i = k|\boldsymbol{\phi}^{(q)})$.

– **(M)** Mise à jour des paramètres $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_K)$ de la distribution $P_{\tilde{\mathbf{z}}^{(q)}}(\mathbf{z}|\boldsymbol{\phi})$ et des paramètres $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K)$ des densités $f(.|\theta_k)$:

$$\boldsymbol{\phi}^{(q)} = \underset{\boldsymbol{\phi}}{\operatorname{argmax}} \sum_{i \in S} \sum_{k=1}^{K} \tilde{t}_{ik}^{(q)} \log \tilde{\pi}_{ik}, \tag{4.33}$$

$$\text{et } \boldsymbol{\theta}^{(q)} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \sum_{i \in S} \sum_{k=1}^{K} \tilde{t}_{ik}^{(q)} \log f(y_i|\theta_k). \tag{4.34}$$

L'algorithme NREM est un algorithme d'estimation de paramètres des modèles de champ de Markov caché sous approximation de type champ moyen et non pas un algorithme de classification. Sous l'approximation en champ moyen (équation 4.30), on est ramené à un modèle de mélange indépendant. Le MAP et le MPM conduisent à choisir en chaque site i la classe la plus probable connaissant l'observation y_i.
Par exemple, pour le MPM :

$$\forall i \in S, z_i^{mpm} = \underset{z_i}{\operatorname{argmax}}\, P_{\tilde{\mathbf{z}}}(z_i|y_i) = \underset{z_i}{\operatorname{argmax}}\, \tilde{\pi}_{iz_i} f(y_i|\theta_{z_i}) = \underset{k}{\operatorname{argmax}}\, \tilde{t}_{ik}. \tag{4.35}$$

Les $\tilde{t}_{ik}$ obtenus en sortie de l'algorithme NREM permettent donc d'obtenir directement la classification voulue.

4.4 Critère BIC de sélection de modèle

La sélection du modèle est un problème central dans les analyses statistiques. Il s'agit de sélectionner le modèle qui présente un bon compromis entre complexité et adéquation aux données. Il existe dans la littérature, plusieurs méthodes de sélection de modèles comme la validation croisée (Durand, 2003), le Multifold Cross Validation proposé par Zhang (1993) ou encore le Repeated Learning Testing (Burman, 1989). Ces critères sélectionnent bien les modèles pour leurs qualités prédictives. Mais ils ne peuvent pas s'appliquer dans le cadre des champs de Markov puisque les données ne sont plus indépendantes contrairement aux modèles de mélange pour lesquels ils ont été développés. Nous avons choisi d'utiliser le

critère **BIC** (*Bayesian Information Criterion*) (Schwarz, 1978) qui est un des critères les plus répandus et qui s'adapte dans le cadre des champs de Markov. Le principe de ce critère est de sélectionner parmi l'ensemble des modèles $\mathcal{M}$ étudiés, celui qui maximise la quantité :

$$BIC_{\mathcal{M}} = 2 \log L(\psi_{\mathcal{M}}^{MV}; \mathbf{y}) - \kappa_{\mathcal{M}} \log N, \tag{4.36}$$

où $\kappa_{\mathcal{M}}$ est le nombre de paramètres du modèle $\mathcal{M}$ et L la valeur de la vraisemblance, calculée pour les estimateurs du maximum de vraisemblance $\mathbf{\Psi}_{\mathcal{M}}^{MV}$ et les observations $\mathbf{y}$.
Le critère BIC se décompose en deux termes : le terme de vraisemblance $2 \log L(\mathbf{\Psi}_{\mathcal{M}}^{MV}; \mathbf{y})$ favorisant la sélection d'un modèle complexe et le terme de pénalité $\kappa_{\mathcal{M}} \log n$, fonction croissante du nombre de paramètres, favorisant la sélection d'un modèle parcimonieux. Ce critère a montré son intérêt pratique lors d'expérience dans le cas de l'estimation du nombre de classes d'un modèle de mélange (Fraley and Raftery, 1998). Il a été observé, cependant, qu'il a tendance à surestimer le nombre de classes lorsque le vrai modèle ne fait pas partie de la famille considérée (Biernacki et al., 2000). Lorsque le modèle $\mathcal{M}$ est celui d'un modèle de mélange indépendant comme défini en section 4.2.1, le critère BIC est donné par :

$$BIC_{\mathcal{M}} = 2 \sum_{i \in S} \log \Big(\sum_{k=1}^{K} \pi_k f(y_i|\theta_k) \Big). \tag{4.37}$$

Lorsque le modèle $\mathcal{M}$ est celui d'un champ de Markov caché comme défini en section 4.2, le critètre BIC ne peut être calculé sans approximation. Deux approximations sont possibles, l'une utilisant l'approximation en champ moyen de la distribution de Gibbs P_G et l'autre utilisant l'approximation en champ moyen de la fonction de partition W.

– **Critère BIC par approximation de la distribution de Gibbs**
Soit $\mathcal{M}$ le modèle de Markov caché, en utilisant l'approximation en champ moyen $P_{\tilde{\mathbf{z}}}$ de la distribution $P_G(\mathbf{z}|\mathbf{y})$ (voir section 4.2), le critère BIC est alors approché par le critère $BIC_{\mathcal{M}}^{G}$:

$$BIC_{\mathcal{M}}^{G} = 2 \sum_{i \in S} \log \sum_{k=1}^{K} f(y_i|\theta_k) P_{\tilde{\mathbf{z}}}(Z_i = k|\boldsymbol{\phi}) - \kappa_{\mathcal{M}} \log N.$$

– **Critère BIC par approximation de la fonction de partition**
Soit $\mathcal{M}$ le modèle de Markov caché de distribution définie en section 4.2.2.1. Remar-

quons que :

$$\begin{aligned} P(\mathbf{y}|\mathbf{\Psi}) = \frac{P_G(\mathbf{y},\mathbf{z}|\mathbf{\Psi})}{P_G(\mathbf{z}|\mathbf{y},\mathbf{\Psi})} &= \frac{f(\mathbf{y}|\mathbf{z},\boldsymbol{\theta})P_G(\mathbf{z}|\boldsymbol{\phi})}{P_G(\mathbf{z}|\mathbf{y},\mathbf{\Psi})} \\ &= \underbrace{\frac{f(\mathbf{y}|\mathbf{z},\boldsymbol{\theta})\exp\left(-H(\mathbf{z}|\boldsymbol{\phi})\right)}{\exp\left(-H(\mathbf{z}|\mathbf{y},\mathbf{\Psi})\right)}}_{=1}\frac{W(\mathbf{y},\mathbf{\Psi})}{W(\boldsymbol{\phi})} \\ &= \frac{W(\mathbf{y},\mathbf{\Psi})}{W(\boldsymbol{\phi})} \end{aligned} \tag{4.38}$$

Il s'ensuit à partir de l'équation (4.36) que le critère BIC s'écrit sous la forme :

$$BIC_{\mathcal{M}} = 2\log W(\mathbf{y},\mathbf{\Psi}^{\mathbf{MV}}) - 2\log W(\boldsymbol{\phi}^{MV}) - \kappa_{\mathcal{M}}\log N. \tag{4.39}$$

Une autre alternative à $BIC_{\mathcal{M}}$ est d'approcher les constantes de partition $W(\mathbf{y},\mathbf{\Psi}^{MV})$ et $W(\boldsymbol{\phi}^{MV})$. Le critère BIC s'écrit dans ce cas là :

$$\begin{aligned} BIC^{W}_{\mathcal{M}} = &\, 2\log W^{(\boldsymbol{\phi})}_{\tilde{\mathbf{z}}} + 2\mathbb{E}_{P_{\tilde{\mathbf{z}}}}[-H(\mathbf{Z},\boldsymbol{\phi}) + H_{\tilde{z}}(\mathbf{Z},\boldsymbol{\phi})] \\ &- 2\log W_{\tilde{\mathbf{z}}}(\mathbf{y},\mathbf{\Psi}) - 2\mathbb{E}_{P_{\tilde{\mathbf{z}}}(.|\mathbf{y},\mathbf{\Psi})}[-H(\mathbf{Z},\boldsymbol{\phi}) + H_{\tilde{\mathbf{z}}}(\mathbf{Z},\boldsymbol{\phi})] \\ &- \kappa_{\mathcal{M}}\log N, \end{aligned} \tag{4.40}$$

avec $P_{\tilde{\mathbf{z}}}$ l'approximation en champ moyen de la distribution gibbsienne P_G et $W_{\tilde{\mathbf{z}}}$ sa constante de normalisation.
Il existe un lien entre ces deux approximations. En effet, l'approximation $BIC^{G}_{\mathcal{M}}$ (donnée par l'équation (4.4)) peut s'écrire aussi comme :

$$BIC^{G}_{\mathcal{M}} = 2\log W_{\tilde{\mathbf{z}}}(\mathbf{y},\mathbf{\Psi}) - 2\log W_{\tilde{\mathbf{z}}}(\boldsymbol{\phi}) - \kappa_{\mathcal{M}}\log N.$$

Cela correspond au développement à l'ordre 1 de $BIC^{W}_{\mathcal{M}}$. Il est prouvé que l'approximation $BIC^{W}_{\mathcal{M}}$ est plus fine que l'approximation $BIC^{G}_{\mathcal{M}}$ (Forbes and Peyrard, 2003).

CHAPITRE 5

Cartographie du risque épidémiologique à l'aide des champs de Markov cachés

Sommaire

Nous abordons la cartographie du risque en épidémiologie comme un problème de classification à l'aide de modèles de Markov cachés discrets et de modèles de mélange de Poisson. Rappelons qu'en épidémiologie animale, on a essentiellement besoin de la classification pour définir clairement les zones dans lesquelles on pourra décider d'imposer des mesures de lutte. Le modèle de Markov caché qu'on propose est une variante du modèle de Potts (4.1.3), où le paramètre d'interaction dépend des classes de risque.

Les paramètres de ce modèle sont inconnus et doivent être estimés. Plusieurs algorithmes d'estimation existent dans la littérature. Les plus utilisés sont les algorithmes de type MCMC

(voir section 3.1.5) lorsque l'on est dans un cadre bayésien. Dans un contexte de données manquantes, c'est l'algorithme EM avec des approximations qui est largement utilisé (voir section 4.3). Afin d'estimer les paramètres de notre modèle, nous utilisons l'algorithme EM combiné à une approche variationnelle de type champ moyen (voir 4.3). Comme mentionné en 4.3, cette approche permet d'appliquer l'algorithme EM dans un cadre spatial et présente une alternative efficace aux méthodes d'estimation basées sur des simulations de type MCMC.
Nous nous intéressons dans ce travail aux maladies rares, ce qui se traduit par le fait que les taux de risque attendus sont petits. Ces taux affectent les moyennes des classes du modèle de mélange de Poisson. Cela signifie que les classes du modèle sont, par conséquent, mal séparées.
Pour appréhender ce problème, nous proposons une nouvelle stratégie d'initialisation appropriée aux modèles de mélange de Poisson quand les classes sont mal séparées.
Dans ce chapitre, nous présentons le modèle proposé en section 5.2, l'algorithme EM champ moyen appliqué à ce modèle en section 5.3 et la stratégie d'initialisation proposée en section 5.4.

5.1 Motivations et objectifs de la cartographie

L'objectif principal de la cartographie du risque en épidémiologie est de mettre en évidence l'hétérogénéité spatiale, les zones géographiques les plus à risque et l'influence de certains facteurs, ce qui permet de mieux comprendre les mécanismes des épidémies. Il s'agit d'estimer le risque de contamination, à partir des données disponibles, pour chaque unité géographique et de pouvoir comparer ce risque entre les unités.
En épidémiologie, on s'intéresse en particulier à la situation géographique des zones à risque élevé ou faible, mais aussi aux "différences de risque" entre ces régions. Ces différences de risque, appelées contrastes, sont des informations capitales pour les autorités sanitaires qui doivent en fonction de cela décider (ou non) de la prise de mesures de protection locales ou nationales.
Les modèles de cartographie présentés au chapitre 3 sont tous basés sur des approches bayésiennes hiérarchiques qui produisent une estimation du risque pour chaque site. Rappelons que les estimations produites pour chaque unité géographique i, par ces modèles, permettent de tracer des cartes de risque continues. Ces cartes ne sont pas toujours très lisibles en particulier si le nombre d'unités est grand. Dans la plupart de ces modèles, à part les modèles de Green and Richardson (2002) ou encore de Alfo et al. (2009) qui peuvent fournir des classes de risque, la classification des risques, doit être effectuée *a posteriori* par l'utilisateur. Cela

peut être fait d'une manière empirique ou en appliquant une méthode de classification sur les estimations de risque obtenues. Cette étape *a posteriori* pose un problème pour les épidémiologistes quant au choix de la méthode de classification "pertinente" ou encore du nombre de classes si cela est effectué d'une manière empirique. Une bonne alternative, pour les épidémiologistes, est de proposer une méthode intégrant cette étape de classification.
Dans ce travail, nous nous orientons vers les champs de Markov cachés discrets. Ils ne produisent pas une estimation du risque pour chaque unité géographique mais ils travaillent sur des classes (ou groupes) de risque permettant la classification des unités géographiques en groupes ou zones de risque bien délimitées. De plus, un algorithme d'estimation EM approché permet d'obtenir directement la classification requise.

5.2 Champ de Markov caché pour la classification du risque

Les modèles de champs de Markov cachés et plus spécifiquement les modèles de mélange font partie des méthodes les plus utilisées en classification. Dans ce travail, nous abordons le problème de cartographie comme un problème de classification.
Un problème de classification est spécifié en terme d'un ensemble de sites S et un ensemble d'étiquettes $\mathcal{L}$. Un site peut représenter un objet, un point ou une région. Ici, un site représente une unité géographique de la zone étudiée. Une étiquette est un événement qui peut se produire en un site. Dans le problème de cartographie en épidémiologie, un événement typique est l'affectation d'une unité à un certain niveau de risque. Nous ne considérons, dans ce travail, que le cas où le nombre de niveaux de risque est fini, c'est-à-dire les étiquettes prennent des valeurs discrètes dans un ensemble de K étiquettes. Dans la suite, nous considérons $\mathcal{L}$ comme un vecteur de dimension K, $\mathcal{L} = \{e_1, \ldots, e_K\}$ où chaque e_k a toutes ses composantes égales à 0 à part la $k^{\text{ième}}$ composante qui est égale à 1.
L'objectif alors est d'affecter chaque unité géographique à un niveau de risque parmi les K niveaux de risque possibles qui sont eux même inconnus et qui doivent être estimés. De plus, comme mentionné auparavant, nous prenons en compte la dépendance spatiale entre les nombres de cas dans les différentes régions en vue d'obtenir une estimation consistante du risque. En général, les risques sont supposés être semblables pour les régions voisines. L'idée est d'exploiter l'information provenant des régions voisines pour fournir une estimation plus fiable du risque pour chacune de ces régions. Rappelons que nous nous plaçons dans un cadre où les données sont agrégées par unités géographiques. On dispose en épidémiologie de deux types de données :

– les observations concernant les effectifs de cas (nombre de malades, de morts, etc.) dis-

ponibles pour chaque unité géographique, que l'on dénote par : $y_i, i \in S = \{1, \ldots, N\}$,
- les effectifs de populations cibles (données démographiques), que l'on dénote par : $n_i, i \in S = \{1, \ldots, N\}$.

La supposition commune à beaucoup de modèles en épidémiologie est que, pour chaque unité $i \in S = \{1, \ldots, N\}$, le nombre de cas y_i est une réalisation d'une distribution de Poisson dont la moyenne dépend du niveau de risque associé à la région. Rappelons que la loi de Poisson est adaptée pour modéliser les maladies non contagieuses.

Étant donné que notre objectif est de classer les unités géographiques dans différents niveaux de risque alors l'affectation d'une unité géographique quelconque à un niveau de risque et les niveaux de risque représentent la partie cachée et les paramètres du modèle.

Les données sont, donc, naturellement divisées en données observées $\mathbf{y} = \{y_1, \ldots, y_N\}$ et données non observées (ou manquantes) $\mathbf{z} = \{z_1, \ldots, z_N\}$ qui sont considérées comme des variables aléatoires dénotées par : $\mathbf{Z} = \{Z_1, \ldots, Z_N\}$. Le champ $\mathbf{z}$ représente une classification des unités géographiques.

Quand les variables $\{Z_i, i \in S\}$ sont indépendantes, le modèle est réduit à un modèle de mélange (section 4.2.1). Par contre, quand elles sont dépendantes les relations entre les sites sont représentées par un graphe. Deux sites voisins représentent les deux noeuds d'un graphe reliés par une arête (voir section 4.1). En cartographie, on considère le plus souvent la plus simple des structures de graphe connectant les voisins contingents. On considère que les unités i et j sont voisines si et seulement si elles sont contigues spatialement.

5.2.1 La structure cachée du modèle

Les dépendances spatiales entre les variables aléatoires voisines $\{Z_i, i \in S\}$ sont prises en compte en supposant que la distribution jointe de $\{Z_1, \ldots, Z_N\}$ est un champ aléatoire de Markov discret (section 4.1) défini sur ce graphe avec une fonction d'énergie (4.4) qui s'écrit dans ce cas comme :

$$H(\mathbf{z}; \boldsymbol{\beta}) = \sum_{i \in S} V_i(z_i; \boldsymbol{\beta}) + \sum_{\substack{i,j \\ i \sim j}} V_{ij}(z_i, z_j; \boldsymbol{\beta}),$$

L'ensemble des paramètres $\boldsymbol{\beta}$ est composé des paramètres $\boldsymbol{\alpha}$ et $\mathbb{B}$. On pourrait considérer que les fonctions potentiels d'ordre 2, V_{ij}, dépendent de z_i et z_j mais aussi de i et j.

Comme les z_i ne prennent qu'un nombre fini K de valeurs, on peut définir, pour chaque i et j, une matrice $\mathbb{B}_{ij} = (\mathbb{B}_{ij}(k,l))_{1 \leq k,l \leq K}$ de dimension $K \times K$ telle que $V_{ij}(z_i, z_j; \beta) = -\mathbb{B}_{ij}(k,l)$ si $z_i = e_k$ et $z_j = e_l$. En utilisant la notation vectorielle et en notant z_i^t la transposée du vecteur z_i, il est équivalent d'écrire $V_{ij}(z_i, z_j; \boldsymbol{\beta}) = -z_i^t \mathbb{B}_{ij} z_j$. Pareillement on considère les fonctions

potentiels V_i qui peuvent dépendre de z_i et i. Si on considère α_i comme un vecteur de dimension K, on peut écrire $V_i(z_i, \boldsymbol{\beta}) = -\alpha_i(k)$ si $z_i = e_k$, où $\alpha_i(k)$ est la $k^{\text{ième}}$ composante de α_i. De façon équivalente on peut écrire $V_i(z_i, \boldsymbol{\beta}) = -z_i^t \alpha_i$. Ce vecteur α_i peut être interprété comme le vecteur des poids associés à chacune des K classes. Lorsque α_i est nul, aucun niveau de risque n'est favorisé, c'est-à-dire tous les niveaux de risque sont équiprobables. Si en plus, pour tout i et j, $\mathbb{B}_{ij} = b \times I_K$ où b est un scalaire réel et I_K est la matrice d'identité de dimension $K \times K$, les paramètres $\boldsymbol{\beta}$ se réduisent à un seul paramètre d'interaction b et on retrouve le traditionnel modèle de Potts (section 4.1.3) usuellement utilisé en segmentation d'image.

Comme mentionné au chapitre 4, le modèle de Potts est souvent le plus approprié pour la classification parce qu'il favorise les voisins qui sont dans la même classe (c'est-à-dire. qui ont le même niveau de risque). Cependant, ce modèle pénalise les paires qui ont des niveaux de risque différents avec la même pénalité quelles que soient les valeurs de ces niveaux de risque. Contrairement à certaines applications en segmentation d'images, dans le contexte de la cartographie du risque l'ordre des classes en lien avec leur situation géographique et leur interprétation en terme de gravité du risque, est très important. Donc il serait plus approprié, d'un point de vue épidémiologique, d'introduire cette idée de gradation des risques et de transitions progressives entre les classes dans le modèle.

Pour les cas où $\mathbb{B}_{ij}$ n'est pas égale à $b \times I_K$, la construction de cette matrice permet l'intégration d'interactions plus fines entre voisins. La matrice $\mathbb{B}_{ij}$ apporte une grande flexibilité à la modélisation des dépendances spatiales entre les unités. Cette construction représente la flexibilité de la modélisation offerte par notre modèle.

En pratique ces paramètres peuvent être réglés selon la connaissance *a priori* des experts, ou ils peuvent être estimés à partir des données.

Dans notre modèle, on propose de modéliser $\mathbb{B}$ par une matrice diagonale définie, pour un certain scalaire positif et réel b, par :

$$\begin{aligned} \mathbb{B}(k,k) &= b \quad \text{pour tout } k = \{1 \dots K\} \\ \mathbb{B}(k,l) &= b/2 \quad \text{pour tout } (k,l) \text{ tel que } |k-l| = 1 \\ \mathbb{B}(k,l) &= 0 \quad \text{sinon.} \end{aligned} \tag{5.1}$$

Ce modèle que l'on propose est appelé par la suite le *semi-graduel* et peut encore s'écrire : $\mathbb{B}(k,l) = b\,[1 - |k-l|/2]_+$.

L'idée de cette modélisation est de favoriser en premier les voisins qui sont dans la même classe et en second ceux qui sont dans des classes voisines. Tous les autres voisins sont traités avec un poids égal. Cette manière de modéliser $\mathbb{B}$ est la structure "non stantard" de $\mathbb{B}$ la plus

simple qui permet d'introduire des gradations des niveaux de risque. La probabilité conditionnelle $p(z_i|\mathbf{z}_{\mathcal{N}_i};\boldsymbol{\beta})$ de la variable z_i sachant les valeurs de ses voisins $\mathbf{z}_{\mathcal{N}_i}$ et les paramètres $\boldsymbol{\beta}$ du champ est facilement calculable par la formule suivante :

$$P(z_i|\mathbf{z}_{\mathcal{N}_i};\boldsymbol{\beta}) = \frac{\exp\left(z_i^t(\boldsymbol{\alpha} + \mathbb{B}\sum_{j\in\mathcal{N}_i} z_j)\right)}{\sum_{k=1}^{K}\exp\left(\alpha(k) + e_k^t\mathbb{B}\sum_{j\in\mathcal{N}_i} z_j\right)}. \tag{5.2}$$

5.2.2 Le modèle d'observation pour les données de comptage

La section précédente décrit la partie cachée du modèle. Pour que le modèle soit complètement défini, il faut spécifier la modélisation des observations. La distribution $P(y_i|z_i)$ est, en général, une distribution standard. Typiquement en cartographie du risque pour les maladies rares et non contagieuses, c'est une distribution de Poisson qui est utilisée $Pois(y_i; n_i\lambda_{z_i})$ où n_i est l'effectif de la population cible pour l'unité géographique i et l'indice z_i de λ_{z_i} montre que le paramètre de la distribution dépend de la valeur z_i qui détermine le niveau du risque et par conséquent la valeur du risque parmi un nombre fini de valeurs $\{\lambda_1, \dots \lambda_K\}$.
Avec notre notation vectorielle, nous pouvons écrire $z_i^t\boldsymbol{\lambda}$ pour λ_{z_i} avec $\boldsymbol{\lambda} = (\lambda_1, \dots \lambda_K)^t$ et la distribution conditionnelle s'écrit :

$$P(Y_i = y_i|Z_i = z_i;\boldsymbol{\lambda}) = Pois(y_i;\ n_i\ z_i^t\boldsymbol{\lambda}) = \exp(-n_i\ z_i^t\boldsymbol{\lambda})\ \frac{(n_i\ z_i^t\boldsymbol{\lambda})^{y_i}}{y_i!}\ . \tag{5.3}$$

Notons qu'en pratique, les épidémiologistes préférent considérer le risque relatif (RR) (section 2.3.2). Toutefois, dans le cas d'une population unique sans structure, le risque relatif est équivalent au risque absolu. Nous avons choisi de construire un modèle général en considérant le risque absolu tout en précisant que le passage du risque absolu, dans notre modèle, au risque relatif est possible. Il suffit de remplacer les populations n_i par les nombres de cas attendus E_i.
L'hypothèse habituelle d'indépendance conditionnelle pour les variables $\mathbf{y}$ étant donnée la classification $\mathbf{z}$ conduit à la loi jointe :

$$P(\mathbf{Y} = \mathbf{y}|\mathbf{Z} = \mathbf{z};\boldsymbol{\lambda}) = \prod_{i\in S} Pois(y_i; n_i z_i^t\boldsymbol{\lambda}). \tag{5.4}$$

Il s'ensuit que l'énergie du champ caché $\mathbf{z}$ étant donné le champ observé $\mathbf{y}$ s'exprime comme :

$$H(\mathbf{z}|\mathbf{y};\boldsymbol{\lambda},\boldsymbol{\beta}) = H(\mathbf{z};\boldsymbol{\beta}) - \sum_{i\in S}\log Pois(y_i; n_i z_i^t\boldsymbol{\lambda}),$$

et sa distribution conditionnelle est :

$$P(\mathbf{z}|\mathbf{y};\boldsymbol{\lambda},\boldsymbol{\beta}) = W(\boldsymbol{\beta})^{-1}\exp(-H(\mathbf{z};\boldsymbol{\beta}) + \sum_{i\in S}\log Pois(y_i; n_i z_i^t\boldsymbol{\lambda})).$$

Les paramètres inconnus du modèle et qui doivent être estimés sont dénotés par $\mathbf{\Psi} = (\boldsymbol{\lambda}, \boldsymbol{\alpha}, \mathbb{B})$. Pour les problèmes de classification, les approches d'estimations se divisent en deux catégories. La première catégorie se focalise sur le problème de trouver la meilleure configuration pour $\mathbf{z}$ en se basant sur un principe de décision bayésien comme le Maximum à Posteriori (MAP) ou le Maximum des Probabilités Marginales (MPM) qui sont présentés en section 4.2.1.4. Ces critères utilisent explicitement la distribution marginale $P(\mathbf{z}|\mathbf{y})$ et utilise la propriété markovienne du champ conditionnel $\mathbf{Z}|\mathbf{Y} = \mathbf{y}$.
Les méthodes les plus populaires de cette catégorie sont l'algorithme Iterated Conditional Mode (ICM) (Besag, 1986) et l'algorithme de recuit simulé (Geman and Geman, 1984). Ils diffèrent dans leur manière de traiter la probabilité marginale $P(\mathbf{z}|\mathbf{y})$ qui n'est pas calculable.
Le second type d'approches est lié à un concept de données manquantes. Ce type d'algorithmes se focalise, pour l'estimation des paramètres, sur le principe du maximum de vraisemblance avec les données manquantes (les z_i dans notre cas). L'algorithme de référence dans cette catégorie est l'algorithme EM (voir section 4.2.1.2). En plus d'une estimation des paramètres inconnus, l'algorithme EM fournit aussi une classification $\mathbf{z}$ en offrant la possibilité de restaurer les données manquantes.
Comme mentionné au Chapitre 4, l'algorithme EM ne peut pas être appliqué directement aux champs aléatoires de Markov cachés et nécessite des approximations. On a présenté en section 4.3 différentes variantes possibles de l'algorithme EM quand il ne peut pas être applicable directement. Pour estimer les paramètres $\mathbf{\Psi}$ de notre modèle proposé pour la cartographie du risque, nous choisirons d'utiliser l'approximation de l'algorithme EM de type champ moyen pour ses performances et ses bons résultats en pratique.

5.3 Estimation des cartes de risque à l'aide de l'algorithme EM champ moyen

En cartographie du risque, l'objectif est de restaurer l'image $\mathbf{z}$ interprétée comme une classification du risque en K classes de risque, représentées par K étiquettes. Ces étiquettes sont inconnues et doivent être traitées comme des données manquantes. Nous devons affecter chaque unité géographique à une des K classes de risque. Pour cela, on considère le principe de (MPM) (voir 4.2.1.4) qui consiste à affecter chaque région i à la classe k qui permet de maximiser la probabilité $P(Z_i = k|\mathbf{y}, \mathbf{\Psi})$. De telles maximisations dépendent des paramètres $\mathbf{\Psi}$ du modèle qui sont, la plupart du temps, inconnus (ou partiellement inconnus quand une connaissance *a priori* est introduite) et doivent être estimés. Dans ce travail, nous

utilisons l'algorithme EM avec des approximations de type champ-moyen (voir section 4.4). Rappelons brièvement que le principe de ces algorithmes est de remplacer la distribution markovienne non calculable par une distribution factorisable pour laquelle l'algorithme EM peut être utilisé. Celles ci correspondent à l'une des plus simples approximations variationnelles et permettent de prendre en compte la structure markovienne tout en préservant les bonnes caractéristiques de l'algorithme EM.
Notons qu'en pratique, ce genre d'algorithme doit être étendu pour incorporer l'estimation de la matrice $\mathbb{B}$ et aussi pour inclure la structure irrégulière des graphes qui ne sont pas les grilles régulières de pixels de l'analyse d'image.

5.3.1 Mise en œuvre de l'algorithme EM champ moyen pour la cartographie du risque

Lorsqu'on est en présence d'un modèle de champ de Markov caché, deux difficultés apparaissent lors du calcul de l'espérance de la vraisemblance complète requise par l'algorithme EM. La constante de normalisation $W(\boldsymbol{\beta})$ (équation (4.3)) ainsi que les probabilités conditionnelles $P(z_i|\mathbf{y};\boldsymbol{\Psi})$ et $P(z_i, z_j|\mathbf{y};\boldsymbol{\Psi})$, pour j dans le voisinage $\mathcal{N}_i$ de i, ne peuvent pas être calculées directement. Comme précisé en section 4.3, l'approximation de type champ-moyen vise à approcher ces probabilités en négligeant les fluctuations des sites voisins du point i considéré. Cela peut être fait en fixant ses voisins $j \in z_{\mathcal{N}_i}$ à leur valeur moyenne. En général, nous parlons d'approximation de type champ moyen lorsque les valeurs z_i ne dépendent pas des valeurs z_j pour $j \in \mathcal{N}_i$ qui sont toutes fixées à des constantes (pas forcément à leur valeur moyenne). Le choix de ces constantes dénotées par $\tilde{\mathbf{z}} = \{\tilde{z}_1, \ldots, \tilde{z}_N\}$ n'est pas arbitraire mais satisfait certaines conditions de cohérence (voir section 4.3.1).
En appliquant l'approximation de type champ-moyen à notre modèle, la probabilité $P(z_i|\mathbf{y};\boldsymbol{\Psi})$ est approchée par la probabilité $P(z_i|\mathbf{y}, \tilde{\mathbf{z}}_{\mathcal{N}_i};\boldsymbol{\Psi})$ avec :

$$P(z_i|\mathbf{y}, \tilde{\mathbf{z}}_{\mathcal{N}_i};\boldsymbol{\Psi}) \propto Pois(y_i; n_i z_i^t\, \boldsymbol{\lambda})\; P(z_i|\tilde{\mathbf{z}}_{\mathcal{N}_i};\boldsymbol{\beta}), \tag{5.5}$$

et :

$$P(z_i|\tilde{z}_{\mathcal{N}_i};\boldsymbol{\beta}) \propto \exp\left(z_i^t(\boldsymbol{\alpha} + \mathbb{B}\sum_{j\in\mathcal{N}_i}\tilde{z}_j)\right). \tag{5.6}$$

La constante de normalisation n'est pas spécifiée mais son calcul est simple car il n'inclut qu'une somme sur K termes. Alors, pour tout $j \in \mathcal{N}_i$, la probabilité $P(z_i, z_j|\mathbf{y};\boldsymbol{\Psi})$ est approchée par le produit $P(z_i|\mathbf{y}, \tilde{\mathbf{z}}_{\mathcal{N}_i};\boldsymbol{\Psi})P(z_j|\mathbf{y}, \tilde{\mathbf{z}}_{\mathcal{N}_j};\boldsymbol{\Psi})$. Ces deux approximations sont faciles à calculer (équations (5.5) et (5.6)).

La forme générale de l'algorithme EM de type champ moyen est donnée en section 4.4. Rappelons que cet algorithme consiste à alterner deux étapes. L'étape (**NR**), qui consiste en un choix des valeurs du champ de voisins $\tilde{\mathbf{z}}$ et l'étape (**EM**) d'estimation des paramètres du modèle.
L'étape (**EM**) de l'algorithme EM de type champ moyen s'écrit pour notre modèle de la manière suivante :

– Étape (**E**) Calcul, en utilisant (5.5), des probabilités *a posteriori* approximées :

$$\text{Pour tout } i \in S \text{ et } k \in \{1, \ldots, K\}, \quad \tilde{t}_{ik}^{(q)} = P(Z_i = k | \mathbf{y}, \tilde{\mathbf{z}}_{\mathcal{N}_i}^{(q)}; \boldsymbol{\Psi}^{(q-1)})$$

– Étape (**M**) Mise à jour des paramètres selon les équations :

$$\text{Pour tout } k \in \{1, \ldots, K\}, \quad \lambda_k^{(q)} = \frac{\sum_{i \in S} \tilde{t}_{ik}^{(q)} y_i}{\sum_{i \in S} n_i y_i}, \tag{5.7}$$

et

$$\boldsymbol{\beta}^{(q)} = \arg\max_{\boldsymbol{\beta}} \sum_{i \in S} \sum_{k=1}^{K} \tilde{t}_{ik}^{(q)} \log \tilde{\pi}_{ik}^{(q)}(\boldsymbol{\beta}), \tag{5.8}$$

où $\tilde{\pi}_{ik}^{(q)}(\boldsymbol{\beta}) = P(Z_i = k | \tilde{\mathbf{z}}_{\mathcal{N}_i}^{(q)}; \boldsymbol{\beta})$.
Dans ce travail, nous allons nous concentrer sur la variante EM **champ-moyen**. Nous avons bien sûr examiné les performances des deux autres variantes de EM, l'EM **champ-modal** et l'EM **champ-simulé** présentés à la section 4.4. Contrairement à d'autres études, par exemple celles de (Celeux et al., 2003; Alfo et al., 2009), la seule variante qui mène à des résultats raisonnables pour le type de données considérées est l'EM **champ-moyen**. Ceci est probablement dû au fait que cette variante tend à plus lisser les données. Ce surlissage représente un avantage car il permet de mieux restaurer la structure spatiale des données que l'on traite.
Rappelons que pour l'algorithme EM champ-moyen, le champ des voisins $\tilde{\mathbf{z}}^{(q)}$ de l'Étape **(E)** est fixé à l'estimation en champ moyen de l'espérance de la distribution conditionnelle $P_G(\mathbf{z}|\mathbf{y}, \boldsymbol{\Psi}^{(q-1)})$.
Les mises à jour des paramètres λ_k sont disponibles sous forme analytique (équation (5.7)). En revanche, les paramètres *a priori* du champ aléatoire de Markov doivent être trouvés numériquement en utilisant la formule (5.8).
En calculant le gradient pour (5.8) avec $\boldsymbol{\beta} = (\boldsymbol{\alpha}, \mathbb{B})$; $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)$ et $\mathbb{B}$ de la forme définie en (5.1), les solutions $\boldsymbol{\alpha}$ et b de l'équation (5.8) satisfont :

$$\text{Pour } k = \{1, \ldots, K\}, \quad \sum_{i \in S} \tilde{\pi}_{ik}^{(q)}(\boldsymbol{\alpha}, b) = \sum_{i \in S} \tilde{t}_{ik}^{(q)} \ ,$$

et

$$\sum_{i\in S}\sum_{k=1}^{K}\left(\tilde{t}_{ik}^{(q)}-\tilde{\pi}_{ik}^{(q)}(\boldsymbol{\alpha},b)\right)\left(\sum_{j\in\mathcal{N}(i)}\tilde{z}_{j}^{(q)}(k)+1/2\,(\tilde{z}_{j}^{(q)}(k-1)+\tilde{z}_{j}^{(q)}(k+1))\right)=0\ .$$

Une fois que l'on a estimé les paramètres comme décrit ci-dessus, on peut calculer facilement les approximations $P(Z_i = k|\mathbf{y};\boldsymbol{\Psi})$ nécessaires à la classification de chaque région en utilisant le principe du MPM (4.4). Selon ce principe, l'unité géographique i est affectée à la classe k pour laquelle cette probabilité est la plus élévée.

5.3.2 La stratégie "Chercher/Lancer/Sélectionner"

La fonction de vraisemblance possède généralement plusieurs points stationnaires de différentes natures (incluant des maximas locaux ou encore des minimas). Par conséquence, la convergence de l'algorithme EM vers le maximum global peut dépendre fortement des valeurs initiales.

Pour appréhender ce problème d'initialisation, nous incluons l'algorithme EM décrit en section 5.3.1 dans une procédure plus générale. À l'exemple de Biernacki et al. (2003), nous adaptons une stratégie à trois étapes qui a pour objectif d'identifier les valeurs initiales permettant d'obtenir la plus grande vraisemblance en un intervalle de temps assez raisonnable. Ces étapes sont les suivantes :

Chercher. Cette étape vise à adopter une méthode qui permet de générer des ensembles de M valeurs initiales. Ces ensembles peuvent être générés aléatoirement ou en utilisant une stratégie d'initialisation. Nous recommandons d'utiliser, en particulier, la stratégie décrite en section 5.4 qui permet une exploration plus efficace de l'espace où vivent les paramètres.

Lancer. Pour chaque valeur initiale obtenue à l'issue de l'étape **Chercher**, on fait tourner l'EM **champ-moyen** décrit en 5.3.1 jusqu'à l'atteinte d'un critère de convergence, qui est décidé par l'utilisateur en fonction de son usage. Ce critère de convergence peut être, par exemple, un nombre fixe d'itérations ou la stabilisation de la log-vraisemblance.

Sélectionner. Retenir la solution qui donne la plus grande vraisemblance parmi l'ensemble des résultats obtenus en initialisant l'EM **champ-moyen** avec les M différentes valeurs initiales de l'étape **Chercher**.

Dans la section suivante, nous nous concentrons sur l'étape **Chercher** et décrivons la stratégie d'initialisation que nous proposons afin d'aider l'EM **champ-moyen** à converger rapidement vers la bonne solution.

Motivations pour la recherche de valeurs initiales pour l'EM champ-moyen : Le problème d'initialisation de l'algorithme EM pour les modèles de mélange, se pose tout de suite au praticien lors de l'application aux données. Quand les classes du mélange sont mal séparées, l'algorithme a encore plus de mal qu'usuellement à les estimer correctement. La notion de classes séparées pour les mélanges de Poisson est encore moins intuitive qu'elle ne l'est pour les autres modèles de mélanges, notamment les mélanges gaussiens. Cela est dû au fait que la moyenne d'une loi de Poisson correspond aussi à sa variance. En effet, regardons l'exemple de mélanges de Poisson présenté en figure 5.1 avec 5 classes dont les moyennes respectives ne sont pas très différentes. On voit que la distinction entre ces classes n'est pas

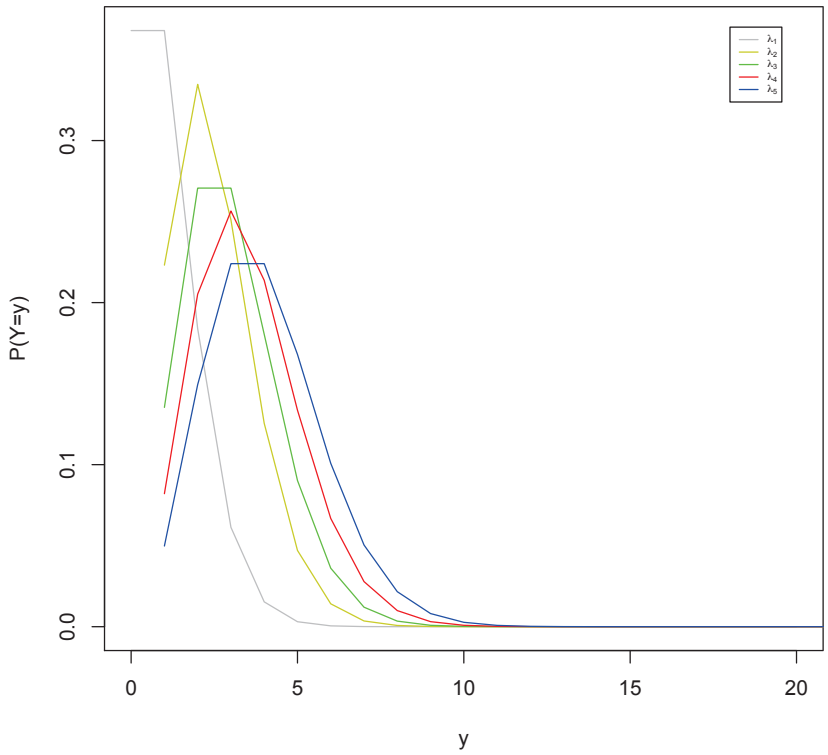


FIGURE 5.1 – . Exemple de mélanges de Poisson à 5 classes avec : $\lambda_1 = 1, \lambda_2 = 1.5, \lambda_3 = 2, \lambda_4 = 2.5$ et $\lambda_5 = 3$.

très visuelle. Il est assez difficile de pouvoir distinguer entre les différentes classes qui se chevauchent et qu'on a déjà du mal à séparer visuellement.
On peut noter que les valeurs considérées dans l'exemple présenté à la figure 5.1 sont très loin des petites valeurs qu'on retrouve dans le cas des maladies rares (voir chapitre 6).
On peut en déduire que si pour des valeurs assez grandes, comparées à celles attendues en épidémiologie animale, on n'arrive pas à séparer les classes, cela sera encore plus délicat pour

ces petites valeurs. Pour les mélanges de Poisson, il est important (plus qu'il ne l'est pour les mélanges gaussiens) d'avoir une stratégie capable de traiter les cas peu séparés. Cette stratégie doit pouvoir fournir des valeurs initiales raisonnables pour que l'algorithme EM puisse converger rapidement vers la bonne solution. Dans le cas des mélanges de Poisson, où la moyenne correspond aussi à la variance, il n'est pas très aisé de trouver de telles valeurs initiales.

5.4 Procédure proposée pour initialiser l'EM champ-moyen

La sensibilité de l'algorithme EM aux valeurs initiales est un problème connu et documenté dans la littérature. Pour appréhender ce problème, différentes stratégies d'initialisation ont été proposées et examinées pour les modèles de mélange gaussiens (McLachlan and Peel (2000) Chapitre2, Biernacki et al. (2003)).
Pour les mélanges Gaussiens, Biernacki et al. (2003) proposent de lancer dans un premier temps plusieurs fois les algorithmes Classification EM (CEM) (voir section 4.2.1.3), Stochastic EM (SEM) (voir section 4.2.1.3) à partir de différentes positions initiales aléatoires et de retenir l'ensemble des valeurs fournissant la plus grande log-vraisemblance. Une autre stratégie serait de lancer plusieurs fois l'algorithme EM lui même de positions aléatoires, de l'arrêter après un petit nombre d'itérations et de retenir l'ensemble des valeurs maximisant la log-vraisemblance. L'ensemble des valeurs retenues pour ces différentes stratégies servent comme valeurs initiales pour démarrer l'algorithme EM.
D'autres stratégies ont été proposées et évaluées par Karlis and Xekalaki (2003) pour, à la fois, les mélanges gaussiens et de Poisson. Dans cet article, les auteurs comparent un ensemble de stratégies basées sur des initialisations par des valeurs de paramètres. Ils présentent en particulier une variante de la méthode proposée par Finch et al. (1989). L'idée de cette technique est d'initialiser uniquement les proportions du mélange étant donné que les estimations des paramètres en découlent automatiquement. Les valeurs initiales des autres paramètres du mélange sont calculées selon une équation prenant en compte les valeurs initiales des proportions obtenues, la moyenne et la variance des données considérées. Ils reportent que cette méthode est efficace pour les mélanges de Poisson à deux composantes et recommandent d'utiliser, en pratique, une méthode alternative pour les mélanges de Poisson à 4 composantes ou plus.
La plupart des stratégies d'initialisation proposées peuvent être divisées en deux catégories : celles qui sont basées sur l'initialisation par les valeurs de paramètres et celles qui sont basées sur une initialisation par des partitions des données.

Pour la première catégorie, comme mentionné par Biernacki (2004) pour le cas de mélange gaussien standard non spatial, les étapes (**E**) et (**M**) produisent, à chaque itération, des valeurs estimées qui ne sont pas arbitraires mais qui sont liées par certaines équations. L'ensemble des estimations obtenues correspondent à une trajectoire de l'algorithme EM dans l'espace des paramètres. Si on tire des valeurs initiales aléatoirement, il se pourrait que ces valeurs ne soient pas dans la trajectoire de l'algorithme EM et cela peut mener à des calculs intensifs inutiles puisque la solution du maximum de vraisemblance appartient forcément à une des trajectoires de l'algorithme EM.
En ce qui concerne la deuxième catégorie, une partition est utilisée pour l'initialisation. Quand on utilise une partition aléatoire des données en K groupes, les valeurs initiales sont obtenues en estimant les paramètres pour chaque groupe.
Dans notre contexte, on estimera le niveau de risque pour chaque groupe qui est le rapport entre le nombres de cas observés du groupe et l'effectif de la population du même groupe. Ces valeurs sont, par construction, dans l'espace des trajectoires du EM mais cette manière d'initialiser a tendance à produire des valeurs très proches et n'explore pas efficacement l'espace des paramètres (voir section 5.4.2).
Dans le contexte de la cartographie du risque, le problème d'initialisation n'a pas été, à notre connaissance, abordée réellement. Alfo et al. (2009) lancent en premier l'algorithme CEM à partir de 500 positions initiales aléatoires. En second lieu, ils lancent leur EM champ moyen initialisé par la solution, parmi les 500, donnant la plus grande vraisemblance obtenue avec CEM. Ils rapportent que les résultats obtenus de cette manière sont satisfaisants. Pourtant Biernacki et al. (2003) évoquent que l'algorithme CEM est encore plus sensible à l'initialisation que l'algorithme EM lui même. Ils expliquent que l'initialisation avec l'algorithme CEM est généralement considérée comme une stratégie peu stable. Nous soupçonnons, que l'exemple présenté par Alfo et al. (2009) ne pose pas de problème d'initialisation et que n'importe quelle stratégie d'initialisation pourrait produire un résultat satisfaisant. Mais ce n'est qu'un cas particulier qui ne doit pas représenter le comportement usuel de l'algorithme.
Dans ce travail, nous abordons le problème d'initialisation en suivant l'idée développée par Biernacki (2004) et basée sur les trajectoires de l'algorithme EM.
Notre tâche est plus compliquée, comparée au travail de Biernacki (2004), parce qu'il nous faut en plus de valeurs initiales pour les paramètres du mélange, des valeurs initiales pour les paramètres du champ de Markov.
Nous présentons par la suite la procédure que nous proposons pour initialiser notre algorithme EM **champ-moyen**.

5.4.1 Initialisation des paramètres de risque à l'aide des trajectoires de l'algorithme EM

Nous proposons une adaptation de la stratégie d'initialisation prenant en compte les trajectoires de l'algorithme EM proposée par Biernacki (2004) pour les mélanges gaussiens. L'idée de cette stratégie est de s'assurer que les valeurs initiales explorent efficacement l'espace des trajectoires de l'algorithme EM. Il est intéressant de noter que quelque soit le modèle pour l'*a priori* spatial, une équation similaire à celle donnée dans Biernacki (2004), qui relie les valeurs des risques λ_k entre elles, peut être trouvée.
Soit $n = \sum_{i \in S} n_i$ l'effectif total de la population. À chaque itération (q) de l'algorithme EM, on dénote par $n_k^{(q)}$ la quantité :

$$n_k^{(q)} = \frac{\sum_{i \in S} \tilde{t}_{ik}^{(q)} n_i}{n},$$

qui peut être interprétée comme la proportion de la population appartenant au $k$ième niveau de risque. Il s'ensuit facilement que $\sum_{k=1}^{K} n_k^{(q)} = 1$.
En utilisant l'équation (5.7) pour l'estimation courante du risque, on a :

$$\sum_{k=1}^{K} n_k^{(q)} \lambda_k^{(q)} = \frac{\sum_{i \in S} y_i}{n} = \bar{\lambda}. \tag{5.9}$$

$\bar{\lambda}$ peut être interprété comme un risque moyen. Il a la propriété de ne dépendre que des données. À chaque itération de l'algorithme, l'estimation courante des paramètres $\lambda_k^{(q)}$ satisfait cette équation. Par conséquent, toutes les trajectoires de l'algorithme EM sont incluses dans l'espace défini par cette équation. L'idée est alors de tirer des valeurs pour les paramètres λ_k dans cet espace.
Une façon simple de le faire est de suivre les étapes de la procédure suivante :

Étape 1. On tire des valeurs de $n_k^{(0)}$ selon une distribution de Dirichlet

$$\mathcal{D}(\pi, \ldots, \pi),$$

avec tous les paramètres égaux à $\pi = 1$.

Étape 2. On tire, ensuite, uniformément et sans répétition des valeurs pour $\lambda_l^{(0)}$ dans l'échantillon $\{y_1/n_1, \ldots, y_N/n_N\}$ sauf pour une des composantes k du mélange (qui est tirée au hasard) qui vérifie l'équation :

$$\lambda_k^{(0)} = \frac{\bar{\lambda} - \sum_{l \neq k} n_k^{(0)} \lambda_l^{(0)}}{n_k^{(0)}}. \tag{5.10}$$

Ces étapes génèrent un vecteur de valeurs de paramètres aléatoires. Le nombre de valeurs initiales obtenues de cette facon sont décidées par l'utilisateur. Il faut néanmoins remarquer que comme dans Biernacki (2004) pour les paramètres des gaussiennes, l'équation (5.10) à l'étape 2 ne garantit en rien la positivité des paramètres $\lambda_k^{(0)}$. Si ce n'est pas le cas, l'échantillon tiré est écarté et on relance l'étape 1.

5.4.2 Illustration de la stratégie d'initialisation à l'aide des trajectoires de l'algorithme EM

Avant de présenter la procédure complète de recherche de valeurs initiales des paramètres de notre modèle, nous présentons dans cette partie un exemple pour illustrer la différence entre cette stratégie d'initialisation et les autres stratégies utilisant des valeurs de paramètres aléatoires ou encore des partitions aléatoires.
On va un peu plus détailler les stratégies décrites dans la partie introductive de la section 5.4 et on va les comparer dans cette section sur un exemple qu'on présentera juste après. Les trois stratégies d'initialisation comparées ici sont les suivantes :

1. La première stratégie consiste à initialiser l'algorithme EM par des valeurs de paramètres. Elle fait partie de la première catégorie (voir le début de la section 5.4) des méthodes d'initialisation proposées dans la littérature. Comme le montre la figure 5.2, l'idée générale de cette stratégie est de générer un ensemble de valeurs, à partir des données, qui servira par la suite pour initialiser l'algorithme EM. Plusieurs façons d'obtenir ces valeurs ont été proposées dans la littérature. Celles qui sont les plus communément utilisées reposent sur l'idée de lancer en un premier temps l'algorithme CEM, ou l'algorithme EM lui même pour un petit nombre d'itérations ou encore un autre algorithme. Ensuite, récupérer les valeurs de paramètres obtenues qui serviront comme point de départ pour l'algorithme EM. Une autre idée serait d'obtenir ces valeurs initiales aléatoirement à partir des données.
   Étant donnée la sensibilité du CEM et du EM comme discuté en (section 5.4), nous avons choisi pour notre exemple d'illustration d'obtenir ces valeurs initiales aléatoirement. Dans la suite de notre travail, cette stratégie d'initialisation sera dénotée S_r.

2. La deuxième stratégie d'initialisation est celle que l'on propose à la section 5.4.1. Elle fait partie de la première catégorie des méthodes d'initialisation (section 5.4). La forme générale de cette stratégie est donnée à la figure 5.3. Comme le montre cette figure, et contrairement à la stratégie décrite au premier point, les valeurs initiales des paramètres

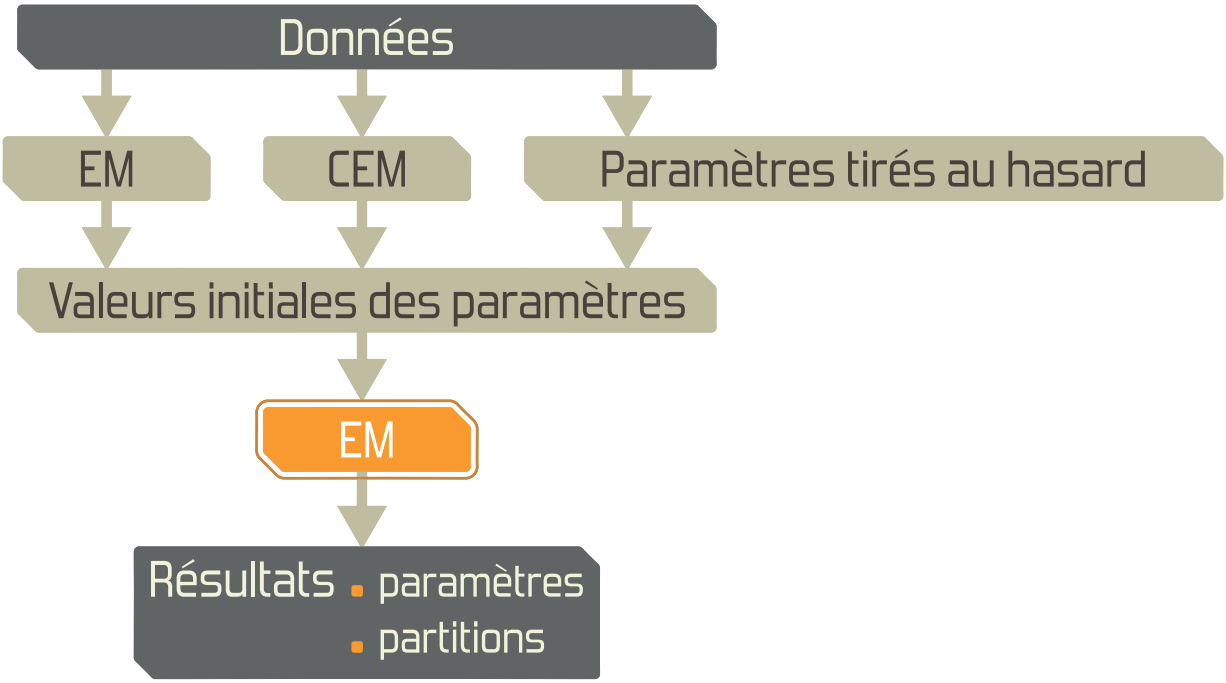


FIGURE 5.2 – Forme générale des stratégies d'initialisation de l'algorithme EM basées sur l'initialisation par des valeurs de paramètres.

sont obtenues directement à partir des données. Par la suite, cette stratégie sera notée S_t.

3. La troisième stratégie comparée dans cette section est une stratégie appartenant à la deuxième catégorie des stratégies d'initialisation. Contrairement aux deux autres stratégies décrites ci-dessus, et comme le montre la figure 5.4, l'idée générale de cette stratégie est d'initialiser l'algorithme EM avec des partitions obtenues d'une manière quelconque. Dans notre exemple d'illustration, les partitions pour l'initialisation sont obtenues au hasard. Nous appellons cette stratégie S_p.

Pour illustration, on prend un exemple simple de mélange de Poisson à deux classes pour lesquelles les vraies moyennes sont $\lambda_1 = 0.1$, $\lambda_2 = 0.2$ et les proportions de ces classes sont $\pi_1 = \pi_2 = 0.5$. On considère une centaine de sites $N = 100$ et on crée des valeurs pour les effectifs de la population en échantillonnant sans répétition parmi des entiers entre 10 et 109. Ensuite nous simulons, les nombres de cas y_i selon un mélange de Poisson à deux composantes à l'aide de l'équation (5.3). L'histogramme des valeurs de y_i obtenues est montré à la figure 5.5. La valeur correspondante de $\bar{\lambda}$ est aux alentours de 0.141.
Dans le cas de deux classes, il résulte de l'équation (5.9) que les trajectoires de l'algorithme

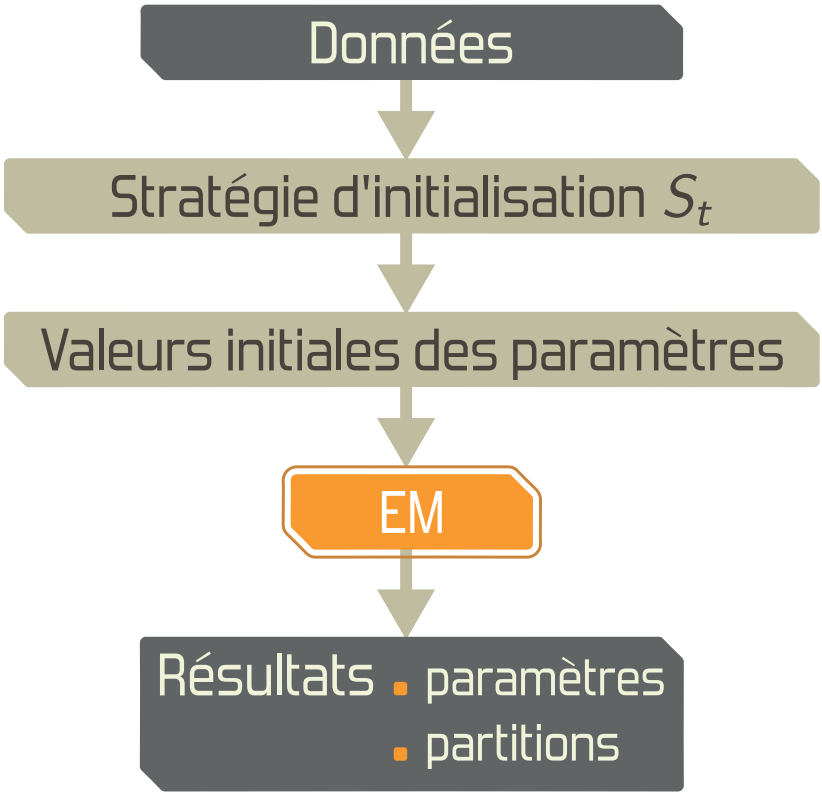


FIGURE 5.3 – Forme générale de la stratégie d'initialisation S_t proposée pour l'algorithme EM.

EM et donc les estimations du maximum de vraisemblance de λ_1 et de λ_2 sont contraintes à vivre dans la région grise de la figure 5.6 (a). Cette région est délimitée par les lignes verticales et horizontales définies respectivement par $\lambda_1 = \bar{\lambda}$ et $\lambda_2 = \bar{\lambda}$. Le point correspondant aux vraies valeurs $\lambda_1 = 0.1$ et $\lambda_2 = 0.2$ est marqué sur la figure 5.6 (a) par "X".
Nous montrons sur la figure 5.6 (a) les 100 valeurs initiales de λ_1 et λ_2 obtenues avec la stratégie S_r en tirant aléatoirement des valeurs selon la loi uniforme entre 0 et 0.4 et ordonnées de manière à ce que $\lambda_1 < \lambda_2$. On peut voir que sur les 100 points présentés à la figure 5.6 (a), il y en a seulement 42 qui se situent dans l'espace des trajectoires de l'algorithme EM. La plupart des points restants sont plus concentrés à la droite de cette région et en dehors de la délimitation. Il y a aussi quelques points qui se situent en bas de la région grise. On peut dire alors que les valeurs initiales obtenues de cette manière sont réparties d'une manière aléatoire dans l'espace et ne sont pas forcément que dans l'espace des trajectoires de l'algorithme EM.
Pour comparaison, nous montrons respectivement à la figure 5.6 (b) les 100 valeurs obtenues en utilisant notre stratégie S_t décrite en (section 5.4.1) et les 100 valeurs initiales obtenues

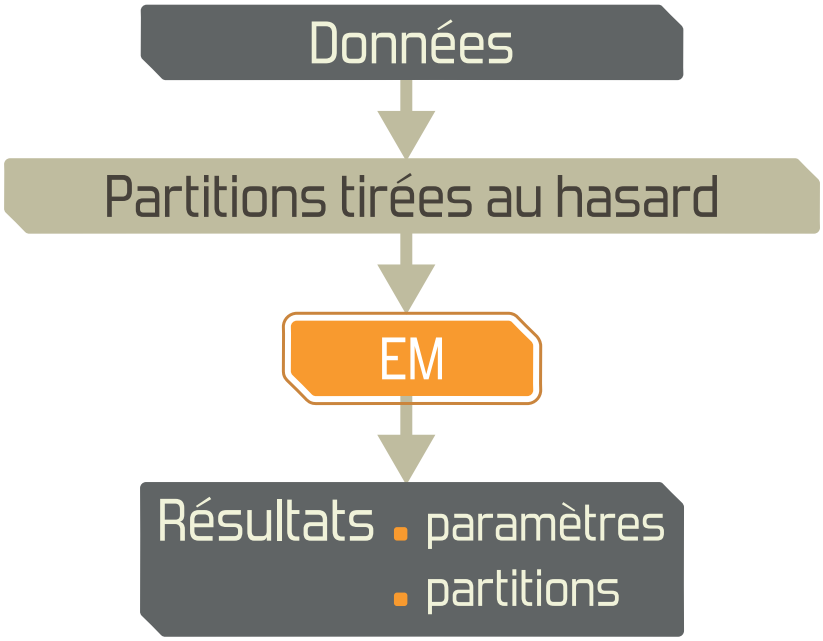


FIGURE 5.4 – Forme générale des stratégies d'initialisation de l'algorithme EM avec des partitions au hasard.

par l'initialisation S_p à la figure 5.6 (c). Comme prévu, les initialisations obtenues avec S_p ont tendance à produire des valeurs proches de $\bar{\lambda}$ et qui sont concentrées autour de ce point. Ces valeurs obtenues n'arrivent pas à explorer efficacement l'espace des paramètres. Nous pouvons voir qu'avec la stratégie S_p, on produit plusieurs fois les mêmes valeurs des paramètres. Cela est illustré par le fait que plusieurs points sont confondus sur la figure 5.6 (c).
Contrairement aux deux stratégies S_r et S_p, notre stratégie S_t explore mieux l'aire grise et on peut voir sur la figure 5.6 (b) que les points sont plus concentrés autour de la vraie valeur $(0.1, 0.2)$. On peut constater aussi que les 100 valeurs obtenues sont situées dans l'aire des trajectoires de l'algorithme EM et qu'aucun point n'est en dehors de cette région.
Nous avons constaté que ce phénomène est encore plus frappant quand on augmente le nombre des valeurs pour l'initialisation.
Pour mieux comprendre le comportement de l'algorithme EM pour les mélanges de Poisson vis-à-vis de l'initialisation et le besoin de bien choisir judicieusement le point de départ pour cet algorithme, nous montrons à la figure 5.7 le couple de valeurs (λ_1, λ_2) obtenues après une itération de l'algorithme EM initialisé par les trois différentes stratégies d'initialisation S_r, S_t et S_p.
Pour mieux visualiser les différences entre les différentes méthodes, on ne montre que 20 des

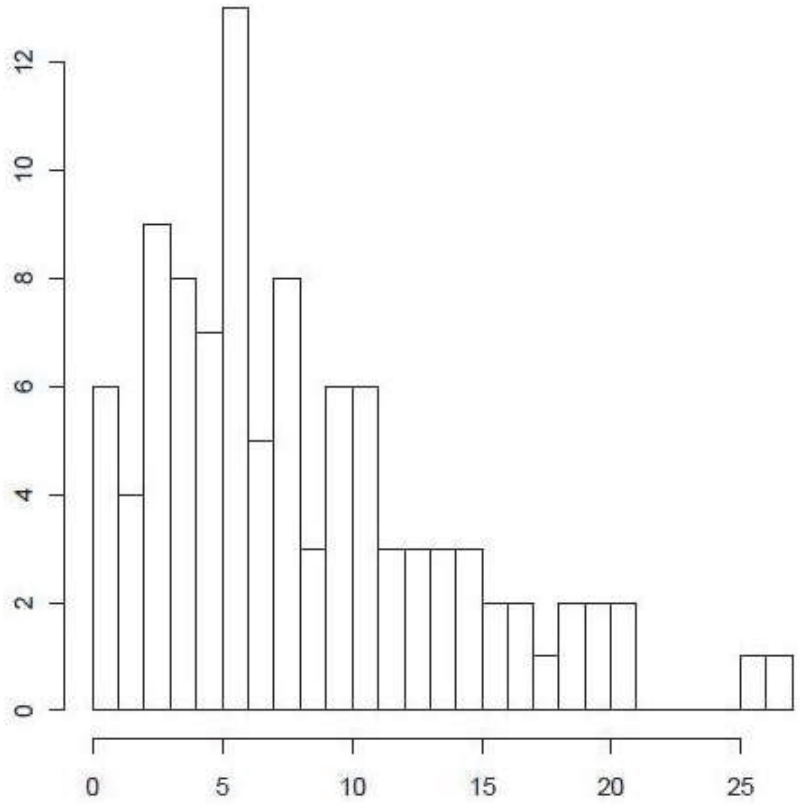


FIGURE 5.5 – Histogramme des 100 valeurs obtenues pour le mélange de Poisson à deux classes avec $\lambda_1 = 0.1$, $\lambda_2 = 0.2$, $b = 0$ et $\alpha_1 = \alpha_2 = 0.5$.

valeurs obtenues. La première remarque qu'on peut faire, concernant ces valeurs, est qu'elles sont toutes dans l'aire grise pour les différentes stratégies. Rappelons que cette aire grise représente l'espace des trajectoires de l'algorithme EM. Cela est normal que toutes ces valeurs soient, par construction, dans cette région. Ce qui nous importe, avec cette comparaison, c'est de voir si les initialisations obtenues par les différentes stratégies permettront à l'algorithme EM de converger vers la solution désirée ou pas. Sur la figure 5.7(a), on peut voir que les valeurs obtenues après une itération de l'algorithme EM, avec la stratégie S_r sont plus dispersées que celles obtenues par notre stratégie S_t comme présenté dans la figure 5.7(b). En effet, ces valeurs (obtenues par S_t) sont plus proches de la vraie valeur $(0.1, 0.2)$. On peut noter aussi, que sur les 20 points présentés, il y a moins de valeurs qui peuvent être considérées comme "loin" de la vraie valeur avec notre S_t qu'avec S_r.
Quant aux valeurs obtenues avec la stratégie S_p, on peut notamment remarquer sur la figure 5.6 (c) qu'elles sont concentrées autour du mauvais point. Ceci n'est pas très surprenant si on réexamine les valeurs initiales obtenues avec cette stratégie et présentées à la figure 5.6 (c).

Pour les deux stratégies S_r et S_p, on pourrait penser que parmi les 20 points, il y en a peu qui pourront aider l'algorithme EM à converger vers le solution désirée. Cet exemple illustre l'aptitude de notre stratégie S_t à produire des valeurs initiales proches des vraies valeurs. Ces valeurs initiales pourront ainsi aider l'algorithme EM à converger vers la solution désirée et en un temps plus réduit. Cela n'exclut en aucun cas le fait que les deux autres stratégies S_r et S_p sont capables de produire des valeurs initiales "raisonnables". En effet ces deux stratégies pourraient éventuellement, dans certains cas, générer des valeurs d'initialisation proches des vraies valeurs. Cependant, cela prendra certainement plus de temps, pour les cas complexes, qu'avec la stratégie S_t. Il ne faut pas perdre de vue que les données que nous étudions sont plus complexes que l'exemple présenté dans cette section. Les paramètres du mélange sont plus petits et le nombre de composantes est plus grand. Nous avons choisi cet exemple pour montrer que même dans un cas assez simple de mélange de Poisson, une stratégie assez "sophistiquée" peut être nécessaire.

5.4.3 Procédure complète de recherche de valeurs initiales

Pour complèter la procédure de recherche de valeurs initiales "raisonnables" pour les paramètres de notre modèle, il nous faut en plus proposer des valeurs initiales raisonnables des paramètres λ_k, proposer des valeurs initiales pour les paramètres du champ de Markov $\boldsymbol{\beta} = (\boldsymbol{\alpha}, \mathbb{B})$.
Quand la matrice $\mathbb{B}$ est réduite à la valeur b, comme pour le modèle semi-graduel que nous proposons à la section 5.2.1, notre stratégie complète d'initialisation se décompose en deux étapes :

Chercher 1. Générer M valeurs initiales pour $\boldsymbol{\lambda}^{(0)}$ à l'aide des deux étapes de la stratégie S_t décrite en 5.4.1.

Chercher 2 Pour chacune des valeurs $\boldsymbol{\lambda}^{(0)}$ obtenues à l'issue de l'étape **Chercher 1**, on pose $\boldsymbol{\alpha}^{(0)} = 0$ (cela traduit le fait qu'aucune des classes n'est favorisée) et $b^{(0)} = 1$. On fait tourner notre EM **champ-moyen** décrit en 5.3.1 avec la valeur de b fixée à 1 jusqu'à l'atteinte du critère de convergence choisi. Seules les valeurs de $\boldsymbol{\alpha}$ et $\boldsymbol{\lambda}$ sont mises à jour selon les équations (5.7) et (5.8). Nous proposons d'utiliser un critère de convergence basé sur la différence relative entre les vraisemblances calculées lors de deux itérations successives. Ce critère signifie que les deux étapes du EM **champ-moyen** sont répétées jusqu'à ce que la croissance relative de la log-vraisemblance entre deux itérations successives devienne plus petite qu'un certain seuil ε.

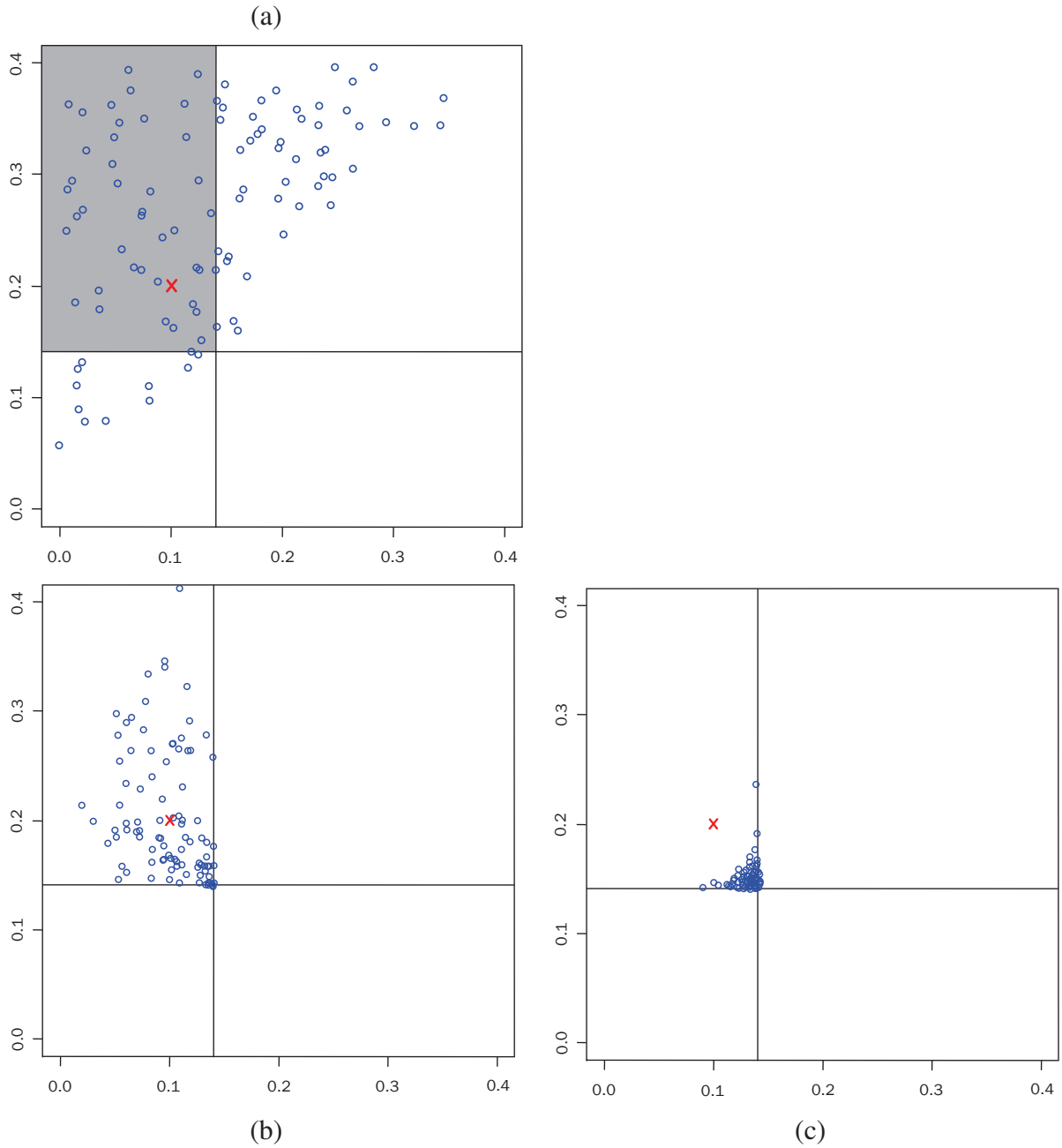


FIGURE 5.6 – Exemple de mélange de Poisson avec 2 classes. (a) 100 valeurs de (λ_1, λ_2) tirées au hasard entre 0 et 0.4 et ordonnées de manière à ce que $\lambda_1 < \lambda_2$; (b) 100 valeurs du couple (λ_1, λ_2) générées par la stratégie utilisant les trajectoires de l'algorithme EM ; (c) 100 valeurs de (λ_1, λ_2) obtenues à partir de partitions au hasard des données.

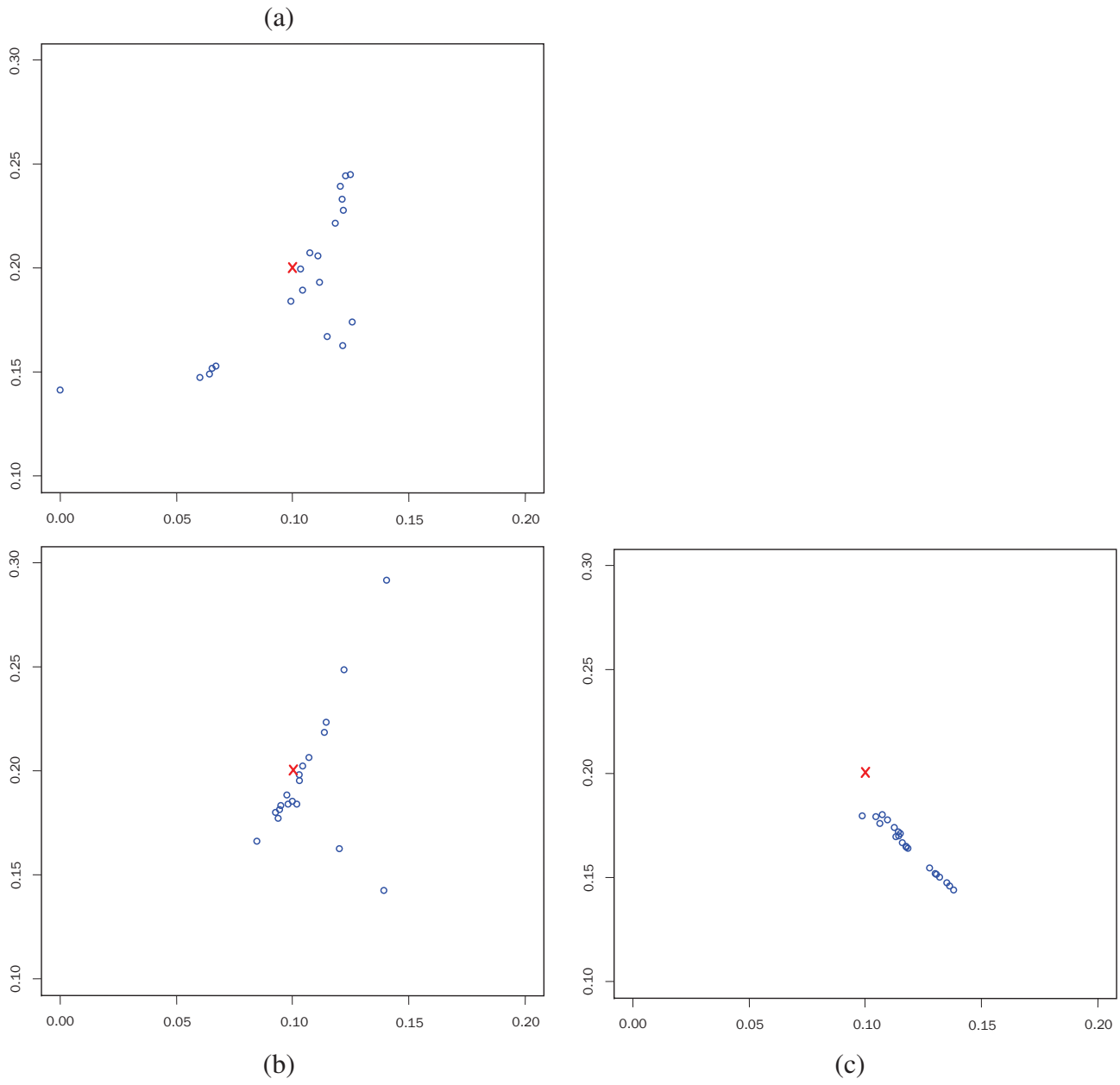


FIGURE 5.7 – 20 valeurs de (λ_1, λ_2) obtenues après 1 itération de l'algorithme EM initialisé avec : (a) : des valeurs de paramètres tirées au hasard selon S_r, (b) : des valeurs générées à partir de la stratégie d'initialisation S_t ; (c) : des valeurs obtenues des partitions au hasard des données avec S_p.

L'idée de l'étape **Chercher 2** est d'empêcher un comportement non désiré de l'algorithme dans le cas de données complexes ou très bruitées. Nous avons observé dans le cas de nos

FIGURE 5.8 – Mise en application de notre stratégie complète d'initialisation S_{tra}.

données simulées (voir la section 6.2) qu'en imposant une certaine quantité de structure spatiale, on peut empêcher l'algorithme de converger vers des solutions aberrantes. Typiquement, cela peut être fait en fixant $b = b^{(0)} = 1$ pour un certain nombre d'itérations avant de laisser tous les paramètres s'estimer avec notre EM **champ-moyen**. Nous avons choisi d'initialiser, à cette étape, le paramètre b à la valeur 1 suite aux résultats d'exploration que nous avons obtenus sur nos données simulées.

Cette stratégie est une solution simple que nous proposons pour contourner le problème des petites valeurs de risque et composantes de mélange de Poisson mal séparées. Il faut noter que

dans le contexte de la cartographie du risque, parler d'un mélange de Poisson est un raccourci à cause de l'introduction des effectifs de la population n_i dans l'équation (5.3). Cette modification fait que l'intuition que l'on a de ces modèles n'est pas aussi directe que celle que l'on a pour les mélanges de Poisson traditionnels.

Dans la suite de ce travail, cette stratégie sera appelée S_{tra}. Pour des cas plus simples, où les composantes du mélange de Poisson sont plus séparées, l'étape **Chercher 2** n'est, en général, pas nécessaire.

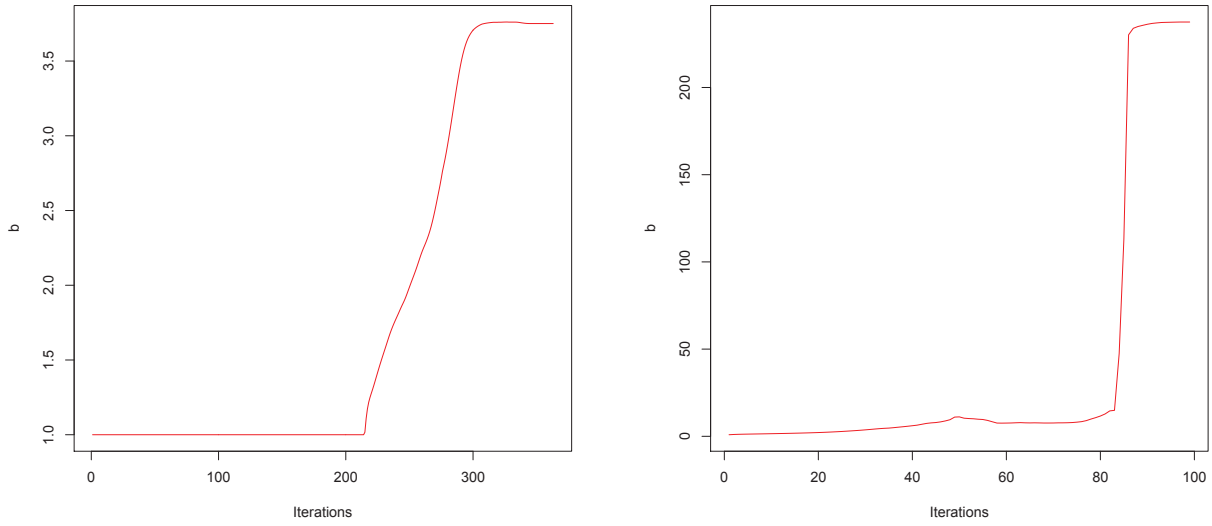
Pour illustrer l'utilité de l'étape **Chercher 2**, nous présentons les résultats obtenus pour un exemple "réaliste" proche des données réelles qui nous intéressent et qui seront présentées à la section 6.3. Nous montrons dans les figures 5.9 (a) et 5.10 (a) , une séquence typique de valeurs pour b et $\boldsymbol{\lambda}$ obtenues respectivement avec notre procédure complète d'initialisation S_{tra} comme décrit ci-dessus.

Nous présentons aussi, pour comparaison, dans les figures 5.9 (b) et 5.10 (b), l'ensemble des valeurs des paramètres b et $\boldsymbol{\lambda}$ obtenues respectivement avec la procédure complète et en omettant l'étape **Chercher 2** de notre procédure complète d'initialisation. Pour ce cas, ces paramètres sont obtenus en lancant notre EM **champ-moyen** en permettant à tous les paramètres de s'estimer sans contraintes. L'initialisation est obtenue comme auparavant grâce à l'étape **Chercher 1**. Pour ces deux procédures comparées, les valeurs initiales des paramètres sont les mêmes.

On observe clairement sur la figure 5.9 (b) que pour la procédure sans l'étape **Chercher 2** le paramètre b augmente rapidement vers de très grandes valeurs. Ce comportement tend à piéger l'algorithme dans une solution insignifiante avec une très grande interaction spatiale. Quant aux valeurs de b obtenues avec notre procédure complète d'initialisation S_{tra}, et comme présentée à la figure 5.9 (a), elles restent raisonnables. On peut remarquer aussi, comme illustré par la figure 5.9, que le nombre d'itérations nécessaires à l'atteinte du critère de convergence que l'on a choisi est plus grand avec notre procédure complète qu'avec celle-là en omettant l'étape **Chercher 2**. En réalité, même si l'étape **Chercher 2** peut être longue, nous avons constaté que la vitesse de convergence de notre EM **champ-moyen** reste très raisonnable pour nos données.

En ce qui concerne les valeurs des paramètres $\boldsymbol{\lambda}$, les deux procédures comparées (la complète et celle sans l'étape **Chercher 2**) produisent des valeurs assez comparables comme présenté à la figure 5.10 (a) ;(b). Cela est dû, probablement, au fait que les valeurs $\boldsymbol{\lambda}$ obtenues par notre stratégie S_t sont assez proches des vraies valeurs. En effet, on a remarqué que ces valeurs ne changent pas subitement dans l'étape **Chercher 2** de notre procédure complète S_{tra}, contrairement aux valeurs de b et $\boldsymbol{\alpha}$.

Les valeurs de $\boldsymbol{\alpha}$ restent également entre -4 et 1 quand la procédure d'initialisation est réali-



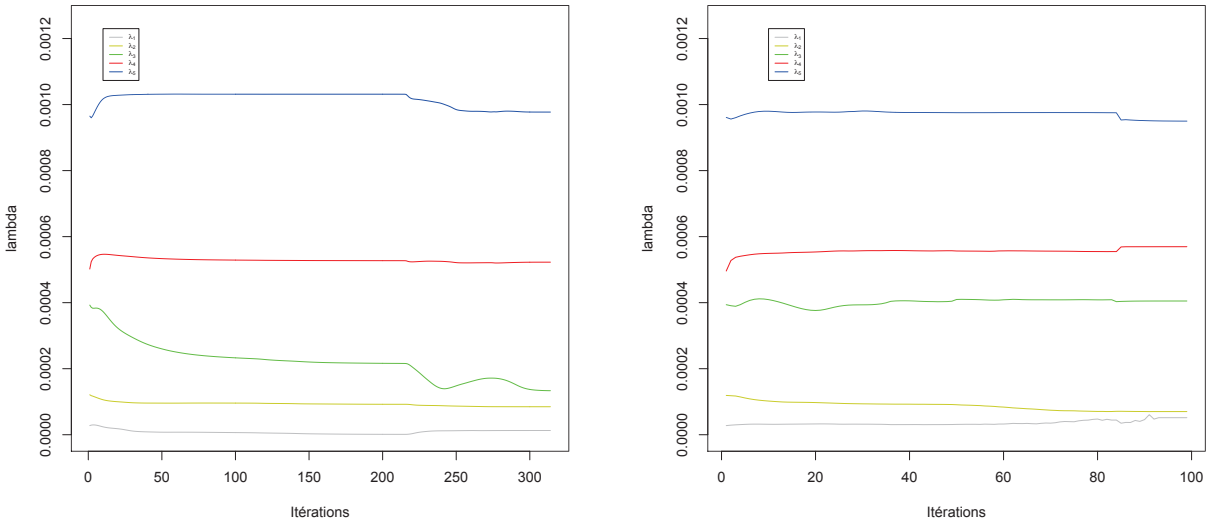
(a) (b)

FIGURE 5.9 – Paramètres obtenus pour un exemple simulé de champ de Markov avec 5 composantes. (a) les valeurs de b obtenues en utilisant notre procédure complète de recherche de valeurs initiales ; (b) : les valeurs de b obtenues en omettant l'étape **Chercher 2** de notre procédure.

sée complètement (avec l'étape **Chercher 2**), alors que dans le cas où cette étape est omise, la valeur $\alpha(1)$ (le paramètre $\boldsymbol{\alpha}$ pour la première classe) baisse rapidement vers une très petite valeur négative comme compensation pour la très grande valeur du paramètre b correspondante.
Notons aussi que ce genre de comportement pathologique est dû aux petites valeurs de risque utilisées dans cet exemple et qui sont en général les valeurs de risque attendus en épidémiologie animale.

5.5 Discussion

Dans ce chapitre nous avons présenté le modèle que l'on propose pour la cartographie du risque, basé sur une approche par champs de Markov cachés discrets. Le modèle proposé est une variante du modèle de Potts, qui est un modèle largement utilisé pour la classification en analyse d'image. La différence entre ce modèle et le semi-graduel que nous proposons réside



(a) (b)

FIGURE 5.10 – Paramètres obtenus pour un exemple simulé de champ de Markov avec 5 composantes. (a) les valeurs de $\boldsymbol{\lambda}$ obtenues en utilisant notre procédure complète de recherche de valeurs initiales ; (b) les valeurs de $\boldsymbol{\lambda}$ obtenues sans l'étape **Chercher 2** de notre procédure.

dans la construction de la matrice $\mathbb{B}$ qui traduit les hypothèses d'interaction entre sites. Pour le modèle semi-graduel, où la matrice $\mathbb{B}$ se réduit à un paramètre b, le paramètre d'interaction dépend de la différence entre classes de risque voisines.

Nous avons présenté aussi la mise en application de l'algorithme EM **champ-moyen** pour l'estimation des paramètres de ce modèle. Cette méthodologie proposée (modèle + algorithme EM champ moyen) nous permet d'obtenir simultanément l'estimation des différents niveaux de risque et la classification de chaque unité géographique à ces différents niveaux sans avoir recours à une étape *a posteriori* pour la classification. Certes, pour notre modèle on a recours à un critère tel que le MAP ou le MPM pour obtenir cette classification, mais cela est obtenu directement grâce aux $\tilde{t}_{ik}$ obtenus à l'issue de l'étape (**E**) du EM **champ-moyen**.

Comme nous l'avons mentionné tout au long de ce chapitre, les algorithmes de type EM donnent des résultats fortement dépendants de l'initialisation. Ce problème d'initialisation est connu par les praticiens et plusieurs travaux ont été conduits pour palier à ce problème. Dans le cadre des mélanges, ce problème est d'autant plus grand quand les composantes sont mal séparées. Pour les mélanges de Poisson, ce problème d'initialisation est encore plus complexe qu'il l'est pour les autres mélanges. Comme illustré dans l'exemple (voir figure 5.1)

de mélange de Poisson, la séparation des composantes avec des moyennes proches n'est pas évidente. Le problème avec la loi de Poisson est que la moyenne de la loi est aussi sa variance. Dans ce cas, il n'est pas très intuitif de décider à partir de quelle valeur on peut considérer que les composantes sont bien séparées ou pas. Nous nous sommes rendus compte lors de la mise en pratique de notre modèle, qu'il y avait un important problème d'initialisation dû sans doute à la surdispersion de la loi de Poisson mais aussi aux petites valeurs de moyennes de classes (niveaux de risque) auxquelles nous sommes confrontés lors des études en épidémiologie animale ou encore en épidémiologie humaine pour les maladies rares.
Nous avons essayé de mettre en place une stratégie d'initialisation capable de prendre en compte ce problème d'initialisation et d'agir en conséquence pour aider notre EM **champ-moyen** à converger vers une solution raisonnable afin de nous fournir une bonne classification et une estimation raisonnable des paramètres de notre modèle.
Avant de proposer la stratégie S_{tra}, nous avons essayé d'initialiser notre EM **champ-moyen** avec les méthodes usuellement utilisées dans la littérature. Nous avons constaté que les initialisations basées sur des partitions de données n'est pas la meilleure stratégie pour nos données. En effet, tous les algorithmes, tels que le k-means (MacQueen, 1967) ou encore le CEM, qu'on a essayé pour obtenir une pré-classification pour initialiser notre EM champ-moyen ont du mal à obtenir des partitions raisonnables. En effet, dûe aux petites valeurs de risques atendues en l'épidémiologie des maladies rares qui résulte en un excés de zéros dans nos données, l'algorithme k-means a du mal à faire la différence entre les vrais zéros et les zéros structurés et donc a du mal à avoir une pré-classification raisonnable. De même pour les algorithmes CEM et EM qui ont tendance à vider des classes et la pré-classification obtenue ne représente en aucun cas une bonne initialisation. En se basant sur ces remarques, nous avons orienté notre recherche de méthodes d'initialisation vers celles basées sur des valeurs de paramètres plutôt que sur des partitions des données. Nous avons essayé d'illustrer cela par un exemple simple présenté à la section 5.4.2.
La stratégie S_{tra} qu'on propose est une adaptation de la stratégie utilisant les trajectoires de l'algorithme EM pour initialiser les mélanges indépendants à un cas de systèmes dépendants que sont les champs de Markov cachés.
Cette stratégie est divisée en deux étapes qui permettent d'avoir un ensemble de paramètres initial complet pour les paramètres du mélange et ceux du champ de Markov. La première étape de S_{tra} consiste à trouver des valeurs initiales pour les paramètres du mélange selon les trajectoires de l'algorithme EM. La deuxième étape permet de trouver des paramètres $\boldsymbol{\alpha}$ et des nouveaux paramètres pour $\boldsymbol{\lambda}$ en prenant en compte la structure spatiale modélisée par le champ de Markov. Dans cette deuxième étape, nous initialisons $b^{(0)} = 1$ et le paramètre $\boldsymbol{\alpha} = 0$. Ce choix de la valeur initiale pour $\boldsymbol{\alpha}$ est justifiée par le fait qu'on a aucune idée sur la

proportion des différentes classes de risque et qu'on ne veut favoriser aucune des classes par rapport à d'autres. Pour la valeur initiale de b, le choix a été décidé au vu des résultats obtenus sur nos données simulées. La stratégie S_{tra} que nous proposons peut être appliquée sur d'autres type de données mais pas forcément avec ces choix de valeurs initiales. Ces valeurs peuvent être ajustées selon les besoins de l'utilisateur et en fonction des données traitées.

Nous avons présenté l'intérêt d'avoir une procédure complète d'initialisation pour les paramètres et avons illustré par un exemple de données simulées, proches des vraies données, l'utilité de l'étape 2 de notre stratégie S_{tra}. Cette étape permet d'avoir une valeur raisonnable du paramètre b qui représente le paramètre d'interaction entre les sites.

Nous avons montré à la section 5.4.2 que la stratégie S_t pour les mélanges indépendants peut explorer efficacement l'espace des paramètres du mélange. On s'attend alors à ce que la stratégie S_{tra}, qui est une extension de la stratégie S_t pour le cadre spatial, produise des valeurs initiales raisonnables pour notre EM **champ-moyen** et à ce qu'elle lui permette de converger vers la bonne solution.

Nous montrons dans le chapitre 6 les résultats obtenus en initialisant l'algorithme EM champ-moyen avec S_{tra} sur des exemples de données dépendantes et comparons ses performances à d'autres stratégies.

CHAPITRE 6

Application aux données

Sommaire

La motivation principale de notre travail, comme expliqué dans les parties précédentes, est de traiter les maladies rares pour lesquelles les nombres de cas enregistrés sont faibles. Pour ce type de maladies, moins de 10 cas sont observés pour une population de quelque milliers. Notre objectif principal, et qui est plus lié au contexte animal qu'humain, est la classification des niveaux de risque de ces maladies. Mais cela n'empêche, en aucun cas, d'utiliser le modèle *semi-graduel* que l'on propose (voir section 5.2) et la stratégie d'initialisation S_{tra} (section 5.4) dans le contexte des maladies humaines rares.
Afin d'illustrer la performance de ce modèle et la procédure d'initialisation proposée, nous avons effectué des tests sur des données simulées qui ont des caractéristiques semblables à la maladie de l'Encéphalopathie Spongiforme Bovine (ESB) en France. Cette maladie rare,

représente l'application sur données réelles pour ce travail. Dans ce chapitre, nous présentons les résultats obtenus par notre méthodologie (modèle + stratégie d'initialisation) sur les données simulées (section 6.2) et les données de l'ESB (section 6.3).
Nous comparons notre modèle *semi-graduel* à d'autres modèles de la même famille à la section 6.2.1, notamment au modèle de Potts standard qui est, comme dit auparavant, communément utilisé pour la classification dans d'autres domaines. Nous comparons aussi notre stratégie d'initialisation S_{tra} à d'autres stratégies usuellement utilisées pour démarrer l'algorithme EM aux sections 6.2.2 et 6.2.3.

6.1 Préliminaires

Dans toutes nos illustrations, que ce soit pour les données simulées ou réelles, la structure sous-jacente est dérivée du territoire Français. En général, les données sont naturellement groupées en unités administratives. Le nombre de voisins est alors très variable et la répartition des unités n'est absolument pas régulière. Pour la carte de la France, le nombre de voisins est très fluctuant, jusqu'à 13 voisins pour les cantons par exemple, ce qui résulte en des frontières de largeurs variables. On choisit de transformer les données en les groupant sur une grille régulière définie. Ainsi nous avons utilisé un maillage de la France en 1264 hexagones de largeur 23 km chacun (450 km^2). La structure de voisinage est basée sur les hexagones adjacents et les unités géographiques ont, généralement, 6 voisins sauf pour les unités du bord. En effet, étant donné qu'on se limite à la carte de la France, les frontières avec la Belgique ou l'Espagne par exemple ont moins de voisins "Français" que les autres unités qui sont plus au centre. Pour chaque hexagone, l'effectif de la population n_i est fixée à la population bovine en France pour les années 2001-2005, qui est considérée comme stable. Les effectifs n_i varient entre 0 et 32039. La carte de la population bovine et l'histogramme des effectifs de population sont présentés à la figure 6.1.
Nous considérons, ensuite, différentes observations $\mathbf{y}$ suivant qu'elles sont générées (voir section 6.2.1) ou observées (section 6.3.1). Pour les données simulées, les nombres de cas seront générés selon la distribution de Poisson (donnée par l'équation (5.3)) et des valeurs de risque λ connues. On considérera des valeurs proches de celles des risque attendues pour l'ESB. Quant aux données réelles, les nombres de cas seront ceux enregistrés par les autorités dans les bases de données.
Pour illustrer la performance de la stratégie d'initialisation S_{tra} présentée en section 5.4.3, nous la comparons avec deux autres stratégies d'initialisation et présentons les résultats obtenus sur données simulées à la section 6.2, pour l'estimation des niveaux de risque et la

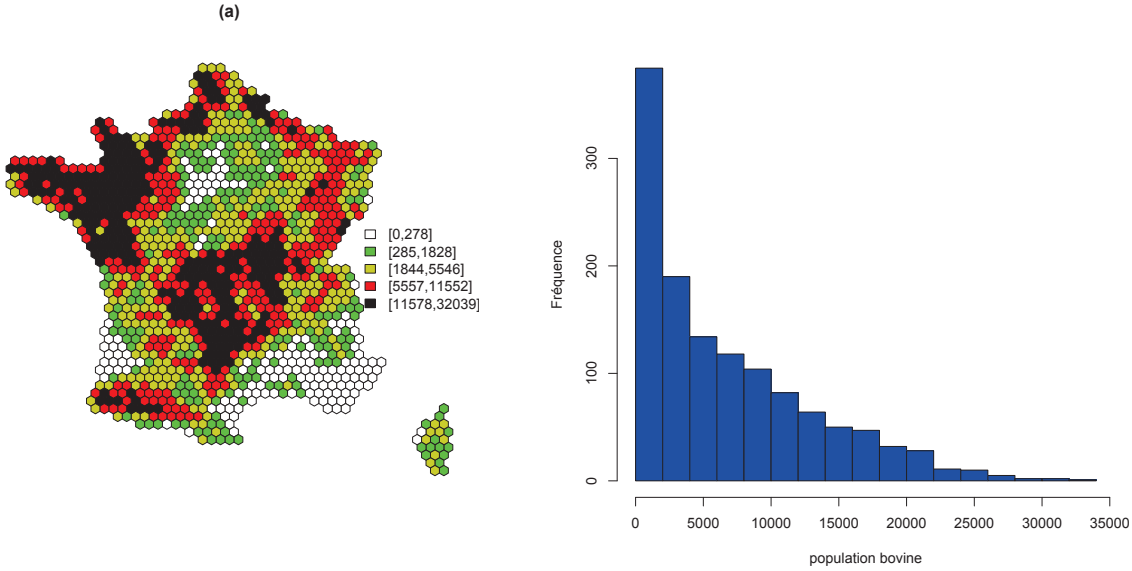


FIGURE 6.1 – Données démographiques. (a) : Carte de la population des vaches en France. (b) : histogramme des effectifs de la population des vaches.

classification des unités dans ces niveaux de risque. Les trois stratégies comparées sont :

- **Stratégie** S_{tra} : on génère M valeurs initiales pour tous les paramètres du modèle en utilisant les propriétés des trajectoires de l'algorithme EM et la procédure complète de recherche de valeurs initiales décrite en section 5.4.3. On lance ensuite notre EM **champ-moyen** initialisé par chaque ensemble de paramètres générés jusqu'à la convergence. L'ensemble des valeurs donnant la plus grande vraisemblance est retenu.
- **Stratégie** S_{rand} : cette stratégie diffère de la stratégie précédente dans la manière de générer les M valeurs initiales. Les M valeurs initiales pour λ sont générées selon une loi uniforme, typiquement entre les valeurs 0 et 1, qui correspondent aux valeurs minimum et maximum que peut prendre le rapport y_i/n_i pour notre exemple de cartographie. Quant aux paramètres $\boldsymbol{\alpha}$ et b, on les initialise respectivement aux valeurs 0 et 1 par analogie avec S_{tra} afin que les deux stratégies soient comparables. On lance ensuite notre EM **champ-moyen** initialisé par cet ensemble de paramètres jusqu'à sa convergence. Le résultat correspondant à la plus grande vraisemblance, sur les M résultats obtenus, est retenu.
- **Stratégie** S_{EMM} : on génère M valeurs initiales pour les paramètres λ selon une loi uniforme. les paramètres $\boldsymbol{\alpha}$ sont initialisés, comme pour les deux autres stratégies, à 0, tandis que le paramètre b est fixé à 0 pour cette stratégie. Cette initialisation nous

ramène à un système indépendant pour lequel l'algorithme standard EM pour les mélanges indépendants est lancé pour chaque ensemble de M valeurs initiales jusqu'à sa convergence. Le résultat correspondant à la plus grande vraisemblance est retenu. Les valeurs des paramètres estimés de ce résultat sont retenues et on lance notre EM **champ-moyen** initialisé par ces valeurs.

Les deux stratégies S_{tra} et S_{rand} correspondent à la procédure Chercher/Lancer/Sélectionner décrite en (5.3.2). La différence entre ces deux stratégies réside dans la manière de génèrer les valeurs initiales pour les paramètres λ. Comme décrit ci-dessus, ces paramètres sont obtenus d'une manière aléatoire pour S_{rand} alors qu'ils sont générés selon les trajectoires de l'algorithme EM et une procédure permettant de prendre en compte l'information spatiale présente dans les données pour S_{tra}. Quant à S_{EMM}, elle ne correspond pas tout à fait à la procédure Chercher/Faire-tourner/Choisir. En effet, cette stratégie diffère des deux autres parce que l'étape Chercher fait appel à l'algorithme EM standard pour les mélanges indépendants au lieu du EM **champ-moyen** qui tient compte de la dépendance spatiale entre les sites. Cette stratégie représente une des méthodes communément utilisées pour apréhender le problème d'initialisation. En particulier, elle ressemble à celle utilisée par Alfo et al. (2009) où l'algorithme CEM non spatial (voir section 4.2.1.3) est utilisé au lieu de l'algorithme EM non spatial dans S_{EMM}. Nous choisissons d'utiliser l'algorithme EM car CEM est connu pour être lui même moins stable que l'algorithme EM. Ceci est documenté par Biernacki et al. (2003) et nous l'avons également constaté en testant l'algorithme CEM sur nos données simulées.
Ces trois stratégies comparées font partie de la première catégorie des méthodes d'initialisations (voir section 5.4) qui est illustrée par la figure 5.2 en remplaçant l'algorithme EM par l'EM **champ-moyen**. Notre choix de présenter et comparer ces trois stratégies n'est pas aléatoire mais dû au fait que les autres méthodes d'initialisation usuellement utilisées dans la littérature (voir section 5.3.2) ont du mal à obtenir des initialisations raisonnables sur nos données.

6.2 Données simulées

Dans cette section nous présentons les données simulées qu'on a utilisées pour illustrer la performance du modèle *semi-graduel* et de la stratégie d'initialisation S_{tra} comparés à d'autres modèles et d'autres stratégies décrites à la section 6.1.

6.2.1 Description des données

Nous considérons dans cette partie, deux cartes de risque synthétiques avec respectivement 3 et à 5 classes (voir figure 6.2). Pour l'exemple avec à 3 classes, les niveaux de risque

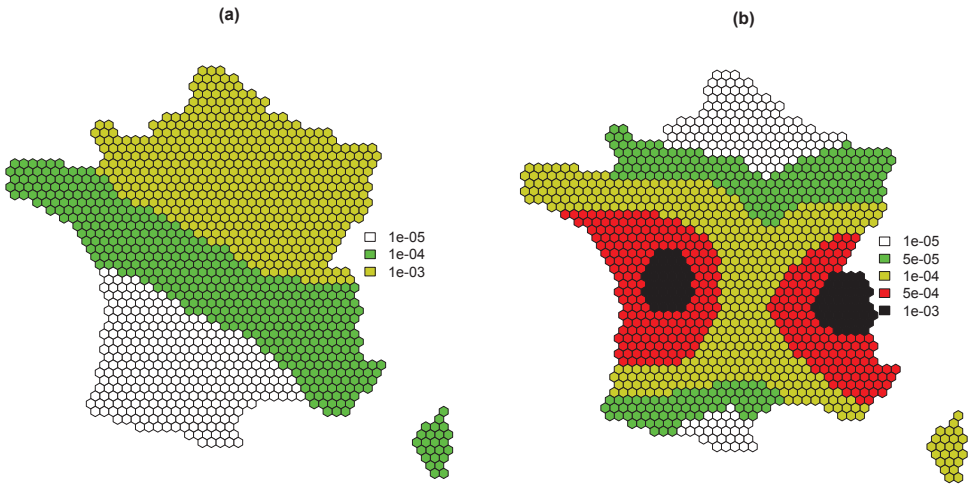


FIGURE 6.2 – Cartes de risque pour la simulation de données. (a) : "vraie" carte de risque pour l'exemple à 3 classes ; (b) : "vraie" carte de risque pour l'exemple à 5 classes.

sont choisis égaux à : $\lambda_1 = 1e^{-5}$ (faible), $\lambda_2 = 1e^{-4}$ (moyen) et $\lambda_3 = 1e^{-3}$ (fort). Pour l'exemple avec à 5 classes, les niveaux de risque sont : $\lambda_1 = 1e^{-5}$ (très faible), $\lambda_2 = 5e^{-5}$ (faible), $\lambda_3 = 1e^{-4}$ (moyen), $\lambda_4 = 5e^{-4}$ (fort) et $\lambda_5 = 1e^{-3}$ (très fort).
À partir des effectifs de la population n_i, des "vraies" valeurs de risque ci-dessus et des cartes de risque avec des classes connues, on peut simuler facilement des nombres de cas y_i selon la distribution de Poisson donnée par l'équation (5.3). Des exemples de nombres de cas simulés de cette façon pour les exemples à 3 classes et à 5 classes sont présentés à la figure 6.3.

6.2.2 Comparaison entre différentes stratégies d'initialisation pour l'exemple à 3 classes

Dans cette partie nous comparons les résultats obtenus avec les différentes stratégies d'initialisation S_{tra}, S_{rand} et S_{EMM} sur les données simulées présentées à la figure 6.3 (a). La performance de chaque stratégie est évaluée en considérant deux indicateurs concernant d'une part la classification obtenue et d'autre part l'estimation des valeurs de risque.
Pour le premier indicateur, nous considérons, pour chaque classe, le coefficient de similarité

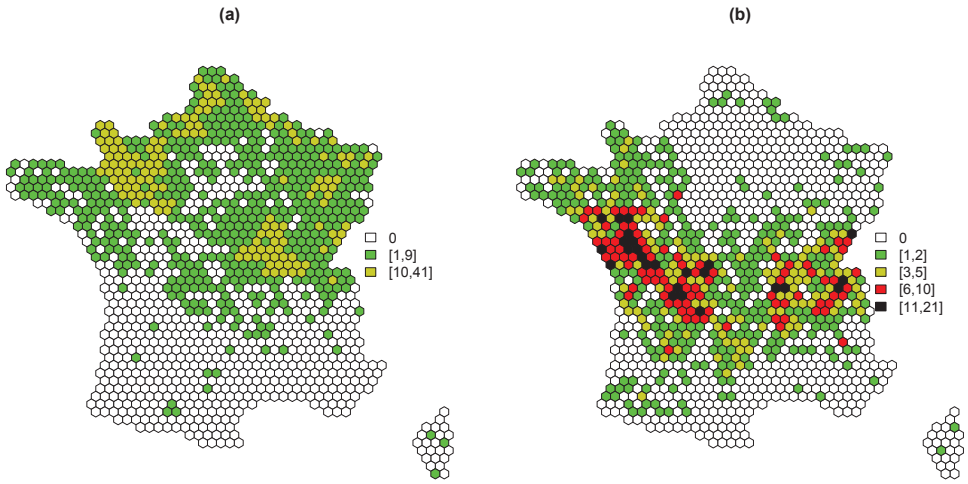


FIGURE 6.3 – Cartes de données simulées . (a) exemple de nombre de cas simulés pour l'exemple à 3 classes ; (b) exemple de nombre de cas simulés pour l'exemple à 5 classes.

de Dice (DSC) (Dice, 1945). Ce coefficient mesure le recouvrement entre la vraie classification et la segmentation qui résulte du modèle utilisé. Le DSC est donné par :

$$d_k = \frac{2TP_k}{2TP_k + FN_k + FP_k} \tag{6.1}$$

avec :

- TP_k (Vrais Positifs) : le nombre des sites qui sont estimés dans la classe k dans la segmentation résultante et qui sont effectivement dans la classe k dans la vraie classification (ceux qui ont été bien classés par le modèle).
- FP_k (Faux Positifs) : les sites qui sont dans la classe k dans la segmentation estimée mais qui ne sont pas dans la classe k dans la vraie segmentation.
- FN_k (Faux Négatifs) : les sites qui ne sont pas dans la classe k dans la segmentation estimée alors qu'ils le sont dans la vraie segmentation.

Le coefficient d_k prend ses valeurs dans $[0, 1]$. Si ce coefficient est égal à 1 alors le recouvrement est parfait. Cela implique que tous les sites sont bien classés.

6.2.2.1 Estimation des valeurs de risque et classifications pour un exemple simulé

Pour toutes les comparaisons présentées dans cette section et la suivante, nous avons utilisé $M = 1000$ valeurs initiales pour démarrer notre EM **champ-moyen**. Le résultat présenté pour chaque stratégie est le meilleur résultat, sur les 1000, obtenu en terme de vraisemblance. En ce qui concerne le nombre de classes K, nous le considérons, dans cette section, fixe au

cours de l'algorithme et égal à sa vraie valeur. Mentionnons que nous avons aussi appliqué un critère de sélection du nombre de classes et les résultats de ce choix sont présentés à la section 6.2.4.

Nous montrons à la table 6.1, le risque estimé et les valeurs du DSC calculées selon l'équation (6.1) pour chaque stratégie et pour chacune des classes de risque.

On peut remarquer que pour les unités à risque faible que la stratégie S_{rand} sous-estime le

Vraie valeur du risque	Strategie	DSC	Risque estimé
faible	S_{rand}	0.97	6.86 e^{-06}
$1e^{-05}$	S_{EMM}	0.71	3.07 e^{-05}
	S_{tra}	0.84	1.11 e^{-05}
moyen	S_{rand}	0.97	9.61 e^{-05}
$1e^{-04}$	S_{EMM}	0.75	9.33 e^{-05}
	S_{tra}	0.86	9.12 e^{-05}
fort	S_{rand}	0.99	1.02 e^{-03}
$1e^{-03}$	S_{EMM}	0.99	1.02 e^{-03}
	S_{tra}	1	9.87 e^{-04}

TABLE 6.1 – Exemple à 3 classes . Le coefficient de similarité de Dice (DSC) et le risque estimé pour chaque classe en utilisant notre modèle et l'algorithme **EM champ-moyen** initialisé par S_{rand}, S_{EMM} et S_{tra}.

risque ($\hat{\lambda}_1 = 6.86e^{-06}$), tandis qu'avec la stratégie S_{EMM}, le risque est surestimé avec une valeur estimée à $\hat{\lambda}_1 = 3.03e^{-05}$. La stratégie S_{tra} estime la valeur de risque à $\hat{\lambda}_1 = 1.11e^{-05}$ qui est légèrement supérieure à la vraie valeur $\lambda_1 = 1e^{-05}$. On peut dire, pour cet exemple, que la stratégie S_{tra} estime mieux le niveau de risque pour les régions à risque faible.

Quant aux régions à risque moyen, correspondant à la valeur $\lambda_2 = 1e^{-04}$, les trois stratégies sous-estiment légèrement le risque. On peut voir, sur la table 6.1, que S_{tra} sous-estime un peu plus le risque que les deux autres stratégies avec $\hat{\lambda}_2 = 9.12e^{-05}$ contre $\hat{\lambda}_2 = 9.33e^{-05}$ pour S_{EMM} et $\hat{\lambda}_2 = 9.61e^{-05}$ pour S_{rand}. Pour les régions considérées à risque fort, les stratégies S_{rand} et S_{EMM} estiment le risque à la même valeur ($\hat{\lambda}_3 = 1.02e^{-03}$) qui est légèrement au dessus de la vraie valeur ($\lambda_3 = 1e^{-03}$). Quant à S_{tra}, la valeur estimée ($\hat{\lambda}_3 = 9.87e^{-04}$) est un peu en dessous de la vraie.

En ce qui concerne la classification obtenue pour chaque stratégie et illustrée par la figure 6.4, on peut voir "visuellement" sur la figure 6.4 (b) que la stratégie S_{rand} recouvre presque parfaitement la vraie classification. On le voit surtout pour la classe moyenne où la région du

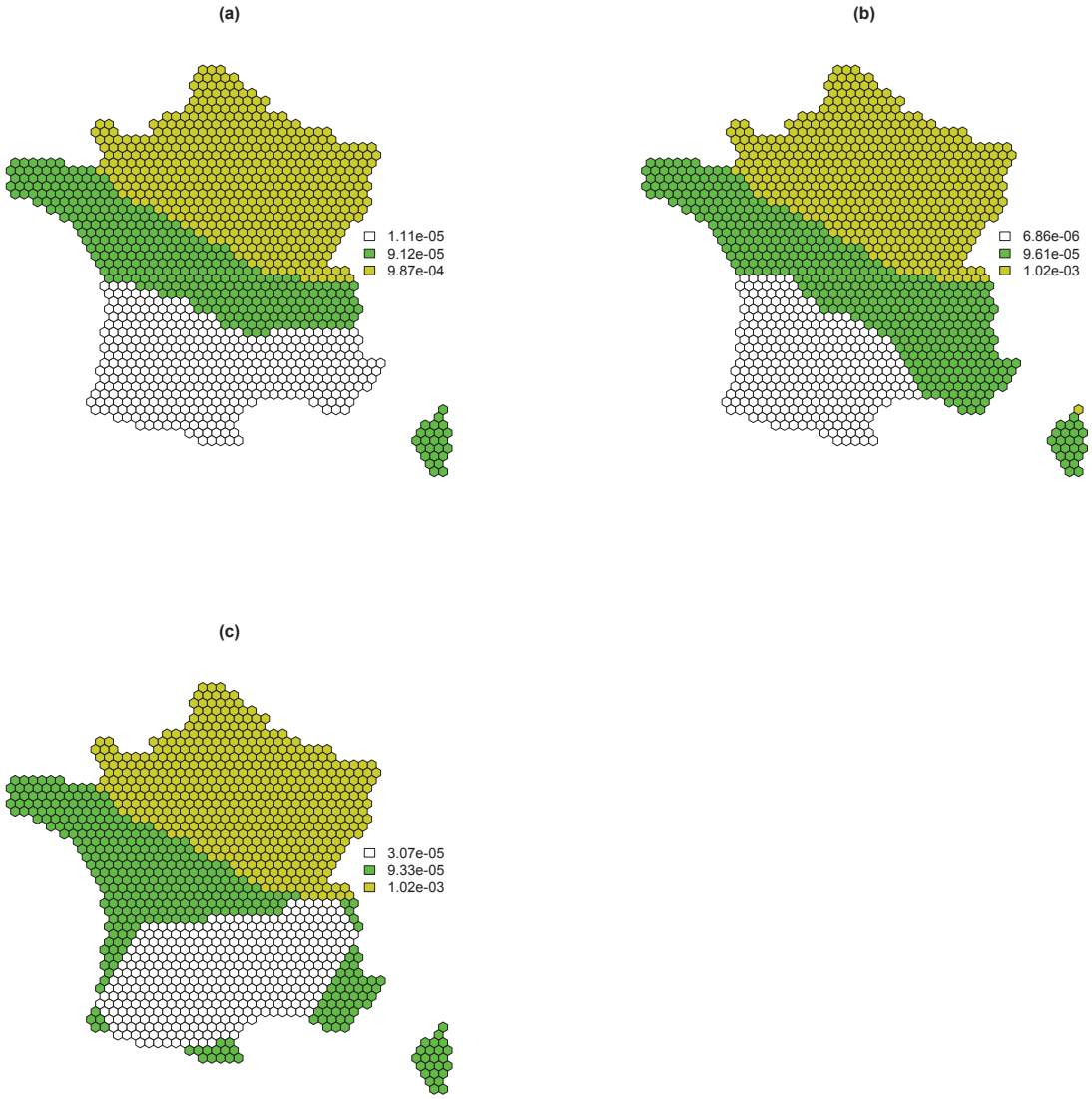


FIGURE 6.4 – Cartes de risque obtenues pour l'exemple à 3 classes avec : (a) la stratégie S_{tra} ; (b) la stratégie S_{rand} et (c) la stratégie S_{EMM}.

Sud-Est de la France est recouverte contrairement aux deux autres stratégies. En effet avec la stratégie S_{tra} (voir figure 6.4 (a)), cette région est affectée à la classe à risque faible. Quant à la classification obtenue avec la stratégie S_{EMM} et illustrée par la figure 6.4 (c), le modèle hésite entre l'affecter à la classe à risque faible ou moyen et elle est divisée, en conséquence, en deux parties : une appartenant à la classe à risque faible et l'autre à celle à risque moyen.

On peut aussi voir que c'est une tendance de cette stratégie pour toutes les régions du bord.
Le critère DSC pour cet exemple, présenté à la table 6.1, vient confirmer ces remarques "visuelles". En effet, les valeurs de ce critère obtenues par S_{rand} pour les trois classes de risque sont proches de 1. Ce qui signifie que la segmentation obtenue est très proche de la vraie. Pour la stratégie S_{EMM}, ce critère est moins bon pour les classes à risques faible et moyen avec un DSC autour de la valeur 0.7. Pour ces niveaux de risque, le DSC obtenu par S_{tra} est aux alentours de 0.8. En ce qui concerne les régions à risque fort, la valeur du DSC est la même que celle obtenue par S_{rand} et égale à 0.99. Quant à la stratégie S_{tra}, la classification résultante, en terme de DSC, est significativement meilleure que celle obtenue par S_{EMM} mais moins bonne que S_{rand} pour les classes faible et moyenne.
Ce que l'on peut tirer, comme conclusion, pour cet exemple est que les trois stratégies d'initialisation comparées donnent des résultats satisfaisants pour les différents niveaux de risque mais avec un net avantage pour S_{rand} et S_{tra} par rapport à S_{EMM} surtout pour les niveaux de risque faible et moyen.

6.2.2.2 Estimation des valeurs de risque et classification pour un jeu de 100 simulations

Les petites valeurs de risque causent quelques difficultés en l'interprétation des résultats et sont responsables d'instabilité du comportement de l'algorithme EM. Un exemple de ces comportements pathologiques, où la valeur du paramètre d'interaction b devient vite très grande et piége l'algorithme dans une solution insignifiante, est illustré par la figure 5.9 (b). Du fait de ce genre de problèmes, on ne peut pas tirer de conclusions générales en se basant sur un seul jeu de données. Afin d'approfondir les comparaisons sur l'effet de ces stratégies sur notre EM **champ-moyen** pour ce type de données, nous les comparons sur 100 jeux de données simulées avec les mêmes valeurs de risque et la même carte des niveaux de risque utilisées pour l'exemple à 3 classes (voir section 6.2).
Nous montrons dans le tableau 6.2 la moyenne et l'écart-type des valeurs du DSC sur l'ensemble des 100 jeux de données simulées de l'exemple à 3 classes. On montre aussi dans ce tableau, les moyennes et écart-types des risques estimés.
Pour les risques faibles, la stratégie S_{EMM} fournit une valeur moyenne du risque égale à $\hat{\lambda}_1 = 4.12e^{-05}$ et de variance $3.11e^{-06}$. Cela, en quelque sorte, confirme la conclusion de la section 6.2.2.1, concernant la surestimation des risques faibles par cette stratégie. Les stratégies S_{rand} et S_{tra} donnent des valeurs moyennes de risque ($\hat{\lambda}_1 = 1.02e^{-05}$ et $\hat{\lambda}_1 = 1.49e^{-05}$) plus proches de la vraie valeur ($\lambda_1 = 1e^{-05}$) mais avec une variance plus petite pour S_{rand} que pour S_{tra}. Cela signifie que les estimations obtenues par S_{rand}, pour ces niveaux de risque, sont plus stables que ceux obtenues par S_{tra} qui représentent une plus grande variabilité.

En ce qui concerne le niveau de risque moyen, les deux stratégies S_{rand} et S_{tra} fournissent respectivement une estimation moyenne de risque ($\hat{\lambda}_2 = 9.82e^{-05}$ et $\hat{\lambda}_2 = 1.15e^{-04}$) assez proches de la vraie valeur ($\lambda_2 = 1e^{-04}$) et de variances égales, respectivement, à $6.06e^{-06}$ et $6.84e^{-05}$. La stratégie S_{EMM} a tendance à surestimer le risque avec une grande variabilité sur l'ensemble des 100 jeux de données.
Quant aux régions à risque fort, les trois stratégies produisent en moyenne des résultats comparables avec un léger avantage pour S_{rand} et S_{tra} pour lesquelles les estimations moyennes sont de variances plus petites que celles obtenues par S_{EMM}. Pour pouvoir mieux visualiser

Vraie valeur du risque	Strategie	DSC	Risque estimé
faible	S_{rand}	0.84 (0.25)	$1.02e^{-05}(3.31e^{-06})$
$1e^{-05}$	S_{emm}	0.53 (0.33)	$4.12e^{-05}(3.11e^{-06})$
	S_{tra}	0.79 (0.25)	$1.49e^{-05}(1.48e^{-05})$
moyen	S_{rand}	0.88 (0.20)	$9.82e^{-05}(6.06e^{-06})$
$1e^{-04}$	S_{emm}	0.44 (0.41)	$2.19e^{-04}(2.13e^{-04})$
	S_{tra}	0.77 (0.30)	$1.15e^{-04}(6.84e^{-05})$
fort	S_{rand}	0.99 (0.09)	$9.94e^{-04}(1.71e^{-05})$
$1e^{-03}$	S_{emm}	0.93 (0.18)	$9.99e^{-04}(2.57e^{-05})$
	S_{tra}	0.96 (0.10)	$9.97e^{-04}(1.74e^{-05})$

TABLE 6.2 – Résultats de l'ensemble des 100 données simulées à 3 classes. Moyenne et écart-type du coefficient de similarité de Dice (DSC), moyenne et écart-type des valeurs estimées du risque pour chaque classe en utilisant différentes stratégies d'initialisation.

la précision des estimations du risque par les différentes stratégies, on présente à la figure 6.5 les "boîtes à moustaches" des erreurs relatives des risques estimés données par la formule : $RE = \frac{|\lambda - \hat{\lambda}|}{\lambda}$, avec $\hat{\lambda}$ est l'estimateur de λ.
Les figures 6.5 (a) ; (b) montrent que les valeurs médianes des erreurs relatives sont proches pour S_{tra} et S_{rand}. Cela signifie que les estimations du risque obtenues par ces deux stratégies ne sont pas très différentes. Quand à S_{EMM}, l'erreur relative pour les risques faibles estimés est plus élevée que celle obtenue par S_{rand} et S_{tra}. Lorsqu'il s'agit des risques moyen et fort, les RE sont comparables à celles des deux autres.
En général, pour les trois stratégies, les valeurs des erreurs relatives montrent que l'estimation des risques forts est plus précise que pour les risques faibles. Pour la classification, on peut dire que la performance de S_{EMM}, en terme du DSC (voir table 6.1), est moins bonne pour les risques faible et moyen avec un DSC moyen égal à 0.44 et 0.53 contre 0.84 et 0.88

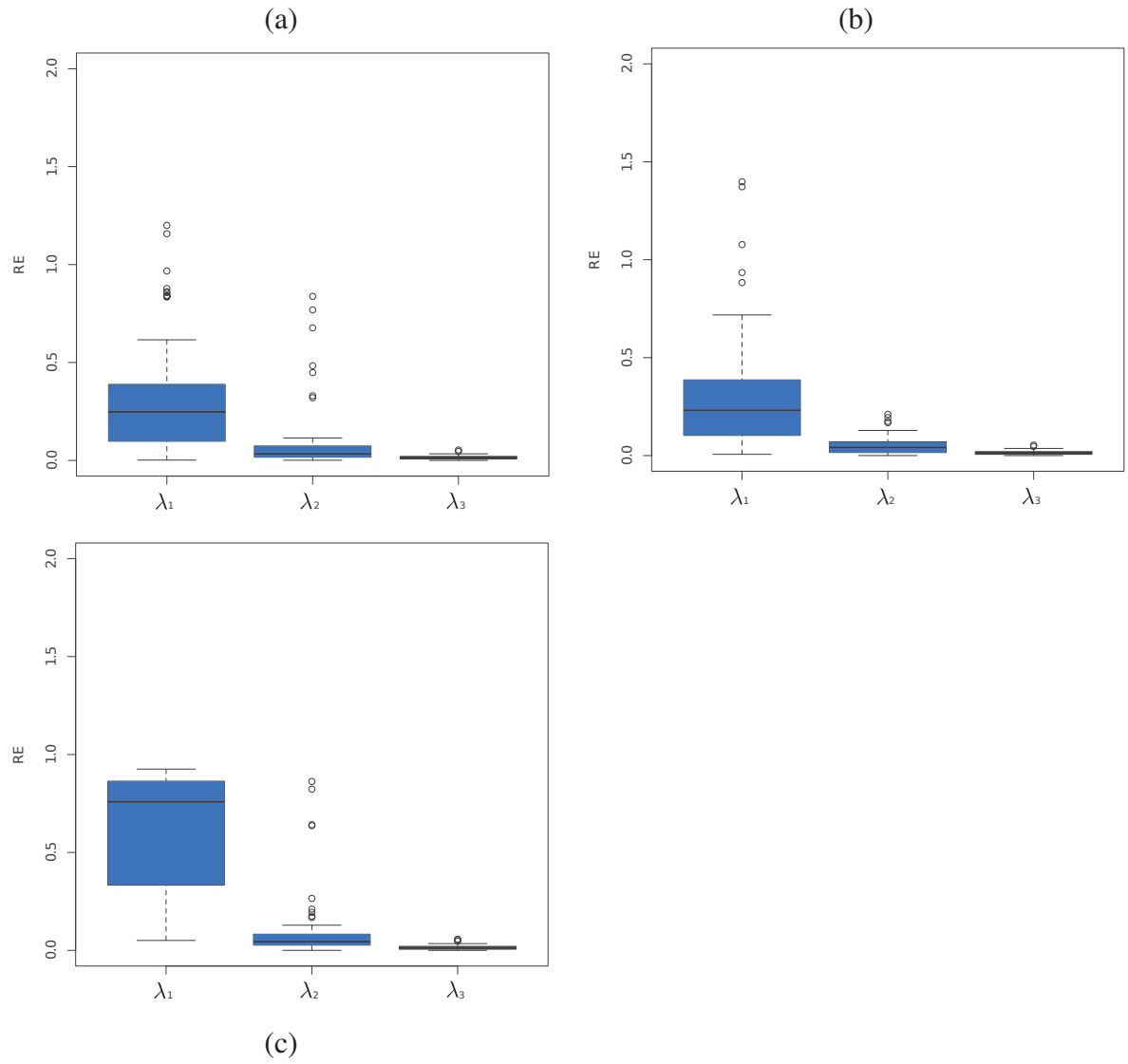


FIGURE 6.5 – L'erreur relative des paramètres de risque obtenus pour les 100 simulations de l'exemple à 3 classes avec les trois différentes stratégies : (a) S_{tra}, (b) S_{rand} et (c) S_{EMM}.

pour S_{rand} et 0.79 et 0.77 pour S_{tra}. Pour le risque fort, la performance des trois stratégies est comparable avec un léger avantage pour S_{rand} dont le DSC moyen est égal à 0.99 contre 0.96 pour S_{tra} et 0.93 pour S_{EMM}.

Conclusion sur l'exemple à 3 classes : En se basant sur les résultats obtenus, on peut dire que les trois stratégies comparées donnent des résultats raisonnables pour les différents niveaux de risque, pour cet exemple. En terme des valeurs du DSC et de l'estimation des

paramètres. La différence pincipale est observée pour les régions à risques faible et moyen. La stratégie S_{EMM} a tendance à plus surestimer les niveaux de risque faible. Quant aux stratégies S_{rand} et S_{tra}, elles sont comparables en terme d'estimation du risque surtout pour les classes à risque fort. En général, on peut constater que pour les trois stratégies l'estimation des risques forts est plus précise que celles des risques faibles. Il est clair pour la classification, et d'après les cartes obtenues et les valeurs du DSC, que les deux stratégies S_{rand} et S_{tra} donnent des résultats meilleurs que S_{EMM} surtout pour les régions à faible risque.

6.2.3 Comparaison entre différentes stratégies d'initialisation pour l'exemple à 5 classes

Dans cette section, nous présentons les résultats obtenus pour l'exemple à 5 classes. Comme pour l'exemple à 3 classes (voir section 6.2.2), nous comparons les différentes stratégies en terme de DSC (donné par l'équation 6.1) pour la classification et les valeurs de risque estimé. L'objectif de cet exemple est d'évaluer le comportement de notre modèle dans une situation plus complexe que celle présentée par l'exemple à 3 classes.
On présente ici, les résultats obtenus pour un jeu de données (section 6.2.3.1) et pour les simulations intensives sur 100 jeux de données (voir section 6.2.3.3).

6.2.3.1 Estimation des valeurs de risque et classifications pour un exemple simulé

Dans cette section nous présentons les résultats obtenus pour un jeu de données à 5 classes. Le tableau 6.3 montre les résultats du risque estimé et les résultats du critère Dice calculés selon l'équation (6.1) obtenus pour chaque stratégie et pour chacune des classes de risque.
Pour les régions considérées comme à risque très faible, avec une valeur fixée à $1e^{-05}$, on peut remarquer que la stratégie S_{EMM} surestime ce niveau de risque ($\hat{\lambda}_1 = 3.93e^{-05}$). Les deux autres stratégies surestiment aussi le risque avec des valeurs égales, respectivement, à $2.47e^{-05}$ pour S_{rand} et $1.83e^{-05}$ pour S_{tra}. En comparant ces valeurs estimées de risque, on voit que S_{tra} est celle qui surestime le moins ce risque. En ce qui concerne les régions à risque faible ($\lambda_2 = 5e^{-05}$), les trois stratégies surestiment encore le risque et cette fois-ci c'est S_{rand} qui surestime moins que les deux autres le niveau de risque de cette classe avec une valeur estimée à $\hat{\lambda}_2 = 8.95e^{-05}$ contre $\hat{\lambda}_2 = 1.18e^{-04}$ pour S_{EMM} et $\hat{\lambda}_2 = 1.02e^{-04}$ pour S_{tra}. Quant à la classe à risque moyen, on peut noter que S_{rand} estime le risque à une valeur ($\hat{\lambda}_3 = 1.32e^{-04}$) proche de la vraie valeur ($\lambda_3 = 1e^{-04}$). Les stratégies S_{EMM} et S_{tra} surestiment le risque et donnent des valeurs proches et égales, respectivement, à $4.99e^{-04}$ et $4.83e^{-04}$. Comme pour

Vraie valeur du risque	Strategie	DSC	Risque estimé
très faible	S_{rand}	0.59	2.47 e^{-05}
$1e^{-05}$	S_{EMM}	0.54	3.93 e^{-05}
	S_{tra}	0.62	1.83 e^{-05}
faible	S_{rand}	0.39	8.95 e^{-05}
$5e^{-05}$	S_{EMM}	0.05	1.18 e^{-04}
	S_{tra}	0.24	1.02 e^{-04}
moyen	S_{rand}	0	1.32 e^{-04}
$1e^{-04}$	S_{EMM}	0.09	4.99 e^{-04}
	S_{tra}	0	4.83 e^{-04}
fort	S_{rand}	0.84	4.93 e^{-04}
$5e^{-04}$	S_{EMM}	0.76	5.14 e^{-04}
	S_{tra}	0.91	7.99 e^{-04}
très fort	S_{rand}	0.72	9.08 e^{-04}
$1e^{-03}$	S_{EMM}	0.87	1.03 e^{-03}
	S_{tra}	0.96	1.83 e^{-03}

TABLE 6.3 – Exemple à 5 classes. Le coefficient de similarité de Dice (DSC) et le risque estimé pour chaque classe en utilisant notre modèle et l'algorithme *EM champ-moyen* initialisé par S_{rand}, S_{EMM} et S_{tra}.

les deux premières classes (risques faible et très faible), S_{EMM} est celle qui surestime le plus le niveau de risque.

On peut voir, pour les régions à risque fort, que S_{tra} est la stratégie qui surestime le plus ce niveau avec une valeur estimée à $\hat{\lambda}_4 = 7.99e^{-04}$ contre la vraie valeur $\lambda_4 = 5e^{-04}$. Quant aux stratégies S_{rand} et S_{EMM}, les valeurs estimées sont proches de la vraie valeur, avec une valeur supérieure pour S_{EMM} ($\hat{\lambda}_4 = 5.14e^{-04}$) et une valeur inférieure pour S_{rand} ($\hat{\lambda}_4 = 4.93e^{-04}$). Enfin pour les classes à risque très fort, S_{tra} est celle qui surestime le niveau de risque contrairement aux stratégies S_{rand} et S_{EMM} qui estiment presque correctement la vraie valeur ($\lambda_5 = 1e^{-03}$) en l'évaluant respectivement à $\hat{\lambda}_5 = 9.08e^{-04}$ et $\hat{\lambda}_5 = 1.03e^{-03}$.

Pour la classification, on peut voir "visuellement" sur la figure 6.6 qu'avec la stratégie S_{EMM}, on retrouve à 5 classes comme dans la vraie segmentation. En revanche, on remarque que cela est dû à la division de la vraie classe à risque fort en deux classes qui correspondent à pratiquement les mêmes valeurs de risque avec $\hat{\lambda}_3 = 4.99e^{-04}$ et $\hat{\lambda}_4 = 5.14e^{-04}$ comme montré dans le tableau 6.3. Les deux premières classes de la vraie segmentation (correspondant aux va-

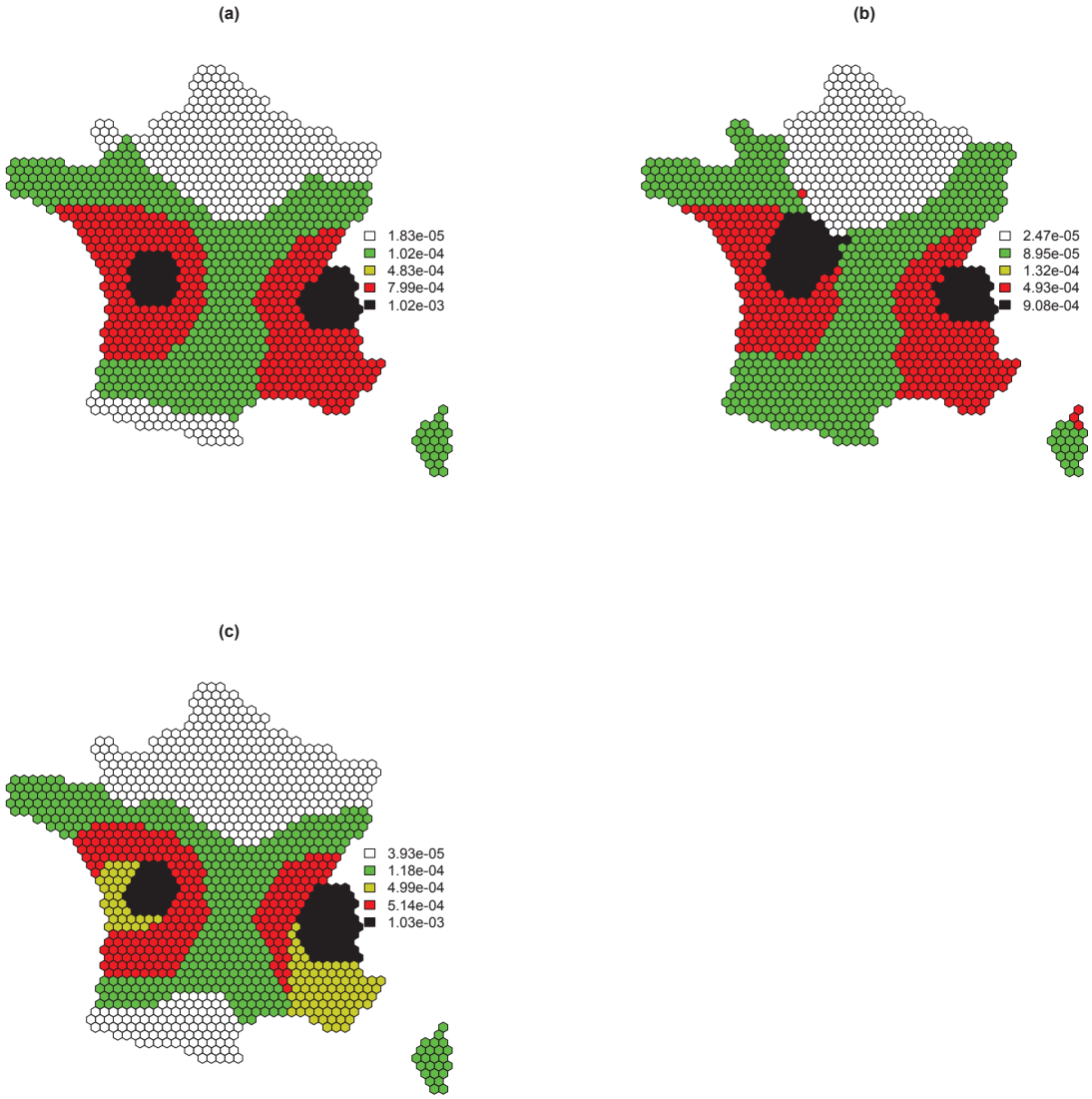


FIGURE 6.6 – Cartes de risque obtenues pour l'exemple à 5 classes avec : (a) la stratégie S_{tra} ; (b) la stratégie S_{rand} et (c) la stratégie S_{EMM}.

leurs $1e^{-05}$ et $5e^{-05}$) sont fusionnées en une classe dont le risque est estimé à $\hat{\lambda}_1 = 3.93e^{-05}$. La troisième classe de la vraie segmentation (avec la valeur de risque $\lambda_3 = 1e^{-04}$) correspond à la deuxième classe de la classification obtenue par S_{EMM}. Quant à la classe à risque très fort ($\lambda_5 = 1e^{-03}$), la classe obtenue ne correspond pas exactement à la vraie. La partie est de cette classe est plus grande qu'en réalité et la partie ouest est décalée. Malgré ses 5 classes

bien identifiées cette segmentation est donc loin d'être satisfaisante.
En ce qui concerne la classification résultante de la stratégie S_{rand}, on peut voir, dans la figure 6.6 (b), qu'une classe est perdue et qu'on ne retrouve que 4 classes. C'est la classe moyenne estimée à $\hat{\lambda}_3 = 1.32e^{-04}$, qui disparaît. La première classe obtenue correspond à une fusion de la vraie classe à risque très faible et une partie de la classe à risque faible. La partie, au sud, de la vraie classe à risque très faible est classée dans la deuxième classe de la segmentation obtenue. Cela explique la valeur surestimée du risque, $\hat{\lambda}_1 = 2.47e^{-05}$, pour cette classe. Quant à la classe à risque faible obtenue, elle est plus grande qu'en réalité et c'est le résultat de la fusion de la vraie classe à risque faible et de celle à risque moyen. La classe à risque fort est recouverte correctement même si cette stratégie a tendance à surlisser cette région affectant la partie sud-est de la segmentation. Enfin, la classe à risque très fort n'est pas complètement correctement identifiée.
Pour le résultat obtenu par S_{tra}, on peut voir sur la figure 6.6 (c) que, comme pour S_{rand}, cette stratégie a aussi du mal à séparer les régions à risques faible et moyen. Comme pour S_{rand}, la classe perdue est la classe moyenne. La classe à risque très faible est mieux identifiée, "visuellement", que par les deux autres stratégies. En effet, la partie sud de la vraie classe à risque très faible est recouverte même si elle est plus grande qu'elle l'est en réalité. La classe à risque faible obtenue est une fusion des vraies classes à risque faible et moyen. Cette classe correspond à la classe moyenne dans la vraie classification, ce qui explique sa valeur estimée à $\hat{\lambda}_2 = 1.02e^{-04}$ qui est proche de la vraie valeur de risque de la classe moyenne ($\lambda_3 = 1e^{-04}$). En ce qui concerne la classe à risque fort, on voit qu'elle est mieux identifiée avec S_{tra} qu'avec S_{rand} et S_{EMM}. Le problème du bord au sud-est persiste avec cette stratégie, comme avec S_{rand}, et cette région est affectée à la classe à risque fort alors qu'elle est dans la classe moyenne de la vraie classification. Quant à la classe à risque très fort, S_{tra} réussit à la recouvrir presque parfaitement.
Les valeurs obtenues par le critère DSC (tableau 6.3) confirment ces remarques. En effet, pour les risques très faibles, S_{tra} se comporte, légèrement, mieux que S_{rand} avec une valeur de DSC égale à 0.62 contre 0.59 pour S_{rand}. S_{EMM} est celle qui produit le plus mauvais résultat de classification, pour ces niveaux de risque, comparés aux deux autres avec un DSC égal à 0.52. Pour la classe à risque faible, les trois stratégies ont du mal à l'identifier clairement. Les valeurs du DSC pour cette région sont de 0.5 pour S_{EMM}, 0.39 pour S_{rand} et est égal à 0.24 pour S_{tra}. En accord avec ce qui a été observé à la figure 6.6, le DSC est nul pour S_{rand} et S_{tra}, ce qui est expliqué par la disparition de cette classe dans les classifications obtenues par ces stratégies. Le recouvrement de cette classe par S_{EMM} n'est pas bon non plus avec un DSC très faible de 0.09. Quand aux régions à risques fort et très fort, en accord avec ce qui est observé dans la figure 6.6, le DSC est respectivement égal à (0.91 et 0.96) pour S_{tra} contre

(0.84 et 0.72) pour S_{rand} et (0.76 et 0.87) pour S_{EMM}.
Au vu de ces résultats, on peut dire que la stratégie S_{tra} permet d'améliorer la classification dans les régions à risque élevé.

6.2.3.2 Estimation des valeurs de risque et classifications pour un exemple initialisé avec $M = 10$ valeurs

Afin de mieux voir la différence entre les stratégies S_{rand}, S_{EMM} et S_{tra}, nous réduisons le nombre de valeurs initiales de $M = 1000$ à $M = 10$. Cela peut être typiquement nécessaire si le temps ou les ressources de calcul doivent être limités. L'idée est de voir laquelle de ces stratégies est capable de recouvrir la vraie classification avec un nombre limité de valeurs initiales.
Rappelons que la stratégie S_{tra} est la stratégie S_t présentée à la section 5.4.1 avec l'ajout d'une étape de recherche des valeurs initiales pour le couple $(\boldsymbol{\alpha}, b)$ selon la procédure complète décrite à la section 5.4.3. En principe, cette comparaison devrait profiter à la stratégie S_{tra} qui est plus efficace pour explorer l'espace des paramètres (comme illustré par la figure 5.7 (b) pour S_t).
On montre à la figure 6.7 les différentes classifications obtenues par chacune des trois stratégies. On observe qu'elles peuvent toutes détecter les régions à risque élevé. En effet, elles réussissent toutes à recouvrir la classe à risque fort et arrivent à identifier les deux grandes parties de cette région (à l'est et à l'ouest). Pour la classe à risque très fort, S_{EMM} ne détecte que sa partie se trouvant à l'est (voir figure 6.7 (c)). Elle n'arrive pas à retrouver la partie de cette classe localisée à l'ouest. Quant à S_{tra}, elle réussit à identifier les deux parties de cette région, même si cette région n'est pas correctement recouverte, comme illustré par la figure 6.7 (a). En ce qui concerne S_{rand}, elle n'identifie que la moitié de cette classe localisée à l'ouest comme montré à la figure 6.7 (b).
Quant à la classe à risque fort, les trois stratégies arrivent à détecter les deux moitiés de cette région se trouvant à l'est et à l'ouest. Pour la classe moyenne, seule la stratégie S_{EMM} en identifie une mais elle ne correspond pas à la vraie région de niveau moyen. Cette stratégie a tendance à produire une classe plus grande qu'elle ne l'est en réalité. Avec S_{rand} et S_{tra}, comme déjà observé en section 6.2.3.1, cette classe disparaît.
Pour les classes à risque très faible et faible, les deux stratégies S_{rand} et S_{tra} arrivent à détecter qu'il y a deux classes différentes même si elles sont mal identifiées et ne sont pas recouvertes correctement. La stratégie S_{tra} a tendance à surestimer la classe à risque très faible et à affecter plus de sites à cette classe que ceux qui le sont dans la vraie classification. S_{rand} surestime aussi ce niveau de risque et affecte les sites du bord du sud-ouest à cette classe alors qu'ils

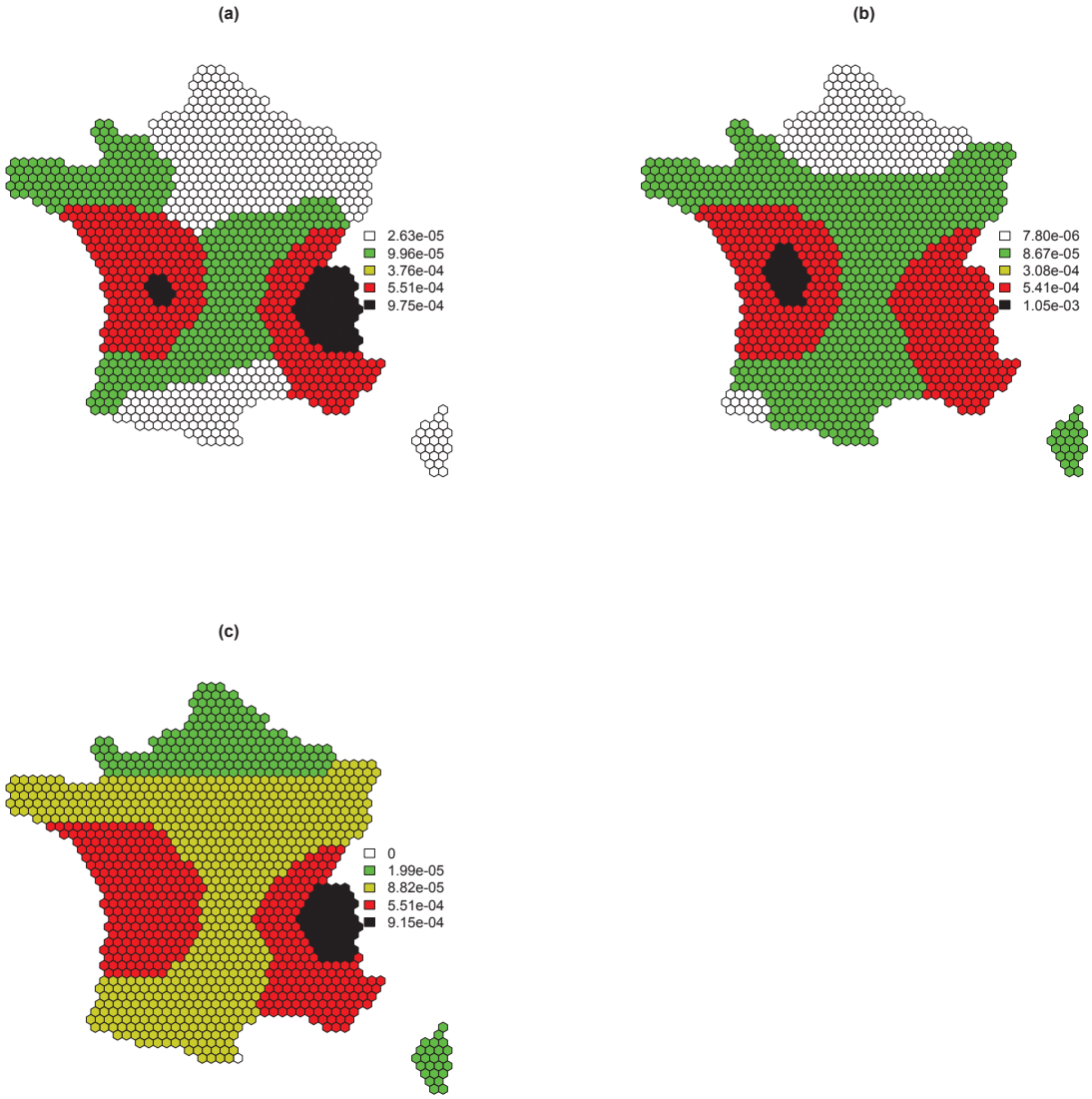


FIGURE 6.7 – Cartes de risque obtenues, pour l'exemple à 5 classes, en lançant le EM **champ-moyen** à partir de $M = 10$ positions initiales. (a), (b), (c) : Cartes de risque obtenues respectivement par les stratégies S_{tra}, S_{rand} et S_{EMM}.

sont dans la classe de risque faible dans la vraie segmentation.
Cet exemple confirme les conclusions faites à la section 6.2.3.1 concernant le fait que la stratégie S_{tra} aide à améliorer la classification des sites dans les régions à risque élevé. Ceci est positif pour la stratégie S_{tra} car ceux sont plus spécifiquement ces régions qui sont potentiel-

lement critiques et d'intérêt en épidémiologie. Quant aux estimations des risques, elles sont similaires pour les trois stratégies (voir figure 6.7).

6.2.3.3 Estimation des valeurs de risque et classification pour un jeu de 100 simulations

Nous présentons dans le tableau 6.4, comme pour l'exemple à 3 classes (voir section 6.2.2.2), la moyenne et l'écart-type des valeurs estimées des risques et du coefficient de similarité DSC pour les trois stratégies d'initialisation. Nous montrons aussi les "boîtes à moustaches", à la figure 6.8, des erreurs relatives des risques estimés pour les 100 jeux de données de l'exemple à 5 classes.

On peut noter, en examinant le tableau 6.4, que pour les régions à risque très faible (corres-

Vraie valeur du risque	Strategie	DSC	Risque estimé
très faible	S_{rand}	0.42 (0.29)	$2.17e^{-05}(2.15e^{-05})$
$1e^{-05}$	S_{EMM}	0.36 (0.24)	$2.58e^{-05}(2.98e^{-06})$
	S_{tra}	0.56 (0.20)	$2.07e^{-05}(1.53e^{-05})$
faible	S_{rand}	0.29 (0.19)	$7.99e^{-05}(7.53e^{-05})$
$5e^{-05}$	S_{EMM}	0.22 (0.18)	$5.43e^{-04}(3.49e^{-05})$
	S_{tra}	0.29 (0.17)	$9.62e^{-05}(4.39e^{-05})$
moyen	S_{rand}	0.38 (0.25)	$1.74^{-04}(1.57e^{-04})$
$1e^{-04}$	S_{EMM}	0.16 (0.21)	$3.03e^{-04}(2.06e^{-04})$
	S_{tra}	0.09 (0.18)	$3.33e^{-04}(1.37e^{-04})$
fort	S_{rand}	0.51 (0.33)	$4.58e^{-04}(1.97e^{-05})$
$5e^{-04}$	S_{EMM}	0.55 (0.33)	$5.74e^{-04}(5.86e^{-05})$
	S_{tra}	0.66 (0.38)	$5.57e^{-04}(1.05e^{-04})$
très fort	S_{rand}	0.44 (0.18)	$8.71e^{-04}(4.27e^{-04})$
$1e^{-03}$	S_{EMM}	0.65 (0.34)	$9.78e^{-04}(1.76e^{-04})$
	S_{tra}	0.83 (0.17)	$1.05e^{-03}(7.66e^{-05})$

TABLE 6.4 – Résultats pour 100 jeux de données de l'exemple à 5 classes. Moyennes et écart-types du coefficient de similarité de Dice (DSC) et moyennes et écart-types des valeurs de risque estimées pour chaque classe en utilisant différentes stratégies d'initialisation.

pondantes à $\lambda_1 = 1e^{-05}$), les trois stratégies ont toutes tendance à surestimer en moyenne le risque. S_{EMM} est celle qui surestime le plus, en moyenne, le risque avec une valeur estimée à $\hat{\lambda}_1 = 2.58e^{-05}$ contre $\hat{\lambda}_1 = 2.17e^{-05}$ pour S_{rand} et $\hat{\lambda}_1 = 2.07e^{-05}$ pour S_{tra}. Quant à la

variance de ces estimations, c'est S_{EMM} qui présente la plus petite variance et S_{rand} la plus grande. Pour les régions à risque faible ($\lambda_2 = 5e^{-05}$), les trois stratégies surestiment aussi le niveau de risque. S_{tra} est celle qui surestime le plus le risque, cette fois ci, avec une valeur moyenne estimée à $\hat{\lambda}_2 = 9.62e^{-05}$. La stratégie S_{EMM} est celle qui surestime le moins le risque en moyenne avec $\hat{\lambda}_2 = 5.43e^{-05}$ contre $\hat{\lambda}_2 = 7.99e^{-05}$ pour S_{tra}. Comme pour les régions à risque très faible, c'est la stratégie S_{rand} qui présente la plus grande variabilité des estimations et S_{EMM} la plus petite.
En ce qui concerne la classe moyenne ($\lambda_3 = 1e^{-04}$), les trois stratégies surestiment le risque. Pour ce niveau de risque, c'est S_{tra} qui le surestime le plus en moyenne avec $\hat{\lambda}_3 = 3.33e^{-04}$ et avec la plus petite variance. Quant à S_{rand}, c'est celle qui surestime le moins le risque pour cette région avec une estimation moyenne égale à $\hat{\lambda}_3 = 1.74e^{-04}$ et avec une variance, légèrement, plus grande que celle de S_{tra}.
Quant à la classe à risque fort, les trois stratégies produisent des valeurs moyennes de risque proches de la vraie valeur ($\lambda_4 = 5e^{-04}$). Mais c'est S_{rand} qui fournit la valeur moyenne de risque la plus proche avec $\hat{\lambda}_4 = 4.58e^{-04}$ contre $\hat{\lambda}_4 = 5.54e^{-04}$ pour S_{tra} et $\hat{\lambda}_4 = 5.74e^{-04}$ pour S_{EMM}. C'est S_{rand} qui présente le moins de variabilité pour l'estimation moyenne du risque contre une plus grande variance pour l'estimation obtenue par S_{tra}.
Enfin pour la classe à risque très fort, S_{EMM} et S_{tra} produisent des estimations de risque proches, en moyenne, de la vraie valeur ($\lambda_5 = 1e^{-03}$) avec $\hat{\lambda}_5 = 9.78e^{-04}$ pour S_{EMM} et $\hat{\lambda}_5 = 1.05e^{-03}$ pour S_{tra}. Quant à S_{rand}, le risque est estimé à $\hat{\lambda}_5 = 8.71e^{-04}$. Pour cette classe, la plus grande variance est donnée par S_{rand} et la plus petite par S_{tra}.
Nous montrons également à la figure 6.8, les "boîtes à moustaches" des erreurs relatives des moyennes estimées. On peut voir que la médiane de l'erreur relative pour l'estimation moyenne de $\hat{\lambda}_1$ est presque la même avec S_{tra} (voir figure 6.8 (a)) et S_{rand} (voir figure 6.8 (b)), avec une variance légèrement plus petite pour S_{tra}. Avec S_{EMM}, cette erreur est plus grande pour ce niveau de risque. Pour $\hat{\lambda}_2$, la médiane de l'erreur relative est plus petite pour S_{rand} et S_{EMM} que pour S_{tra}. Quant à $\hat{\lambda}_3$, la valeur médiane de l'erreur relative est plus grande pour S_{tra} et elle est très petite pour S_{rand}. En ce qui concerne $\hat{\lambda}_4$ et $\hat{\lambda}_5$, les médianes des erreurs relatives des estimations obtenues par S_{tra} et S_{EMM} sont très petites comparées à celles obtenues avec S_{rand}.

Conclusion pour l'exemple à 5 classes : Pour cet exemple où le nombre de classes $K = 5$ est supérieur à celui de l'exemple à 3 classes étudié en section 6.2.2, les trois stratégies ont du mal à séparer les régions à risque très faible et faible et ont tendance à perdre une classe. Ceci n'est pas tout à fait vrai pour S_{EMM}, qui arrive à obtenir les à 5 classes demandées.

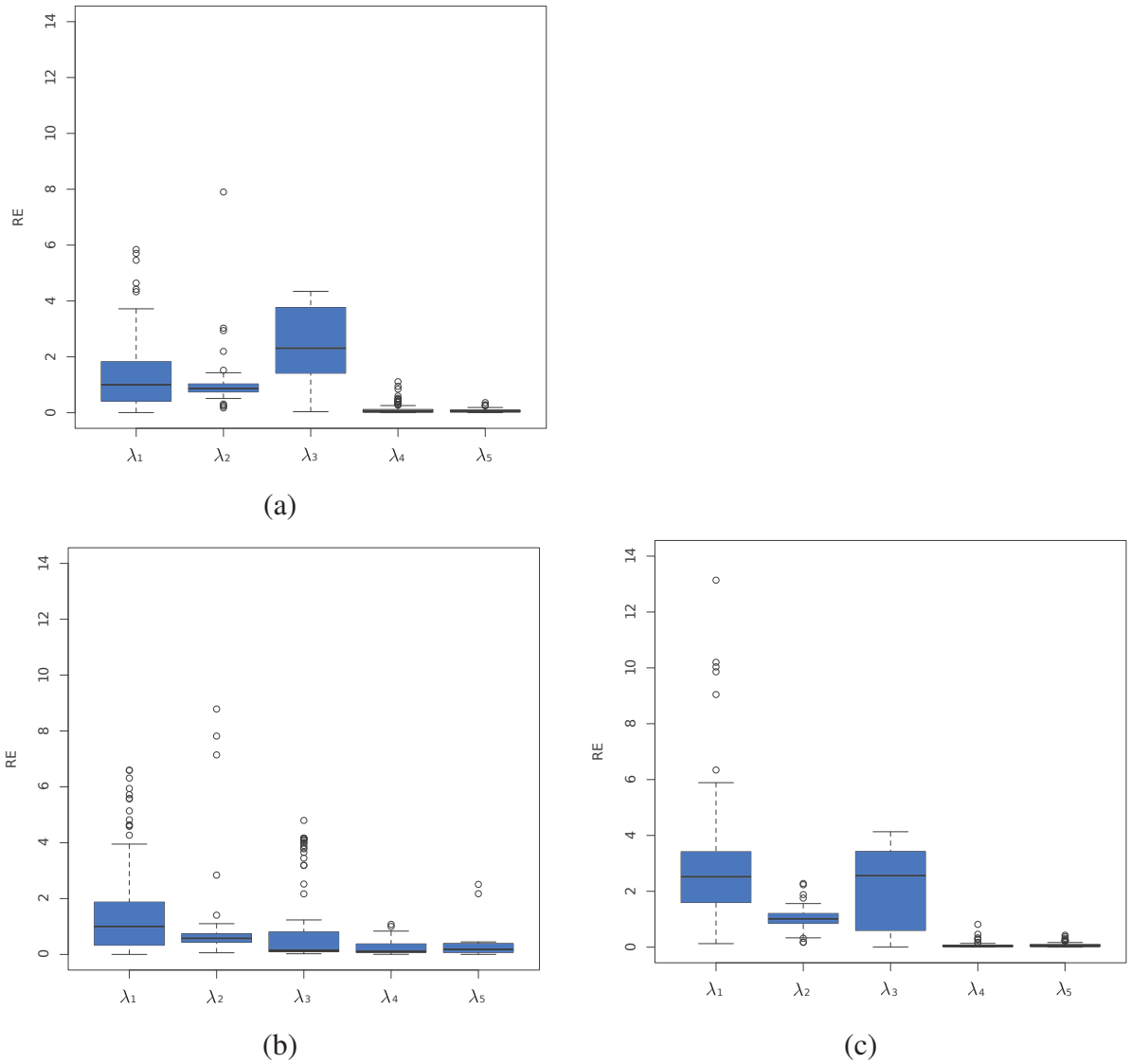


FIGURE 6.8 – L'erreur relative des paramètres de risque obtenus pour les 100 jeux de données simulées pour l'exemple à 5 classes avec les trois différentes stratégies : (a) S_{tra}, (b) S_{rand} et (c) S_{EMM}.

Toutefois, cette stratégie retrouve le bon nombre de classes en scindant une classe en deux avec des risques estimés à des valeurs similaires pour ces deux classes. Ceci n'est en pratique pas satisfaisant et est une tendance générale de la stratégie S_{EMM}.
Pour ce qui est de l'estimation du risque, les trois stratégies ont tendance à surestimer les niveaux de risques faibles et à estimer correctement les risques élevés. Toutefois, S_{tra} semble fournir une estimation du risque meilleure pour les classes à risque élevé, ce qui est une

propriété intéressante dans le contexte épidémiologique.

6.2.4 Choix du nombre de classes

L'un des inconvénients majeurs des méthodes de classification est la spécification en générale requise du nombre de classes. Or, dans le cas de données réelles, ce nombre de classes est rarement connu.
Pour le déterminer, on ne peut se baser uniquement sur la vraisemblance, ou la vraisemblance complétée qui augmentent avec le nombre de paramètres et donc de classes. Plus il y a de classes, plus la partition est ajustée aux données, et donc plus la vraisemblance est grande. Ainsi, chercher à déterminer le nombre de classes en maximisant la vraisemblance nous conduit irrémédiablement à une partition en N classes, où chaque site serait dans sa propre classe, ce qui ne nous apporte aucune information. La vraisemblance seule n'est pas un "bon" critère pour le choix du nombre de classes.
C'est ainsi qu'on a, en général, recours à des critères de choix du modèle permettant de choisir le nombre de classes le plus probable au vu des données.
Nous avons choisi de prendre comme critère de choix du nombre de classes le Bayesian Information Criterion (BIC) approché (voir section 4.4).
Avant d'appliquer notre modèle et obtenir les résultats présentés aux sections 6.2.2 et 6.2.3, nous avons testé plusieurs nombres de classes, qui restaient fixés au cours de l'algorithme. Nous avons alors choisi le modèle avec la plus petite valeur de BIC. Les choix donnés par le BIC pour les 100 jeux de données des exemples à 3 classes et à 5 classes sont présentés ici.
Pour l'exemple 3 classes, chacun des 100 jeux de données, nous lançons notre EM **champ-moyen** initialisé par la stratégie S_{tra} pour $K = 2$ et $K = 3$. Nous avons observé que pour $K \geq 4$, l'algorithme perd systématiquement une classe et ces valeurs de K ne sont jamais choisies. Pour la stratégie S_{tra}, le critère choisit $K = 3$, 75 fois sur les 100 jeux de données et $K = 2$, 25 fois.
De même, pour l'exemple 5 classes, nous avons calculé le BIC approché pour sélectionner une valeur de K entre 2 et 7. La valeur $K = 5$ a été choisie 46 fois, $K = 4$ a été choisie 42 fois, $K = 3$, 12 fois et $K = 2, 6, 7$ n'ont jamais été sélectionnés. Des résultats semblables ont été observés pour les autres stratégies.
Ces résultats confirment, et spécialement pour l'exemple à 5 classes, que les données que nous devons traiter ne correpondent pas à un cas simple où les classes sont bien séparées.

Discussion sur les stratégies d'initialisation. Globalement, nous observons que les trois stratégies que nous avons comparé ont tendance à restaurer plus facilement les régions à risque fort que celles à risque faible.
La stratégie S_{EMM} ne semble pas très adaptée à ce type de données vu qu'elle a tendance à produire des classifications artificielles menant à des résultats peu satisfaisants. Il nous semble que cette stratégie n'est pas la plus appropriée pour initialiser notre EM **champ-moyen**.
La stratégie S_{rand} fournit des résultats assez similaires à ceux obtenus par S_{tra} même si pour l'exemple à 3 classes, elle se comporte plutôt mieux que S_{tra}. Ce qui peut paraître surprenant si l'on considére qu'il devrait être facile de faire mieux qu'une initialisation aléatoire. Il faut signaler toutefois que cette initialisation aléatoire ne l'est pas totalement, du fait que l'initialisation des paramètres du modèle du champ de Markov ($\boldsymbol{\alpha}$ et b) est imposé. On a remarqué lors de nos tests que si l'on initialisait les paramètres α et b aléatoirement aussi comme les λ, le EM **champ-moyen** avait du mal à obtenir des résultats raisonnables. De plus, le nombre de valeurs initiales utilisées pour les deux exemples est assez élevé (M=1000 initialisations pour 1264 données) ce qui permet d'explorer correctement l'espace des valeurs de λ.
En ce sens, la stratégie S_{tra} fournit des résultats assez satisfaisants comparés aux autres stratégies. En particulier, la proportion d'hexagones correctement classés est améliorée en utilisant cette stratégie. Aussi, cette stratégie fournit des résultats satisfaisants avec une meilleure exploration de l'espace, si l'on dispose de ressources calculatoires limitées, typiquement lorsque M est réduit à quelques initialisations.
Au vu de ces comparaisons, nous décidons, dans la suite du travail, d'initialiser notre EM **champ-moyen** avec la stratégie S_{tra}.

6.2.5 Comparaison entre différentes formes de $\mathbb{B}$

Afin de poursuivre notre étude d'évalution de la performance du modèle que nous proposons pour la cartographie du risque épidémiologique, nous nous intéressons dans cette partie à comparer notre variante du modèle de Potts avec d'autres spécifications pour la partie cachée du modèle. Nous considérons ainsi trois spécifications. Notre EM **champ-moyen** est initialisé, dans les trois cas, avec la stratégie S_{tra}. Nous comparons le modèle *semi-graduel* que nous proposons pour lequel la matrice $\mathbb{B}$ est donnée par $\mathbb{B} = b\,[1 - |k-l|/2]_+$ au modèle de Potts standard qui correspond à $\mathbb{B} = b \times I_K$ et à une autre variante du modèle de Potts qui dépend de la distance des classes les unes par rapport aux autres, et qui traduit l'hypothèse de gradation de risque (comme discuté en section 5.2.1) pour des données semblables à celles étudiées. Cette nouvelle variante est définie par $\mathbb{B} = b(1 - |k-l|/(K-1))$ et on l'appellera

dans la suite le modèle *graduel*.
Si le nombre de classe est $K = 2$, la matrice $\mathbb{B}$ du modèle *graduel* se réduit à $\mathbb{B} = b \times I_K$, ce qui conduit au modèle de Potts standard. En revanche, si $K = 3$, la matrice $\mathbb{B}$ devient $\mathbb{B} = b\ [1 - |k - l|/2]_+$, ce qui correspond au modèle *semi-graduel*. Notons que le modèle *graduel* ne devient différent des modèles de Potts et du *semi-graduel* que lorsque le nombre de classes est $K > 3$.
Dans la section suivante, nous montrons les résultats de comparaison pour des exemples simulés de à 3 classes et à 5 classes et pour un ensemble de 100 jeux de telles données. Rappelons que ces résultats sont obtenus en fixant le nombre de classes à sa vraie valeur.

6.2.5.1 Résultats pour l'exemple à 3 classes

Comme dans les sections 6.2.2 et 6.2.3, nous présentons, dans le tableau 6.5, les estimations du risque et les valeurs du critère DSC obtenues par chaque modèle. À la figure 6.4, nous montrons les cartes de risque produites par les trois différents modèles pour l'exemple à 3 classes. Notons que dans ce cas, seuls deux jeux de résultats sont disponibles du fait que les modèles *semi-graduel* et *graduel* sont équivalents lorsque $K = 3$.
En examinant le tableau 6.5, on peut voir que pour la région à risque faible (avec $\lambda_1 = 1e^{-05}$), le modèle de Potts surestime son niveau de risque ($\hat{\lambda}_1 = 6.52e^{-05}$) alors que les deux modèles équivalents (le *graduel* et le *semi-graduel*), donnent une valeur de risque, $\hat{\lambda}_1 = 1.11e^{-05}$, proche de la vraie valeur.
Quant à la région à risque moyen ($\lambda_2 = 1e^{-04}$), le modèle de Potts surestime aussi ce niveau avec la valeur $\hat{\lambda}_2 = 5.21e^{-04}$ alors que les deux autres modèles équivalents sous-estiment légèrement le niveau de risque avec $\hat{\lambda}_2 = 9.12e^{-05}$.
En ce qui concerne la région à risque fort, correspondante à $\lambda_3 = 1e^{-03}$, la valeur de risque est estimée correctement ($\hat{\lambda}_3 = 1.00e^{-03}$) avec le modèle de Potts. Quant à la valeur estimée ($\hat{\lambda}_3 = 9.87e^{-04}$) par les modèles *semi-graduel* et *graduel*, elle est légèrement inférieure à la vraie valeur.
Sur la figure 6.9, nous montrons les classifications obtenues par les modèles *semi-graduel* et *graduel* (figure 6.9 (a)) et le modèle de Potts (figure 6.9 (b)).
On peut remarquer que la carte de risque obtenue par les modèles *semi-graduel* et *graduel* ressemblent plus, visuellement, à la vraie carte (donnée par la figure 6.2 (a)) alors qu'avec le modèle de Potts une des trois classes est perdue.
Les valeurs du critère DSC, montrées à la table 6.5, confirment cette remarque visuelle. En effet, le DSC pour la région faible est de 84% avec les modèles *graduel* et *semi-graduel* contre 58% avec le modèle de Potts. La classe moyenne disparaît avec le modèle de Potts, d'où la

Résultats pour un exemple à 3 classes			
Modèle pour $\mathbb{B}$	Niveau de risque	DSC	Risques estimés
Modèle de Potts	faible	0.58	$6.52e^{-05}$
$\mathbb{B} = b\ I_K$	moyen	0	$5.21e^{-04}$
	fort	0.99	$1.00e^{-03}$
Modèles *semi-graduel*	faible	0.84	$1.11e^{-05}$
et *graduel*	moyen	0.86	$9.12e^{-05}$
$\mathbb{B}(k,l) = b\ (1 - \lvert k-l\rvert/(K-1))$	fort	1	$9.87e^{-04}$
Résultats pour les 100 simulations à 3 classes			
Modèle pour $\mathbb{B}$	Niveau de risque	DSC	Risques estimés
Modèle de Potts	faible	0.40 (0.34)	$3.35e^{-05}(3.34e^{-05})$
$\mathbb{B} = b\ I_K$	moyen	0.43 (0.37)	$2.17e^{-04}(2.46e^{-04})$
	fort	0.92 (0.22)	$9.84e^{-04}(2.25e^{-04})$
Modèles *semi-graduel*	faible	0.79 (0.25)	$1.49e^{-05}(1.48e^{-05})$
et *graduel*	moyen	0.77 (0.30)	$1.15e^{-04}(6.84e^{-05})$
$\mathbb{B}(k,l) = b\ (1 - \lvert k-l\rvert/(K-1))$	fort	0.96 (0.10)	$9.97e^{-04}(1.74e^{-05})$

TABLE 6.5 – Le coefficient de similarité de Dice (DSC) et le risque estimé pour chaque classe en utilisant différents modèles pour $\mathbb{B}$ et S_{tra} comme méthode d'initialisation pour l'EM **champ-moyen**.

valeur 0 pour le critère DSC. Par contre, les deux variantes du modèle de Potts (*semi-graduel* et *graduel*) montrent une bonne performance pour cette classe avec un DSC de 86%. Pour la classe à risque fort, le DSC est égal à 0.99 pour le modèle de Potts et à 1 pour les modèles *semi-graduel* et *graduel*.

Nous comparons aussi ces trois modèles, sur l'ensemble des 100 jeux de données pour l'exemple à 3 classes et présentons les résultats obtenus à la table 6.5. L'estimation moyenne du risque $\hat{\lambda}_1 = 3.35e^{-05}$ obtenue avec le modèle de Potts pour les régions à risque faible est supérieure à la vraie valeur $(\lambda_1 = 1e^{-05})$. Quant aux modèles *semi-graduel* et *graduel*, l'estimation $\hat{\lambda}_1 = 1.49e^{-05}$ est aussi supérieure à la vraie valeur mais reste plus proche que celle produite par le modèle de Potts. Les variances des estimations moyennes résultantes des deux modèles sont grandes et sont de l'ordre des moyennes obtenues.

Pour la classe à risque moyen, la valeur moyenne estimée $\hat{\lambda}_2 = 2.17e^{-04}$ par le modèle de Potts est aussi supérieure à la vraie valeur $(\lambda_2 = 1e^{-04})$. Alors que celle obtenue par les modèles *semi-graduel* et *graduel* $(\hat{\lambda}_2 = 1.15e^{-04})$ reste aux alentours du vrai niveau de risque

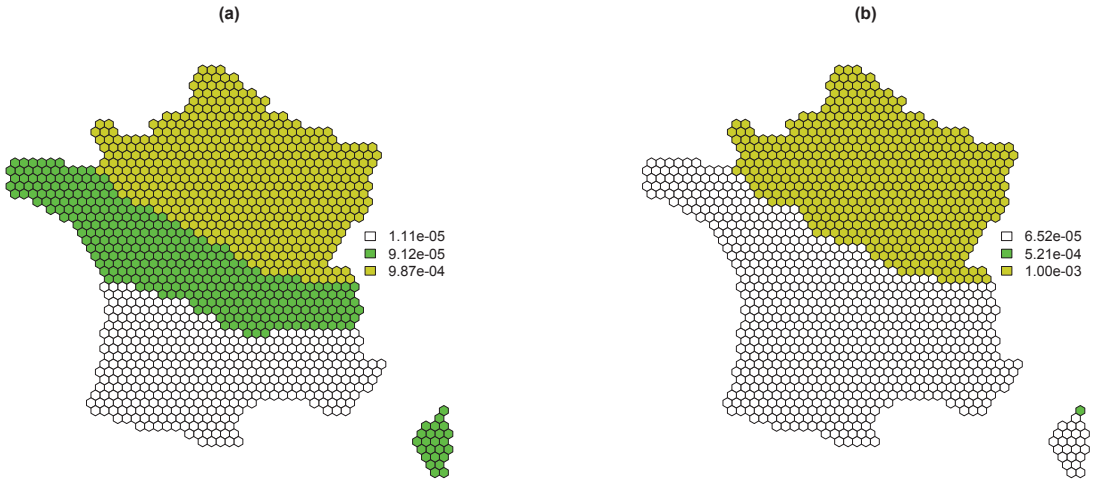


FIGURE 6.9 – Résultats de la classification pour l'exemple à 3 classes obtenus par : (a) les modèles *semi-graduel* et *graduel* et (b) le modèle de Potts.

λ_2. Les variances des estimations moyennes résultantes du modèle de Potts est plus grande que celles des modèles *semi-graduel* et *graduel*.
En ce qui concerne la classe à risque fort, les trois modèles produisent des estimations moyennes de risque comparables avec $\hat{\lambda}_3 = 9.84e^{-04}$ pour Potts et $\hat{\lambda}_3 = 9.97e^{-04}$ pour les deux autres et proches de la vraie valeur $\lambda_3 = 1e^{-03}$ avec une variance plus petite pour les modèles *semi-graduel* et *graduel*.
Pour la classification, en terme de critère DSC, la performance des modèles *semi-graduel* et *graduel* est meilleure avec des coefficients de DSC moyens égaux à 0.79 contre 0.40 avec le modèle de Potts ; pour la classe à risque faible, à 0.77 contre 0.43 pour la classe à risque moyen et à 0.96 contre 0.92 pour celle à risque fort.

6.2.5.2 Résultats pour l'exemple à 5 classes

Nous présentons, dans le tableau 6.6, les estimations de risque et les valeurs du critère DSC obtenues par chaque modèle. À la figure 6.10, nous montrons les cartes de risque produites par les trois modèles pour l'exemple à 5 classes. Notons que pour cet exemple, le nombre de classes est $K = 5$ et les deux modèles *semi-graduel* et *graduel* ne sont pas équivalents.

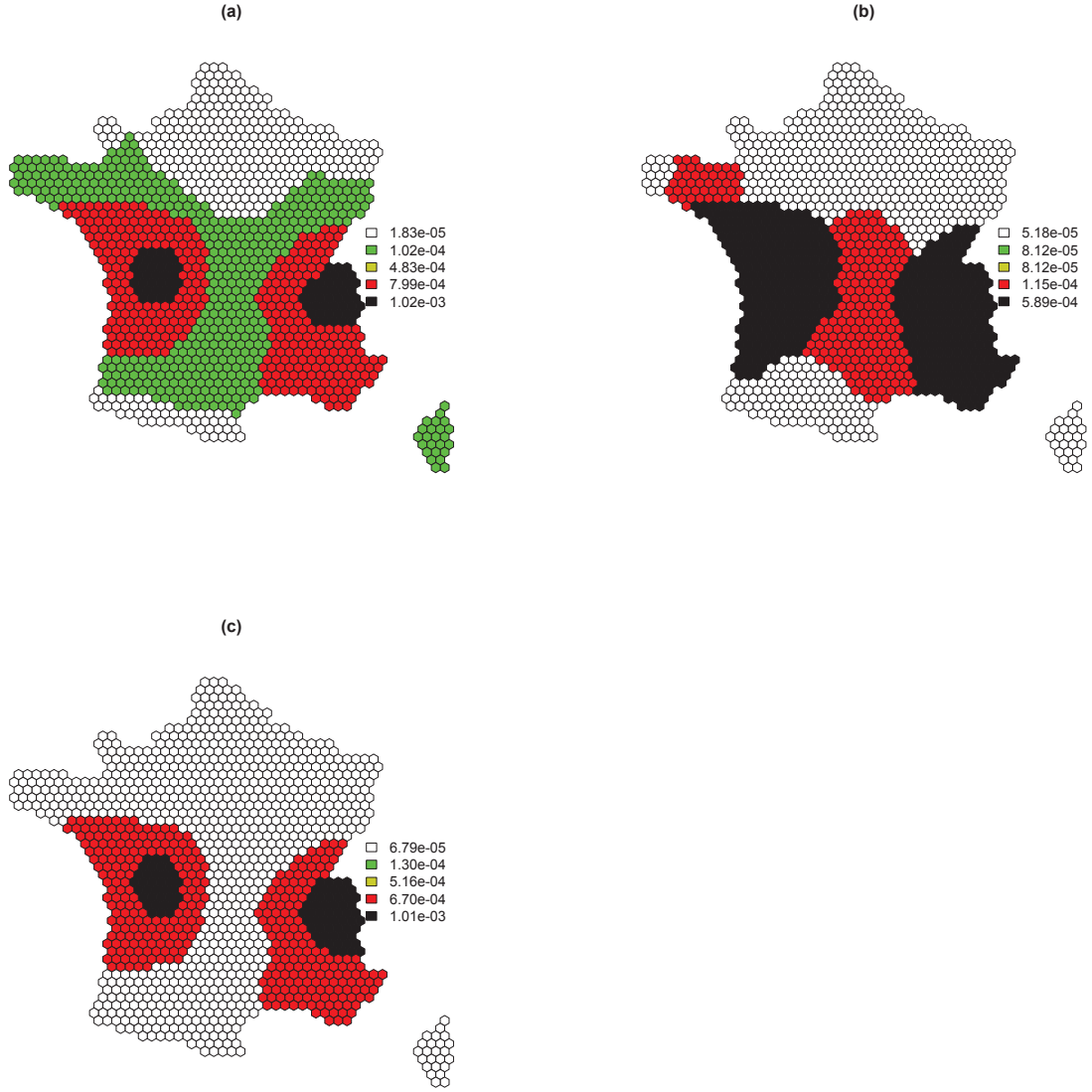


FIGURE 6.10 – Les résultats de la classification pour l'exemple à 5 classes. Cartes de risque obtenues avec : (a) le modèle *semi-graduel*, (b) le modèle de Potts et (c) le modèle *graduel*.

Nous pouvons voir dans le tableau 6.6 que les trois modèles surestiment les niveaux de risque pour les régions à risques faible et très faible. Même si les modèles de Potts et le *graduel* surestiment plus ces niveaux avec des valeurs égales respectivement à $\hat{\lambda}_1 = 5.18e^{-05}$ et $\hat{\lambda}_1 = 6.79e^{-05}$ contre une surestimation ($\hat{\lambda}_1 = 1.83e^{-05}$) plus petite pour le *semi-graduel*. Le modèle de Potts estime les deux niveaux de risque faible et moyen à la même valeur

$\hat{\lambda}_2 = 8.12e^{-05}$. Ceci explique la disparition de ces deux classes, dans la segmentation obtenue, et leur fusion avec la classe à risque très faible comme montrée à la figure 6.10 (b). On observe à la figure 6.10 (c), le même phénomène avec le modèle *graduel* où les trois premiers niveaux de risque sont surestimés et les deux autres classes (à risques faible et moyen) disparaîssent et sont fusionnées avec la classe à risque très faible. Pour la région à risque fort, le niveau de risque est estimé, avec le modèle de Potts, à la valeur $\hat{\lambda}_4 = 1.15e^{-04}$ qui reste proche de la vraie valeur $\lambda_4 = 1e^{-04}$. Mais en examinant la figure 6.10 (b), on voit que cette classe est mal identifiée et ne correspond pas à la classe à risque fort de la vraie classification (voir figure 6.2 (b)). Quant aux modèles *semi-graduel* et *graduel* même s'ils surestiment les niveaux de risque avec des valeurs égales respectivement à $\hat{\lambda}_4 = 7.99e^{-04}$ et $\hat{\lambda}_4 = 6.70e^{-04}$, les classes résultantes correspondent à la vraie classification comme illustrées par les figures 6.10 (a) ; (c). En ce qui concerne la classe à risque très fort, le modèle de Potts sous-estime (avec $\hat{\lambda}_5 = 5.89e^{-04}$) le vrai niveau de risque. Par contre, le *semi-graduel* et le *graduel* estiment correctement le risque avec respectivement $\hat{\lambda}_5 = 1.02e^{-03}$ et $\hat{\lambda}_5 = 1.01e^{-03}$. Pour le modèle de Potts, la figure 6.10 (b) montre que cette classe n'est pas la vraie classe à risque très fort mais correspond plutôt à la classe à risque fort dans la vraie classification (montrée à la figure 6.2 (b)). Quant aux modèles *semi-graduel* et *graduel*, les classes correspondent à celles de la vraie segmentation.
Les valeurs du DSC confirment ces remarques. En effet, pour le modèle de Potts ce coefficient est égal à 0 pour les deux classes à risques faible et moyen qui disparaîssent alors qu'il est égal à 0.44 pour la région à risque très faible et à 0.03 et à 0.31 pour les deux classes à risques fort et très fort qui sont mal identifiées. Pour les modèles *semi-graduel* et *graduel*, ce coefficient est élevé pour les classes à risques fort et très fort avec des valeurs égales respectivement à (0.89 et 0.92) contre (0.91 et 0.96). Comme pour le modèle de Potts, le coefficient DSC pour la classe moyenne est égal à 0 pour les modèles *semi-graduel* et *graduel*. Par contre, pour la classe à risque faible il n'y a que le *semi-graduel* qui a une bonne performance comparée aux deux autres avec un coefficient égal à 0.24. Pour les régions à risque très faible, le *semi-graduel* fournit des résultats meilleurs avec un DSC égal à 0.62 contre 0.33 pour le modèle *graduel* et 0.44 pour celui de Potts.

Conclusion sur la comparaison entre les trois formes de $\mathbb{B}$: Ces expériences montrent que les modèles *semi-graduel* et le *graduel* améliorent les résultats que l'on peut obtenir avec le modèle de Potts et suggérent qu'il est moins probable de perdre les classes de risque les plus problèmatiques avec le modèle *semi-graduel* qu'avec le modèle *graduel* comme illustré par la figure (6.10) pour l'exemple à 5 classes. De même pour l'exemple à 3 classes, où le

Résultats pour l'exemple à 5 classes			
Modèle pour $\mathbb{B}$	Niveau de risque	DSC	Risques estimés
Modèle de Potts	très faible	0.44	$5.18e^{-05}$
$\mathbb{B} = b\ I_K$	faible	0	$8.12e^{-05}$
	moyen	0	$8.12e^{-05}$
	fort	0.03	$1.15e^{-04}$
	très fort	0.31	$5.89e^{-04}$
Modèle *graduel*	très faible	0.33	$6.79e^{-05}$
$\mathbb{B}(k,l) = b\ (1 - \lvert k-l\rvert/(K-1))$	faible	0	$1.30e^{-04}$
	moyen	0	$5.16e^{-04}$
	fort	0.89	$6.70e^{-04}$
	très fort	0.92	$1.01e^{-03}$
Modèle *semi-graduel*	très faible	0.62	$1.83e^{-05}$
$\mathbb{B}(k,l) = b\ [1 - \lvert k-l\rvert/2]_+$	faible	0.24	$1.02e^{-04}$
	moyen	0	$4.83e^{-04}$
	fort	0.91	$7.99e^{-04}$
	très fort	0.96	$1.02e^{-03}$
Résultats pour les 100 simulations à 5 classes			
Modèle pour $\mathbb{B}$	Niveau de risque	DSC	Risques estimés
Modèle de Potts	très faible	0.19 (0.21)	$2.45e^{-05}(2.69e^{-05})$
$\mathbb{B} = b\ I_K$	faible	0.25 (0.21)	$7.47e^{-05}(7.01e^{-05})$
	moyen	0.32 (0.25)	$1.82e^{-04}(1.53e^{-04})$
	fort	0.48 (0.32)	$4.15e^{-04}(1.81e^{-04})$
	très fort	0.57 (0.25)	$8.47e^{-04}(3.95e^{-04})$
Modèle *graduel*	très faible	0.40 (0.24)	$4.21e^{-05}(2.45e^{-05})$
$\mathbb{B}(k,l) = b\ (1 - \lvert k-l\rvert/(K-1))$	faible	0.19 (0.19)	$1.33e^{-04}(1.01e^{-04})$
	moyen	0.14 (0.28)	$2.64e^{-04}(2.00e^{-04})$
	fort	0.75 (0.32)	$4.98e^{-04}(2.02e^{-05})$
	très fort	0.89 (0.07)	$1.01e^{-03}(6.09e^{-05})$
Modèle *semi-graduel*	très faible	0.56 (0.20)	$2.07e^{-05}(1.53e^{-05})$
$\mathbb{B}(k,l) = b\ [1 - \lvert k-l\rvert/2]_+$	faible	0.29 (0.17)	$9.62e^{-05}(4.39e^{-05})$
	moyen	0.09 (0.18)	$3.33e^{-04}(1.37e^{-04})$
	fort	0.66 (0.38)	$5.57e^{-04}(1.05e^{-04})$
	très fort	0.83 (0.17)	$1.05e^{-03}(7.66e^{-05})$

TABLE 6.6 – La moyenne et l'écart-type du coefficient de similarité de Dice (DSC) et le risque estimé pour chaque classe en utilisant différents modèles pour $\mathbb{B}$ et S_{tra} comme méthode d'initialisation pour le EM **champ-moyen**.

Résultats pour l'exemple à 3 classes			
	Niveau de risque	DSC	Risques estimés
Le modèle BYM	faible	0.60	$1.61e^{-04}$
	moyen	0.19	$1.08e^{-03}$
	fort	0.05	$3.30e^{-03}$
Résultats pour l'exemple à 5 classes			
	Niveau de risque	DSC	Risques estimés
Le modèle BYM	très faible	0.54	$8.42e^{-05}$
	faible	0.25	$1.26e^{-04}$
	moyen	0.41	$1.84e^{-04}$
	fort	0.74	$6.18e^{-04}$
	très fort	0.07	$1.97e^{-03}$

TABLE 6.7 – Le coefficient de similarité de Dice (DSC) et le risque estimé pour chaque classe en utilisant différents modèles pour $\mathbb{B}$ et S_{tra} comme méthode d'initialisation pour le EM **champ-moyen**.

modèle de Potts perd une classe alors que les modèles *semi-graduel* et *graduel* retrouvent correctement les à 3 classes (voir figure 6.9).

6.2.6 Comparaison du modèle *semi-graduel* avec le BYM pour les exemples à 3 classes et à 5 classes

Nous comparons notre modèle combiné à la stratégie d'initialisation S_{tra} au modèle proposé par Besag et al. (1991) et noté ici BYM (section 3.2.2.1) qui est le modèle le plus utilisé en épidémiologie animale. Ce modèle produit des valeurs de risque estimées différentes en chaque hexagone. Une étape supplémentaire est donc nécessaire pour obtenir une segmentation en un nombre fini K de niveaux de risque. Ceci est obtenu en appliquant une procédure de classification sur les valeurs continues estimées. Une méthode qui est communément utilisée est l'algorithme EM pour les mélanges Gaussiens. C'est cette mèthode que nous utilisons pour obtenir les cartes de risque présentées à la figure 6.11 pour les exemples à 3 classes et à 5 classes. Nous présentons aussi dans le tableau 6.7 les estimations du risque obtenues et les valeurs du critère DSC pour ces deux exemples.

Pour l'exemple à 3 classes, le BYM surestime les valeurs des trois niveaux de risque. On

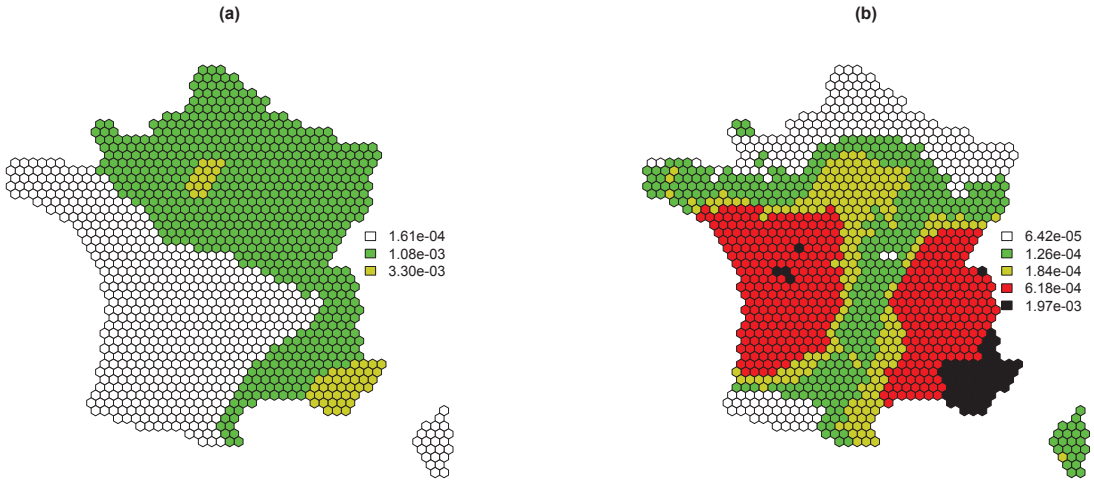


FIGURE 6.11 – Cartes de risque obtenues avec le BYM pour les exemples : (a) à 3 classes et (b) à 5 classes.

peut voir dans le tableau 6.7 que ce sont spécialement les risques faibles qui ont tendance à être surestimés le plus. En effet, le risque pour la première classe est estimé à la valeur $\hat{\lambda}_1 = 1.61e^{-04}$ alors que la vraie valeur de risque, pour cette région, est égale à $\lambda_1 = 1e^{-05}$. De même pour la deuxième classe, la valeur de risque est estimée à $\hat{\lambda}_2 = 1.08e^{-03}$ tandis que la vraie valeur est égale à $\lambda_2 = 1e^{-04}$. Enfin, le niveau de risque de la troisième classe est estimé à $\hat{\lambda}_3 = 3.03e^{-03}$ alors qu'il l'est à $\lambda_3 = 1e^{-03}$. En comparant ces estimations avec celles obtenues par le modèle *semi-graduel* (voir tableau 6.5), on peut voir que les niveaux de risque sont mal estimés par le BYM alors qu'ils sont très proches des vraies valeurs pour le modèle *semi-graduel*.

On voit sur la figure 6.11 (a) que les classes ne sont pas proprement identifiées. En effet, le bord sud-est est affecté à la classe à risque fort alors qu'il est dans la classe à risque faible dans la vraie classification (voir figure 6.2 (a)). Les régions à risque fort de la vraie segmentation sont affectées à la deuxième classe dans la segmentation obtenue. La classe à risque fort dans la segmentation obtenue correpond aux régions où l'effectif de la population est petit (par exemple le sud-est de la France). Les valeurs du critère DSC montrent des coefficients de recouvrement petits qui sont égales à 0.05 pour la classe à risque fort et à 0.19 à risque moyen. Par contre, pour la classe à risque faible, ce coefficient est plus élevé et est égal à 0.60. Ces coefficients de recouvrement montrent que la performance du modèle BYM reste moins

bonne que le modèle *semi-graduel* pour lequel les valeurs de ce coefficient vont de 0.84 pour la classe à risque faible à 1 pour celle à risque fort (voir tableau 6.5).
Pour l'exemple à 5 classes, on peut faire les mêmes remarques concernant les estimations des niveaux de risque. Les risques des trois premières classes sont surestimés et sont évalués à $\hat{\lambda}_1 = 6.42e^{-05}$ au lieu de $\lambda_1 = 1e^{-05}$, à $\hat{\lambda}_2 = 1.26e^{-04}$ pour $\lambda_2 = 5e^{-05}$ et $\hat{\lambda}_3 = 1.84e^{-04}$ pour $\lambda_3 = 1e^{-04}$. Pour la classe à risque fort, le risque est estimé à $\hat{\lambda}_4 = 6.18e^{-04}$ qui est supérieur à la vraie valeur $\lambda_4 = 5e^{-04}$. De même pour la région à risque très fort, le risque est encore surestimé et sa valeur est égale à $\hat{\lambda}_5 = 1.97e^{-03}$ au lieu de $\lambda_5 = 1e^{-03}$. Quant à la classification, on peut voir sur la figure 6.11 (b) que le BYM arrive à retrouver les 5 classes de la vraie carte (figure 6.2 (b)). On peut faire les mêmes remarques, que pour l'exemple à 3 classes, et voir dans la figure 6.11 (b) que les régions qui sont détectées à risque très fort dans la segmentation sont celles où l'effectif de la population est petit (encore le sud-est de la France). On peut faire les mêmes remarques concernant la structure des classes et voir que le BYM n'identifie pas correctement les classes. En examinant le tableau 6.7, on peut voir que la performance de ce modèle, en terme de DSC, n'est pas très bonne pour les classes à risque très fort avec un coefficient égal à 0.07 (contre 0.96 pour le *semi-graduel*). Pour la classe à risque fort, ce coefficient est raisonnable avec une valeur égale à 0.74 mais reste inférieure à la valeur obtenue par le modèle *semi-graduel* qui est égale à 0.91. Quant à la classe moyenne, la performance du BYM est meilleure avec une valeur égale à 0.41 contre 0 pour le *semi-graduel*. Pour les régions à risques faible et très faible, la performance des deux modèles reste comparable avec des valeurs de DSC égales respectivement à (0.54 et 0.25) pour le BYM contre (0.62 et 0.24) pour le *semi-graduel*. En comparant aussi les niveaux de risque obtenues par le BYM et le *semi-graduel* (voir tableau 6.5) pour le même exemple, on peut voir que les deux modèles fournissent des estimations comparables.

Conclusion sur la comparaison avec le modèle BYM : Ces résultats suggèrent que le modèle usuel, le modèle BYM, dans sa version la plus simple, n'est pas vraiment adapté aux maladies rares pour des populations inhomogènes. En effet, il tend à estimer des risques élevés dans des régions avec des petits effectifs de population. On le voit clairement pour la région sud-est de la France, pour laquelle l'effectif de la population est petit. Le BYM l'affecte à la plus forte classe de risque pour les deux exemples à 3 classes et à 5 classes. Il a aussi tendance à produire des cartes de risque avec des frontières moins clairement délimitées que celles produites par le modèle *semi-graduel*. Au vu de ces résultats, on peut conclure que notre modèle (le *semi-graduel*) est plus adapté que le BYM pour nos données en terme de

localisation précise des régions, surtout celles à fort risque.

6.3 Données d'Encéphalopatie Spongiforme Bovine (ESB)

Dans cette section, nous nous concentrons sur les données réelles qui sont l'application principale de notre travail. Nous présentons les données de l'ESB à la section 6.3.1 et montrons les résultats obtenus par le modèle *semi-graduel* et le modèle BYM à la section 6.3.2.

6.3.1 Description des données

L'encéphalopathie spongiforme bovine (ESB) est une infection dégénérative du système nerveux central des bovins. C'est une maladie mortelle, causée par un agent infectieux moléculaire appelé la protéine prion. Une épizootie d'ESB a touché l'Europe et en particulier le Royaume-Uni entre 1988 et 2008. Le total des cas observés était d'environ 200000 animaux durant cette période. En France, un peu moins de 1000 têtes ont été déclarées atteintes. Cette épidémie trouve son origine dans l'utilisation de farines de viande pour l'alimentation des bovins, obtenues à partir de parties non consommées des carcasses bovines et de cadavres d'animaux. L'épidémie a pris une tournure particulière quand les scientifiques se sont apercus en 1996 de la possibilité de transmission de la maladie à l'homme par le biais de la consommation de produits carnés. Plusieurs analyses spatiales de cette maladie ont été publiées (Abrial et al., 2005; Allepuz et al., 2007; Paul et al., 2007).
Les données observées, dont on dispose, se composent d'un nombre de cas infectés pour chaque hexagone du territoire Français et de la démographie bovine entre le 1 juillet 2001 et le 31 décembre 2005. La figure 6.12 est une cartographie du nombre des cas observés. On peut voir qu'il y a peu de cas contaminés et le maximum des cas observés ne dépasse pas les 5 cas.

6.3.2 Résultats obtenus pour l'ESB

Sur les données décrites ci-dessus, nous appliquons le modèle *semi-graduel* avec l'EM **champ-moyen** initialisé par la stratégie S_{tra} décrite en section 5.4. En ce qui concerne le nombre de classes, le BIC approché (voir section 4.4) suggère de choisir $K = 3$. Pour comparaison, nous considérons le modèle BYM qui est largement utilisé en épidémiologie.
La figure 6.13 montre les cartes de risque obtenues avec avec le modèle *semi-graduel* (figure

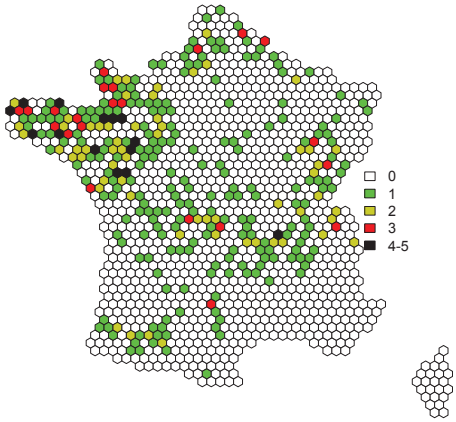


FIGURE 6.12 – Les nombres d'ESB enregistrés entre le 1 Juillet 2001 et le 31 Décembre 2005 en France.

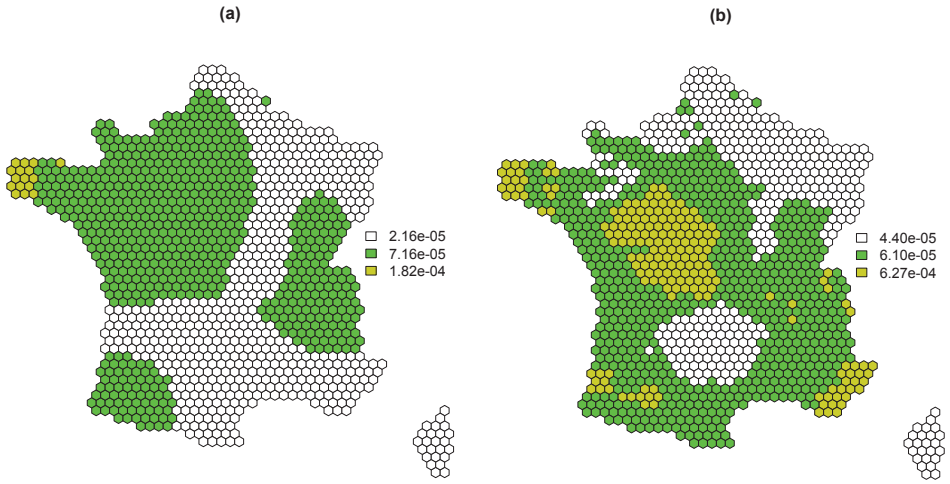


FIGURE 6.13 – Cartes de risque estimées en utilisant : (a) le modèle *semi-graduel* et (b) le modèle BYM.

6.13 (a)) et le modèle BYM (figure 6.13 (b)). Selon les connaissances *a priori* des experts de l'ESB en France, les régions à risque élevé sont localisées en Bretagne, dans le centre, dans le sud-ouest de la France et les Alpes.
En ce qui concerne la carte du risque estimé par le modèle BYM (voir figure 6.13 (b)), des

régions à fort risque supplémentaires à celles définies par les experts sont mises en évidence. Les frontières ne sont pas clairement délimitées incluant des régions connues pour leur risque faible et on observe qu'il y a des hexagones isolés à l'intérieur de régions de risque homogène. Typiquement, la région Rhône-Alpes (connue pour être une région à fort risque) n'est pas clairement identifiée mais fusionnée avec la région du sud-est considérée, par les experts, comme une région à risque peu élevé par rapport aux régions Rhône-Alpes et sud-ouest. Dans cette région urbaine le nombre de cas observés et l'effectif de la population de vaches sont petits.

Quant à la cartographie obtenue par le modèle *semi-graduel* (voir figure 6.13 (a)), trois régions sont clairement délimitées et correspondent à celles attendues par les experts. Contrairement au résultat obtenu par le modèle BYM, la Côte d'Azur est identifiée par le modèle *semi-graduel* comme une région à faible risque. La Bretagne est identifiée comme une région à risque plus fort que le centre, le sud-ouest et la région Rhône-Alpes qui sont considérées comme à risque fort par le *semi-graduel*. On peut observer que les régions obtenues, par ce modèle, sont homogènes et peu d'hexagones sont isolés à l'intérieur de ces régions.

En ce qui concerne les estimations du risque, les deux modèles estiment les deux premières classes à des valeurs comparables avec $\hat{\lambda}_1 = 2.16e^{-05}$ et $\hat{\lambda}_2 = 7.16e^{-05}$ pour le *semi-graduel* contre $\hat{\lambda}_1 = 4.40e^{-05}$ et $\hat{\lambda}_2 = 6.10e^{-05}$ pour le modèle BYM. En ce qui concerne la classe à fort risque, le modèle BYM a tendance à produire une valeur estimée ($\hat{\lambda}_3 = 6.27e^{-04}$) plus grande que celle fournie ($\hat{\lambda}_3 = 1.82e^{-04}$) par le *semi-graduel*.

On peut ainsi dire que les résultats obtenus par le modèle *semi-graduel* sont plus fiables que ceux du modèle BYM. Nous soupçonnons que la mauvaise estimation donnée par le modèle BYM traduit essentiellement la force de l'*a priori* spatial en l'absence d'une information suffisante sur la structure spatiale dans les données. Nous supposons que ce comportement ressort pour les données de l'ESB car le nombre de cas observés est très petit. La carte résultante utilisant le modèle BYM refléterait donc principalement l'information *a priori* plutôt que celle présente dans les observations. Nous avons aussi comparé les résultats obtenus par les trois modèles décrits en section 6.2.5 en utilisant la même stratégie d'initialisation S_{tra}. En appliquant le critère BIC pour le choix du nombre de classes, nous avons retenu aussi $K = 3$ pour le modèle de Potts et le modèle *graduel*, qui est équivalent au *semi-graduel* dans ce cas. En comparant toutes les valeurs du BIC, les meilleurs scores pour ces trois modèles sont équivalents mais les cartes obtenues par le *semi-graduel* et le *graduel* (Figure 6.13(a)) ont clairement plus de sens que la carte obtenue par le modèle de Potts (Figure 6.14). En effet, ce modèle a tendance à affecter la région Rhône-Alpes à la classe à plus fort risque comme c'est le cas pour la Bretagne. En revanche, on retrouve le même problème concernant le sud-est qu'avec le modèle BYM. Cette région est affectée à une classe à fort risque mais

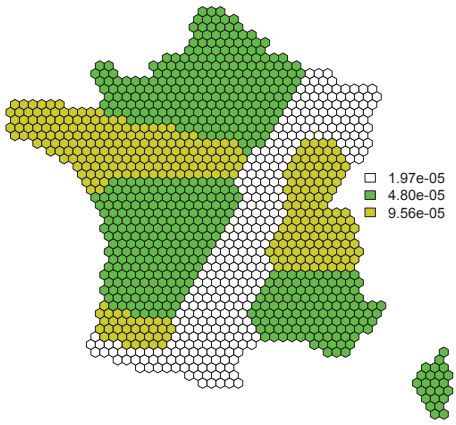


FIGURE 6.14 – Carte de risque estimé en utilisant le modèle de Potts et le EM **champ-moyen** initialisé par la stratégie S_{tra}.

moins fort que celle dans laquelle sont affectées la région Rhône-Alpes et la Bretagne. Quant au sud-ouest de la France, le modèle de Potts divise cette région en deux parties et classe la première moitié dans la classe à plus fort risque et l'autre dans la classe à faible risque. Quant aux valeurs de risque estimées par le modèle de Potts elles sont inférieures à celles obtenues par le *semi-graduel* et par le modèle BYM (voir figure 6.14).

6.4 Discussion

Dans ce chapitre nous avons présenté les résultats, sur les données simulées et les données réelles de l'ESB, des expériences que nous avons conduites pour étudier le comportement du modèle *semi-graduel* et de la stratégie d'initialisation S_{tra}.
Afin d'évaluer la performance de S_{tra}, nous l'avons comparé à deux autres méthodes d'initialisation. Ces deux autres stratégies choisies sont celles qui semblaient donner des résultats "raisonnables" sur nos données simulées en terme de classification et d'estimation des niveaux de risque. Pour faire cette comparaison, nous avons effectué des tests sur des données simulées qui ont des caractéristiques semblables à celles de l'ESB. Nous avons choisi de simuler deux situations qui sont représentées par les exemples à 3 classes et à 5 classes. L'exemple à 3 classes représente un jeu de donnée plus facile que l'exemple à 5 classes avec moins de classes et avec des frontières plus faciles à séparer que celles avec le à 5 classes.
En ce qui concerne l'exemple à 3 classes, les trois stratégies comparées sont capables, plus ou moins, de restaurer la vraie segmentation et S_{tra} fournit des résultats comparables aux deux

autres. Pour l'exemple à 5 classes, S_{tra} semble fournir des résultats meilleurs surtout pour les niveaux de risque élevés spécialement en terme de classification.
Signalons, toutefois, que les données étudiées sont complexes et qu'il est difficile de restaurer les structures qui y sont sous-jacentes. Cela explique, en partie, le fait que le comportement des stratégies n'est pas très différent pour les niveaux à faible risque, qui représentent les classes les plus problématiques pour notre modèle. Par contre, pour les niveaux de risque qui sont plus élevés, on voit une différence avec une meilleure performance pour S_{tra}.
Nous avons aussi comparé le modèle *a priori* du champ de Markov caché (le *semi-graduel*) au modèle de Potts et à une autre variante de ce modèle (le modèle *graduel*). Rappelons que la différence entre ces trois modèles réside dans la construction de la matrice $\mathbb{B}$. Les résultats obtenus montrent que, pour le type de données que l'on veut traiter, le *semi-graduel* a une performance meilleure que les deux autres modèles.
Pour finir notre étude, nous avons comparé notre modèle de champ de Markov caché discret (le *semi-graduel*) au modèle BYM, qui est le plus utilisé en épidémiologie animale. Nous avons montré que le modèle *semi-graduel* fournit des cartes de risque plus fiables que les cartes obtenues par le traditionnel modèle BYM, avec moins d'erreurs de classification et plus de régions clairement délimitées en zones de risque.
L'application de notre méthodologie (modèle *semi-graduel* et stratégie S_{tra}) sur les données réelles de l'ESB fournit des résultats satisfaisants et en accord avec la connaissance des experts sur cette épidémie. Nous nous contentons de cette comparaison à défaut d'avoir un critère de sélection de modèle approprié. En effet, pour les modèles Bayésiens c'est le DIC qui est le plus adapté et dans notre cas c'est le BIC qu'on utilise.

CHAPITRE 7

Conclusion et Perspectives

Sommaire

7.1 Conclusion

Nous avons développé dans ce travail une méthodologie pour aborder le problème de la classification en niveaux de risque d'infection pour les maladies non contagieuses et rares en épidémiologie.
La première partie de cette méthodologie repose sur une approche probabiliste fondée sur une modélisation markovienne pour prendre en compte les dépendances entre les unités géographiques de la région étudiée. Le modèle que l'on propose est une variante du modèle de Potts, qui est largement utilisé dans les problèmes de segmentation d'images, pour capter cette dépendance tout en intégrant l'hypothèse épidémiologique de la gradation du risque. Ceci se fait essentiellement par le remplacement du paramètre d'interaction, un simple scalaire dans le cas du modèle de Potts, par une matrice d'interaction $\mathbb{B}$. Un premier avantage de cette approche est la grande flexibilité qu'apporte la matrice $\mathbb{B}$ à la modélisation des dépendances spatiales entre les individus. Un deuxième avantage est que les approches basées sur les champs de Markov cachés permettent une modélisation flexible des phénomènes tout en considérant les connaissances biologiques qu'ont les experts sur les données traitées.
La deuxième partie de notre méthodologie concerne l'estimation des paramètres. Nous avons choisi d'utiliser des approximations de type champ moyen et l'algorithme NREM, originellement proposé par Celeux et al. (2003) pour le modèle classique de champ de Markov caché à bruit indépendant. Cet algorithme représente une alternative aux méthodes d'estimation basées sur des simulations intensives de type Markov Chain Monte Carlo (MCMC) qui sont

utilisées largement pour l'inférence dans les modèles hiérarchiques bayésiens proposés pour la cartographie du risque en épidémiologie. Le premier avantage de l'algorithme EM de type champ moyen est que la classification est obtenue directement à partir des probabilités *a posteriori* calculées à l'étape (**E**) et aucune méthode de classification *a posteriori* n'est requise. Un deuxième avantage est la rapidité de mise en œuvre de l'algorithme EM par rapport aux algorithmes de type MCMC qui restent assez coûteux en temps de calcul.
L'algorithme EM et ses versions approchées ne sont pas pour autant sans défauts. Aussi, dans une troisième partie de notre méthodologie, nous avons considéré le problème de l'initialisation de l'algorithme EM. Nous avons proposé une stratégie complète d'initialisation pour les paramètres du modèle. Cette stratégie est décomposée en deux étapes. La première est adaptée aux modèles de mélange de Poisson quand les classes sont mal séparées. La seconde étape permet de produire de nouvelles valeurs d'initialisation en prenant en compte et les données et l'*a priori* imposé sur la structure spatiale des données.
Nous avons appliqué notre modèle sur des données simulées réalistes et des données réelles pour en évaluer la performance. Nous avons comparé d'une part le modèle proposé à d'autres modèles avec différentes formes de matrice d'interaction $\mathbb{B}$. Les résultats obtenus montrent que pour les données considérées, notre construction de la matrice $\mathbb{B}$ donnée par une matrice bi-diagonale convenait. Nous avons d'autre part aussi comparé la performance de notre stratégie d'initialisation à d'autres utilisées usuellement. Les résultats obtenus suggèrent que cette technique d'initialisation améliore la classification des unités à risque élevé, ce qui est un point important du fait que ces unités sont d'un intérêt majeur en épidémiologie.
L'approche que nous avons proposée est différente des modèles existants dans la littérature pour la cartographie du risque en épidémiologie. En particulier, notre approche est basée en priorité sur une idée de classification plutôt que sur une estimation du risque seul. Le but de cette méthodologie n'est toutefois pas de remplacer les approches précédentes, mais de proposer et de rendre disponible un cadre d'étude différent relativement simple à appréhender pour les praticiens. Dans certaines applications et pour des questions assez spécifiques, notamment en épidémiologie animale, ce cadre permet d'apporter des informations complèmentaires sur la structure sous-jacente du risque en fournissant une classification des unités géographiques avec une bonne délimitation des zones à différents niveaux de risque.

7.2 Perspectives

La méthodologie que nous proposons semble produire des résultats satisfaisants sur les données traitées qui nous intéressaient en priorité dans ce travail. Toutefois, le modèle pourrait

être complété ou amélioré afin de répondre à d'autres questions que pourraient se poser les épidémiologistes, par exemple sur un autre type de données ou dans un autre cadre d'étude. Nous esquissons dans cette partie quelques pistes qui nous semblent intéressantes.

Un système de voisinage plus sophistiqué. Le système de voisinage dans notre modèle est basé sur la proximité géographique. Il serait intéressant d'introduire des relations de dépendance plus complexes. Ces relations peuvent représenter des similarités entre les unités géographiques qui sont dûes à des facteurs environnementaux par exemple. Typiquement, pour l'exemple de l'ESB, les sites peuvent être considérés comme voisins s'ils ont le même fournisseur de farines animales. On pourrait aussi, éventuellement, introduire des interactions plus complexes pour prendre en compte un facteur écologique comme la dissémination par le vent.

Graphe non régulier au lieu de graphe régulier. Un voisinage plus sophistiqué peut éventuellement résulter en une structure non régulière du graphe sous-jacent. Or dans notre étude, nous n'avons pas explicitement testé l'influence de la régularite du graphe. En effet, nous avons effectué une transformation de notre carte de France afin d'obtenir une structure régulière et nous avons effectué tout nos tests sur ce type de réseau. Le passage d'un graphe régulier à un graphe non régulier ne pose en principe pas de problème méthodologique mais nous soupçonnons que dans un tel cas, une pondération des voisins en fonction de leur nombre ou de leur influence pourrait être pertinente et intéressante à étudier.

Amélioration de la forme de la matrice d'interaction $\mathbb{B}$. Sur un autre plan, la nature de la matrice $\mathbb{B}$ (diagonale ou pleine) reflète les influences relatives des différentes classes à travers le réseau. Cette matrice, comme mentionné auparavant, offre une grande flexibité. La forme de cette matrice peut également dépendre des données disponibles et des connaissances d'experts et permettre d'introduire une autre forme de pondération sur les classes cette fois. Une extension simple que nous voyons pour cette matrice est de considérer deux paramètres d'interaction b_1 et b_2 différents au lieu d'un seul paramètre b. Dans ce cas la matrice B serait définie comme suit :

$$\begin{aligned}
\mathbb{B}(k,k) &= b_1 \quad \text{pour tout } k = \{1 \dots K\} \\
\mathbb{B}(k,l) &= b_2 \quad \text{pour tout } (k,l) \text{ tel que } |k-l| = 1 \\
\mathbb{B}(k,l) &= 0 \quad \text{sinon.}
\end{aligned}$$

L'idée de cette extension est de permettre de détecter d'avantage les discontinuités en évitant le surlissage qui peut se produire dans le cas d'un seul paramètre du fait qu'il n'y a pas de distinction entre les termes diagonaux et hors diagonale. Une autre extension naturelle serait de considérer un contexte bayésien pour ces deux paramètres et de leur assigner des lois *a priori*.

Introduction de covariables dans le modèle. Afin de mieux comprendre les mécanismes qui régissent la propagation d'une maladie, il est possible d'introduire des covariables à différents niveaux de la hiérarchie de notre modèle sans trop changer sa structure. En effet, la seule différence consisterait à remplacer par exemple le risque relatif λ_{z_i} par $\lambda_{z_i} \exp^{\sum_j x_{ij} c_j}$ avec x_{ij} représentant les covariables en question mesurées en i et c_j les coefficients qui leur sont associés. Dans le cas où les covariables considérées ont un effet, leur introduction dans le modèle peut améliorer la classification obtenue et augmenter la précision des estimations des niveaux de risque. L'utilisation de l'algorithme EM champ-moyen pour l'inférence se généralise facilement à ce cas.

Critère de choix du modèle. Pour pouvoir comparer notre modèle, sur les données réelles, à un modèle hiérarchique bayésien, il serait important de mener une comparaison plus quantitative que celle qui consiste à observer la plus ou moins bonne adéquation des cartes produites aux attentes des épidémiologistes. Pour ce faire, il faudrait utiliser un critère de sélection de modèle. Nous avons montré que pour notre modèle le BIC est approprié. Pour les modèles bayésiens, c'est le DIC qui est le plus utilisé. Il serait donc intéressant de réfléchir à un critère qui pourrait permettre cette comparaison.

Extension du modèle au contexte spatio-temporel. La plupart des données de comptage géolocalisées en épidémiologie sont collectées dans le temps et l'espace. La prise en compte de la dimension temporelle afin de mieux comprendre l'évolution des maladies est un enjeu important en épidémiologie. Plusieurs méthodes ont été proposées dans la littérature afin de temporaliser les cartes spatiales. En général, ces modèles sont des extensions des modèles spatiaux présentés en section 3.2. Dans ce cas, la supposition commune à tous les modèles est :

$$Y_{ij} \sim Pois(E_{ij}\lambda_{ij})$$

pour l'unité i au temps $j, \{j \in 1, \ldots, T\}$. Encore une fois la différence entre les modèles existants réside dans la loi *a priori* qu'on donne à λ_{ij}. Bernadinelli et al. (1995) proposent

d'étendre la formule (3.3) au spatio-temporel en introduisant des termes modélisant séparément les interactions spatiales, temporelles ou encore spatio-temporelles. Waller et al. (1997), Xia and Carlin (1998) proposent un modèle différent où ils introduisent des effets aléatoires décrivant respectivement l'hétérogénéité structurée et l'hétérogénéité non structurée et qui peuvent varier dans le temps. Ceci représente en quelque sorte une extension des modèles de type BYM. D'autres exemples plus récents de modélisation spatio-temporelle concernent les modèles de mélange comme dans Boehning et al. (2000) où il examine la possibilité de traiter les aspects temporels séparément et sans interaction. D'autres modèles alternatifs ont été proposés par Besag and Tantrum (2003) qui proposent une approche spatio-temporelle auto-logistique.

Au vu de ce grand intérêt pour le spatio-temporel, nous avons pensé à étendre notre modèle spatial au spatio-temporel afin d'avoir des cartes comparables dans le temps pour aider les épidémiologistes à encore mieux comprendre les mécanismes d'une épidémie en visualisant son évolution dans le temps. Une extension naturelle du modèle bidimensionnel proposé est un modèle tridimensionnel dans lequel le voisinage est défini par un graphe prenant sinultanément les voisins spatiaux et temporels pour chaque unité géographique du domaine étudié. Dans ce cas, le paramètre d'interaction b ne peut plus être interprété comme un paramètre d'interaction spatiale. Cette première façon de prendre en compte l'aspect temporel dans notre modèle spatial est facile à mettre en œuvre vu que la structure de notre graphe de voisinage est assez flexible pour tenir compte de ce nouveau voisinage.

Nous avons effectué quelques tentatives dans ce sens, que nous n'avons pas montrés dans ce travail, étant donné que nos interrogations sur les questions auxquelles on veut répondre restent nombreuses. Nous croyons toutefois que la comparaison de notre modèle avec et sans temporalisation pourrait apporter une information, bien que très partielle, sur l'importance de prendre en compte des relations temporelles. Un rafinement consiste ensuite à séparer les composantes spatiale et temporelle sont séparées avec deux paramètres d'interaction b_s pour le spatial et b_t pour le temporel. Les dépendances temporelles peuvent être modélisées de la même manière que les dépendances spatiales (le modèle semi-graduel). Ce modèle sera intéressant s'il y a une forte corrélation temporelle entre les cartes spatiales. C'est le cas notamment des maladies contagieuses pour lesquelles on s'attend à avoir une forte évolution de l'épidémie entre le temps t et $t + 1$.

Extension du modèle aux maladies transmissibles ou contagieuses. Le modèle étudié dans cette thèse, comme les modèles bayésiens hiérarchiques usuels présentés au chapitre 3, sont adaptés aux maladies non transmissibles. Par exemple, l'application la plus classique

en maladie humaine concerne les cancers. Or la plus grande partie des maladies humaines ou animales sont transmissibles ou contagieuses. Il serait donc intéressant d'étendre les méthodes de cartographie du risque dans ce cadre. La loi de Poisson qui modélise les données ne serait alors plus adaptée en raison de son absence de mémoire. En effet, en raison de la diffusion, le fait d'abserver au moins un certain nombre de cas influe sur la probabilité d'en observer encore plus. Il faudrait donc utiliser une autre distribution que la Loi de Poisson pour modéliser les cas observés. Dans un contexte spatio-temporel, la complexité du voisinage à considérer serait certainement plus importante que pour des maladies non transmissibles. En effet, la transmission d'une unité en un temps donné à sa voisine au temps suivant nécessite la prise en compte dans le voisinage d'une unité des unités alentours aux temps suivant et précédent.

Annexe

Les résultats expérimentaux présentés dans cette thèse ont été réalisés à l'aide du logiciel SpaCEM^3 (Spatial Clustering with EM and Markov models) développé au sein de l'équipe Mistis à l'INRIA Rhône-Alpes.

Le logiciel est développé en C++. La dernière version de SpaCEM^3 (spacem3 2.0) propose une interface graphique développée avec la librairie QT. Le logiciel est disponible en téléchargement à l'adresse (http ://spacem3.gforge.inria.fr/) pour les systèmes d'exploitation Linux (package .deb et .rpm), Windows et MacOs, ainsi qu'une documentation sous forme de tutorial. Le logiciel peut être utilisé de deux façons :

- sans interface graphique en lançant directement l'exécutable
- avec interface graphique permettant de visualiser les données, d'effectuer leur classification et de visualiser les résultats.

SpaCEM^3 accepte les données sous forme de fichier .txt ou .dat en texte ou en binaire. Chaque individu est représenté par une ligne et chaque dimension ou chaque variable de cet individu par une colonne. Les dépendances entre les individus sont modélisées sous la forme d'un graphe de voisinage. Deux types de graphes peuvent être utilisés : le type Image correspondant aux N plus proches voisins dans une grille régulière, et le type Structure pour un graphe non régulier (la liste des voisins est alors à fournir dans une fichier text). Le logiciel SpaCEM^3 peut également construire certains graphes classiques (graphe de Delaunay, graphe de Gabriel, graphe de voisinage relatif, graphe des ε-voisins, graphe des k-voisins réciproques).

Ce logiciel propose une variété d'algorithmes pour la classification, supervisée ou non supervisée d'individus en interaction, sur lesquels sont mesurés des données uni ou multidimensionnelles. Ceci inclut la segmentation d'images, avec comme structure de dépendance sous-jacente des grilles régulières de pixels. Plus généralement, les algorithmes implémentés permettent de classer des données multimodales et dépendantes du fait de leur localisation spatiale ou du fait d'autres types de relations décrites par des structures graphiques quelconques.

L'approche principale se fonde sur l'utilisation de l'algorithme EM (ou sur approximation en champ moyen NREM) pour une classification floue et sur les modèles de champs de Markov pour la modélisation des dépendances. Les fonctionnalités de SpaCEM^3 incluent les points suivants :

- Classification non supervisée d'individus, basée sur une description des dépendances à l'aide d'un graphe non nécessairement régulier et un traitement basé sur les champs de Markov cachés. Les modèles markoviens disponibles incluent diverses extensions du

modèle de Potts standard, notamment avec la possibilité d'utiliser des modèles d'interaction plus généraux.
- Classification supervisée d'individus, basée sur la famille de modèles de Markov triplets avec des phases d'apprentissage et de test.
- Critère de sélection de modèles (BIC, ICL et leurs approximations en champ moyen) permettant de sélectionner "le meilleur" modèle de champs de Markov cachés en fonction des données.
- Simulation de modèles de champs de Markov généraux avec interactions d'ordre 1 et 2 (modèle de Potts et ses diverses extensions).
- Simulation de modèles de champs de Markov triplets généraux avec interactions d'ordre 1 et 2.
- Simulation de modèles de champs de Markov caché à bruit indépendant.
- Simulation de modèles de champs de Markov triplet à bruit indépendant.
- Possibilité de traiter le cas de données de très grande dimension dans un cadre markovien.
- Possibilité de traiter le cas d'observations manquantes avec imputation off-line (par les KNN, la moyenne), on-line (au cours de l'algorithme) ou sans imputation.

Algorithmes de classification. Les algorithmes disponibles dans SpaCEM^3 peuvent être répertoriés dans deux grandes classes :
- Les algorithmes dits usuels qui sont : ICM, Kmeans et l'algorithme EM et ses différentes versions (CEM, NEM et NCEM).
- Les algorithmes basés sur une approche variationnelle de l'algorithme EM sous modélisation markovienne : l'algorithme NREM décrit en chapitre 3 et ses variantes pour modèles triplets (Blanchet, 2007) et avec données manquantes.

Utilisation des algorithmes en pratique. Pour l'utilisation pratique, trois points sont importants : le choix de l'initialisation, le choix du critère d'arrêt et la sélection de modèle.

- **Techniques d'initialisation.** Dans SpaCEM^3, trois techniques d'initialisation de la classification sont proposées : une initialisation aléatoire, une initialisation par Kmeans, ou encore une initialisation fixée par l'utilisateur via un fichier texte.
- **Critères d'arrêt.** SpaCEM^3 peut calculer trois critères permettant de s'assurer de la bonne convergence des algorithmes de segmentation :
    1. un critère basé sur la différence des vraisemblances complétées entre deux itérations.

2. un critère basé sur la plus grande différence entre les probabilités conditionnelles de classification de chaque individu, entre deux itérations successives.
3. un critère basé sur la proportion d'individus pour lesquels la classification a changé entre deux itérations.

Pour arrêter les algorithmes, on peut également fixer un nombre d'itérations à effectuer. SpaCEM^3 permet en outre de visualiser l'évolution de ces critères et le comportement des paramètres au cours des itérations.

– **Sélection de modèle.** Les critères disponibles dans SpaCEM^3 sont Le Bayesian Information Criterion (BIC) qui est certainement le plus répandu et le critère Integrated Completed Likelihood (ICL) (Biernacki et al., 2000) qui permet de tenir compte de la pertinence de la classification obtenue.

Exemple d'utilisation de SpaCEM^3 avec l'interface graphique

Choix de données

Le type des données doit être spécifié par l'utilisateur. Deux choix sont possibles, "Image" ou "Structure". Pour le type Image, deux systèmes de voisinage sont disponibles : (i) le système de voisinage de premier ordre (chaque unité a 4 voisins). (ii) le système de voisinage de second ordre (chaque unité a 8 voisins). Pour le type Structure, un fichier .nei de la liste de voisins doit être fourni (voir figure 7.1 pour illustration).

Choix de modèle

Deux sortes de modèles sont disponibles dans SpaCEM^3 qui sont les modèles de mélanges indépendants "IID" et les modèles de champs de Markov cachés "HMRF". Le logiciel SpaCEM^3 traite le cas des champs de Markov discrets en se limitant aux potentiels sur les cliques d'ordre 1 et 2. Les modèles considérés sont des extensions du modèle de Potts. Le logiciel SpaCEM^3 permet de considérer quatre cas de matrice d'interaction $\mathbb{B}$: matrice pleine, matrice pleine avec composants diagonaux identiques, matrice diagonale et matrice proportionnelle à l'identité. Il en résulte 8 modèles possibles, selon ou non qu'on inclut des paramètres de champ externe α (voir (Blanchet et al., 2009) pour plus de détails).
La figure 7.2 illustre un exemple de choix de modèle pour les données simulées de l'exemple à 3 classes étudié à la section 6.2.2. Le paramètre à spécifier en premier est le nombre de

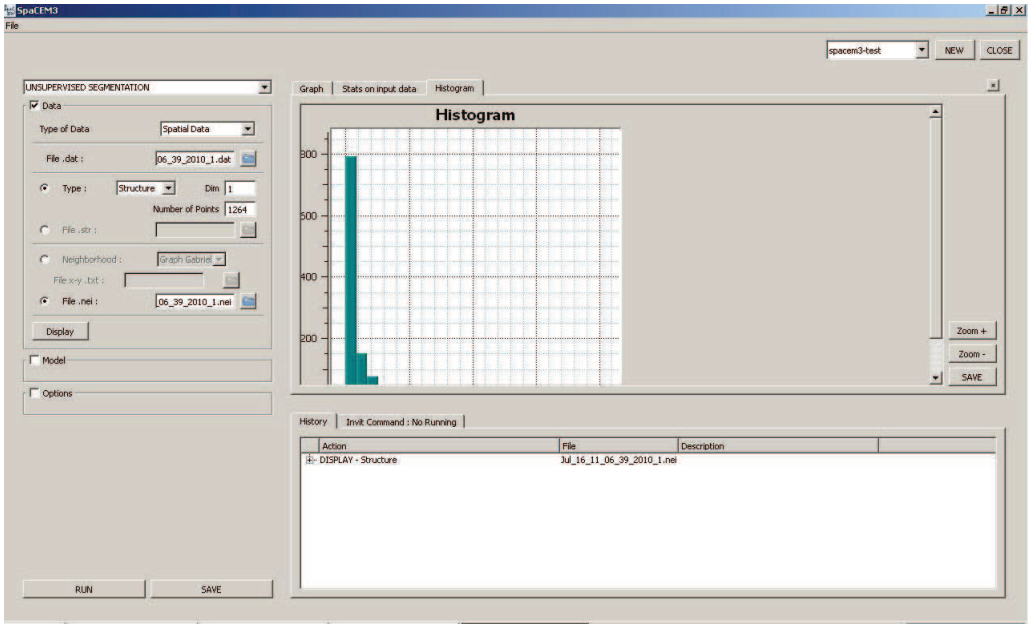


FIGURE 7.1 – Spécification des données de type Structure dans $SpaCEM^3$.

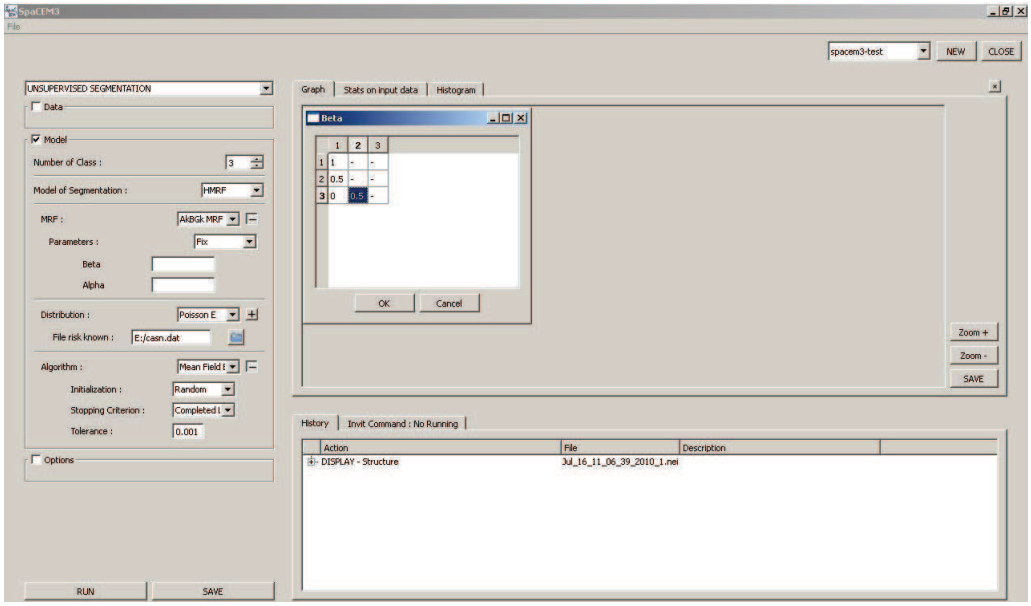


FIGURE 7.2 – Exemple de choix de modèle et d'algorithme d'estimation pour un jeu de données à 3-classes.

classes. Les paramètres $\boldsymbol{\alpha}$ et b sont à estimer ou à considérer comme connus. Dans l'état actuel de l'interface graphique, le seul choix proposé c'est ou d'estimer tous les paramètres ou

de les fixer tous. Dans cet exemple présenté, la matrice $\mathbb{B}$ est définie comme pour le modèle *semi-graduel* et le paramètre $\boldsymbol{\alpha}$ peut être spécifié si l'on dispose d'une connaissance *a priori*. Pour le modèle d'observations, la distribution PoissonE est choisie pour cet exemple. Cette distribution est spécifique à la cartographie du risque et dépendant du nombre de cas attendus E qui doit être fourni dans un fichier .dat. L'algorithme d'estimation choisi pour cet exemple est l'algorithme EM **champ-moyen**. Dans les options, on peut demander d'afficher les valeurs du BIC ou du critère ICL.

Exemple d'utilisation de SpaCEM^3 sans l'interface graphique

En l'état actuel des choses, le modèle *semi-graduel* que nous proposons n'est pas vraiment implémenté dans l'interface graphique. En effet, dans notre modèle les paramètres $\boldsymbol{\alpha}$ ne pas fixés et doivent être estimés et la matrice $\mathbb{B}$ est définie comme dans l'exemple illustré par la figure 7.2. La stratégie d'initialisation S_{tra} décrite en section 5.4.3 n'est pas implémentée non plus dans l'interface graphique. On donne ci-après un exemple d'utilisation pour l'estimation et la classification d'observations éventuellement manquantes en utilisant notre modèle *semi-graduel* combiné à notre procédure complète de recherche de valeurs initiales pour l'algorithme EM **champ-moyen**.

// Crée des données spatiales 'sdat' par lecture du fichier des observations et du graphe de voisinage :

```
Spatial_Data *sdat = new Spatial_Data();
sdat -> ReadFromFile("mes donnees");
```

// Retourne le graphe de voisinage dans 'nei'

```
Neighborhood_System *nei = new Neighborhood_System();
nei = sdat -> Get_NS();
```

// Retourne le nombre de sites N and la dimension D des données

```
uint N = sdat->Get_N();
uint D = sdat->Get_D();
```

// Lecture de fichier de cas attendus E

```
 string file_param_poisson= "cas attendus E"
```

// Crée un vecteur "PoissonE" de $K = 3$ poisonniennes

```
uint K =3;
vector<Distribution*> poisson(K);
    for (uint k=0;k<K;k++){
      poisson[k] = new Poisson_E(D);
dynamic_cast<Poisson_E *>(poisson[k])
->readRFromFile(file_param_poisson);

      osModel << "R" <<k+1<<" "<< D << " ";

       vector<float> R(1,Rvals[k]);
       dynamic_cast<Poisson_E *>(poisson[k])->Set_R(R);
   }
```

// Crée un champ de Markov (modèle *semi-graduel*) 'mrf' à K=3 classes et le graphe de voisinage 'nei' des données

```
 AkBGfix_MRF *mrf = new  AkBGfix_MRF(nei,K);
```

// Crée un champ de Markov caché 'hmrf'

```
  HMRF *hmrf = new HMRF(mrf,poisson);
```

// Crée un algorithme de segmentation 'algo' (**champ-moyen**)

```
 Mean_Field_EM *algo = new Mean_Field_EM(hmrf,sdat);
```

// Procédure complète de recherche de valeurs initiales S_{tra} avec un seuil $\varepsilon = 1e^{-03}$ pour le critère de convergence.

```
sprintf(parname,"M valeurs initiales de parametres", trialnum);
hmrf->ReadFromFile(parname);
 {
```

```
      mrf->FixBeta(true);
     }

 algo->Init_Density(cout);
 algo->Set_Stopping_Rule1(1,0.001);
    cvg=false;
    iter=0;
    while (!cvg){
     osModel << endl;
     osModel << iter<< " ";

      algo->Run(1,osModel);
      hmrf->WriteToFile(std::string(iteration_params_out));

      algo->Update_Stopping_Rules();
      cvg=algo->HasConverged();
      iter++;
      algo->Display_Stopping_Rule(osalgo);
        }

 mrf->FixBeta(false);


 algo->Set_Stopping_Rule1(1,0.001);
     cvg=false;
     iter=0;
     while (!cvg){
      osModel << endl;
      osModel << iter<< " ";

cout<<"final"<<endl;

        algo->Run(1,osModel);
        hmrf->WriteToFile(std::string(iteration_params_out));

        algo->Update_Stopping_Rules();
```

```
        cvg=algo->HasConverged();
        iter++;
        algo->Display_Stopping_Rule(osalgo);
        }
```

// Retourne la classification MAP dans le vecteur 'maplabels'

```
 vector<uint> map_labels;
algo->Compute_MAP_Labels(map_labels);
```

// Écrit à l'écran la valeur du critère BICw et le sauvegarde dans le fichier des paramètres

```
double p=algo->BICw();
cout<<algo->BICw()<<endl;
```

// Sauve les paramètres du modèle obtenus avec un ensemble de valeurs initiales 'trialfolder' parmi les M valeurs dans le fichier 'ressim.par'

```
char file_par_out[300];
sprintf(file_par_out,"%s/ressim.par",trial_folder);
hmrf->WriteToFile(string(file_par_out));
```

Bibliographie

D. Abrial, D. Calavas, N. Jarrige, and C. Ducrot. Poultry, pig and the risk of BSE following the feed ban in France : A spatial analysis. *Veterinary Research*, 36 :615–628, 2005. 134

D. Agarwal, A. Gelfand, and S. Citron-Pousty. Zero-inflated models with application to spatial count data. *Environmental and Ecological Statistics*, 162(2) :195–209, 2002. 38

M. Alfo, L. Nieddu, and D. Vicari. Finite Mixture Models for Mapping Spatially Dependent Disease Counts. *Biometrical Journal*, 52 :84–97, 2009. 76, 83, 87, 106

A. Allepuz, A. Lopez-Quilez, A. Forte, G. Fernandez, and J. Casal. Spatial analysis of bovine spongiform encephalopathy in Galicia, Spain(2002-2005). *Preventive Veterinary Medicine.*, 79(2-4) :174–185, 2007. 134

V. Arora and P. Lahiri. On the superiority of the bayesian method over the blup in small area estimation problems. *Statistica Sinica*, 7 :1053–63, 1997. 37

G. A. Avruskin, G. M. Jacquez, J. R. Meliker, M. J. Slotnick, A. Kaufmann, and J. O. Nriagu. Visualization and exploratory analysis of epidemiologic data using a novel space time information system. *International Journal of Health Geographics*, 3(26), 2004. 1

J. Berger and J. Bernardo. On the development of the reference prior method. In *Bayesian statistics 4*, pages 35–60. London : Oxford university Press, 1992. 28

L. Bernadinelli, D. Clayton, C. Pascutto, C. Montomoli, M. Ghislandi, and M. Songini. Bayesian analysis of space-time variation in risk. *Statistics in Medicine*, 14 :2433–43, 1995. 41, 142

J. Besag. Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society*, 35 :192–236, 1974. 40, 42, 50, 54, 58

J. Besag. Statistical analysis of non-lattice data. *The Statistician*, 24 :179–95, 1975. 42

J. Besag. On the statistical analysis of dirty pictures. *Journal of the Royal Statistical Society Series B*, 48(3) :259–302, 1986. 58, 81

J. Besag and J. Tantrum. *Likelihood analysis of binary data in space and time*. New York : Oxford University Press, 2003. 43, 143

J. Besag, J. York, and A. Mollié. Bayesian image restoration, with two applications in spatial statistics. *Annals of the Institute of Statistical Mathematics*, 43(1) :1–59, 1991. 39, 40, 131

N. Best, S. Richardson, and A. Thomson. A comparison of bayesian spatial models for disease mapping. *Statistical Methods in Medical Research*, 14 :35–59, 2005. 39, 41, 45

N. G. Best, R. A. Arnold, A. Thomas, L. A. Waller, and E. M. Conlon. Bayesian model for spatially correlated disease and exposure data. In J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, editors, *Bayesian Statistics*, volume 6, pages 131–56. Oxford University Press, 1999. 40

C. Biernacki. Initializing EM Using the Properties of its Trajectories in Gaussian Mixtures. *Statistics and Computing*, 14(3) :267–279, 2004. 87, 88, 89

C. Biernacki, G. Celeux, and G. Govaert. Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22 :719–725, 2000. 73, 147

C. Biernacki, G. Celeux, and G. Govaert. Choosing starting values for the EM algorithm for getting the highest likelihood in multivariate Gaussian mixture models. *Computational Statistics and Data Analysis*, 41 :561–575, 2003. 84, 86, 87, 106

JF. Bithel. The choice of test for detecting raised disease risk near a point source. *Statistics in Medicine*, 14(21-22) :2309–22, 1995. 17

JF. Bithel and RA. Stone. On statistical methods for analysing the geographical distribution of cancer cases near nuclear installations. *Journal of Epidemiology Community Health*, 43 (1) :79–85, 1989. 17

JF. Bithel, SJ. Dutton, GJ. Draper, and NM. Neary. Distribution of childhood leukaemias and non-hodgkin's lymphomas near nuclear installations in england and wales. *BIOMIRROR Journal*, 309(6953) :501–5, 1994. 17

J. Blanchet. Modèles markoviens et extensions pour la classification de données complexes. In *PHD thesis, Université Joseph Fourier, Grenoble I*. 2007. 146

J. Blanchet, F. Forbes, S. Chopart, and L. Azizi. Le logiciel SpaCEM^3 pour la classification de données complexes. *La revue Modulad*, 40 :147–66, 2009. 147

D. Boehning, E. Dietz, P. Schlattmann, L. Mendonca, and U. Kirchner. The zero-inflated poisson model and the decayed, missing and filled teeth index in dental epidemiology. *Journal of the Royal Statistical Society Series A*, 162(2) :195–209, 1999. 38

D. Boehning, E. Dietz, and P. Schlattmann. Space-time mixture modelling of public health data. *Statistics in Medicine*, 19 :2333–44, 2000. 143

P. Burman. A comparative study of ordinary cross-validation, v-fold cross-validation and the repeated learning-testing methods. *Biometrika*, 76(3) :503–14, 1989. 72

P. Caragea and M. Kaiser. Covariates and time in the autologistic model. *Technical Report, Iowa State University, Department of Statistics*, (19), 2006. 43

B. Carlin, J. Clark, and A. Gelfand. Elements of hierarchical bayesian inference. In *Hierarchical modelling for the environmental sciences : statistical methods and applications*, pages 3–24. Oxford University Press, USA, 2006. 47

G. Celeux and J. Diebolt. The SEM algorithm : a probabilistic teacher algorithm derived from the mixture problem. *Computational Statistics Quaterly*, 2(1) :73–82, 1985. 63

G. Celeux and G. Govaert. A classification EM algorithm for clustering and two stochastic versions. *Computational Statistics and Data Analysis*, 14(3) :315–32, 1992. 63

G. Celeux, F. Forbes, and N. Peyrard. EM procedures using mean field-like approximations for Markov model-based image segmentation. *Pattern Recognition*, 36 :131–144, 2003. 67, 70, 71, 83, 139

B. Chalmond. An iterative technique for reconstruction of m-ary images. *Pattern Recognition*, 22(6) :747–761, 1989. 67

D. Chandler. *Introduction to Modern Statistical Mechanics*. Oxford University Press, 1987. 67

D. Clayton and L Bernadinelli. Bayesian methods for mapping disease risk. *Geographical and Environment Epidemiology : Methods for Small Area Studies, eds. P.Elliot, J.Cuzik, D.English, and R.Stern, Oxford, UK :Oxford University Press*, pages 205–220, 1992. 41

D. G. Clayton and J. Kaldor. Empirical bayes estimates of age-standardized relative risks for use in disease mapping. *Biometrics*, 43 :671–681, 1987. 35, 36, 37

N. A. C. Cressie. *Statistics for Spatial Data (revised ed.)*. New York : Wiley, 1993. 40, 41

L. Cucala. A flexible spatial scan test for case event data. *Computational Statistics and Data Analysis*, 53(8) :2843–50, 2009. 16

A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society Series B*, 39(1) :1–38, 1977. 61, 62

L. R. Dice. Measures of the amount of ecologic association between species. *Ecology*, 26 : 297–302, 1945. 108

P. G. Diggle, J. Tawn, and R. Moyeed. Model-based geostatistics. *Journal of the Royal Statistical Society Series C*, 47 :299–350, 1998. 41

E. Durand. Modèles statistiques pour la structure génétique des populations : organisation spatiale et liens de parenté. In *PHD thesis, Institut Polytechnique de Grenoble*. 2009. 46

J. B. Durand. Modèles à structure cachée : inférence, sélection de modèles et applications. In *PHD thesis, Université Joseph Fourier, Grenoble I*. 2003. 72

P. Elliott, J. C. Wakefield, N. G. Best, and D. J. Briggs. Spatial epidemiology : methods and applications. Oxford University Press, 2000. 15

P. R. Epstein. Climate and health. *Science*, 285(5426), 1999. 1

P. Erdos and A. Rényi. On random graphs. *Publicationes Mathematicae*, 6 :290–97, 1959. 52

C. Fernandez and P. J. Green. Modelling spatially correlated data via mixtures : a Bayesian approach. *Journal of the Royal Statistical Society Series B*, 64(4) :805–826, 2002. 37, 45

J. Ferreira, D. Denison, and C. Holmes. *Partition modelling*. New York : CRC Press, 2002. 45

S. Finch, N. Mendell, and H. Thode. Probabilistic measures of adequacy of a numerical search for a global maximum. *Journal of the Royal Statistical Society Series B*, 39 :1–38, 1989. 86

F. Forbes and N. Peyrard. Hidden Markov random field selection criteria based on mean field-like approximations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25 : 1089–1101, 2003. 74

C. Fraley and A. E. Raftery. How many clusters ? which clustering method ? answers via model-based cluster analysis. *The Computer Journal*, 41 :578–88, 1998. 73

J. French and M. Wand. Generalized additive models for cancer mapping with incomplete covariates. *Biostatistics*, 5 :177–91, 2004. 44

A. Froment. Une approche écoanthropologique de la santé publique. *Nature Sciences Sociétés*, 5 :5–11, 1997. 1

A. Gelman, J. B. Carlin, and H.S. Stern. Bayesian data analysis, second edition. Chapman and Hall/CRC, 2003. 34

S. Geman and D. Geman. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transaction on Pattern analysis and Machine Intelligence*, 6(6) : 721–741, 1984. 50, 81

S. Geman and C. Graffigne. Markov random fiels image models and their applications to computer vision. In *International Congress of Mathematicians*, pages 1496–1517, 1987. 59

M. Ghosh, K. Natarajan, T.W.F. Stroud, and B. P. Carlin. Generalized linear models for small area estimation. *J. Amer. Statist. Assoc.*, 93 :273–82, 1998. 37

M. Ghosh, K. Natarajan, L. A. Waller, and D. Kim. Hierarchical GLMs for the analysis of spatial data : An application to disease mapping. *Journal of Statistical Planning Inference.*, 75 :305–318, 1999. 40

S. Ghosh, P. Mukhopadhay, and J. C. Lu. Bayesian analysis of zero-inflated regression models. *Journal of Statistical planning and Inference*, 136 :1360–75, 2006. 38

W. R. Gilks, S. Richardson, D. J. Spiegelhalter, N. G. Best, L. D. McNeil, L. D. Sharples, and A. J. Kirby. Modelling complexity : Applications of gibbs sampling in medicine. *Journal of the Royal Statistical Society Series B*, 55 :39–52, 1993. 29

W. R. Gilks, S. Richardson, and D. J. Spiegelhalter. *Markov chain Monte Carlo in Practice*. London : Chapman and Hall, 1996. 29

P. J. Green. Reversible jump mcmc computation and bayesian model determination. *Biometrika*, 82 :711–32, 1995. 32, 37

P. J. Green and S. Richardson. Hidden markov models and disease mapping. *Journal of the American Statistical Association*, 97 :1055–70, 2002. 37, 45, 46, 76

M. L. Gumpertz, J. M. Graham, and J. B. Ristaino. Autologistic model of spatial pattern of phytophtora epidemic in bell pepper : Effects of soil variables on disease presence. *Journal of Agricultural, Biological and Environmental Statistics*, 2 :131–56, 1997. 42, 43

W. K. Hastings. Monte carlo sampling methods using markov chains and their applications. *Biometrika*, 57(1) :97–109, 1970. 30

S. H. Heisterkamp, G. Doornobs, and N. J. D. Nagelkerke. Assessing the impact of environmental pollution sources using space-time models. *Statistics in Medicine*, 19 :2569–78, 2000. 41

R. Henderson, S. Shimakura, and D. Gorst. Modelling spatial variation in leukemia survival data. *Journal of the American Statistical Association*, 97 :965–72, 2002. 41

J. A. Hoeting, M. Leecaster, and D. Bowden. An improved model for spatially correlated binary responses. *Journal of Agricultural, Biological and Environmental Statistics*, 5 : 102–14, 2000. 42

L. Huang, LW. Pickle, and B. Das. Evaluating spatial methods for investigationg global clustering and cluster detection of cancer cases. 27(25) :5111–42, 2008. 17

E. Ising. Beitrag zur theorie des ferromagnetismus. *Zeitschrift fur Physik*, 31 :253–58, 1925. 55

N. Jean-Pierre and G. Mina. Physique statistique des phénomènes collectifs en sciences économiques et sociales. *Mathématiques et sciences humaines*, 172 :67–89, 2005. 56

D. Karlis and E. Xekalaki. Choosing initial values for the EM algorithm for finite mixtures. *Computational Statistics and Data Analysis*, 41 :577–590, 2003. 86

G. Kauermann and J.D. Opsomer. Local likelihood estimation in generalized additive models. *Scandinavian Journal of Statistics*, 30 :317–37, 2003. 25

J. Kelsall and J. Wakefield. Modelling spatial variation in disease risk : A geostatistical approach. *Journal oh the American Statistical Association*, 97 :692–701, 2002. 41

L. Knorr-Held and G. Rasser. Bayesian detection of clusters and discontinuities in disease maps. *Biometrics*, 56 :13–21, 2000. 41, 45

S. Kullback and R. A. Leibler. On information and sufficiency. *The annals of Mathematical Statistics*, 22(1) :79–86, 1951. 46

M. Kulldorff. A spatial scan statistic. *Communication in Statistics-Theory and Methods*, 26 (6) :1481–96, 1996. 16

M. Kulldorff. Tests of spatial randomness adjusted for an inhomogeneity : a general framework. *Journal of the American Statistical Association*, 101(475) :1289–305, 2006. 17

M. Kulldorff and N. Nagarwalla. Spatial disease clusters : detection and inference. *Statistics in Medicine*, 14(8) :799–810, 1995. 16

H. Kunsch. Intrinsic autoregressions and related models on the two-dimensional lattice. *Biometrika*, 74 :517–24, 1987. 40

D. Lambert. Zero-inflated poisson regression, with application to defects in manufacturing. *Technometrics*, 34 :1–14, 1992. 38

P. M. Lankford. Regionalization : theory and alternative algorithms. *Geographical Analysis*, 1 :196–212, 1969. 52

S. Lauritzen, A. Dawid, B. Larsen, and H. G. Leimer. Independence properties of directed markov fields. *Networks*, 20 :491–505, 1990. 53

A. B. Lawson and F. L. R. Williams. *An Introductory Guide to Disease Mapping*. New York : Wiley, 2001. 36

S. Z. Li. *Markov random field modeling in image analysis*. Springer, 2001. 67

B.G. Lindsay. Composite likelihood methods. *Contemporary Mathematics*, 80 :221–238, 1988. 25

Y. Macnab. Spline smoothing in bayesian disease mapping. *Environmetrics*, 18 :727–44, 2007. 44

J. B. MacQueen. Some methods for classification and analysis of multivariate observations. In *5th Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley, University of California Press, 29 October-2 November 1967. 101

J. M. Marin and C. Robert. *Bayesian Core : A Practical Approach to computational Bayesian Statistics*. New York : Springer, 2007. 38

P. McCullagh and J. A. Nelder. London : Chapman and Hall, 1989. 45

G.J. McLachlan and D. Peel. *Finite Mixture Models*. New York : John Wiley and Sons, 2000. 60, 86

A. J. McMichael. Human culture, ecological change, and infectious disease : are we experiencing history's fourth great transition ? *Ecosystem Health*, 7(2) :107–15, 2001. 1

N. Metropolis, A. Rosenbluth, M. Rosenbluth, A. Teller, and E. Teller. Equation of state calculations by fast computing machines. *Jouranl of Chemical Physics*, 21(6) :1087–92, 1953. 30

A. Mollie. Bayesian and empirical bayes approaches to disease mapping. In A. Lawson, A. Biggeri, and D. Boehning, editors, *Disease mapping and risk assessment for public health*, pages 15–29. Wiley, 1999. 41

S. E. Morris and J. C. Wakefield. Assessment of disease risk in relation to a pre-specified source. In P. Elliott, J. C. Wakefield, N. G. Best, and D. J. Briggs, editors, *Spatial epidemiology : methods and applications*, pages 153–84. Oxford University Press, 2000. 16

D. J. Nott and T. Rydén. Pairwise likelihood methods for inference in image models. *Biometrika*, 86 :661–76, 1999. 25

M. Paul, D. Abrial, N. Jarrige, S. Rican, M. Garrido, D. Calavas, and C. Ducrot. Bovine spongiform encephalopathy and spatial analysis of the feed industry. *Emerging Infectious Diseases.*, 13(6) :867–872, 2007. 134

N. Peyrard. Approximations de type champ moyen des modèles de champ de Markov pour la segmentation de données spatiales. In *PHD thesis, Université Joseph Fourier, Grenoble I.* 2001. 70

W. Pieczynski. Champs de markov cachés et estimation conditionnelle itérative. *Traitement du Signal*, 11 :141–153, 1994. 67

W. Qian and D. M. Titterington. Estimation of parameters in hidden Markov models. *Philosophical Transactions of the Royal Society of London*, Series A 337 :407–428, 1991. 67

C. Robert. *L'analyse statistique bayésienne*. Economica, 1992. 24

C. Robert and G. Casella. *Monte Carlo statistical methods*. Springer, 1999. 58

B. Scherrer, M. Dojat, F. Forbes, and C. Garbay. Locus : Local cooperative unified segmentation of mri brain scans. In *10th International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI07)*, Brisbane, Australia, 29 October-2 November 2007. 56

P. Schlattman and D. Boehning. Mixture models and disease mapping. *Statistics in Medicine*, 12 :1943–50, 1993. 37

G. Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2) :131–34, 1978. 73

D. J. Spiegelhalter, N. G. Best, B. P. Carlin, and A. van der Linde. Bayesian deviance, the effective number of parameters and the comparison of arbitrarily complex models. *Journal of the Royal Statistical Society Series B*, 64 :583–640, 2002. 47

D.J. Spiegelhalter, N.G. Best, W.R. Gilks, and H. Inskip. Hepatitis b : a case study in MCMC methods. In W.R. Gilks, S. Richardson, and D.J. Spielgelhalter, editors, *Markov chain Monte Carlo in Practice*. London : Chapman and Hall, 1996. 27

M. Stevenson, R. Morris, A. B. Lawson, J. Wilesmith, J. M. Ryan, and R. Jackson. Area level risks for bse in british cattle before and after the july 1988 meat and bone meal feed ban. *Preventive Veterinary Medicine*, 69, 2005. 34

D. Strauss. Clustering on coloured lattices. *Journal of Applied Probability*, 14 :135–43, 1977. 57

D. Sun, R. K. Tsutakawa, and P. L. Speckman. Posterior distribution of hierarchical models using CAR(1) distributions. *Biometrika*, 86 :341–50, 1999. 40

D. Sun, R. K. Tsutakawa, H. Kim, and Z. He. Spatio-temporal interaction with disease mapping. *Statistics in Medicine*, 19 :2015–35, 2000. 41

T. Tango. A class of tests for detecting 'general' and 'focused' clustering for rare diseases. *Stat Med*, 14(21-22) :2323–34, 1995. 17

T. Tango and K. Takahashi. A flexibly shaped spatial scan statistic for detecting clusters. *International Journal of Health Geographics*, 4 :11, 2005. 16

R. Thibshirani and T. Hastie. Local likelihood estimation. *Journal of the American Statistical Association*, 82 :559–568, 1987. 25

G. Toussaint. The relative neighborhood graph of a finite planar set. *Pattern Recognition*, 12 : 261–68, 1980. 52

C. Varin, G. Host, and O. Skare. Pairwise likelihood inference in spatial generalized linear mixed models. *Computational Statistics and Data Analysis*, 49 :1173–91, 2005. 25

H. Wackernagel. *Multivariate Geostatistics (3rd ed.)*. New York : Springer, 2003. 41

J. Wakefield. Disease mapping and spatial regression with count data. *Biostatistics*, 8(2) : 158–83, 2007. 18, 40

J. C. Wakefield, J. E. Kelsall, and SE. Morris. Clustering, cluster detection, and spatial variation in risk. In P. Elliott, J. C. Wakefield, N. G. Best, and D. J. Briggs, editors, *Spatial epidemiology : methods and applications*, pages 128–52. Oxford University Press, 2000. 17

L. Waller and C. Gotway. Applied spatial statistics for public health data. New York : Wiley, 2004. 34, 41

L. A. Waller, B. P. Carlin, H. Xia, and A. Gelfand. Hierarchical spatio-temporel mapping of disease rates. *Journal of the American Statistical Association*, 92 :607–17, 1997. 41, 143

R. Washino and B. Wood. Application of remote sensing to arthropod vector surveillance and control. *American Journal of Tropical Medicine and Hygiene*, 50(6) :134–44, 1994. 1

G. Wei and M. A. Tanner. A monte carlo implementation of the em algorithm and the poor man's data augmentation algorithms. *Journal of the American Statistical Association*, 85 (411) :699–704, 1990. 67

C. K. Wikle. Spatial modelling of count data : A case study in modelling breeding bird survey data on large spatial domains. In A. B. Lawson and D. G. T. Denison, editors, *Spatial Cluster Modelling*. London : CRC Press, 2002. 41

R. L. Wolpert and K. Ickstadt. Poisson-Gamma random field models for spatial statistics. *Biometrika*, 85 :251–67, 1998. 39

C. Wu. On the convergence properties of the em algorithm. *Annals of Statistics*, 11 :95–103, 1983. 61

C. H. Wu and P. C. Doerschuk. Cluster expansions for the deterministic computation of bayesian estimators based on markov random fields. *IEEE transactions On Pattern Analysis and Machine Intelligence*, 17(3) :275–93, 1995. 68

H. Wu and F. W. Huffer. Modeling the distribution of plant species using the autologistic regression model. *Environmental and Ecological Statistics*, 4 :49–64, 1997. 42

H. Xia and B. P. Carlin. Spatio-temporal models with errors in covariates : mapping ohio lung cancer mortality. *Statistics in Medicine*, 17 :2025–43, 1998. 41, 143

L. Younes. Estimation and annealing for gibbsian fields. *Annales de l'Institut Henri Poincaré (B), Probabilités et Statistique*, 24 :269–94, 1988. 67

J. Zhang. The mean field theory in EM procedure for Markov random fields. *IEEE Transactions on Signal Processing*, 40(10) :2570–2583, 1992a. 67

J. Zhang. The Mean field theory in EM procedures for markov random fields. *IEEE Transactions on signal processing*, 40 :2570–2583, 1992b. 67

P. Zhang. Model selection via multifold cross validation. *Annals of Statistics*, 21(1) :299–313, 1993. 72

S. Zhang, D. Sun, C. He, and M. Schootman. A bayesian semi-parametric model for colorectal cancer incidences. *Statistics in Medicine*, 25 :285–309, 2006. 44