Modélisation probabiliste et inférence bayésienne pour l'analyse de la dynamique des mélanges de fluides géologiques : détection des structures et estimation des paramètres

Christophe Reype

UNIVERSITÉ DE LORRAINE

École doctorale IAEM Lorraine

Modélisation probabiliste et inférence bayésienne pour l'analyse de la dynamique des mélanges de fluides géologiques : détection des structures et estimation des paramètres.

THÈSE

présentée et soutenue publiquement le 14 décembre 2022

pour l'obtention du

Doctorat de l'Université de Lorraine

(mention mathématiques appliquées)

par

Christophe Reype

Composition du jury

Présidente :	Anne Gegout-Petit	Université de Lorraine
Rapporteurs :	Madalina Olteanu	Paris Dauphine-PSL
	Aila Särkkä	Chalmers University of Technology and University of Gothenburg
Examinateurs :	Nicolas Desassis	Mines Paris-PSL
	Jorge Mateu	Universitat Jaume I
	Antonin Richard	Université de Lorraine
Invités :	Daniele Luigi Pinti	Université du Québec à Montréal
	Jacques Pironon	CNRS
Direction :	Madalina Deaconu	Inria
	Radu Stoica	Université de Lorraine

Institut Élie Cartan de Lorraine : UMR 7502

Remerciements

Cette thèse représente la quintessence de plus de deux décennies. Ces prochaines lignes sont dédiées à toutes les personnes qui m'ont aidé à accomplir ce travail.

Les premiers à remercier sont ceux qui m'ont guidé tout au long de ces trois incroyables dernières années : Radu, Madalina et Antonin. Vous avez été (et êtes toujours) des exemples pour moi, tant sur le plan scientifique que sur le plan humain. Votre patience, votre disponibilité, vos conseils et votre soutien m'ont grandement aidé à mûrir, mais aussi tenir pendant mes périodes de doutes. Merci pour les longues séances de travail sous les étoiles, merci pour tous les conseils malgré les emplois du temps surchargés et merci pour les contextualisations des travaux de cette thèse.

Je remercie chaleureusement Madalina Olteanu et Aila Särkkä pour avoir accepté de rapporter ma thèse, ainsi que tous les autres membres du jury, Nicolas Desassis, Anne Gegout-Petit, Jorge Mateu, Daniele Luigi Pinti et Jacques Pironon.

Cette thèse a été possible grâce à l'Université de Lorraine, le projet LUE DEEPSURF ainsi que les laboratoires Institut Elie Cartan de Lorraine, Inria et GeoRessources. Je remercie également toutes les personnes qui m'ont permis de travailler dans les meilleures conditions : Nathalie Benito, Elodie Cunat, Laurence Quirot, Paola Schneider et Isabelle Blanchard. Je remercie Didier Gemmerlé qui m'a permis de garder l'optimisme sans perdre le nord pendant l'écriture de mon code C++. Un énorme merci à Pauline Collon et Xavier Antoine pour les conseils précieux pendant les comités de suivi et tout le long de cette aventure.

Un grand merci aussi à ceux qui m'ont entouré ces dernières années. Ceux qui ne sont plus au laboratoire aujourd'hui comme Clémence, Pierre-Adrien et Johan, mais aussi ceux qui sont toujours là : Rémi et son énergie, Sara et sa bienveillance et Coralie et son expérience. Mention spéciale à tous les doctorants qui ont donné vie à nos pauses : Anouk, Benjamin, Jocelyn, Nathan, Raphaël, Rodolphe, Valentin, Victor, Vincent, Yann,... Un grand merci à mon co-bureau Alexis pour ces trois ans. Évidemment, je suis obligé de mentionner Thomas malgré sa passion contre-nature pour une série de livres et des films associées. Toi aussi, mon contraire, je te remercie d'avoir partagé ce rush final, le stress de la soutenance et nos inavouables idées sur le monde. Merci Pierre.

Je tiens à remercier toutes les personnes qui m'ont enseigné et donné envie de continuer :

Eric Galard, Thomas Respaud, Armelle Guillou et Laurent Gardes. Merci également à Lionel Lenôtre pour m'avoir fait confiance et m'avoir encouragé dans cette voie. Un merci à ceux qui m'ont donné de merveilleux souvenirs de prépa : mes colocs Aristée, Arthur et Léo ainsi que mes comparses Aurélien, Robin et Yann. Le meilleur groupe de TIPE Fabien et Luc et celui qui m'a fait découvrir tout un monde Mike.

Je remercie aussi Nadim pour toutes les games de trop, Théo et Charlotte pour les séances de jeux, Elisa malgré son rôle de commanditaire dans une sombre histoire de remorque qui me hante encore, Mallaury pour sa joie de vivre et son énergie et Martin pour ces dernières années pleines de discussions et de débats passionnant. Je ne peux pas oublier ma triforce personnelle : la sagesse avec Julie qui a réveillé ma passion des mathématiques, le courage avec Victor qui m'a inspiré dans son aisance à essayer de nouvelles choses et à rencontrer les gens et la force avec Fabien qui m'a montré qu'avec du temps et des efforts, on peut tout apprendre. Cette thèse est aussi le résultat de tous les efforts et tous les sacrifices qu'a faits ma mère pour me permettre de poursuivre ma voie et pour cela, je ne la remercierai jamais assez. Merci à toi aussi Kévin pour ton soutien et pour m'avoir montré que nous ne sommes pas condamnés à répéter les erreurs. Merci à toi Stéphane, mon pilier, mon modèle, mon exemple et mon meilleur ami depuis le plus lointain de mes souvenirs. Merci à vous Fabien et Lisa, j'espère que je suis le grand frère que vous rêvez d'avoir. Enfin merci à toi Lucie, pour ces années. Je ne peux pas faire une liste exhaustive de tout ce que je veux te dire, mais on a encore tout le temps pour en parler.

A vous tous et à ceux que j'ai malheureusement oubliés, sachez que sans vous, je ne serai pas là aujourd'hui donc mille fois merci.

Sommaire

Table des figures

Liste des tableaux

Chapitre 1

Introduction. État de l'art

1.1 Présentation du problème

L'activité dans la croûte terrestre est intimement liée au déplacement de fluides en surface et en profondeur [Pirajno, 2009, Yardley and Bodnar, 2014]. Les interactions fluides-fluides et fluides-roches provoquent des modifications importantes des caractéristiques physico-chimiques de la croûte. L'étude de ces processus est donc primordiale pour renforcer la compréhension du sol et du sous-sol.

Les interactions fluides-fluides et fluides-roches sont difficiles à observer. Pour palier à ce manque, il est nécessaire de développer des modèles conceptuels et quantitatifs à partir des données issues de prélèvements sur le terrain.

Dans cette thèse, les processus étudiés sont les mélanges de fluides. Les mélanges de fluides de sources différentes sont des processus couramment observés, par exemples dans les systèmes composés d'une rivière et de ses affluents ou les systèmes hydrothermaux impliquant des mélanges des fluides profonds le long de failles. Le nombre exact et la composition des sources de fluides dans un système impliquant des mélanges entre différentes sources ne sont pas toujours connus a priori. Les sources de fluides en surface et subsurface sont d'origine multiple, mais on peut les regrouper en cinq catégories principales : les eaux météoriques (dérivées de l'eau de pluie), les eaux marines, les eaux de bassin (circulant dans la porosité des sédiments), les eaux magmatiques (libérées lors de la cristallisation de magma) ou encore les eaux métamorphiques (libérées lors du métamorphisme des roches) (Figure 1.1).

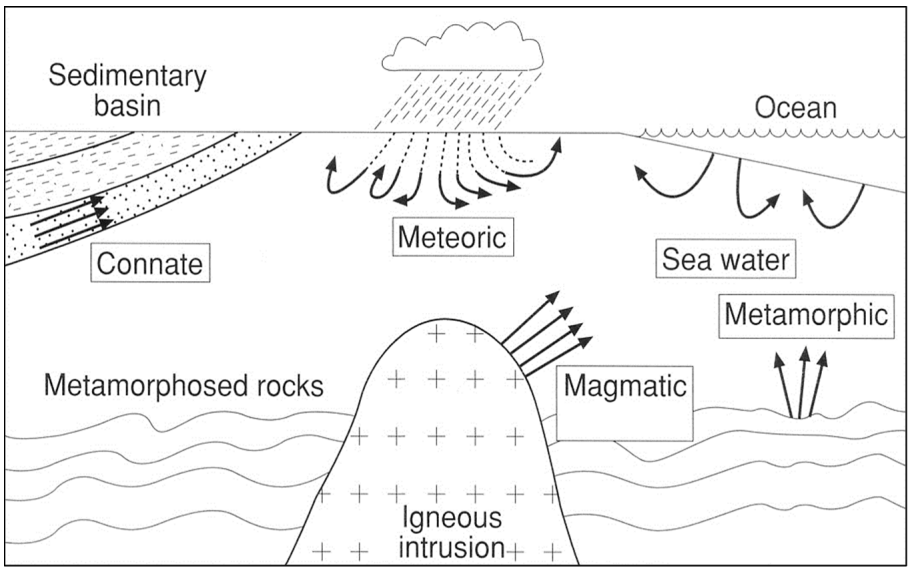


FIGURE 1.1 – Schéma conceptuel de la croûte continentale supérieure de la Terre. Les principales sources d'eaux de la surface et du sous-sol, contribuant à la formation des eaux souterraines par des processus de mélanges, sont représentées [Robb, 2005].

Dans cette thèse, nous considérons les données dites hydrogéochimiques. Ces données sont des mesures de composition de $K \in \mathbb{N}$ paramètres hydrogéochimiques (concentration en ions ou en molécules, composition isotopique, ...) réalisées sur $m \in \mathbb{N}$ prélèvements. Ces données sont généralement associées à des incertitudes de mesure. Ces prélèvements sont obtenus soit par mesure directe, soit par échantillonnage et analyse d'eaux de surface ou de subsurface dans des forages, soit par échantillonnage et analyse d'inclusions fluides (reliques de fluides géologiques anciens, piégés dans les minéraux). Les modèles ont pour objectif de déterminer les interactions fluides-fluides et fluides-roches qui ont créé les prélèvements. Plus particulièrement, le modèle de cette thèse est un modèle de détection de sources dans des mélanges de fluides : les interactions étudiées sont donc entre les fluides. La création de ce modèle est classique : un bilan des connaissances et des méthodes d'analyse des mélanges de fluides, une recherche des outils théoriques pour la modélisation du problème et enfin une présentation et une analyse du modèle et de ses résultats.

Étant construit autour de l'analyse de données hydrogéochimiques, le modèle développé dans cette thèse, trouve par exemple des applications dans les domaines de suivi des contaminations de rivières et de nappes souterraines, de la géothermie, du stockage souterrain des déchets et de la métallogénie. Plus largement, ce

modèle est potentiellement utile pour l'analyse de jeux de données géochimiques multidimensionnels impliquant des interactions entre plusieurs pôles de compositions (magmas, solutions, solides,...). Au-delà du problème particulier de la caractérisation des mélanges de fluides géologiques, l'estimation des compositions des sources peut être utilisée pour retracer les origines de certains matériaux ou des sources de pollution [Longman et al., 2018], la formation de nappes phréatiques, de filons métallifères, la vitesse de fonte des glaciers [Arendt et al., 2015] ou l'optimisation du colmatage d'anciens puits de gaz ou de pétrole [Skuce et al., 2015]. Dans d'autres domaines que les géosciences, la détection des sources peut être appliquée, par exemple à la caractérisation du régime alimentaire d'une espèce peut aussi se déduire de cette étude [Phillips and Gregg, 2001,Parnell et al., 2010].

Les jeux de données hydrogéochimiques sont visualisés par des points dans un espace dont les dimensions représentent les compositions en les différents paramètres hydrogéochimiques. Cet espace est appelé "espace des données". Dans cet espace, la position d'un point indique la composition chimique d'un prélèvement et la distance entre des points la différence de composition chimique entre les prélèvements.

Si aucune réaction chimique (ex : précipitation) n'a lieu au cours du mélange, alors les points représentant les prélèvements sont considérés comme des barycentres des sources [Faure, 1997]. Ainsi en notant $d_j = (d_{j;1}, \dots, d_{j;K}) \in \mathbb{R}^K$ le prélèvement numéro $j \in [\![1; m]\!]$ et $\mathbf{s} = (s_1, \dots, s_n)^\top \in \mathbb{R}^{n \times K}$ l'ensemble des $n \in \mathbb{N}$ sources où $s_i = (s_{i;1}, \dots, s_{i;K}) \in \mathbb{R}^K$ est la position de la source numéro $i \in [\![1; n]\!]$ alors :

$$d_{j;k} = \sum_{i=1}^{n} \gamma_{j;i} s_{i;k}, \tag{1.1}$$

avec $0 \leq \gamma_{j;i} \leq 1$ la contribution de la source i pour le prélèvement j et $\sum_{i=1}^{n} \gamma_{j;i} = 1$. Chaque prélèvement est donc à l'intérieur de l'enveloppe convexe formée par les sources. Cette formule est étendue pour le vecteur de données $\mathbf{d} = (d_1, \dots, d_m)^\top$ et la matrice de contribution $\gamma = (\gamma_1, \dots, \gamma_m)^\top \in \mathbb{R}^{m \times n}$ avec $\gamma_j = (\gamma_{j;1}, \dots, \gamma_{j;n})$ le vecteur de contribution pour la donnée j :

$$\mathbf{d} = \gamma \mathbf{s}. \tag{1.2}$$

Ce type de mélange est dit linéaire. L'hypothèse sur le caractère ponctuel des sources (le fait qu'une source soit représentée par un point) et des données est une approximation. En effet, la composition d'un fluide n'est pas uniforme dans tout

le fluide : la composition du fluide devrait être représentée par un point moyen équipé d'une zone d'incertitude ou de variabilité. En pratique, l'incertitude est supposée suivre une loi de probabilité, souvent gaussienne, qui dépend aussi de la méthode de mesure. Dans la suite, nous considérons que les sources et les données sont ponctuelles et représentées par la composition moyenne du fluide.

Si une ou plusieurs précipitations chimiques ont lieu au cours du mélange, les valeurs des paramètres hydrogéochimiques sont modifiées et les données ne peuvent plus être visualisées comme des barycentres des sources. Pour cette raison, nous poserons l'hypothèse que les mélanges considérés dans notre modèle sont conservatifs, c'est-à-dire sans réaction chimique.

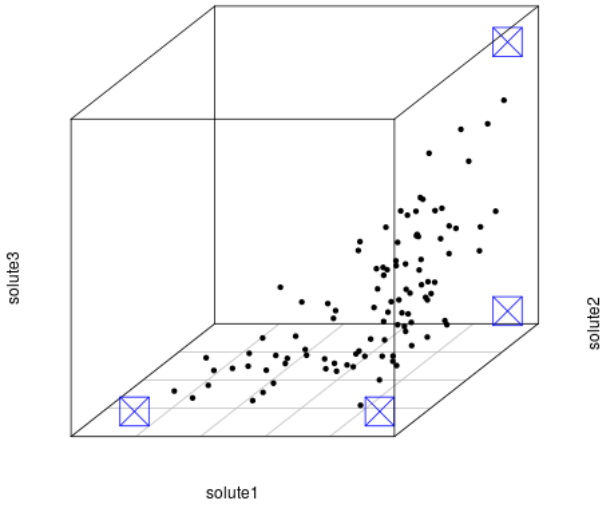
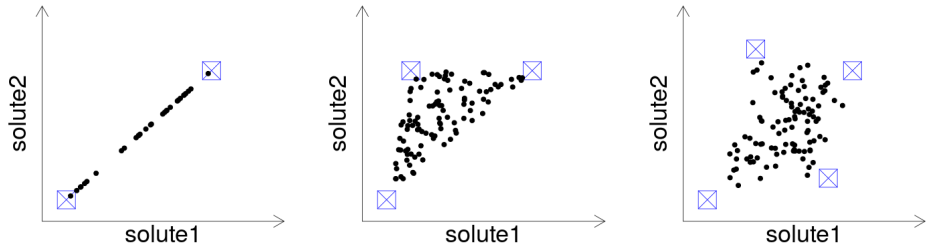


FIGURE 1.2 – Projection d'un jeu de données synthétique, issu d'un système de mélange à quatre sources dans un espace à trois dimensions. Les axes représentent la concentration en un "soluté1", "soluté2" et "soluté3". Les symboles bleus représentent les sources, les points noirs les données.

Le plus souvent, l'analyse hydrogéochimique considère les projections des données dans un plan en considérant uniquement deux paramètres hydrogéochimiques. Lorsque deux sources ponctuelles sont considérées, les données sont distribuées sur le segment reliant ces sources. Lorsque plus de sources ponctuelles sont considérées, les données sont distribuées dans le polygone convexe qui contient toutes les sources et dont les sommets sont des sources. La Figure 1.3 représente des systèmes de mélange à deux (a), trois (b) et quatre (c) sources

dans le plan formé par les paramètres hydrogéochimiques “soluté1” et “soluté2”.



(a) (b) (c)

FIGURE 1.3 – Projection d’un jeu de données synthétique, issu d’un système de mélange à deux (a), trois (b) et quatre (c) sources dans un espace à deux dimensions. En abscisse est représentée la concentration en un “soluté1” et en ordonnée la concentration en un “soluté2”. Les symboles bleus représentent les sources, les points noirs les données.

L’analyse de ces données a deux objectifs : détecter les sources, c’est-à-dire estimer leur nombre et leur composition, et, lorsqu’elles sont connues, estimer leur contribution dans chaque donnée.

Ce manuscrit présente une méthode non supervisée de détection de sources pour des systèmes de mélange conservatifs dans des espaces de données multidimensionnels. Cette méthode considère l’ensemble des mesures de tous les prélèvements et donc l’espace des données complet et pas uniquement des projections sur des plans définis par deux paramètres. La détection est basée sur les connaissances physiques, mais aussi géologiques utilisées pour la détection graphique. La modèle s’inspire des méthodes de détection de structure dans des données spatialisées, utilisé dans des domaines comme l’astronomie, l’épidémiologie ou encore l’analyse d’image. En théorie, ce modèle pourra être appliqué à tous les systèmes de mélanges qui considèrent les données comme des barycentres de sources à détecter, donc des données qui peuvent être visualisées comme dans la Figure 1.3.

1.2 Détection de sources à partir de données hydrogéochimique

En utilisant les notations de l'équation (1.2), le jeu de données s'écrit :

$$\mathbf{d} = \begin{pmatrix} d_{1;1} & \dots & d_{1;K} \\ \vdots & \ddots & \vdots \\ d_{j;1} & \dots & d_{j;K} \\ \vdots & \ddots & \vdots \\ d_{m;1} & \dots & d_{m;K} \end{pmatrix} \tag{1.3}$$

avec en ligne les coordonnées ou composantes en les paramètres hydrogéochimiques, de chaque donnée et en colonne les valeurs pour la coordonnée ou composante en un paramètre hydrogéochimique.

Le mélange linéaire de n sources ponctuelles à l'origine du jeu de données s'écrit :

$$\begin{pmatrix} d_{1;1} & \dots & d_{1;K} \\ \vdots & \ddots & \vdots \\ d_{j;1} & \dots & d_{j;K} \\ \vdots & \ddots & \vdots \\ d_{m;1} & \dots & d_{m;K} \end{pmatrix} = \begin{pmatrix} \gamma_{1;1} & \dots & \gamma_{1;n} \\ \vdots & \ddots & \vdots \\ \gamma_{j;1} & \dots & \gamma_{j;n} \\ \vdots & \ddots & \vdots \\ \gamma_{m;1} & \dots & \gamma_{m;n} \end{pmatrix} \begin{pmatrix} s_{1;1} & \dots & s_{1;K} \\ \vdots & \ddots & \vdots \\ s_{i;1} & \dots & s_{i;K} \\ \vdots & \ddots & \vdots \\ s_{n;1} & \dots & s_{n;K} \end{pmatrix}.$$

La détection de sources, ou analyse inverse [Weltje, 1997], peut se faire de deux principales manières.

La première est la détection graphique qui repose presque exclusivement sur l'œil de l'expert. Les données sont projetées dans les plans formés par tous les couples de dimensions parmi les K possibles : il y a donc $L = K(K-1)/2$ plans d'études différents possibles. Pour chaque plan numéro $l \in [\![1; L]\!]$, il existe deux dimensions k_{l_1} et k_{l_2} de $[\![1; K]\!]$ (avec $k_{l_1} < k_{l_2}$) tel que le jeu de données projeté

dans le plan se lise dans la matrice (1.3) en lisant les colonnes k_{l_1} et k_{l_2} :

$$\mathbf{d}_{\{l\}} = \begin{pmatrix} d_{1;k_{l_1}} & d_{1;k_{l_2}} \\ \vdots & \vdots \\ d_{j;k_{l_1}} & d_{j;k_{l_2}} \\ \vdots & \vdots \\ d_{m;k_{l_1}} & d_{m;k_{l_2}} \end{pmatrix}. \tag{1.4}$$

Sur chaque plan, l'enveloppe convexe des données est supposée esquisser un polygone dont les sommets représentent la position des sources. Le nombre de sources, et donc de sommets du polygone, est défini par l'utilisateur à partir de la forme du nuage de points, formé par les données, ainsi que des connaissances sur les données. Cette méthode permet de détecter très rapidement les sources. Les sources détectées sur chaque plan ont l'avantage de corroborer avec les connaissances sur les mélanges (mélange linéaire, conservatifs,...) mais aussi sur les données (nombre de sources, types d'interactions fluides-fluides possible,...). En effet, les sources détectées expliquent les données, c'est-à-dire que les données sont à l'intérieur du polygone convexe formé par les sources. De plus, ces sources intègrent dans leur construction les connaissances du milieu tel que les sources déjà détectées ou la plage de valeurs que peuvent prendre les sources.

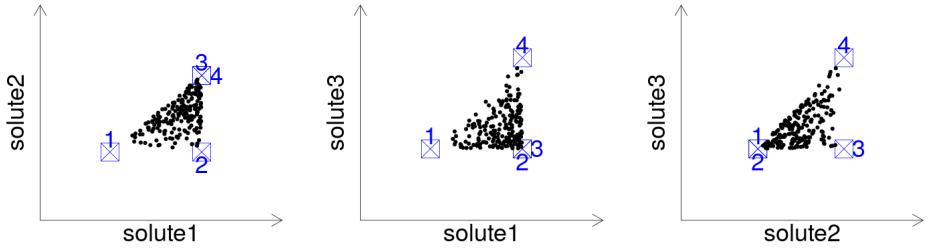
Cependant, cette méthode présente trois désavantages majeurs :

— l'estimation des sources donc de leur nombre et de leur position dépend de l'utilisateur : deux utilisateurs peuvent détecter un nombre différent de sources ou des sources de composition différentes,

— cette méthode n'est pas automatique, l'utilisateur doit estimer personnellement les sources sur les $K(K-1)/2$ plans,

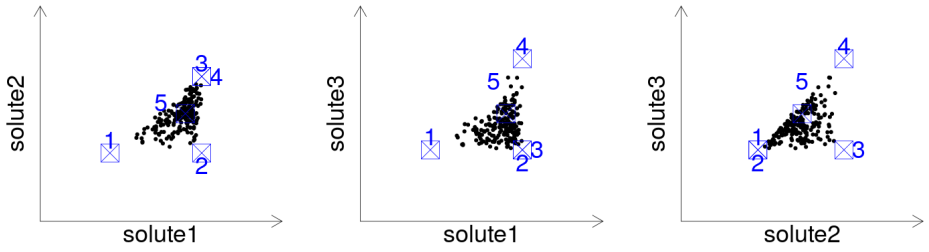
— rien ne garantit que toutes les sources soient un sommet de l'enveloppe convexe des sources sur tous les plans.

Le dernier désavantage est le plus problématique. Il est en effet possible que la projection sur un plan d'une source puisse soit se superposer à celle d'une autre source (Figure 1.4), soit se situer à l'intérieur de l'enveloppe convexe des sources (Figure 1.5). Il est alors difficile de reconstruire les sources dans l'espace des données à partir des plans de projections. Il est donc nécessaire d'utiliser des méthodes numériques afin de supprimer certains de ces désavantages.



(a) (b) (c)

FIGURE 1.4 – Projection d'un jeu de données synthétique, issu d'un système de mélange à quatre sources dans un espace à trois dimensions, dans les trois plans de projection d'études. Sur chaque plan, le mélange semble être issu d'un système de mélange à trois sources, la dernière étant superposée à une des autres sources.



(a) (b) (c)

FIGURE 1.5 – Projection d'un jeu de données synthétique, issu d'un système de mélange à cinq sources dans un espace à trois dimensions, dans les trois plans de projection d'études. Sur chaque plan, le mélange semble être issu d'un système de mélange à trois sources, la quatrième étant superposée à une des autres sources et la cinquième étant à l'intérieur de l'enveloppe convexe des sources.

La seconde manière d'estimer les sources est de considérer le problème géométriquement : il faut trouver le polyèdre convexe qui explique les données. Pour ce faire, il faut relever trois challenges : sélectionner le nombre de sources (le nombre de sommets du polyèdre), générer des polyèdres convexes qui contiennent les données et trouver un critère pour sélectionner le "bon" polyèdre. Le critère de sélection est souvent de garder le polyèdre qui minimise la différence entre son volume et le volume de l'enveloppe convexe des données. Ce critère découle de l'hypothèse selon laquelle les données esquissent le polyèdre convexe formé

par les sources. En d'autres termes, les données décrivent bien le mélange des sources.

Cette méthode a l'avantage de ne nécessiter l'intervention d'un utilisateur uniquement lors de l'initialisation. De plus, cette méthode reposant sur un modèle mathématique : les sources détectées sont les mêmes lorsque l'initialisation est la même. Enfin, cette méthode peut être appliquée à des données multidimensionnelles, réduisant ainsi les effets de projection. Cependant, cette méthode de détection est d'autant plus difficile et calculatoire lorsque le nombre de dimensions des données est grand. En effet, les notions d'enveloppes convexes, de points à l'intérieur de l'enveloppe convexe et de volume d'enveloppes convexes, nécessite des algorithmes dont la complexité dépend du nombre de dimensions.

Dans le cas de données en deux dimensions, détecter les sources revient à trouver un polygone convexe qui contienne les données. Le critère de sélection du polygone choisi dans [Pinti et al., 2020] est la minimisation de l'aire du polygone qui contient les données. Dans cet article, les données sont supposées le résultat d'un mélange de trois sources. Ainsi, le modèle présenté dans cet article recherche le plus petit triangle, en terme d'aire, qui contient les données. Les sources sont alors positionnées aux sommets de ce triangle. Dans le cas de données multidimensionnelles, cette méthode doit être appliquée sur chacun des plans d'études : se pose alors le problème de la reconstruction des sources en multidimensions. Dans cet article, l'objectif est d'étudier les éléments que sont le chlore (Cl) et le brome (Br) afin de mieux comprendre les processus de formation de la Terre. En effet, l'évolution de leurs compositions isotopiques permet de comprendre les processus d'accrétion et de différentiation chimiques. Les données utilisées sont des mesures de fluides géothermales prélevés à Cerro Prieto, Las Tres Virgenes et Los Azufres au Mexique. Ces données sont supposées être le résultat du mélange entre trois sources : le manteau, la croûte et une zone de subduction. Ce jeu de données servira de test de la méthode proposée dans ce manuscrit sur des données réelles.

L'analyse en composantes principales (ACP) peut être utilisée pour estimer le nombre de sources [Christophersen and Hooper, 1992]. La première étape de cette analyse est de diagonaliser la matrice $\mathbf{d}^\top\mathbf{d}$. Cette étape est possible, car cette matrice est symétrique semi-définie positive et admet donc des valeurs propres positives. Les vecteurs propres associés à ces valeurs propres forment une base de l'espace des données. La quantité d'information contenue dans un sous-espace formé en prenant un sous ensemble des vecteurs propre, donc le

pourcentage de la variance contenu dans cet espace, est obtenue en calculant le rapport entre la somme des valeurs propres associées à ces vecteurs propres et la somme de toutes les valeurs propres. L'ACP peut donc être utilisée pour visualiser les données dans un espace de plus faible dimension que l'espace des données tout en conservant un maximum d'information. Le choix du nombre de dimensions de cet espace est laissé à l'expert, c'est à lui de chercher le rapport optimal entre nombre de dimensions et inertie de l'espace.

L'estimation des contributions des sources, appelée analyse directe [Weltje, 1997], utilise généralement des modèles basés sur les analyses de mélanges des membres extrémaux (*End-Member Mixing Analysis (EMMA)*) [Christophersen et al., 1990]. Ce type d'approche a pour but d'estimer γ dans l'équation (1.2). Lorsque les sources sont connues, le problème devient une régression linéaire. La matrice de contribution peut alors être estimée par l'estimateur des moindres carrées $\gamma = (\mathbf{s}^\top \mathbf{s})^{-1}\mathbf{s}^\top \mathbf{d}$. Cette approche ne garantit pas $0 \leq \gamma_{j;i} \leq 1$ et $\sum_{i=1}^{n} \gamma_{j;i} = 1$. Il faut alors considérer le problème comme une régression linéaire sous contraintes. De plus, les sources sont considérées ici comme ponctuelles. Cependant, les sources sont généralement connues au travers d'une loi de probabilité *a priori*. Pour résoudre ce problème, des méthodes existent basées sur les modèles de mélange bayésien d'isotope stable (SIMMs) [Parnell et al., 2013]. La probabilité d'avoir la matrice des contributions γ sachant les données $\mathbf{d}$ est :

$$\mathbb{P}(\gamma|\mathbf{d}) = \frac{L(\mathbf{d}|\gamma)p(\gamma)}{\sum_\gamma L(\mathbf{d}|\gamma)p(\gamma)}$$

avec $L(\mathbf{d}|\gamma)$ la vraisemblance des données sachant la matrice des contributions γ et $p(\gamma)$ la loi *a priori* de choisir γ.

Les modèles de mélanges bayésiens sont adaptés aux différents jeux de données. Dans [Arendt et al., 2015], la méthode Bayesian Monte Carlo permet de sélectionner les contributions générées par une loi *a priori* en fonction de la vraisemblance des données. Dans [Gholami et al., 2019], c'est la méthode Mean Absolute Fit, une autre méthode bayésienne, qui est utilisée pour quantifier les contributions des sources.

Le modèle MixSir (*Mixing Sampling-Importance-Resampling*) [Moore and Semmens, 2008] permet de calculer la distribution des contributions à partir d'une modélisation bayésienne. Les sources sont connues par une loi *a priori* gaussienne. Les contributions des sources sont générées par une loi *a priori* bêta.

Le modèle SIAR (*Stable Isotope Analysis in R*) [Parnell et al., 2010] est un modèle de mélange bayésien sous forme d'un package R qui prend en compte les variations naturelles et les incertitudes pour estimer les contributions.

Le modèle MixSIAR (*Mixing Stable Isotope Analysis in R*) [Stock et al., 2018] est un modèle de détection de sources développé par les écologues qui combine les modèles MixSir et SIAR. Ce modèle peut aussi être adapté dans le cas de l'étude de la formation des gisements de minerai [Holley and Phillips, 2022].

Le modèle D-MixSIAR (*Deconvolutional-MixSIAR*) [Blake et al., 2018] ajoute au modèle MixSIAR la possibilité de prendre en compte la structure hiérarchique du jeu de données : les données sont considérées aussi en fonction du temps et du lieu de leur formation.

Dans la pratique, le nombre de sources est souvent considéré petit (souvent inférieur à 10). Une source peut ne pas être détectée si elle n'a pas influencé suffisamment le jeu de données, c'est-à-dire si sa contribution dans chaque prélèvement est très faible, voire presque nul.

La détection des sources se doit de considérer la totalité des données afin d'éviter des effets de projection. En effet, une source qui forme l'enveloppe convexe des sources sur un plan d'étude, peut ne plus le faire dans un autre plan. Dans la Figure 1.5, toutes les sources ne font pas partie de l'enveloppe convexe sur tous les plans d'études. Ainsi, en considérant les plans de manière individuelle, certaines sources peuvent ne pas être détectées. Il est donc nécessaire de développer des modèles qui donnent la composition des sources dans toutes les dimensions.

1.3 Détection de structures dans des données spatiales

Les données sont vues comme des données spatialisées. La détection de sources est équivalente à la détection de structure dans des données spatiales ou encore la détection d'objets, c'est-à-dire l'estimation de la position et des paramètres des objets présents dans l'espace analysé.

La détection d'objets, de structures ou de motifs est un problème connu de l'analyse d'image. Les modèles bayésiens sont des modèles probabilistes d'analyse d'images : les images sont supposées la réalisation d'un champ aléatoire

markovien. Ce champ peut être simulé par des méthodes de Monte-Carlo par chaîne de Markov (MCMC). L'image est une grille de $n \in \mathbb{N}$ pixels colorés notée $\mathbf{s} = (s_1, \ldots, s_n)^\top$ parmi l'ensemble de toutes les configurations de pixels Ω où s_i est la couleur du pixel numéro $i \in [\![1; n]\!]$. La répartition des couleurs suit une loi de probabilité définie par une fonction $H : \Omega \to \mathbb{R}$, dite d'énergie, déterminant un champ de Gibbs [Geman and Geman, 1984]. La réalisation d'une image $\mathbf{s} \in \Omega$ est contrôlée par la loi d'un champ de Gibbs de fonction énergie $H(\mathbf{s})$ est :

$$p(\mathbf{s}) = \frac{\exp\left(-H(\mathbf{s})\right)}{\sum_{\mathbf{y}\in\Omega} \exp\left(-H(\mathbf{y})\right)}. \tag{1.5}$$

L'énergie d'une image est la somme de fonctions d'énergie, chacune de ces fonctions contrôlant des aspects particuliers de l'image. L'image la plus probable est celle à l'énergie la plus basse [Winkler, 2003]. La fonction énergie peut contrôler la tendance de l'image à être partitionnée en zones homogènes [Geman et al., 1990]. Le modèle d'Ising est un exemple qui peut être utilisé à ce propos. Dans ce modèle, les pixels $\mathbf{s}$ peuvent prendre deux couleurs : la couleur noire de valeur 1 et la couleur blanche de valeur -1. La fonction énergie du modèle d'Ising s'écrit :

$$H(\mathbf{s}) = -\beta \sum_{i\sim j} s_i s_j,$$

avec $\beta \in \mathbb{R}$ un paramètre, $s_i \in \{-1, 1\}$ la valeur du pixel i et $i \sim j$ qui indique que le pixel numéro j est voisin du pixel numéro i (les pixels ont une face en commun). Le signe de β indique la tendance du modèle à générer des images très partitionnées. Lorsque β est négatif, des pixels voisins de même couleur ($s_i = s_j$) diminuent la fonction énergie et donc augmentent la probabilité de l'image. À l'inverse, lorsque β est positif, des pixels voisins de couleurs différentes ($s_i = -s_j$) augmentent la fonction énergie et donc diminuent la probabilité de l'image. En se basant sur ce modèle, des structures plus complexes peuvent être étudiées, par exemple en augmentant le nombre de couleurs que peut prendre un pixel ou en définissant des structures particulières (des nouveaux types de voisinages, des alternances de couleurs...).

Ces méthodes sont liées à l'utilisation d'une grille. Dans une image, la grille est naturellement délimitée par les pixels. Dans le cas de la détection de sources, il est nécessaire de définir une grille et donc d'ajouter de la dépendance à l'utilisateur. Pour cette raison, un autre modèle mathématique est préféré : les processus ponctuels.

Les processus ponctuels distribuent des configurations de points $\mathbf{s}$ parmi l'ensemble de toutes les configurations de points Ω dans un espace. L'exemple classique est le processus ponctuel de Poisson homogène de paramètre $\rho \in \mathbb{R}_+$ qui distribue de manière uniforme et indépendante dans l'espace un nombre de points n tiré selon une loi de Poisson de paramètre ρ. La densité de probabilité d'un processus ponctuel de Gibbs se définit par une fonction énergie [Møller and Waagepetersen, 2004] et a donc a une forme similaire à celle de l'équation (1.5) [Stoyan et al., 2013] :

$$p(\mathbf{s}) = \frac{\exp\left(-H(\mathbf{s})\right)}{\int_{\mathbf{y}\in\Omega} \exp\left(-H(\mathbf{y})\right) d\mu(\mathbf{y})}, \tag{1.6}$$

avec Ω l'ensemble de toutes les configurations de points possibles et μ la mesure de référence du processus ponctuel de Poisson.

Dans ces processus ponctuels, il s'agit de modéliser la présence d'objets/ d'individus/ d'évènements/de caractéristiques dans l'espace et le temps [M. N. M. van Lieshout, 2000]. L'éventail des domaines d'études est très varié. Par exemple, pour modéliser les cambriolages dans la ville de Los Angeles, un processus ponctuel auto-excité a été étudié [Mohler et al., 2011]. Dans ce modèle, l'occurrence d'un cambriolage, représentée par un point, favorise l'occurrence d'autres cambriolages. Un autre exemple est l'utilisation d'un processus ponctuel de Markov dans la modélisation des séismes [Ogata, 1998]. Les séismes, représentés par des points, produisent des répliques et augmentent la probabilité d'apparition d'autres séismes. Un dernier exemple est un processus ponctuel non stationnaire log-gaussien de Cox utilisé pour modéliser la propagation d'une maladie comme la gastro-entérite au Royaume-Uni [Diggle et al., 2005]. Ce modèle se décompose en trois facteurs : le premier, déterministe, contrôle la propagation spatiale de la maladie, le second, déterministe aussi, contrôle l'évolution temporelle de la maladie et le dernier, stochastique, les foyers d'apparition de la maladie.

Les processus ponctuels marqués permettent d'étudier la distribution de caractéristiques dans l'espace. Un point possède une marque, c'est-à-dire un ensemble de caractéristiques qualitatives et/ou quantitatives. Un processus de Poisson à pulsation rectangulaire est utilisé dans la prédiction des précipitations [Rodriguez-Iturbe et al., 1987]. À chaque évènement pluvieux, représenté par un point, est associé une marque : un rectangle dont la longueur représente la durée de l'évènement et la largeur l'intensité de l'évènement. En foresterie, les processus ponctuels sont utilisés pour modéliser la répartition des arbres [Stoyan and Penttinen, 2000]. La marque d'un point contient des informations sur l'arbre comme le diamètre du tronc, la hauteur de l'arbre, son espèce ou encore son état de

santé.

La fonction énergie d'un processus ponctuel permet d'intégrer les connaissances sur le phénomène étudié et sur les données directement dans le modèle. En intégrant dans le processus ponctuel les considérations sur le mode de propagation d'une maladie, il est possible de modéliser sa propagation [Stoica et al., 2007a]. Un processus de Gibbs appelé Candy [Stoica et al., 2004] détecte un réseau routier à partir d'une image satellite. Ce processus distribue des points représentant le milieu d'un segment de route. La marque d'un point est l'ensemble composé de la largeur du segment, de sa longueur et de son orientation. Ce modèle considère le fait que les routes sont connectées et que ces connections se font le plus souvent avec un certain type d'angle. En astronomie, le processus ponctuel Bisous, basé sur le modèle Candy, détecte les filaments cosmiques dans les cartes tridimensionnelles des galaxies [Stoica et al., 2007b].

Ces derniers exemples ont fortement inspiré la démarche adoptée pendant la thèse. Nous nous sommes proposés de construire un processus ponctuel pour être appliqué à la détection de sources. La flexibilité des processus ponctuels sera utilisée pour intégrer dans le modèle des critères physiques et géologiques sur les mélanges (sources connues, plages de valeurs des sources,...).

1.4 Construction de notre modèle de détection de sources

La configuration de sources n'est pas directement observable à partir des données. Ainsi, elle est supposée être la réalisation d'un processus ponctuel de Gibbs qui distribue les sources dans l'espace des données. La fonction énergie de ce processus considère les contraintes physiques et géologiques des mélanges. Ce modèle, développé au cours de cette thèse, est nommé "HUG" câlin en anglais en raison de la tendance des sources à "enlacer" les données.

Le modèle HUG prend en compte le caractère multidimensionnel des données, les contraintes physiques d'un mélange et des critères géologiques d'un système de mélange à plusieurs sources. Le système de mélange est de la forme (1.4). Les critères géologiques sont ceux considérés pour la détection graphique des sources :

1) les sources sont relativement proches des données,

2) les sources “expliquent” les données, c’est-à-dire que les données sont dans l’enveloppe convexe des sources,

3) le nombre de sources doit être minimisé,

4) les sources ne sont pas trop proches les unes des autres.

Il est important de noter que les notions de position et de distance ne sont pas utilisées dans le sens géographique. Il s’agit respectivement de position dans l’espace des données, donc la composition en les différents paramètres hydrogéochimiques, et de la différence de composition des sources.

La fonction énergie du modèle HUG sera la traduction des critères mentionnés auparavant. Sa densité de probabilité s’écrit

$$p(\mathbf{s}|\theta,\mathbf{d}) = \frac{\exp\left[-U(\mathbf{s}|\theta)\right]}{Z(\theta)} = \frac{\exp\left[-U_{\mathbf{d}}(\mathbf{s}|\theta,\mathbf{d}) - U_i(\mathbf{s}|\theta)\right]}{Z(\theta)}, \tag{1.7}$$

où $U(\mathbf{s}|\theta)$ est l’énergie de la configuration $\mathbf{s}$ et $Z(\theta) = \int_\Omega \exp\left[-U(\mathbf{y}|\theta)\right] d\mu(\mathbf{y})$ est la constante de normalisation, et θ le vecteur des paramètres du modèle de dimension I. La fonction énergie est composée de deux termes. Le premier $U_{\mathbf{d}}(\mathbf{s}|\theta,\mathbf{d})$ est le terme d’attache aux données et contrôle la répartition des sources par rapport aux données. Le second terme $U_i(\mathbf{s}|\theta)$ est le terme d’interaction et contrôle la répartition des sources par rapport aux sources mêmes.

La densité de probabilité du modèle HUG est échantillonnée à l’aide de méthodes de Monte-Carlo par chaînes de Markov (MCMC) s’appuyant sur un algorithme de Metropolis-Hastings. Cet algorithme met à jour une configuration de points en ajoutant un nouveau point ou en supprimant un point de la configuration. L’acceptation ou le refus de cette nouvelle configuration fait intervenir l’équation (1.7). La configuration de sources sélectionnée par le modèle est la configuration la plus probable, donc qui maximise cette même équation (1.7). Pour éviter la détection d’un maximum local, une méthode d’optimisation globale, ici l’algorithme de recuit simulé, est appliquée. Le vecteur de paramètre θ connu dans un premier temps grâce à une loi *a priori* est estimé dans la suite par une méthode d’estimation de paramètre : l’algorithme de calcul bayésien approximatif ou “Approximate Bayesian Computation” en anglais (ABC) Shadow, présenté dans l’article [Stoica et al., 2017].

Le modèle HUG est applicable sur les jeux de données issus d’un mélange linéaire des sources conservatif. De plus, les paramètres hydrogéochimiques considérés ne doivent pas induire de courbe de mélange hyperbolique sur les plans de projections [Langmuir et al., 1978], ce qui peut se produire lorsque le paramètre hy-

drogéochimique considéré est un rapport de concentration ou une composition isotopique. Enfin, le jeu de données doit bien représenter le mélange, c'est-à-dire que l'enveloppe convexe des données doit esquisser l'enveloppe convexe des sources. Si malgré l'incertitude sur les données, les points restent à l'intérieur de l'enveloppe convexe des sources, alors les sources détectées devraient être les mêmes, qu'importe la perturbation des données.

Le modèle HUG utilise les mêmes critères que la détection graphique de sources pour faire de la détection dans l'espace des données complet. L'utilisation d'un modèle mathématique permet de diminuer l'effet de l'utilisateur sur la détection. Enfin, l'utilisation d'une méthode d'estimation de paramètres transforme le modèle en une méthode non supervisée de détection de sources.

Chapitre 2

Processus ponctuels

Les processus ponctuels sont utilisés pour modéliser des phénomènes aléatoires de répartition, par exemple d'objets, d'individus, des arbres ou encore des séismes. Un individu est représenté par un point dans un espace. Une réalisation d'un processus ponctuel est une configuration de points, aussi appelée ensemble de points. Chaque configuration de points est obtenue selon la densité de probabilité qui caractérise le processus ponctuel.

Les processus ponctuels permettent de modéliser des phénomènes aléatoires de répartition de caractéristiques chez les individus. Pour cela, chaque individu possède une marque, c'est-à-dire un ensemble de caractéristiques quantitatives et/ou qualitatives. Par exemple, dans le cas de l'étude de la répartition des arbres, une marque possible est le diamètre du tronc, l'espèce et l'état de santé. Les processus ponctuels qui distribuent des individus dits marqués (des individus possédant une marque) sont appelés processus ponctuels marqués. La densité de probabilité d'un tel processus contrôle la distribution des points, mais aussi des marques associées à ces points.

Les processus ponctuels sont aussi utilisés pour détecter des structures particulières dans des données. Le modèle Candy [Stoica et al., 2004], présenté dans le paragraphe 1.3 est un processus ponctuel marqué qui détecte dans des images satellites des réseaux routiers. Les points de ce processus sont les centres de portions rectilignes de route, la marque contient la largeur, la longueur et l'orientation de la portion de route. D'autres exemples d'applications des processus ponctuels en analyse d'image peuvent se trouver dans [Descombes X. (ed.), 2012]. En plus de celles-ci, des nombreuses applications en astronomie et en sciences de l'envi-

ronnement sont décrites d'une manière synthétique dans [Stoica, 2014] et plus détaillée dans les articles cités.

Ce chapitre présente des notions et des résultats fondamentaux concernant les processus ponctuels. Dans cette démarche, nous avons suivi essentiellement les travaux de [Møller and Waagepetersen, 2004] et de [M. N. M. van Lieshout, 2000]. L'objectif de cette présentation est de donner les éléments nécessaires pour la construction d'un processus ponctuel avec interaction capable de modéliser la répartition des sources dans un mélange de fluides.

Le paragraphe 2.1 définit les processus ponctuels. Le paragraphe 2.2 présente des résultats concernant le processus ponctuel de Poisson. La propriété d'indépendance exclut les interactions. Cependant, ce processus peut être utilisé comme mesure de référence pour pouvoir introduire, par rapport à elle, une densité de probabilité caractérisant des modèles plus complexes capables de gérer des interactions 2.3 et 2.4. Le modèle que l'on souhaite proposer est obtenu à partir d'une superposition des interactions. Cette construction doit garantir un modèle bien défini, un modèle intégrable 2.5.

2.1 Processus ponctuels

Soit un espace mesuré $(W, \mathcal{B}, \nu)$, avec W un espace compact de $\mathbb{R}^d$, $\mathcal{B}$ la famille des boréliens et ν la mesure de Lebesgue.

L'ensemble qui contient toutes les configurations de $n \in \mathbb{N}$ points est noté $W_n = \{\mathbf{s} = \{s_1, \ldots, s_n\} \subset W\}$. L'ensemble des configurations de W s'écrit $\Omega = \bigcup_{n=0}^{\infty} W_n$. Cet ensemble est associé au σ-algèbre $\mathcal{F}$ construit à partir de la fonction de comptage :

$$\begin{aligned} n(\cdot) : B \in \mathcal{B} &\longmapsto \mathbb{N} \\ \{s_1, \ldots, s_n\} &\longrightarrow \sum_{i=1}^{n} \mathbb{1}_{\{s_i \in B\}}. \end{aligned}$$

D'où

$$\mathcal{F} = \sigma\Big(\{\mathbf{s} = \{s_1, \ldots, s_n\} \in \Omega;\, \forall B \in \mathcal{B}, \exists m \in \mathbb{N} \quad n(\mathbf{s}_B) = n(\mathbf{s} \cap B) = m\}\Big).$$

Définition 2.1.1 (Processus ponctuel)**.** *Un processus ponctuel X sur W est une*

fonction mesurable d'un espace de probabilité $(\mathcal{S}, \mathcal{A})$ *dans* $(\Omega, \mathcal{F})$*. La loi* $\mathbb{P}$ *de* X *est donnée, pour tout* $F \in \mathcal{F}$*, par la probabilité :*

$$\mathbb{P}(X \in F) = \mathbb{P}(\{\omega \in \mathcal{S} : X(\omega) \in F\}). \tag{2.1}$$

Seuls les processus ponctuels dits simples, c'est-à-dire qui génèrent des ensembles de points composés d'éléments distincts et finis, seront considérés dans la suite.

La variable de comptage N d'un processus ponctuel X est définie par :

$$\forall B \in \mathcal{B} : \ N(B) = n(X_B) = n(X \cap B). \tag{2.2}$$

L'étude des processus ponctuels se fait au travers de deux approches complémentaires : s'intéresser aux points et s'intéresser aux ensembles ne contenant pas de points.

Les évènements vides dans B sont définis par :

$$F_B = \{\mathbf{s} \in \Omega : \ n(\mathbf{s}_B) = 0\}.$$

Théorème 2.1.1. *La loi d'un processus ponctuel est caractérisée par les probabilités de ses événements vides*

$$\forall B \in \mathcal{B} : \ v(B) = \mathbb{P}(N(B) = 0).$$

Les processus ponctuels marqués sont des processus ponctuels dont chacun des points possède une "marque". Une marque m est un ensemble de caractéristiques qui peuvent être qualitatives (couleur, présence d'une maladie, espèce,...), quantitatives (largeur du tronc, nombre d'individus de la même espèce dans un rayon de 10 mètres, âge,...) ou un mélange des deux. Les marques sont à valeur dans un espace M appelé l'espace des marques. Une réalisation d'un processus ponctuel marqué sur $W \times M$ de n points est notée $\{(s_1, m_{s_1}), \ldots, (s_n, m_{s_n})\}$ où $s_i \in W$ dénote la position du point numéro i et $m_{s_i} \in M$ sa marque.

Un processus ponctuel X sur $W \times M$ est dit marqué si la loi marginale des positions des points est un processus ponctuel sur W.

2.2 Processus ponctuel de Poisson

Le processus ponctuel le plus connu est probablement le processus ponctuel de Poisson. Ce processus répartit, de manière indépendante et identiquement distribuée, des points selon une fonction localement intégrable dite fonction d'intensité $\rho : W \to [0, +\infty)$. Le processus ponctuel de Poisson tire son nom de la loi de Poisson qui est utilisée pour définir le nombre de points à générer dans une réalisation. Le paramètre utilisé dans cette loi de Poisson est appelé mesure d'intensité, notée Υ, et est définie au moyen de la fonction d'intensité du processus ponctuel :

$$\forall B \in \mathcal{B} \ : \Upsilon(B) = \int_B \rho(\xi)\, d\nu(\xi).$$

Conditionnellement au nombre de points n, les points générés sont répartis dans B de manière indépendante et identiquement distribués selon la densité $f(x) = \rho(x)/\Upsilon(B)$ pour tout $x \in B$. Ce type de répartition s'appelle processus binomial et est notée $X \sim Binomial(B, n, f)$.

Définition 2.2.1 (Processus ponctuel de Poisson)**.** *Un processus ponctuel de Poisson sur W avec comme fonction d'intensité ρ et mesure d'intensité Υ, noté $X \sim Poisson(W, \rho)$ satisfait les propriétés suivantes :*

(i) $\forall B \in \mathcal{B} : N(B) \sim \mathcal{P}(\Upsilon(B))$
en notant $\mathcal{P}(\Upsilon(B))$ la loi de Poisson de paramètre $\Upsilon(B)$,

(ii) $\forall n \in \mathbb{N}, \forall B \subseteq W :$
$0 < \Upsilon(B) < \infty \Rightarrow \{X_B | N(B) = n\} \sim Binomial(B, n, f)$
avec $f(\xi) = \rho(\xi)/\Upsilon(B)$.

Lorsque $\rho \in \mathbb{R}_+$ est constant, X est appelé processus ponctuel de Poisson homogène d'intensité ou de taux ρ. Ce type de processus modélise une répartition des points complètement aléatoire, c'est-à-dire uniforme dans l'espace. Lorsque ρ est constant et de valeur 1, le processus $X \sim Poisson(W, 1)$ est appelé processus ponctuel de Poisson standard ou processus ponctuel de Poisson de densité unité. Lorsque ρ n'est pas constant, X est appelé processus ponctuel de Poisson, inhomogène ou hétérogène.

La répartition indépendante des points suivant un gradient est modélisée par un processus ponctuel de Poisson inhomogène. Par exemple, la répartition des arbres dans une forêt peut être modélisée par un processus ponctuel de Poisson inhomogène : les zones avec les meilleures conditions (température, humidité, présence de nutriments,...) auront une plus grande probabilité de contenir un arbre tandis que les zones avec des mauvaises conditions auront une probabilité

plus faible.

Définition 2.2.2 (Processus ponctuel de Poisson stationnaire et isotrope)**.** *Un processus ponctuel de Poisson X est :*

stationnaire *s'il est invariant par translation, i.e. si pour tout $\eta \in W$, la loi de $X + \eta = \{\xi + \eta : \xi \in X\}$ est la même que la loi de X.*

isotrope *s'il est invariant par rotation, i.e. si pour toute rotation autour de l'origine de $\mathbb{R}^d$ notée $\mathcal{O}$ la loi de $\mathcal{O}X = \{\mathcal{O}\xi : \xi \in X\}$ est la même que la loi de X.*

Soient un processus ponctuel de Poisson X sur W et les ensembles $B_1, B_2, \ldots$ de W deux à deux disjoints. Les processus $X_{B_1}, X_{B_2}, \ldots$ sont indépendants deux à deux.

La loi d'un processus ponctuel de Poisson $X \sim Poisson(W, \rho)$ de mesure d'intensité Υ est caractérisée par la probabilité des évènements vides :

$$v(B) = \exp(-\Upsilon(B))$$

pour tout $B \subset W$.

La Proposition 2.2.1 est utilisée pour écrire la probabilité d'un processus ponctuel de Poisson.

Proposition 2.2.1. *Soient une fonction densité $\rho : W \to [0, +\infty]$ et Υ sa mesure d'intensité associée :*

(i) $X \sim Poisson(W, \rho)$ si et seulement si $\forall B \subseteq W$, $\forall F \subseteq \mathcal{F}$:

$$\mathbb{P}(X_B \in F) = \sum_{n \in \mathbb{N}} \frac{\exp(-\Upsilon(B))}{n!} \times \int_B \ldots \int_B \mathbb{1}[\mathbf{s} \in F] \prod_{i=1}^{n} \rho(s_i)\, d\nu(s_1) \ldots d\nu(s_n).$$

(ii) si $X \sim Poisson(W, \rho)$ alors $\forall h : \Omega \to \mathbb{R}_+$ et $\forall B \subseteq W$ avec $\Upsilon(B) < \infty$:

$$\mathbb{E}(h(X_B)) = \sum_{n \in \mathbb{N}} \frac{\exp(-\Upsilon(B))}{n!} \int_B \ldots \int_B h(\mathbf{s}) \prod_{i=1}^{n} \rho(s_i)\, d\nu(s_1) \ldots d\nu(s_n).$$

L'absence d'interaction, inhérente aux processus ponctuel de Poisson, fait naître la nécessité de processus ponctuels tenant en compte les interactions. Le processus ponctuel de Poisson standard est la pierre angulaire sur laquelle repose la construction de ces processus.

2.3 Densité de probabilité de processus avec interaction

La construction des processus ponctuels avec interaction se fait à partir du théorème de Radon-Nykodym et du processus ponctuel de Poisson standard de mesure μ.

Théorème 2.3.1 (Radon-Nykodym)**.** *Soient μ et ν des mesures σ-finies de l'espace mesurable $(\Omega, \mathcal{F})$ avec ν qui est absolument continue par rapport à μ. Il existe une fonction mesurable $f : \Omega \to [0, \infty)$, appelée dérivée de Radon-Nykodym, telle que pour tout $B \in \mathcal{F}$:*

$$\nu(B) = \int_B f(\xi) d\mu(\xi).$$

Pour deux densités de ce type, f et g, nous avons $\mu[f \neq g] = 0$.

Pour tout processus ponctuels X de densité f par rapport au processus ponctuel de Poisson standard et d'après la Proposition 2.2.1, pour tout $F \in \mathcal{F}$, sa loi de probabilité s'écrit :

$$\mathbb{P}(X \in F) = \sum_{n \in \mathbb{N}} \frac{\exp(-\nu(W))}{n!} \int_W \dots \int_W \mathbb{1}_{\mathbf{s} \in F} f(\{s_1, \cdots, s_n\}) d\nu(s_1) \dots d\nu(s_n).$$

La densité f est souvent connue uniquement via une fonction $h : \Omega \to [0, \infty)$ qui lui est proportionnelle, notée $f \propto h$. Cela est dû à la constante de normalisation c difficilement calculable qui a cette forme :

$$c = \sum_{n \in \mathbb{N}} \frac{\exp(-\nu(W))}{n!} \int_W \dots \int_W h(\{x_1, \cdots, s_n\})\, d\nu(s_1) \dots d\nu(s_n).$$

L'intensité conditionnelle de Papangelou [Papangelou, 1974] ou intensité de Papangelou ou encore intensité conditionnelle s'écrit :

$$\forall \mathbf{s} \in \Omega, \xi \in W : \lambda^*(\mathbf{s}, \xi) = \frac{f(\mathbf{s} \cup \xi)}{f(\mathbf{s})}.$$

Le terme $\lambda^*(\mathbf{s}, \xi) d\nu(\xi)$ peut être interprété comme la probabilité pour le processus X d'avoir un point dans la région de volume $d\nu(\xi)$.

L'intensité conditionnelle de Papangelou est utilisée pour définir la notion de stabilité des processus ponctuels et donc l'intégrabilité de leur densité de probabilité. En effet, l'intensité de Papangelou est en bijection avec la densité du processus si la densité f est une fonction héréditaire.

Définition 2.3.1 (Fonction héréditaire)**.** *Soit f la densité d'un processus ponctuel défini sur W. La fonction f est une fonction héréditaire si pour toute réalisations $\mathbf{s}, \mathbf{x} \in \Omega$ telles que $\mathbf{s} \subset \mathbf{x}$ alors :*

$$f(\mathbf{x}) > 0 \Rightarrow f(\mathbf{s}) > 0.$$

2.4 Processus ponctuels de Markov

Les processus ponctuels de Markov considèrent les interactions qu'entre "voisins". Le voisinage est défini par une relation réflexive et symétrique que l'on note $\sim$. La relation de voisinage entre deux points $\xi, \eta \in W$ est notée $\xi \sim \eta$. Le voisinage du point $\xi \in W$ est défini par :

$$V_\xi = \{\eta \in W : \eta \sim \xi\}.$$

Le voisinage de $B \subset W$ est $V_B = \{\eta \in W : \exists \xi \in B : \xi \sim \eta\}$. La frontière ou bordure de B est $\partial B = V_B \backslash B$.

Un exemple de relation de voisinage peut se construire en comparant la distance entre deux points à une constante connue $R \in \mathbb{R}_+$: les points sont voisins s'ils sont à une distance plus petite que R. Dans ce cas, le voisinage du point ξ est la boule centrée en ξ et de rayon R notée $b(\xi, R)$.

Définition 2.4.1 (Processus ponctuels de Markov)**.** *Soit un processus ponctuel X de densité $f : \Omega \to [0, \infty)$ par rapport au processus ponctuel de Poisson standard. Ce processus est dit de Markov si pour tout $\mathbf{s} \in \Omega$ tel que $f(\mathbf{s}) > 0$:*

- *f est héréditaire,*
- *pour tout $\xi \in W$, $\lambda^*(\mathbf{s}, \xi)$ ne dépend que de ξ et de ses voisins $V_\xi \cap \mathbf{s}$.*

Définition 2.4.2 (Fonction de Markov)**.** *Une fonction $f : \Omega \to [0, \infty)$ est appelée fonction de Markov si pour tout $\mathbf{s} \in \Omega$ tel que $f(\mathbf{s}) > 0$ et $\xi \in W$ alors $\frac{f(\mathbf{s} \cup \{\xi\})}{f(\mathbf{s})}$ ne dépend que de ξ et de ses voisins dans $\mathbf{s}$.*

Un processus ponctuel dont la densité est une fonction de Markov est un processus ponctuel de Markov.

Définition 2.4.3 (Processus ponctuels de Gibbs)**.** *Un processus ponctuel sur W*

de densité de probabilité non-normalisée $f : \Omega \to [0, \infty)$ par rapport au processus ponctuel de Poisson standard μ est un processus ponctuel de Gibbs si,

$$p(\mathbf{s}) = \frac{\exp[-U(\mathbf{s})]}{\int_\Omega \exp[-U(\mathbf{y})]\, d\mu(\mathbf{y})}. \tag{2.3}$$

avec $U(\mathbf{s}) = -\log(f(\mathbf{s}))$ la fonction d'énergie et la fonction de partition ou constante de normalisation, la quantité $\int_\Omega \exp[-U(\mathbf{y})]\, d\mu(\mathbf{y})$.

Exemple 2.4.1 (Processus ponctuel de Poisson)**.** *Le processus ponctuel de Poisson est un processus ponctuel de Gibbs. En effet, la densité d'un processus ponctuel de Poisson de fonction d'intensité $\rho : W \to [0, +\infty)$ s'écrit :*

$$f(\mathbf{s}|\rho) \propto \prod_{i=1}^{n(\mathbf{s})} \rho(s_i) = \exp\left(\sum_{i=1}^{n(\mathbf{s})} \log(\rho(s_i))\right). \tag{2.4}$$

En particulier, la fonction d'énergie d'un processus ponctuel de Poisson homogène s'écrit :

$$U(\mathbf{s}|\rho) = -n(\mathbf{s}) \log(\rho). \tag{2.5}$$

Le nombre de points $n(\mathbf{s})$ est la statistique suffisante du modèle.

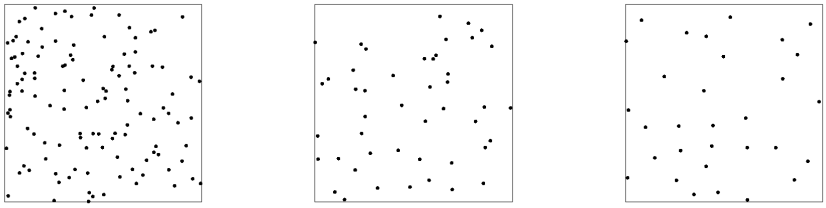
Exemple 2.4.2 (Processus ponctuel de Strauss)**.** *Le processus ponctuel de Strauss est également un processus ponctuel de Gibbs. Sa densité de probabilité par rapport à la référence poissonienne s'écrit :*

$$f(\mathbf{s}|\rho, \gamma) \propto \rho^{n(\mathbf{s})} \gamma^{s_r(\mathbf{s})} = \exp\left(n(\mathbf{s}) \log(\rho) + s_r(\mathbf{s}) \log(\gamma)\right). \tag{2.6}$$

La fonction énergie est

$$U(\mathbf{s}|\rho, \gamma) = -n(\mathbf{s}) \log(\rho) - s_r(\mathbf{s}) \log(\gamma). \tag{2.7}$$

Les statistiques suffisantes du modèle sont $n(\mathbf{s})$ le nombre de points et $s_r(\mathbf{s})$ le nombre de paires de points situées à une distance plus petite que r. Les paramètres du modèle sont $\beta > 0$ et $\gamma \in (0, 1]$. Ce modèle pénalise les configurations de points ayant des paires des points trop rapprochées. Lorsque $\gamma = 1$, le processus est un processus ponctuel de Poisson, tandis que lorsque γ approche 0 le processus tend vers la limite d'un processus dit "hard-core", qui interdit l'interaction entre les paires de points. La Figure 2.1 présente des réalisations d'un processus de Strauss dans le carré unité pour différentes valeurs de γ. Plus la valeur de γ est faible, moins il y a de paires de points à une distance plus petite que r.



(a) (b) (c)

FIGURE 2.1 – Réalisation d'un processus de Strauss dans le carré unité avec $r = 0.1$, $\rho = 100$ et $\gamma = 1$ (a), $\gamma = 0.5$ (b), $\gamma = 0.01$ (c).

2.5 Stabilité des processus ponctuels

La construction d'une densité de processus ponctuel se fait souvent sans avoir accès à sa constante de normalisation. Il faut s'assurer que cette construction définit bien une densité de probabilité. Cette fonction ne doit pas être négative et doit être intégrable. Les résultats suivants sont des outils pour s'assurer que ces propriétés soient vérifiées.

Définition 2.5.1 (Stabilité locale et stabilité de Ruelle)**.** *Soient* $f : \Omega \to [0, \infty)$ *et* $\phi^* : W \to [0, \infty)$ *telle que* $\int_W \phi^*(\xi) d\nu(\xi) < \infty$.

La fonction f *est stable localement (ou localement* ϕ^**-stable) lorsque :*

$$\forall \mathbf{s} \in \Omega, \xi \in W : \ f(\mathbf{s} \cup \xi) \leq \phi^*(\xi) f(\mathbf{s}).$$

La fonction f *est stable au sens de Ruelle (ou* ϕ^**-stable de Ruelle) lorsque :*

$$\forall \mathbf{s} \in \Omega : \ f(\mathbf{s}) \leq \alpha \prod_{\xi \in \mathbf{s}} \phi^*(\xi).$$

La stabilité de Ruelle signifie que f est dominé par une densité de Poisson non normalisée et donc f est intégrable par rapport à un processus ponctuel de Poisson standard. En effet, la stabilité de Ruelle implique qu'il existe $M \in \mathbb{R}_+$ tel que pour tout $\mathbf{s} \in \Omega$ et $\xi \in \Omega \backslash \mathbf{s}$:

$$\lambda^*(\mathbf{s}, \xi) = \frac{f(\mathbf{s} \cup \xi)}{f(\mathbf{s})} \leq M.$$

La stabilité locale implique la stabilité de Ruelle. Il résulte qu'il existe $M \in \mathbb{R}_+$ tel que pour tout $\mathbf{s} \in \Omega$:

$$f(\mathbf{s}) \leq M^{n(\mathbf{s})}.$$

Les processus ponctuels avec interaction décrits par une densité de probabilité sont simulés par des chaînes de Markov dont la distribution à l'équilibre est la loi du processus. L'intensité de Papangelou est nécessaire dans les calculs des taux d'acceptations des transitions des dynamiques de simulation. Plus important encore, les preuves de convergence de ces algorithmes demandent que le modèle processus simulé soit localement stable.

Chapitre 3

Chaînes de Markov propriétés et algorithmes de simulation

Les processus ponctuels de Gibbs avec interaction ont une constante de normalisation qui n'est pas disponible sous forme analytique. La simulation des tels modèles fait appels à des techniques de type Monte-Carlo par chaînes de Markov (MCMC).

Le principe de ces méthodes est de simuler une chaîne de Markov qui a comme distribution d'équilibre la loi du processus ponctuel d'intérêt. Les bonnes propriétés du noyau de transition de la chaîne impliquent implicitement convergence de l'algorithme de simulation, suffisamment rapidement et indépendamment des conditions initiales.

L'algorithme de Metropolis-Hastings (MH) [M. N. M. van Lieshout, 2000, Møller and Waagepetersen, 2004, Stoica, 2014] est notre choix pour la simulation des processus ponctuels. Il s'agit d'un algorithme $\phi-$irréductible, Harris récurent et ergodique géométrique. Lorsque des modèles en très grandes dimensions doivent être simulées, l'échantillonneur de Gibbs peut être utilisée. La simulation des lois conditionnelles se fait soit directement, si possible, sinon elle peut être approchée par un autre échantillonneur, comme l'algorithme MH. C'est la stratégie que l'on adopte pour pouvoir simuler des processus ponctuels avec interaction en très grande dimension, conjointement avec la distribution de leur paramètre.

Le paragraphe 3.1 introduit les chaînes de Markov et présente les propriétés qui permettent à ces chaînes de converger vers une distribution cible. Pour cette pré-

sentation, l'on s'est appuyé sur [Stoica, 2014] et [Winkler, 2003]. L'algorithme MH pour la simulation des processus ponctuels avec interaction, ainsi qu'un algorithme de Gibbs sont décrits dans le paragraphe 3.2. Une analyse des performances de l'algorithme MH pour simuler des processus de Poisson et de Strauss sont analysées dans le paragraphe 3.3.

3.1 Chaînes de Markov

Soit $(\Omega, \mathcal{F}, \mu)$ un espace de probabilité et $X_0, X_1, \ldots, X_n, \ldots$ une suite de variables aléatoires de cet espace noté $(X_n)_{n\in\mathbb{N}}$.

Une chaîne de Markov est une suite de variables aléatoires dont l'état à l'instant n ne dépend que de l'état à l'instant $n-1$ à travers une fonction appelée noyau de transition.

Définition 3.1.1 (Noyau de transition)**.** *Une fonction $P : \Omega \times \mathcal{F} \to [0,1]$ est un noyau de transition, si :*

- $\forall \mathbf{s} \in \Omega$, $P(\mathbf{s}, \cdot)$ *est une mesure de probabilité,*
- $\forall F \in \mathcal{F}$, $P(\cdot, F)$ *est mesurable.*

Définition 3.1.2 (Chaîne de Markov)**.** *Une suite de variables aléatoires $(X_n)_{n\in\mathbb{N}}$ est une chaîne de Markov notée (X_n), si pour un noyau de transition $P : \Omega \times \mathcal{F} \to [0,1]$, un ensemble $F \in \mathcal{F}$ et une suite d'états $(\mathbf{s}_n)_{n\in\mathbb{N}}$ de Ω :*

$$\mathbb{P}(X_{n+1} \in F | X_0 = \mathbf{s}_0, \ldots, X_n = \mathbf{s}_n) = \mathbb{P}(X_{n+1} \in F | X_n = \mathbf{s}_n) = \int_F P(\mathbf{s}_n, d\mathbf{s}).$$

Un noyau de transition P est dit réversible par rapport à une mesure π si pour tout $F_1, F_2 \in \mathcal{F}$:

$$\int_{F_1} \pi(d\mathbf{s}) P(\mathbf{s}, F_2) = \int_{F_2} \pi(d\mathbf{s}) P(\mathbf{s}, F_1).$$

En d'autres termes, un noyau de transition est réversible si la probabilité de passer d'un ensemble F_1 à un ensemble F_2 est la même que de passer de F_2 à F_1.

Si le noyau de transition reste inchangé lorsque n évolue, la chaîne de Markov est dite homogène. Dans le cas contraire, la chaîne de Markov est dite inhomogène.

Une mesure π, $\sigma-$finie sur $(\Omega, \mathcal{F})$, est dite distribution invariante d'une chaîne

de Markov homogène de noyau de transition P si :

$$\forall F \in \mathcal{F} : \pi(F) = \int_{\Omega} \pi(d\mathbf{s}) P(\mathbf{s}, F).$$

Soit une chaîne de Markov (X_n) homogène de noyau de transition P et de distribution invariante π. La distribution invariante est une distribution d'équilibre si et seulement si :

$$\lim_{n\to\infty} P^{(n)}(\mathbf{s}, F) = \pi(F), \quad \forall \mathbf{s} \in \Omega, F \in \mathcal{F}$$

avec $P^{(n)}$ la n-ième application du noyau de transition. Ici l'on a implicitement considéré qu'une topologie définie sur l'espace des mesures de probabilité existe.

Pour simuler une distribution, il faut donc utiliser une chaîne de Markov dont la distribution à l'équilibre est la distribution recherchée. La convergence de la chaîne de Markov est assurée lorsqu'elle est irréductible, récurrente et ergodique.

Une chaîne de Markov est irréductible s'il est possible de passer de n'importe quel point de départ $\mathbf{s}_1 \in \Omega$ à n'importe quel point d'arrivée $\mathbf{s}_2 \in \Omega$ en un nombre fini d'étapes.

Lorsque l'espace des états est un espace continu, la probabilité de passer de $\mathbf{s}_1$ à $\mathbf{s}_2$ est nulle presque partout. Il est donc nécessaire de définir une notion similaire à l'irréductibilité sur les espaces continus : la ϕ-irréductibilité. Ici ϕ est une mesure sur $\mathcal{F}$. Une chaîne est ϕ-irréductible s'il est possible de passer de tout points de départ $\mathbf{s} \in \Omega$ vers tout ensembles d'arrivées $F \in \mathcal{F}$, tels que $\phi(F) > 0$, en un nombre fini d'étapes.

Le temps de retour de la chaîne en $F \in \mathcal{F}$ est donnée par :

$$\tau_F = \min\{n \geq 1 : X_n \in F\}.$$

Si la chaîne n'atteint jamais F alors $\tau_F = \infty$.

Définition 3.1.3 (ϕ-irréductibilité)**.** *Soit ϕ une mesure sur $\mathcal{F}$. La chaîne de Markov (X_n) est dite ϕ-irréductible si pour tout état initial $\mathbf{s} \in \Omega$ et tout ensemble $F \in \mathcal{F}$ tel que $\phi(F) > 0$,*

$$\mathbb{P}_{\mathbf{s}}(\tau_F < \infty) > 0$$

avec $\mathbb{P}_{\mathbf{s}}(\cdot)$ la probabilité conditionnelle de la chaîne sachant l'état initiale $\mathbf{s}$.

Définition 3.1.4. *Soit $P(\cdot, \cdot)$ le noyau de transition d'une chaîne de Markov*

ayant une mesure invariante π. Supposons qu'il existe un ensemble $A \in \mathcal{F}$, une mesure de probabilité ν telle que $\nu(A) = 1$, une constante $\epsilon > 0$ et un entier $n_0 \geq 1$ tel que

$$P^{n_0}(\mathbf{x}, \cdot) \geq \epsilon\nu(\cdot) \quad \forall \mathbf{x} \in A.$$

La chaîne de Markov associée est apériodique si

$$pgcd\{m : \exists \quad \epsilon_m > 0 \quad tel\ que \quad P^m(\mathbf{x}, \cdot) \geq \epsilon_m \nu(\cdot)\} = 1.$$

Si une chaîne de Markov irréductible et apériodique admet une distribution invariante π alors cette distribution est unique et est la distribution à l'équilibre de la chaîne.

Théorème 3.1.1. *Soit (X_n) une chaîne $\pi-$irréductible et apériodique avec π l'unique mesure invariante. Alors, il existe un ensemble $\Omega' \subseteq \Omega$, tel que $\pi(\Omega') = 1$, et pour tout $\mathbf{s} \in \Omega'$:*

$$\sup_{F \in \mathcal{F}} |P^n(\mathbf{s}, F) - \pi(F)| \underset{n \to +\infty}{\longrightarrow} 0$$

La récurrence signifie que les ensembles $F \in \mathcal{F}$, tels que $\phi(F) > 0$, sont visités suffisamment souvent.

Un ensemble $F \in \mathcal{F}$ est Harris récurrent si pour tout $\mathbf{s} \in F$

$$\mathbb{P}_{\mathbf{s}}(\eta_F = \infty) = 1 \tag{3.1}$$

avec η_F le nombre de passages de la chaîne dans F donné par :

$$\eta_F = \sum_{n=1}^{\infty} \mathbb{1}\{X_n \in F\}.$$

Définition 3.1.5 (Récurrence au sens de Harris)**.** *La chaîne de Markov (X_n) est Harris récurrente si elle est ϕ-irréductible et si tout ensemble $A \in F$ tel que $\phi(A) > 0$ est Harris récurrent.*

Il est possible de montrer que la Harris récurrence est une propriété plus forte que la ϕ- irréductibilité. En effet, la Harris récurrence impose $\mathbb{P}_{\mathbf{s}}(\tau_A < \infty) = 1$, alors que la ϕ-irréductibilité ne demande que $\mathbb{P}_{\mathbf{x}}(\tau_A < \infty) > 0$ [Meyn and Tweedie, 2009, Stoica, 2014].

Une chaîne $\phi-$irréductible et Harris récurrente converge vers son unique distribution d'équilibre, indépendamment des conditions initiales. L'ergodicité nous fournit des informations quant à la vitesse de convergence.

Définition 3.1.6. *La norme en variation totale d'une mesure bornée et signée ν sur $(\Omega, \mathcal{F})$ est définie par*

$$\| \nu \| = \sup_{A \in \mathcal{F}} \nu(A) - \inf_{A \in \mathcal{F}} \nu(A).$$

La distance en variation totale entre deux mesures ν_1 et ν_2 est $\| \nu_1 - \nu_2 \|$.

Définition 3.1.7. *Soit (X_n) une chaîne ergodique* **de** *distribution invariante π. La chaîne est dite géométriquement ergodique s'il existe une fonction $M : \Omega \to \mathbb{R}^+$ avec $\pi\|M\| < \infty$ et une constante $\rho \in (0, 1)$ telles que*

$$\| P^n(\mathbf{x}, \cdot) - \pi \| \leq M(\mathbf{x})\rho^n$$

pour tout $\mathbf{x} \in \Omega$. La chaîne est dite uniformément ergodique s'il existe des constantes $M > 0$ et $\rho \in (0, 1)$ telles que

$$\sup_{\mathbf{x} \in \Omega} \| P^n(\mathbf{x}, \cdot) - \pi \| \leq M\rho^n.$$

L'ergodicité géométrique signifie que l'itération répétée du noyau de transition approche la distribution d'équilibre de la chaîne à une vitesse géométrique.

Théorème 3.1.2. *Soit (X_n) une chaîne $\phi-$irréductible et apériodique. La chaîne est géométriquement ergodique s'il existe une fonction $V : \Omega \to [1, \infty)$, des constantes $a < 1$ et $b < \infty$, et un ensemble petit $C \in \mathcal{F}$ telles que*

$$PV(\mathbf{x}) \leq aV(\mathbf{x}) + b\mathbb{1}_C(\mathbf{x}), \quad \forall \mathbf{x} \in \Omega. \tag{3.2}$$

Le résultat suivant permet d'établir l'ergodicité géométrique. Cette propriété implique la Harris récurrence, qui implique la $\phi-$irréductibilité. C'est en utilisant ce résultat que l'on montre que la dynamique Metropolis-Hastings que l'on va présenter est $\phi-$irréductible, apériodique et géométrique ergodique [M. N. M. van Lieshout, 2000, Møller and Waagepetersen, 2004, Stoica, 2014].

3.2 Algorithmes pour la simulation

3.2.1 Algorithme de Metropolis-Hastings

L'algorithme de Metropolis-Hastings simule une chaîne de Markov de noyau P ayant comme loi d'équilibre p. Cela se fait par une procédure de mise à jour des états de la chaîne. A chaque étape, un nouvel état est généré à partir de l'état précédant selon Q le noyau de proposition. Cet état est accepté avec probabilité α, la probabilité d'acceptation. Le schéma décrivant ce procédé est :

- choisir un état initial $\mathbf{s}_0 \in \Omega$, le noyau de proposition $Q(\mathbf{s}, \cdot)$ et la loi d'intérêt $p(\cdot)$
- à l'étape n : générer $\mathbf{s}_{n+1}$ à partir du noyau de proposition $Q(\mathbf{s}_n, d\mathbf{s}_{n+1})$
- accepter le nouvel état avec la probabilité

$$\alpha(\mathbf{s}_n, \mathbf{s}_{n+1}) = \min\left\{1, \frac{p(\mathbf{s}_{n+1})Q(\mathbf{s}_{n+1}, d\mathbf{s}_n)}{p(\mathbf{s}_n)Q(\mathbf{s}_n, d\mathbf{s}_{n+1})}\right\}.$$

Pour simuler un processus ponctuel défini par sa densité de probabilité p, le noyau de transition agit en proposant des "petites" transformations à partir de la configuration initiale $\mathbf{s}$. Avec probabilité p_b on propose de faire une "naissance", c'est-à-dire ajouter un nouveau point à la configuration. Le nouveau point $\eta \in W$ est proposé en utilisant la densité de probabilité $b(\mathbf{s}, \eta)$. Avec probabilité p_d une "mort" est proposé, c'est-à-dire enlever un point de la configuration. Le point η est choisi avec la probabilité $d(\mathbf{s}, \eta)$. Un "changement" est proposé avec la probabilité p_c, c'est-à-dire modifier les paramètres d'un objet de la configuration. Le point η qui a été sélectionné à l'aide de la probabilité $q(\mathbf{s}, \eta)$ pourrait être remplacé par le point ξ choisi en utilisant la loi $c(\mathbf{s}, \eta, \xi)$.

Le noyau de transition résultant, qui assure le passage de la réalisation $\mathbf{s}$ à un

ensemble $F \in \mathcal{F}$, est de la forme :

$$\begin{aligned}
P(\mathbf{s}, F) = p_b \int_\Omega b(\mathbf{s}, \eta)\alpha(\mathbf{s}, \mathbf{s} \cup \{\eta\})\mathbb{1}[\mathbf{s} \cup \{\eta\} \in F]d\nu(\eta) \\
+p_d \sum_{\eta \in \mathbf{s}} d(\mathbf{s}, \eta)\alpha(\mathbf{s}, \mathbf{s} \setminus \{\eta\})\mathbb{1}[\mathbf{s} \setminus \{\eta\} \in F] \\
+p_c \sum_{\eta \in \mathbf{s}} q(\mathbf{s}, \eta) \int_\Omega c(\mathbf{s}, \eta, \xi)\alpha(\mathbf{s}, \mathbf{s} \setminus \{\eta\} \cup \{\xi\})\mathbb{1}[\mathbf{s} \setminus \{\eta\} \cup \{\xi\} \in F]d\nu(\xi) \\
+\mathbb{1}[\mathbf{s} \in F]\Big[1 - p_b \int_\Omega b(\mathbf{s}, \eta)\alpha(\mathbf{s}, \mathbf{s} \cup \{\eta\})d\nu(\eta) \\
-p_d \sum_{\eta \in \mathbf{s}} d(\mathbf{s}, \eta)\alpha(\mathbf{s}, \mathbf{s} \setminus \{\eta\}) \\
-p_c \sum_{\eta \in \mathbf{s}} q(\mathbf{s}, \eta) \int_\Omega c(\mathbf{s}, \eta, \xi)\alpha(\mathbf{s}, \mathbf{s} \setminus \{\eta\} \cup \{\xi\})d\nu(\xi)\Big].
\end{aligned}$$

La réversibilité peut être prouvée obtenue si les conditions d'équilibre local sont respectées. C'est en considérant ces conditions accomplies que l'on peut déterminer les probabilités d'acceptation.

Une proposition "naissance" peut être compensée par une proposition "mort". Pour une configuration $\mathbf{s} \in \Omega$ et un point $\xi \in \Omega \backslash \mathbf{s}$, la première condition d'équilibre local s'écrit

$$p(\mathbf{s})p_b b(\mathbf{s}, \eta)\alpha(\mathbf{s}, \mathbf{s} \cup \{\eta\}) = p(\mathbf{s} \cup \{\eta\})p_d d(\mathbf{s} \cup \{\eta\}, \eta)\alpha(\mathbf{s} \cup \{\eta\}, \mathbf{s}),$$

avec le taux d'acceptation de l'évènement "naissance"

$$r(\mathbf{s}, \eta) = \frac{p_d d(\mathbf{s} \cup \{\eta\}, \eta)}{p_b b(\mathbf{s}, \eta)} \frac{p(\mathbf{s} \cup \{\eta\})}{p(x)}$$

Cela donne pour la probabilité d'acceptation de la proposition "naissance" :

$$\alpha(\mathbf{s}, \mathbf{s} \cup \{\eta\}) = \min\{1, r(\mathbf{s}, \eta)\},$$

et pour la probabilité d'acceptation de l'évènement "mort" est :

$$\alpha(\mathbf{s}, \mathbf{s} \setminus \{\eta\}) = \min\{1, 1/r(\mathbf{s} \setminus \{\eta\}, \eta)\}.$$

L'évènement "changement" peut être inversé par un autre évènement "changement". Pour une configuration $\mathbf{s} \in \Omega$, un point $\eta \in \mathbf{s}$ et un point $\xi \in W$, en

posant $\mathbf{s}' = \mathbf{s} \setminus \{\eta\} \cup \xi$, la seconde condition d'équilibre locale s'écrit :

$$p(\mathbf{s})q(\mathbf{s},\eta)c(\mathbf{s},\eta,\xi)\alpha(\mathbf{s},\mathbf{s}') = p(\mathbf{s}')q(\mathbf{s}',\xi)c(\mathbf{s}',\xi,\eta)\alpha(\mathbf{s}',\mathbf{s}).$$

La probabilité d'acceptation d'un changement est :

$$\alpha(\mathbf{s},\mathbf{s}') = \min\left\{1, \frac{q(\mathbf{s}',\xi)c(\mathbf{s}',\xi,\eta)}{q(\mathbf{s},\eta)c(\mathbf{s},\eta,\xi)}\frac{p(\mathbf{s}')}{p(\mathbf{s})}\right\}.$$

Un Algorithme Metropolis-Hastings pour simuler des processus ponctuel de Gibbs avec interaction décrits par une densité de probabilité p est présenté plus bas.

Algorithme MH : mise à jour de la configuration $\mathbf{s}$ sur $W \subset \Omega$

1) choisir p_b, p_d, p_c avec $p_b + p_d + p_c \leq 1$
2) avec probabilité p_b choisir naissance, avec probabilité p_d choisir mort et avec probabilité p_c choisir changement.

   "naissance" : a) générer un point aléatoire η dans W et poser $\mathbf{s}' = \mathbf{s} \cup \{\eta\}$

   b) calculer $\beta_b = \min\{1, \frac{p_d}{p_b}\frac{p(\mathbf{s}\cup\{\eta\})}{p(\mathbf{s})}\frac{\mu(W)}{n(\mathbf{s})+1}\}$

   "mort" : a) choisir un point η de $\mathbf{s}$ et poser $\mathbf{s}' = \mathbf{s} \setminus \{\eta\}$

   b) calculer $\beta_d = \min\{1, \frac{p_b}{p_d}\frac{p(\mathbf{s}\setminus\{\eta\})}{p(\mathbf{s})}\frac{n(\mathbf{s})}{\mu(W)}\}$

   "changement" : a) choisir un point η de $\mathbf{s}$, générer un point aléatoire ξ avec $c(\mathbf{s},\eta,\xi)$ et poser $\mathbf{s}' = \mathbf{s} \setminus \{\eta\} \cup \{\xi\}$

   b) calculer $\beta_c = \min\{1, \frac{p(\mathbf{s}\setminus\{\eta\}\cup\{\xi\})}{p(\mathbf{s})}\}$
3) la nouvelle configuration $\mathbf{s} = \mathbf{s}'$ est acceptée avec la probabilité correspondante (β_b, β_d ou β_c) ; sinon l'algorithme reste dans l'état $\mathbf{s}$.

Les probabilités d'acceptation présentées plus haut ont été obtenues en utilisant les mécanismes de proposition suivants. La probabilité de générer le point η est :

$$b(\mathbf{s},\eta) = \frac{\mathbb{1}\{\eta \in W\}}{\nu(W)}.$$

La probabilité de sélectionner un point η dans la configuration $\mathbf{s}$ est :

$$d(\mathbf{s},\eta) = q(\mathbf{s},\eta) = \frac{\mathbb{1}\{\eta \in \mathbf{s}\}}{n(\mathbf{s})}.$$

Lors de l'évènement "changement", un point η de la configuration $\mathbf{s}$ est remplacé

par un point ξ généré uniformément dans la boule centrée en η et de rayon connu $r_c \in \mathbb{R}_+$, comme il suit :

$$c(\mathbf{s}, \eta, \xi) = \frac{\mathbb{1}\{\xi \in B(\eta, r_c)\}}{B(\eta, r_c)}.$$

Le noyau de transition induit par ces lois de proposition et ces probabilités d'acceptation forment le noyau de transition d'une chaîne de Markov qui est géométrique ergodique, Harris récurrente, $\phi-$irréductible avec unique distribution invariante, la loi du processus ponctuel p [M. N. M. van Lieshout, 2000, Møller and Waagepetersen, 2004, Stoica, 2014].

3.2.2 Algorithme de l'échantillonneur de Gibbs

L'échantillonneur de Gibbs est un algorithme de type MCMC pour la simulation de densités de probabilité multivariées. Il s'agit d'un cas particulier de l'algorithme de Metropolis-Hastings. Les états successifs sont simulés variable par variable.

Les lois que l'on souhaite simuler sont de la forme $p(\mathbf{s}, \theta_1, \ldots, \theta_I)$ avec $\mathbf{s}$ une configuration de processus ponctuel sur Ω et $\theta = (\theta_1, \ldots, \theta_I)$ un vecteur de $I \in \mathbb{N}$ paramètres à valeur dans un espace Θ. Le passage d'un état $(\mathbf{s}^{(l)}, \theta_1^{(l)}, \ldots, \theta_I^{(l)})$ à un état $(\mathbf{s}^{(l+1)}, \theta_1^{(l+1)}, \ldots, \theta_I^{(l+1)})$ se fait par l'algorithme suivant :

Échantillonneur de Gibbs :

1) pour $i = 1, \ldots, I$
   - générer $\theta_i^{(l+1)} \sim p(\theta_i | \mathbf{s}^{(l)}, \theta_1^{(l+1)}, \ldots, \theta_{i-1}^{(l+1)}, \theta_{i+1}^{(l)}, \ldots, \theta_I^{(l)})$
2) générer $\mathbf{s}^{(l+1)} \sim p(\mathbf{s} | \theta_1^{(l+1)}, \ldots, \theta_I^{(l+1)})$
3) définir le nouvel état $(\mathbf{s}^{(l+1)}, \theta_1^{(l+1)}, \ldots, \theta_I^{(l+1)})$

Les lois conditionnelles peuvent être simples et faciles à simuler. Parfois ceci n'est pas le cas. Dans ce cas, on peut recourir à l'intégration des dynamiques MH dans un échantillonneur de Gibbs.

3.3 Analyse des performances de l'algorithme de Metropolis-Hastings pour la simulation des processus ponctuels de Gibbs

Tous les algorithmes présentés dans cette thèse ont été codés en C++. Les résultats finaux ont été affichés en utilisant le logiciel R.

Ici, nous analysons des performances de l'algorithme MH en considérant la simulation de deux modèles, Poisson et Strauss. Le premier modèle est parfaitement contrôlable analytiquement, il n'a pas besoin d'un échantillonneur MH pour être simulé. Cependant, il peut être utilisé pour bien choisir les différents paramètres de l'algorithme. Ensuite, le modèle de Strauss est considéré. Ce modèle nécessite la simulation à travers une méthode MCMC. L'algorithme MH réglé sur le modèle de Poisson est appliqué sur le modèle de Strauss. Nous étudions la cohérence des résultats obtenus.

L'objectif est avant tout de bien maîtriser la dynamique de simulation sur des modèles "simples" avant de l'appliquer sur des modèles plus complexes comme celui que l'on proposera pour la détection et la caractérisation des sources.

Les paramètres de notre Algorithme MH sont montrés dans le Tableau 3.1. Les valeurs de référence issues à la fin de l'étude y sont également présentées.

Variable	Description	Value
W	domaine d'étude	$[0,1] \times [0,1]$
N	nombre de réalisations	10^3
N_{MH}	nombre d'itérations de MH	200
$p_b; p_d; p_c$	probabilité de naissance ;mort ;changement	$0.2; 0.2; 0.6$
r_c	rayon de changement	0.3
r	rayon d'interaction	0.1

Tableau 3.1 – Présentation des paramètres de l'algorithme MH pour la simulation des processus de Poisson et de Strauss.

Nous établissons à partir des Exemples 3.3.1 et 3.3.2 les observations suivantes. Plus le nombre de mises à jour entre chaque état est important, moins ces états sont corrélés. Clairement, le temps de calcul augmente si N_{MH} augmente, cependant cela améliore la qualité des échantillons. Un compromis doit être réalisé. Il y a beaucoup de liberté quant au choix des valeurs p_b, p_d, p_c pour vu que $0 < p_b + p_d + p_c \leq 1$. Pour des configurations avec un nombre pas trop grand

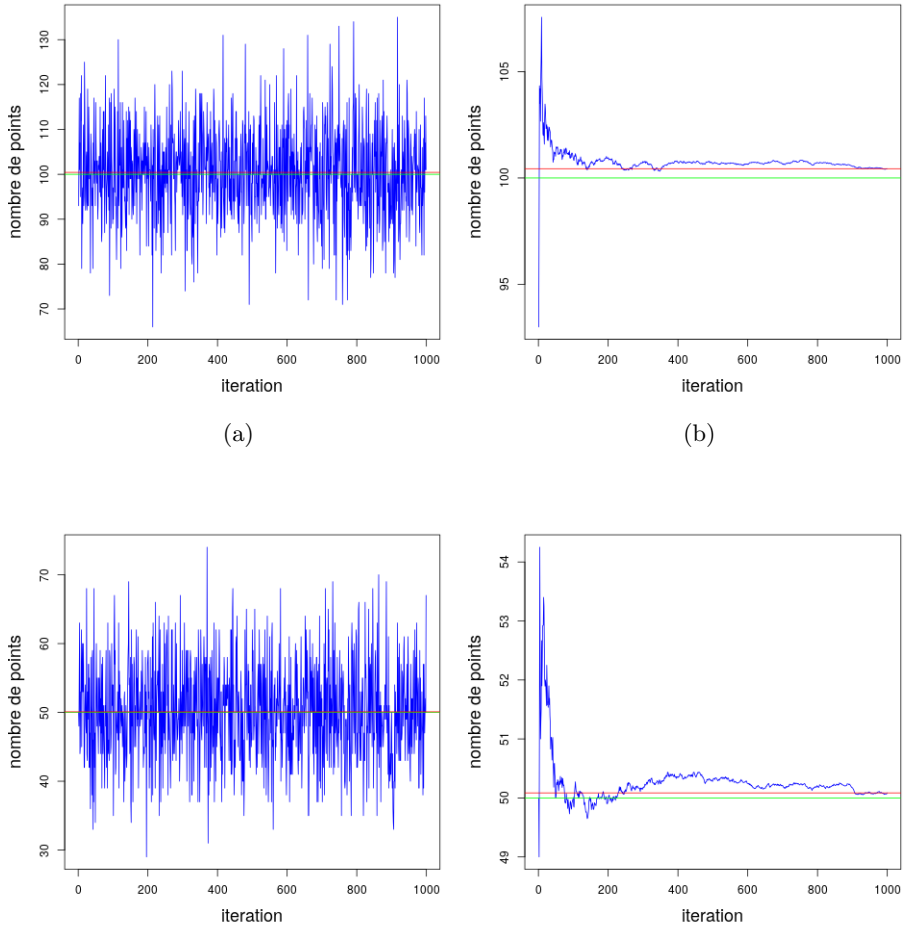
de points, les effets de la variation de ces paramètres peuvent être considérés comme pas très important, hormis une utilisation pathologique de ceux-ci. Dans des nombreuses situations, le nombre de points à l'équilibre varie "peu". Dans ce contexte, la proposition "changement" pourrait être considéré comme celle qui a le plus de chances d'être acceptée. Donc, on pourrait la renforcer par rapport aux autres propositions.

Pour ces expériences, les processus ponctuels sont définis dans le carré unité $W = [0,1] \times [0,1]$. Pour le processus de Strauss, le rayon d'interaction est $r = 0.1$. Le nombre de simulations a été fixé à $N = 1000$, et la loi $c(\mathbf{s}, \eta, \xi)$ est une loi uniforme de paramètre $r_c = 0.3$.

Exemple 3.3.1 (Processus ponctuels de Poisson (suite))**.** *Des processus ponctuels de Poisson, définis comme dans l'Exemple 2.4.1 sont simulés.*

Dans un premier temps, deux processus homogènes sont générés : l'un avec $\theta = -\log(100)$ *et l'autre avec* $\theta = -\log(50)$*. Les nombres moyens de points de ces deux processus sont respectivement* 100 *et* 50*. Pour chacun de ces processus,* $N = 1000$ *réalisations ont été obtenues en générant* 10^5 *réalisations et en gardant une configuration sur* 100*. À chaque itération, l'algorithme MH est appliqué* $N_{MH} = 200$ *fois.*

Les évolutions du nombre de points et de la moyenne cumulée du nombre de points sont représentées dans la Figure 3.1. Le nombre moyen de points est comme attendu très proche des valeurs théoriques 100 *pour le premier processus (sur les Figures (a) et (b)) et* 50 *pour le second (sur les Figures (c) et (d)). Dans chaque cas, la moyenne théorique du nombre de points est représentée en vert tandis que la moyenne empirique est représentée en rouge.*



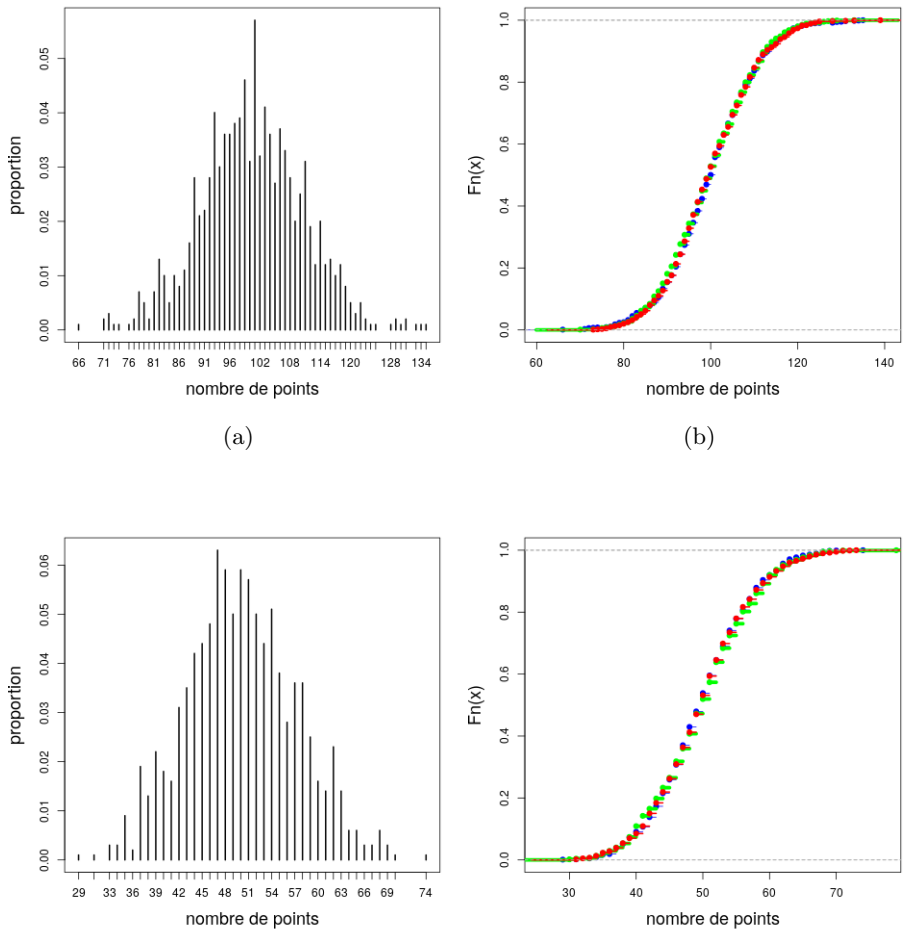
(a) (b)

(c) (d)

FIGURE 3.1 – Évolution du nombre de points et de la moyenne cumulée du nombre de points pour un processus ponctuel de Poisson de paramètre $\theta = -\log(100)$ (respectivement (a) et (b)) et pour un processus ponctuel de Poisson de paramètre $\theta = -\log(50)$ (respectivement (c) et (d)). La valeur moyenne théorique du nombre de points ($\exp(-\theta)$) est représentée en vert, la moyenne empirique en rouge.

Le nombre de points dans une réalisation d'un processus ponctuel de Poisson suit une loi de Poisson. La Figure 3.2 montre l'histogramme du nombre de points et la fonction de répartition du nombre de points pour le premier processus (les Figures (a) et (b)) et pour le second processus (les Figures (c) et (d)). Sur les Figures (b) et (d), la fonction de répartition d'une loi de Poisson de paramètre la moyenne théorique du nombre de points est représentée en vert, celle d'une loi de Poisson de paramètre la moyenne empirique du nombre de points est représentée

en rouge. Les fonctions se superposent dans chacun des cas : le nombre de points dans chaque réalisation semble bien tiré selon une loi de Poisson. Le test de Kolmogorov-Smirnov est appliqué pour vérifier si le nombre de points est bien tiré selon une loi de Poisson de paramètre 100 *pour le premier processus et* 50 *pour le second. Les p-valeurs de ces tests sont respectivement* 0.97 *et* 0.95 *: l'hypothèse est vérifiée pour un seuil* $\alpha = 0.05$.

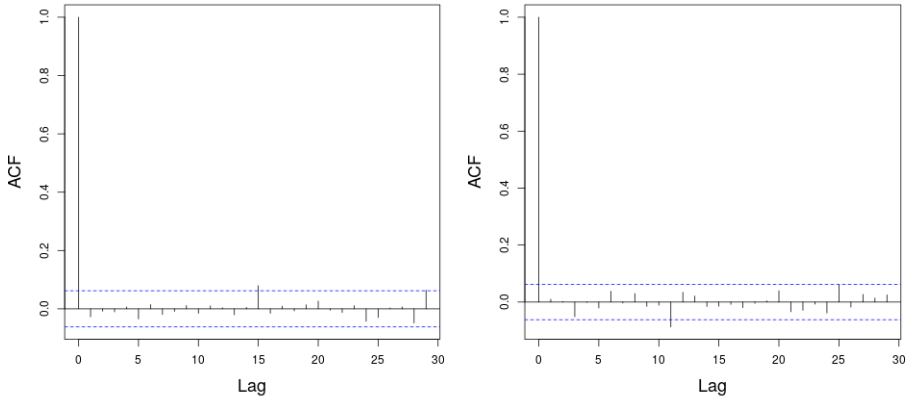


(a) (b)

(c) (d)

FIGURE 3.2 – Histogramme du nombre de points et fonction de répartition du nombre de points pour un processus ponctuel de Poisson de paramètre $\theta = -\log(100)$ (respectivement (a) et (b)) et pour un processus ponctuel de Poisson de paramètre $\theta = -\log(50)$ (respectivement (c) et (d)). La fonction de répartition d'une loi de Poisson de paramètre la moyenne théorique du nombre de points ($\exp(-\theta)$) est représentée en vert, celle d'une loi de Poisson de paramètre la moyenne empirique en rouge. La p-valeur du test de Kolmogorov-Smirnov pour comparer la distribution des réalisations avec une loi de Poisson est 0.97 (c) et 0.95 (d) : dans chaque cas, l'hypothèse est acceptée pour un seuil $\alpha = 0.05$.

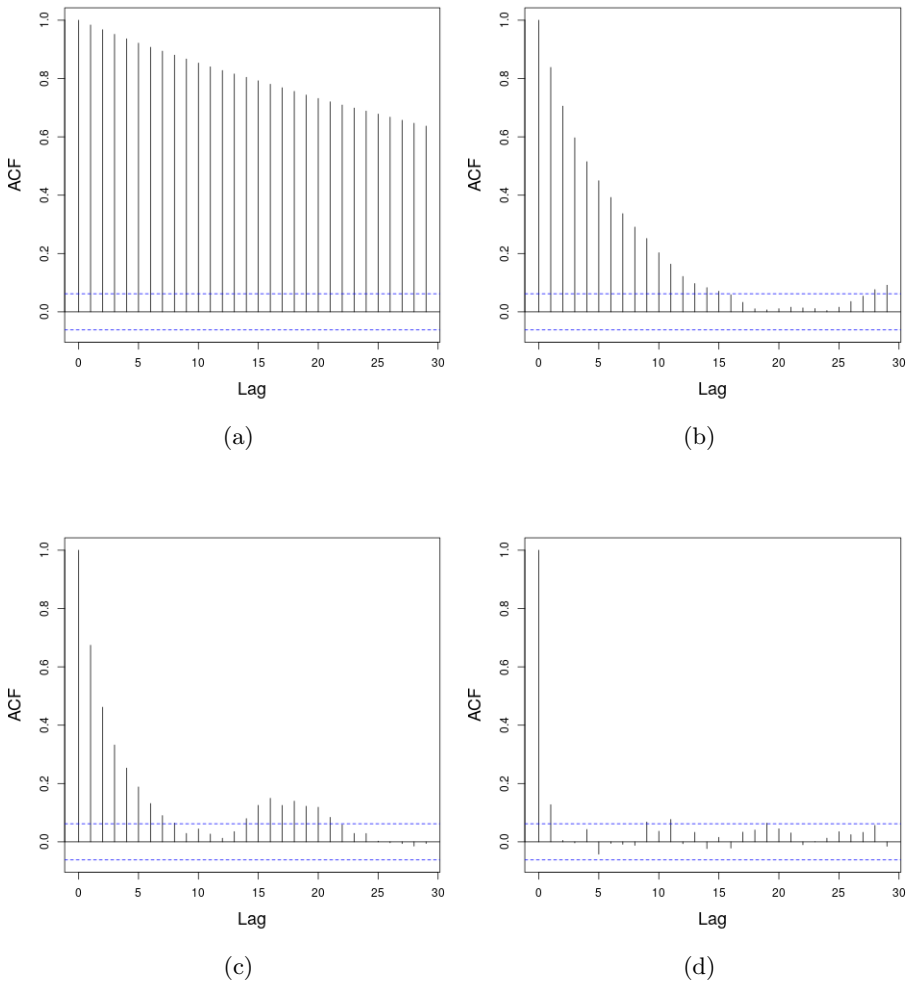
La Figure 3.3 montre que les réalisations successives sont peu corrélés les unes aux autres.



(a) (b)

FIGURE 3.3 – Fonction d'autocorrélation des processus de Poisson de paramètre $\theta = -\log(100)$ (a) et $\theta = -\log(50)$.

Les effets du choix de N_{MH} sont testés sur des processus ponctuels de Poisson de paramètres $\theta = -\log(100)$: un avec $N_{MH} = 10$, un autre avec $N_{MH} = 100$, un autre avec $N_{MH} = 200$ et un dernier avec $N_{MH} = 1000$. Les fonctions d'autocorrélation de ces processus, représentées sur la Figure 3.4, montrent que plus N_{MH} est grand et moins les réalisations sont corrélées entre elles. L'espacement des échantillons joue également un rôle, que l'on peut considérer intégré dans N_{MH} sans perte de généralité.



(a) (b)

(c) (d)

FIGURE 3.4 – Fonction d'autocorrélation du nombre de points d'un processus ponctuels de Poisson de paramètre $\theta = -\log(100)$ simulé par une dynamique MH avec $N_{MH} = 10$ (a), $N_{MH} = 100$ (b), $N_{MH} = 200$ (c) et $N_{MH} = 1000$ (d).

L'effet du choix des probabilités p_b, p_d et p_c est visualisé au travers de quatre processus. Dans le premier cas, l'évènement prépondérant est l'évènement "mort", les probabilités sont $p_b = 0.1$, $p_d = 0.8$ et $p_c = 0.1$. Dans le deuxième cas, l'évènement prépondérant est l'évènement "naissance", les probabilités sont $p_b = 0.8$, $p_d = 0.1$ et $p_c = 0.1$. Dans le troisième cas, l'évènement prépondérant est l'évènement "changement", les probabilités sont $p_b = 0.1$, $p_d = 0.1$ et $p_c = 0.8$. Dans le dernier cas, la probabilité de chacun des évènements est presque égale, les probabilités sont $p_b = 0.3$, $p_d = 0.3$ et $p_c = 0.4$. L'évolution du nombre de points est représentée dans la Figure 3.5, l'évolution de la moyenne cumulée dans la

Figure 3.6, tandis que la boite à moustaches du nombre de points est représentée dans la Figure 3.7.

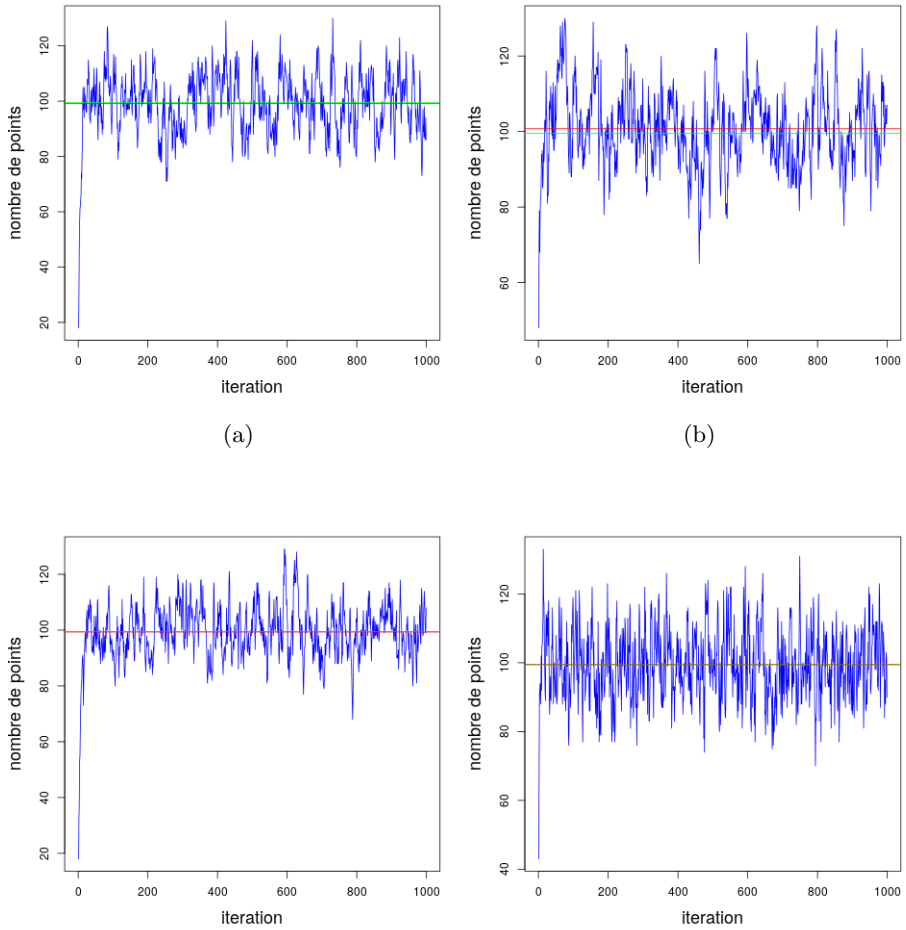
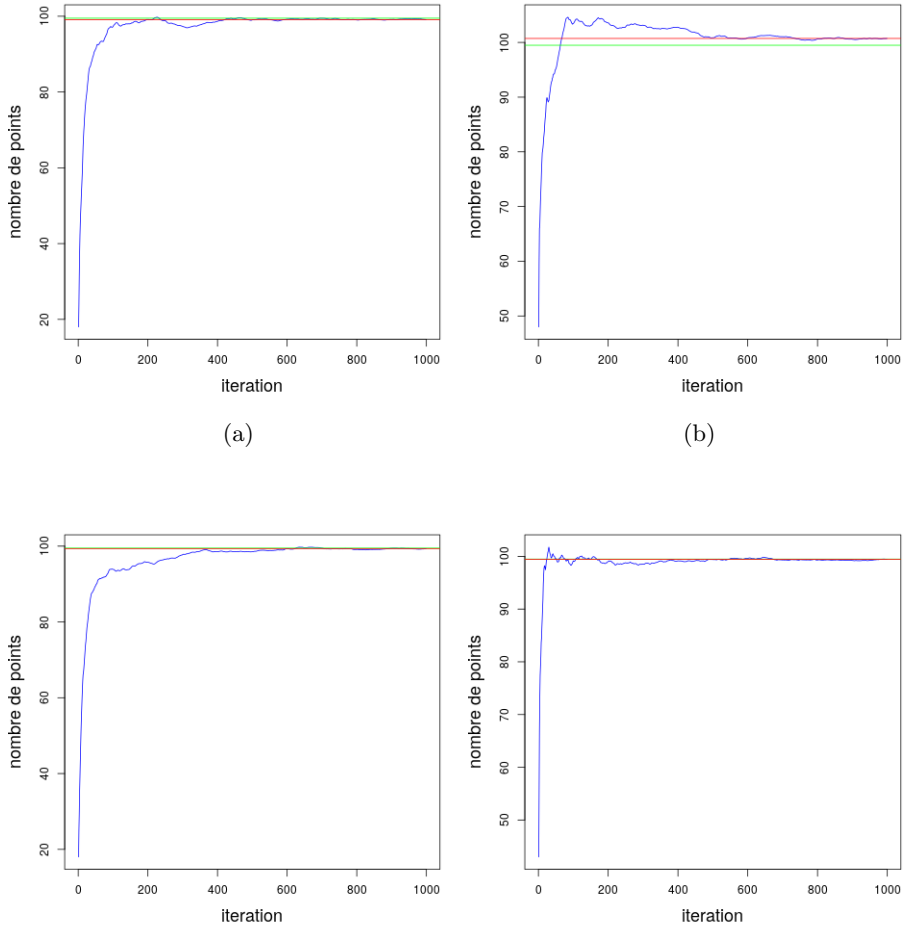


FIGURE 3.5 – Évolution du nombre de points pour un processus ponctuel de Poisson de paramètre $\theta = -\log(100)$ lorsque l'évènement "mort" est prépondérant (a), lorsque l'évènement "naissance" est prépondérant (b), lorsque l'évènement "mort" est prépondérant (c), lorsque l'évènement "changement" est prépondérant (d). La moyenne théorique du nombre de points ($\exp(-\theta)$) est représentée en vert, la moyenne empirique du nombre de points en rouge.



(a) (b)

(c) (d)

FIGURE 3.6 – Évolution de la moyenne cumulée du nombre de points pour un processus ponctuel de Poisson de paramètre $\theta = -\log(100)$ lorsque l'évènement "mort" est prépondérant (a), lorsque l'évènement "naissance" est prépondérant (b), lorsque l'évènement "mort" est prépondérant (c), lorsque l'évènement "changement" est prépondérant (d). La moyenne théorique du nombre de points ($\exp(-\theta)$) est représentée en vert, la moyenne empirique du nombre de points en rouge.

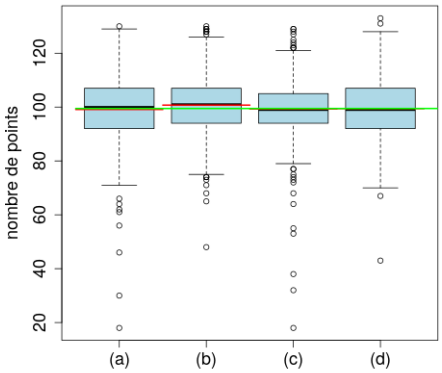
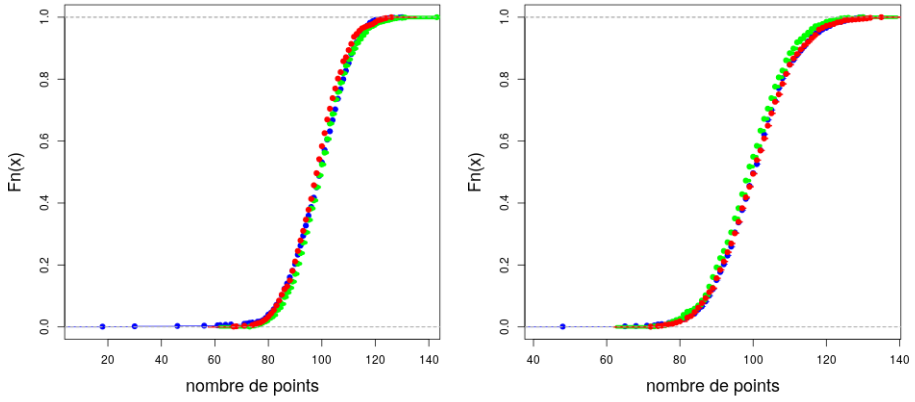


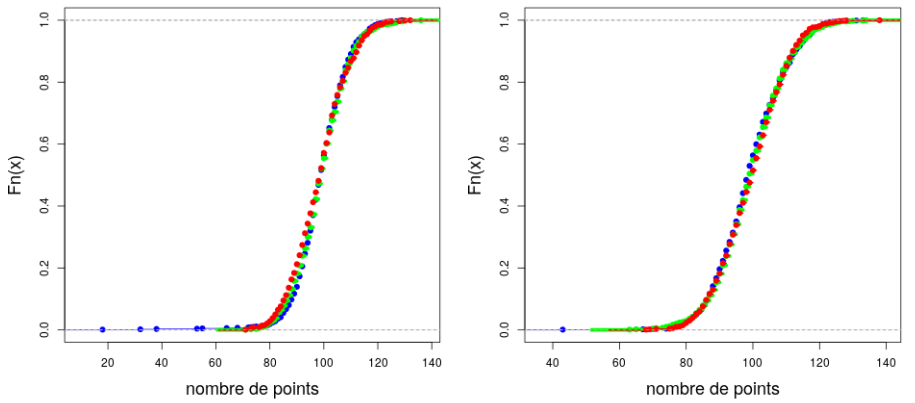
FIGURE 3.7 – Boite à moustaches du nombre de points pour un processus ponctuel de Poisson de paramètre $\theta = -\log(100)$ lorsque l'évènement "mort" est prépondérant (a), lorsque l'évènement "naissance" est prépondérant (b), lorsque l'évènement "mort" est prépondérant (c), lorsque l'évènement "changement" est prépondérant (d). La moyenne théorique du nombre de points ($\exp(-\theta)$) est représentée en vert, la moyenne empirique du nombre de points en rouge.

D'après la Figure 3.8, les nombres de points pour toutes les simulations semblent approcher la loi de Poisson, comme attendu. Le test de Kolmogorov-Smirnov est appliqué pour vérifier si le nombre de points est bien tiré selon une loi de Poisson de paramètre 100. *Les p-valeurs de ces tests sont respectivement* 0.89 *(a),* 0.1 *(b),* 0.16 *(c) et* 0.79 *(d) : dans chaque cas, l'hypothèse est vérifiée pour un seuil* $\alpha = 0.05$.



(a)

(b)



(c)

(d)

FIGURE 3.8 – Fonction de répartition du nombre de points pour un processus ponctuel de Poisson de paramètre $\theta = -\log(100)$ lorsque l'évènement "mort" est prépondérant (a), lorsque l'évènement "naissance" est prépondérant (b), lorsque l'évènement "mort" est prépondérant (c), lorsque l'évènement "changement" est prépondérant (d). La fonction de répartition d'une loi de Poisson de paramètre la moyenne théorique du nombre de points ($\exp(-\theta)$) est représentée en vert, celle d'une loi de Poisson de paramètre la moyenne empirique en rouge. La p-valeur du test de Kolmogorov-Smirnov pour comparer la distribution des réalisations avec une loi de Poisson est 0.89 (a), 0.1 (b), 0.16 (c) et 0.79 (d) : l'hypothèse est donc acceptée pour un seuil $\alpha = 0.05$.

Le choix des probabilités des évènements a un effet sur la vitesse de convergence. D'après ces expériences sur ces modèles, il faut garder un certain équilibre entre p_b et p_d. Dans certaines situations, le nombre de points varie peu quand l'on est

proche du régime d'équilibre. Dans ce cas, il serait préférable d'avoir l'évènement "changement" prépondérant.

Exemple 3.3.2 (Processus ponctuels de Strauss (suite))**.** *Des processus ponctuels de Strauss, définis comme dans l'Exemple 2.4.2 sont simulés.*

L'intérêt de simuler des processus ponctuels de Strauss est de voir si l'algorithme MH prend bien en compte la fonction d'interaction. Les processus ponctuels de Strauss pénalisent en effet les configurations de points en fonction du nombre de paires de points voisins.

Quatre processus de Strauss sont générés. Pour tous les processus, le premier paramètre est $\theta_1 = -\log(100)$*, ensuite les choix pour le paramètre* θ_2 *sont* $0, 0.5, 1$ *et* 100*, respectivement. Pour chacun de ces processus, l'algorithme MH est initialisé selon le Tableau 3.1.*

L'évolution du nombre de points au cours des simulations est représentée par la Figure 3.9, l'évolution de la moyenne cumulée du nombre de points par la Figure 3.10. L'évolution du nombre d'interactions au cours des simulations est représentée par la Figure 3.11, l'évolution de la moyenne cumulée du nombre d'interactions[1] *par la Figure 3.12. Dans ces Figures, le premier processus est représenté sur (a), le deuxième sur (b), le troisième sur (c) et le dernier sur (d).*

1. Une interaction signifie une paire de points ayant une distance entre eux de moins de r.

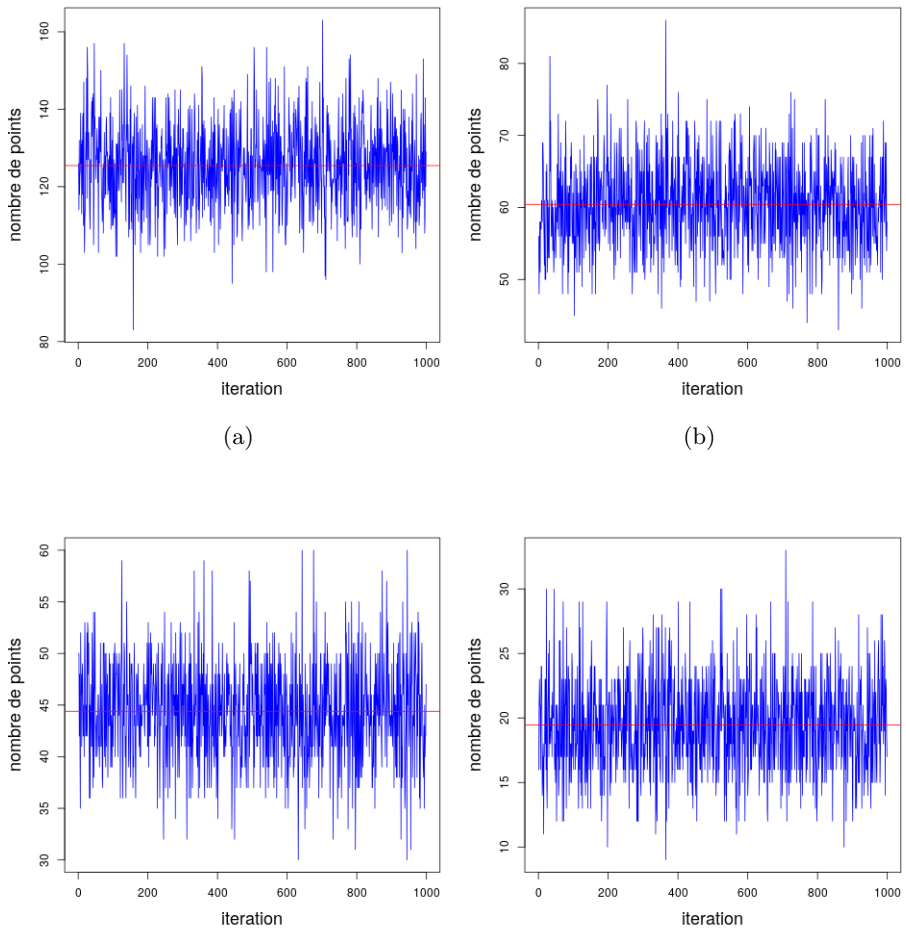


FIGURE 3.9 – Évolution du nombre de points pour des processus ponctuels de Strauss de paramètres $\theta_1 = -\log(100)$ et $\theta_2 = 0$ (a), $\theta_1 = -\log(100)$ et $\theta_2 = 0.5$ (b), $\theta_1 = -\log(100)$ et $\theta_2 = 1$ (c), $\theta_1 = -\log(100)$ et $\theta_2 = 100$ (d). La moyenne empirique du nombre de points est représentée en rouge.

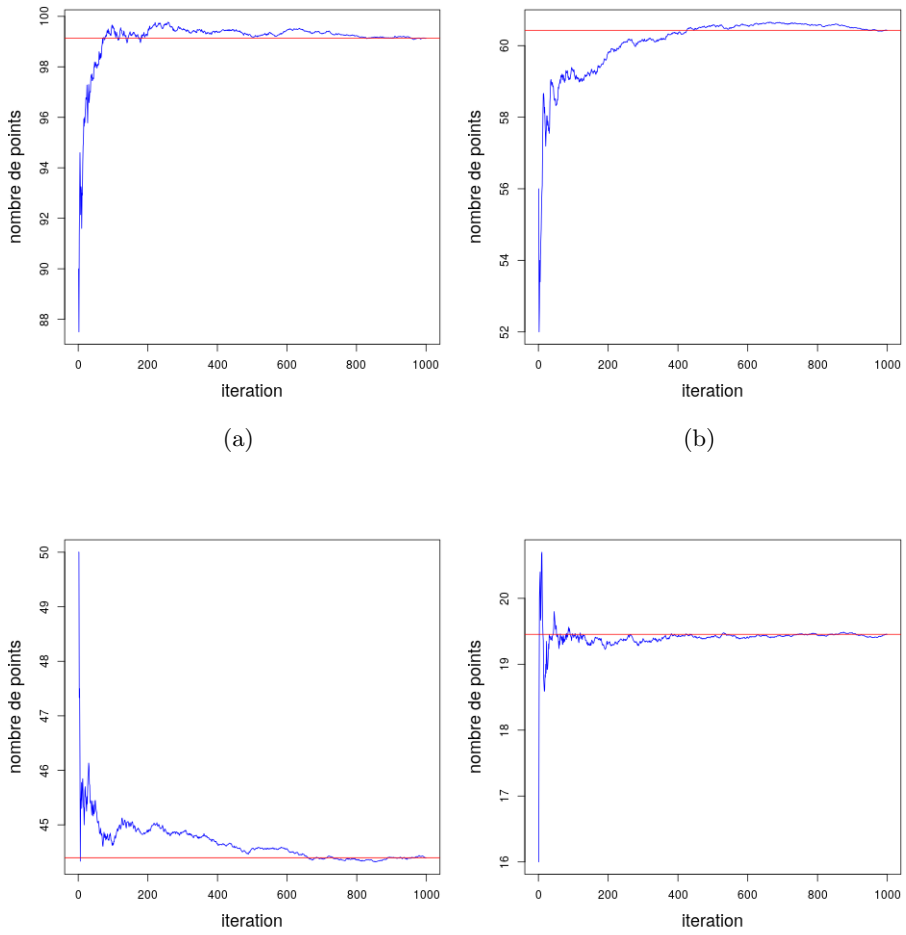


FIGURE 3.10 – Évolution de la moyenne cumulée du nombre de points pour des processus ponctuels de Strauss de paramètres $\theta_1 = -\log(100)$ et $\theta_2 = 0$ (a), $\theta_2 = 0.5$ (b), $\theta_2 = 1$ (c), $\theta_2 = 100$ (d). La moyenne empirique du nombre de points est représentée en rouge.

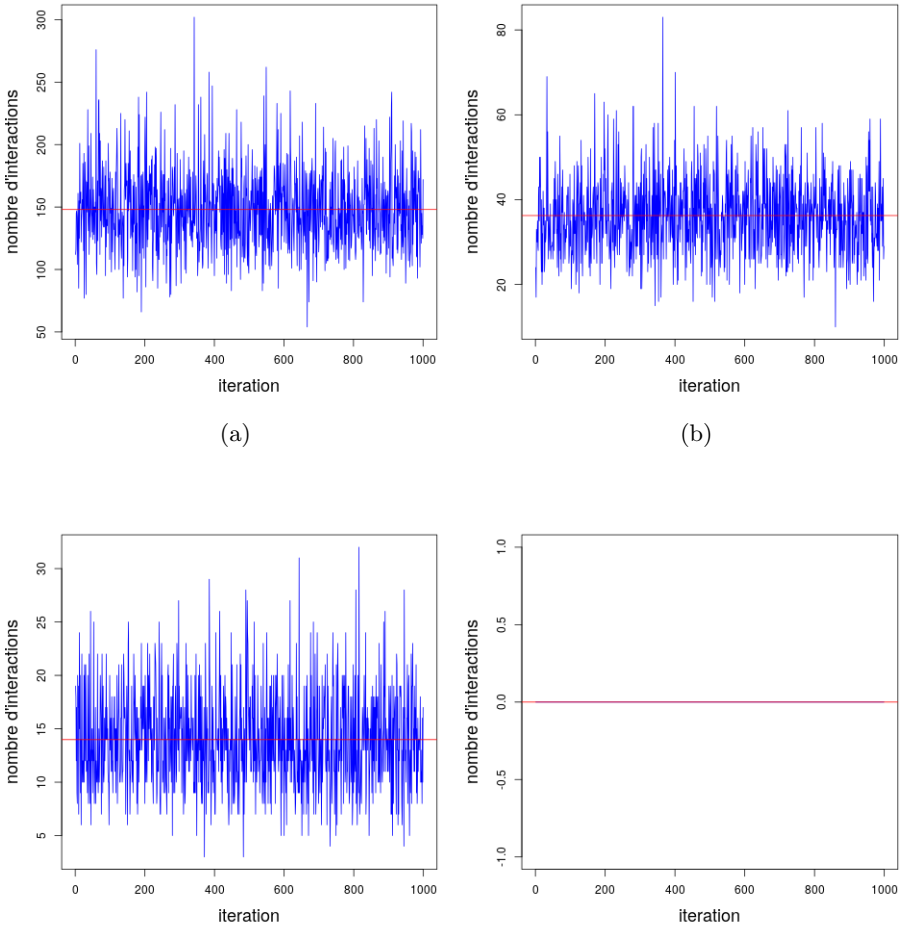


FIGURE 3.11 – Évolution du nombre d'interactions pour des processus ponctuels de Strauss de paramètres $\theta_1 = -\log(100)$ et $\theta_2 = 0$ (a), $\theta_2 = 0.5$ (b), $\theta_2 = 1$ (c), $\theta_2 = 100$ (d). La moyenne empirique du nombre d'interactions est représentée en rouge.

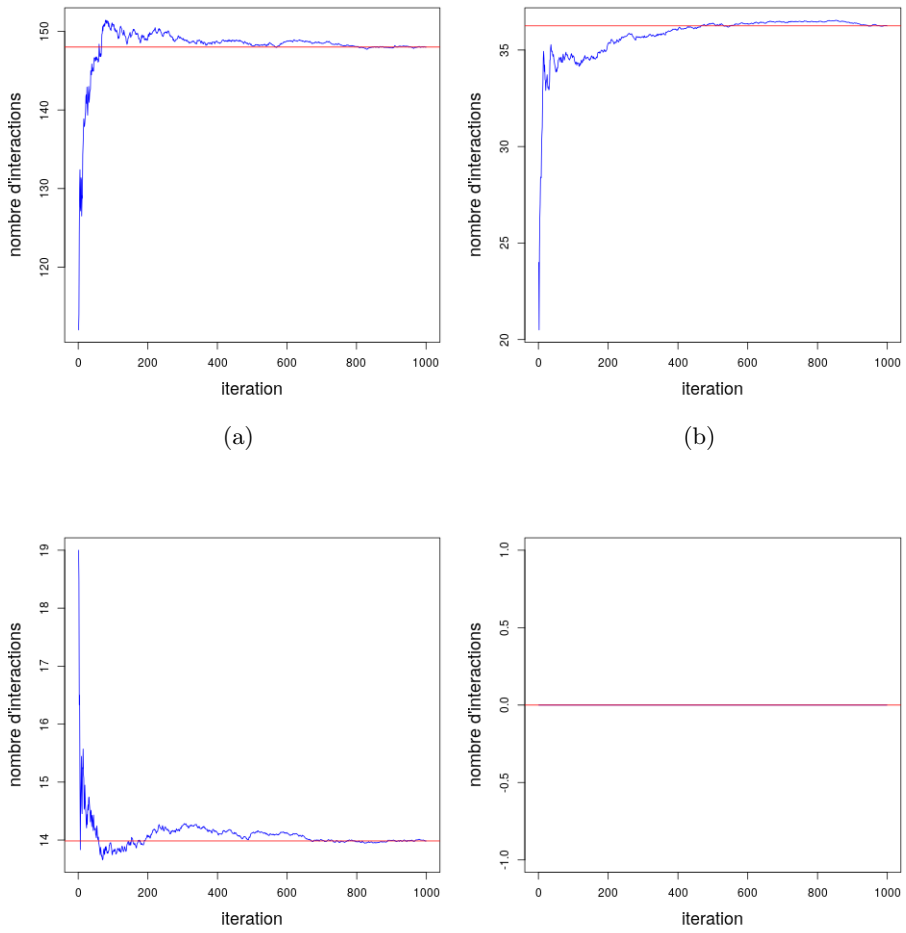
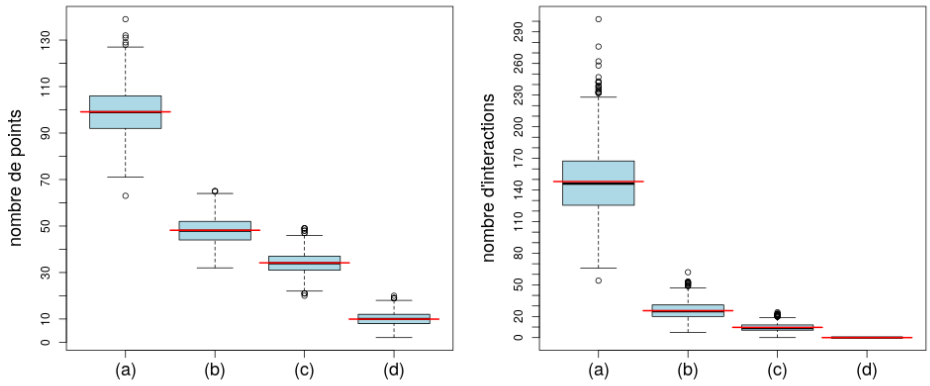


FIGURE 3.12 – Évolution de la moyenne cumulée du nombre d'interactions pour des processus ponctuels de Strauss de paramètres $\theta_1 = -\log(100)$ et $\theta_2 = 0$ (a), $\theta_2 = 0.5$ (b), $\theta_2 = 1$ (c), $\theta_2 = 100$ (d). La moyenne empirique du nombre d'interactions est représentée en rouge.

Comme attendu, le nombre de points et le nombre d'interactions diminue lorsque θ_2 *augmente.*

La Figure 3.13 représente la boîte à moustaches du nombre de points à gauche et du nombre d'interactions à droite pour les processus de Strauss simulés. Sur chacune de ces Figures, (a), (b), (c) et (d) représentent respectivement le premier processus, le deuxième, le troisième et le dernier.



(a) (b)

FIGURE 3.13 – Boîtes à moustaches du nombre de points à gauche et boîtes à moustaches du nombre d'interactions à droite pour les processus de Strauss simulés. Sur chaque Figure, (a) représente le processus ponctuel de Strauss de paramètres $\theta_2 = 0$, (b) $\theta_2 = 0.5$, (c) $\theta_2 = 1$ et (d) $\theta_2 = 100$.

Tous ces résultats montrent un comportement de Strauss attendu. Pour $\theta_2 = 0$*, nous avons un processus de Poisson. Plus* θ_2 *augmente, plus le nombre de paires de points à une distance plus petite que* r *est réduit. Quand* θ_2 *est très grand, cela signifie que les interactions entre points sont presque interdites, car fortement pénalisées, alors la distribution de la statistique suffisante de cette interaction tend à se concentrer vers* 0*.*

Chapitre 4

Inférence statistique

Ce chapitre présente quelques outils d'inférence pour les processus ponctuels de Gibbs. Ces outils seront appliqués à la caractérisation et à la détection des sources via le nouveau modèle que l'on propose : le modèle de HUG.

4.1 Maximisation d'une densité de probabilité : algorithme de recuit simulé

Ici, l'on souhaite maximiser la densité de probabilité d'un processus ponctuel de Gibbs $p(\mathbf{s})$, c'est-à-dire résoudre le problème suivant :

$$\widehat{\mathbf{s}} = \arg\max_{\mathbf{s}\in\Omega} p(\mathbf{s}) = \arg\max_{\mathbf{s}\in\Omega} \frac{\exp\left(-U(\mathbf{s})\right)}{Z} = \arg\min_{\mathbf{s}\in\Omega} U(\mathbf{s}) \tag{4.1}$$

avec Z la constante de normalisation et U la fonction énergie.

La densité de probabilité n'est pas garantie convexe. Cela demande donc une méthode d'optimisation globale. De cette manière, au moins du point de vue théorique, la solution proposée est optimale, car indépendante des conditions initiales.

Le recuit simulé est une méthode d'optimisation globale qui simule d'une manière itérative la loi

$$p_n(\mathbf{s}) \propto p(\mathbf{s})^{\frac{1}{T_n}}, \quad n \geq 0$$

avec T_n un paramètre appelé température. Si T_0 est suffisamment grand et si

$T_n \to 0$ lentement, alors cet algorithme converge en distribution vers la loi uniforme sur le sous-espace de configurations qui maximisent $p(\mathbf{s})$ [Geman and Geman, 1984, van Lieshout, 1994, Stoica et al., 2005].

L'implémentation d'un tel algorithme a besoin d'au moins deux ingrédients : une dynamique de simulation et un schéma de refroidissement. Pour les processus ponctuels, il est possible d'utiliser la dynamique de type Metropolis-Hastings. Quant au schéma de refroidissement [Stoica et al., 2005] prouvent l'existence d'un schéma logarithmique pour des processus ponctuels marqués avec des interactions multiples dont une de type hard-core simulés avec une dynamique de type Metropolis-Hastings. C'est un résultat qui est comparable avec celui obtenu pour les champs de Markov simulés avec un échantillonneur de Gibbs obtenu par [Geman and Geman, 1984]. Un autre résultat de convergence pour le recuit simulé appliqué aux processus ponctuels est présenté dans [van Lieshout, 1994], avec une dynamique de simulation basée sur les processus de type naissance et mort.

En théorie, un schéma de refroidissement doit être trouvé pour chaque dynamique de simulation utilisée. En pratique, les schémas de refroidissement les plus utilisés sont le schéma logarithmique

$$T_i = \frac{T_0}{\log(1+i)},$$

avec T_0 la température initiale et le schéma polynomial

$$T_{i+1} = cT_i \tag{4.2}$$

avec $0 < c < 1$. Le premier schéma est proche de ce que l'on trouve souvent prouvé en littérature, il est suffisamment lent, mais le paramètre T_0 doit être souvent fixé d'une manière empirique. Le deuxième schéma est plus rapide, le risque de se trouver bloquer dans un minimum local est plus grand, mais il permet un aperçu rapide du comportement du modèle autour des températures critiques.

Ainsi, l'algorithme de recuit simulé (ici pour un schéma de refroidissement polynomial) peut s'écrire de la manière suivante :

Algorithme de recuit simulé : Pour la configuration initiale $\mathbf{s}_{(0)}$, la température initiale T_0, le coefficient de refroidissement C et le nombre d'itérations K

pour $k = 1, \ldots, K$

- $T_k = \phi(T_{k-1}) = CT_{k-1}$
- générer $\mathbf{s}_{(k)} \sim p(\mathbf{s})^{1/T_k}$

retourner $\widehat{\mathbf{s}} = \mathbf{s}_{(K)}$

Si la connaissance des paramètres θ du modèle considéré peut se synthétiser sous la forme d'une loi a priori $p(\theta)$ définie sur Θ l'espace des paramètres, alors nous pouvons utiliser le recuit simulé pour résoudre le programme suivant :

$$\widehat{(\mathbf{s}, \theta)} = \underset{\mathbf{s}\in\Omega, \theta\in\Theta}{\arg\max}\, p(\mathbf{s}, \theta) = \underset{\mathbf{s}\in\Omega, \theta\in\Theta}{\arg\max}\, p(\mathbf{s}|\theta)p(\theta).$$

En plus du schéma de refroidissement adéquat, ce problème requiert d'échantillonner la loi jointe $p(\mathbf{s}, \theta)$. La stratégie adoptée dans cette thèse a été de mettre en œuvre un échantillonneur de Gibbs dont certaines lois conditionnelles sont simulées à l'aide d'une dynamique Metropolis-Hastings [Stoica et al., 2004, Stoica et al., 2007b, Stoica, 2014].

La recherche de stratégies d'implémentations optimales pour les algorithmes de type recuit simulé est un problème encore ouvert. Les méthodes dites "simulated tempering" [Møller and Nicholls, 1999, Döge et al., 2004] optent pour la simulation à basse température et tout en réduisant les effets liés aux corrélations entre les réalisations. Différents schémas de refroidissement sont comparés dans [Winkler, 2003]. D'autres schémas de refroidissement plus rapide sont utilisés dans les méthodes "fast simulated annealing" [Szu and Hartley, 1987].

4.2 Outils pour l'analyse des résultats

Les solutions fournies par l'utilisation d'un modèle de détection de structures via le recuit simulée ne sont pas uniques [Heinrich et al., 2012, Stoica, 2014].

Les sorties du recuit simulé convergent en distribution vers la loi uniforme sur le sous-espace des configurations maximisant la loi de probabilité considérée. En plus, l'implémentation de l'algorithme d'optimisation doit s'écarter des condi-

tions d'application théorique, par nécessité. Ainsi, l'on a besoin de "moyenner" les résultats obtenus pour compenser toutes ces incertitudes. Les ensembles de niveau sont un des outils qui permettent atteindre cet objectif.

Soit S le processus ponctuel des sources. Soit $W = \bigcup_{i=1}^{m} \tilde{w}_i$ une décomposition du domaine W en un nombre fini de cellules m telles que $\nu(\tilde{w}_i) = \text{ct.}$, pour tout i. Appelons $\tilde{W}$ l'ensemble de toutes ces cellules.

La probabilité de contact entre le processus des sources et une cellule de la grille est donnée par

$$p(w) = \mathbb{P}(w \cap S \neq \emptyset).$$

Pour $\alpha \in [0, 1]$, considérons l'ensemble de niveau α donné par

$$Q_\alpha = \{w \in \tilde{W} \ : p(w) > \alpha\}.$$

En pratique, le calcul de $p(w)$ ne peut se faire que par des méthodes de type Monte-Carlo, en considérant l'estimateur

$$p_n(w) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{w \cap S_i \neq \emptyset\}}.$$

avec $S_1, S_2, \ldots, S_n$ des réalisations i.i.d de $p(\mathbf{s}|\theta)$.

Alors l'estimateur de l'ensemble de niveau α sera

$$Q_{n,\alpha} = \{w \in \tilde{W} \ : p_n(w) > \alpha\}. \tag{4.3}$$

Les estimateurs de la forme (4.3) sont très intéressants pour notre problème puisqu'ils permettent de rendre le résultat de la méthode que l'on propose plus robuste. En effet, l'estimation par ensembles des niveaux permet de réduire les incertitudes relatives aux données et "moyenner" les effets aléatoires du recuit simulé. Les propriétés de convergence de l'estimateur des ensembles par niveaux ont été étudiées dans [Heinrich et al., 2012]. Cette convergence dépend des deux facteurs, la simulation du modèle nécessaire pour évaluer les probabilités et la décomposition en cellules du domaine. La convergence obtenue est de type L^1.

4.2.1 Reconstruction de la configuration de sources à partir de ses projections

Une autre méthode naturelle pour reconstruire la configuration “moyenne” des sources est d'appliquer un algorithme d'agrégation sur les configurations obtenues suite au recuit simulé. Chaque centre de classe représenterait la position “moyenne” d'une source. Le problème est que du fait du comportement stochastique de l'algorithme de recuit simulé et du fait que la simulation se fasse à partir des projections, le nombre de sources n'est pas connu.

Pour résoudre cet inconvénient, nous proposons un *algorithme d'agrégation de type k-moyennes séquentiel*. Cet algorithme utilise en entrée plusieurs configurations de sources obtenues toutes à la même température d'équilibre pre-fixée.

Le principe de cet algorithme est le suivant. Pour chaque plan de projection, le nombre de classes peut être détecté en s'aidant de l'estimation des ensembles de niveaux des projections des configurations des sources considérées. Le choix du niveau α doit se faire afin de mettre en évidence les régions ayant une forte probabilité de contenir une source. Ainsi, pour ce plan de projection, c'est le nombre de ces régions qui sera le nombre de classes à considérer. Une procédure k-moyennes est lancée pour agréger les sources dans ce plan, en utilisant leurs projections et le nombre de classes obtenu auparavant. Les coordonnées des sources dans ce plan sont remplacées par les coordonnées du centre de la classe à laquelle elles ont été attribuées. Cette procédure est reprise pour chaque plan de projection. Comme l'algorithme de k-moyennes converge dans chaque plan de projection, alors une fois que tous les plans de projections ont été parcourues, le résultat final indique la configuration des sources moyennes.

L'algorithme de k-moyennes séquentiel s'écrit :

Algorithme de k-moyennes séquentiel : Données initiales : le nombre de classes k_l pour chaque plan de projection l et un ensemble de configurations de sources $\mathbf{s}_1, \ldots, \mathbf{s}_e$.

1. pour chaque plan de projection $l = 1, \ldots, L$
   - calculer les centres des classes $\mathbf{x}_k^l$
   - pour chaque source, remplacer les coordonnées dans le plan de projections, par les coordonnées du centre de la classe à laquelle elle appartient
2. retourner le nombre et les centres de classes obtenues

Finalement, le résultat de l'algorithme k-moyennes séquentiel est utilisé pour mettre en place un algorithme de classification dans tout l'espace des données. L'algorithme k-moyenne est sensible aux conditions initiales, donc pour l'algorithme séquentiel, l'ordre dans laquelle les plans de projections sont parcourus compte. Une analyse des performances de cette procédure et des différentes stratégies d'implémentation est faite dans le Chapitre 6.

4.3 Estimation des paramètres

Soit $\mathbf{s}$ une réalisation d'un processus ponctuel marqué que l'on suppose gouverné par un modèle de la famille exponentielle

$$p(\mathbf{s}|\theta) = \frac{\exp\left[-U(\mathbf{s}|\theta)\right]}{Z(\theta)} = \frac{\exp\left(\langle t(\mathbf{s}), \theta\rangle\right)}{Z(\theta)} \tag{4.4}$$

avec le vecteur de paramètres $\theta = (\theta_1, \ldots, \theta_I) \in \Theta \subseteq \mathbb{R}^I$, le vecteur de statistiques suffisante $t(\mathbf{s}) = (t_1(\mathbf{s}), \ldots, t_I(\mathbf{s})) \in \mathbb{R}^I$ et $Z(\theta)$ la constante de normalisation.

La fonction énergie du processus est définie par le produit scalaire :

$$\langle t(\mathbf{s}), \theta\rangle = \sum_{i=1}^{I} \theta_i t_i(\mathbf{s}). \tag{4.5}$$

Le vecteur de paramètres est estimé par

$$\widehat{\theta} = \arg\max_{\theta\in\Theta} p(\mathbf{s}|\theta). \tag{4.6}$$

Plusieurs approches pour résoudre le problème (4.6) existent dans la littérature. Les monographies [van Lieshout and van Zwet, 2001, Møller and Waagepetersen, 2004] et les références y faisant partie proposent une présentation complète du problème. L'estimation la plus efficace du point de vue numérique et du temps de calcul se fait en utilisant les techniques de type pseudo-vraisemblance. L'approche du calcul par maximum de vraisemblance en utilisant des méthodes de type Monte-Carlo est plus intéressant du point de vue mathématique étant donné les garanties théoriques quant à la qualité des solutions obtenues. Le prix à payer est, bien sûr, le coût calculatoire.

Des méthodes récentes de calcul approché bayésien [Stoica et al., 2017, Stoica et al., 2021] permettent de faire une inférence équivalente au maximum de vraisemblance du point de vue qualitatif, l'enrichir par l'échantillonnage de la loi *a posteriori*, sans augmenter le coût calculatoire et être plus stable du point de vue numérique. Ainsi, dans ce qui suit, nous allons présenter d'abord les principes de l'estimation du maximum de vraisemblance par méthodes Monte-Carlo, puis, des techniques récentes de calcul approché bayésien.

4.3.1 Estimation du maximum de vraisemblance

Dans les mêmes conditions que présentées précédemment, soit la fonction de la (log-)vraisemblance

$$\begin{aligned} L(\mathbf{s},\cdot) : \Theta &\to \mathbb{R} \\ \theta &\longmapsto \log(p(\mathbf{s}|\theta)) = U(\mathbf{s}|\theta) - \log(Z(\theta)) = \langle t(\mathbf{s}), \theta\rangle - \log(Z(\theta)). \end{aligned}$$

La fonction de vraisemblance peut s'écrire en introduisant un paramètre connu $\psi \in \Theta$:

$$l(\theta) = L(\mathbf{x},\theta) - L(\mathbf{x},\psi) = \langle t(\mathbf{x}), \theta - \psi\rangle - \log\left(\frac{Z(\theta)}{Z(\psi)}\right). \tag{4.7}$$

L'évaluation de la fonction de vraisemblance demande le calcul du ratio des

constantes de normalisation, comme il suit :

$$\begin{aligned}\frac{Z(\theta)}{Z(\psi)} &= \frac{1}{Z(\psi)}\int_W p(\mathbf{s}|\theta)d\mu(\mathbf{s}) \\ &= \int_W \frac{p(\mathbf{s}|\theta)}{p(\mathbf{s}|\psi)}\frac{p(\mathbf{s}|\psi)}{Z(\psi)}d\mu(\mathbf{s}) \\ &= \mathbb{E}\left[\frac{p(X|\theta)}{p(X|\psi)}\right] = \mathbb{E}[\exp\langle t(X), \theta - \psi\rangle]. \end{aligned} \tag{4.8}$$

avec la dernière espérance calculée par rapport à ψ.

Le calcul de cette intégrale ne peut pas se faire tout le temps d'une manière analytique. En revanche, il est possible de l'approcher par des techniques de type Monte-Carlo. Des réalisations $\mathbf{s}_1, \ldots, \mathbf{s}_n$ du processus de Gibbs de densité $p(\mathbf{s}|\psi)$ sont générées par une chaîne de Markov. L'équation (4.8) est approchée par

$$\frac{Z(\theta)}{Z(\psi)} \simeq \frac{1}{n}\sum_{i=1}^{n}\exp\langle t(\mathbf{s}_i), \theta - \psi\rangle. \tag{4.9}$$

La log-vraisemblance (4.7) est alors approchée par

$$l_n(\theta) = \langle t(\mathbf{s}), \theta - \psi\rangle - \log\left(\frac{1}{n}\sum_{i=1}^{n}\exp\langle t(\mathbf{s}_i), \theta - \psi\rangle\right). \tag{4.10}$$

Sous des conditions assez souples concernant la chaîne de Markov utilisée pour échantillonner le modèle considéré, $\frac{1}{n}\sum_{i=1}^{n}\exp\langle t(X_i), \theta - \psi\rangle$ converge presque sûrement vers $\frac{Z(\theta)}{Z(\psi)}$ [van Lieshout and van Zwet, 2001,Møller and Waagepetersen, 2004]. Par conséquent, $l_n(\theta)$ converge presque sûrement vers $l(\theta)$. Ainsi, en notant $\widehat{\theta}$ le maximum de $l(\theta)$ et $\widehat{\theta_n}$ le maximum de $l_n(\theta)$, alors $\widehat{\theta_n}$ converge presque sûrement vers $\widehat{\theta}$. La fonction de log-vraisemblance est convexe [Monfort, 1997]. Cela suggère, pour la recherche du maximum de cette fonction, une méthode d'optimisation locale.

Proposition 4.3.1. *Pour tout $\mathbf{s} \in \Omega$, la fonction de log-vraisemblance est deux fois différentiable. Le gradient est donné par :*

$$\nabla l(\theta) = t(\mathbf{s}) - \mathbb{E}[t(X)]. \tag{4.11}$$

La matrice hessienne est donnée par :

$$\nabla^2 l(\theta) = -\mathbb{E}[t(X)t(X)^\top] + \mathbb{E}[t(X)]\mathbb{E}[t(X)] \tag{4.12}$$

Dans le cas de la log-vraisemblance approchée, son gradient et sa matrice hessienne sont donnés par :

$$\nabla l_n(\theta) = t(\mathbf{s}) - \frac{\sum_{i=1}^n t(\mathbf{s}_i)\exp\langle t(\mathbf{s}_i), \theta - \psi\rangle}{\sum_{i=1}^n \exp\langle t(\mathbf{s}_i), \theta - \psi\rangle} = t(\mathbf{s}) - \mathbb{E}[n, \theta, \psi]t(\mathbf{s})$$
$$\nabla^2 l_n(\theta) = -\mathbb{E}[n, \theta, \psi]t(\mathbf{s})t(\mathbf{s}_i)t(\mathbf{s}_i)^\top + \mathbb{E}[n, \theta, \psi]t(\mathbf{s})\mathbb{E}[n, \theta, \psi]t(\mathbf{s})^\top.$$

Pour pouvoir effectuer la recherche du maximum, nous devons supposer que notre modèle est décrit par une densité de probabilité identifiable. Une densité de probabilité de p est identifiable si pour tout $\mathbf{s}$ et pour des vecteurs de paramètres $\theta_1, \theta_2 \in \Theta$ tel que $\theta_1 \neq \theta_2$

$$p(\mathbf{s}|\theta_1) \neq p(\mathbf{s}|\theta_2).$$

Nous avons donc tous les ingrédients, pour faire une recherche de maximum par une méthode Newton-Raphson de type Monte-Carlo :

$$\theta^{(k+1)} = \theta^{(k)} - (\nabla^2 l_n(\theta^{(k)}))^{-1}\nabla l_n(\theta^{(k)}). \tag{4.13}$$

Le calcul de l'inverse du Hessien approximé par des méthodes de type Monte-Carlo peut être instable de point de vue numérique. La méthode du gradient stochastique permet d'ignorer le calcul de l'inverse de la matrice hessienne :

$$\theta^{(k+1)} = \theta^{(k)} - \epsilon_k\big(t(\mathbf{s}) - t(\mathbf{s}_k)\big). \tag{4.14}$$

Il est cependant nécessaire de trouver une séquence décroissante optimale ϵ_k et de simuler des réalisations $\mathbf{s}_k$ selon la loi $p(\mathbf{s}|\theta^{(k)})$.

Une manière de combiner les deux techniques est d'utiliser la méthode du gradient à pas optimal :

$$\theta^{(k+1)} = \theta^{(k)} - \rho(\theta^{(k)})\nabla l_n(\theta^{(k)}). \tag{4.15}$$

Le pas de descente $\rho(\theta^{(k)})$ est choisi tel que :

$$\rho(\theta^{(k)}) = \max_{\rho \in \mathbb{R}} l_n(\theta^{(k)} - \rho \nabla l_n(\theta^{(k)})) \tag{4.16}$$

La stabilité numérique de ces méthodes est garantie si ψ est proche de θ. Comme cela ne peut pas être vérifié à l'avance, une manière d'y parvenir est de contrôler la distance entre θ_k et ψ. Quand cette distance est trop grande, un réechantillonnage du modèle avec $\psi = \theta_k$ est nécessaire. Cette opération garantie la convergence de la méthode, mais le coût calculatoire échappe du contrôle [Geyer, 1999].

4.3.2 Méthodes Approximate Bayesian Computation (ABC)

Les méthodes Approximate Bayesian Computation (ABC) ont pour objectif la simulation de la loi *a posteriori* de θ :

$$p(\theta|\mathbf{s}) \propto p(\mathbf{s}|\theta)p(\theta). \tag{4.17}$$

Estimations et tests statistiques impliquant θ peuvent se construire à partir d'un n-échantillon $\theta_1, \ldots, \theta_n$ généré par $p(\theta|\mathbf{s})$.

Ce champ de travail est en plein essor. Cette dynamique est esquissée dans les références de [Stoica et al., 2017, Stoica et al., 2021].

Nous rappelons ici le principe général des méthodes ABC classiques. Afin de générer une réalisation θ_i, un candidat $\psi \in \Theta$ est généré selon la loi *a priori* $p(\psi)$. Une réalisation du modèle $\mathbf{x}$ est alors simulée selon $p(\mathbf{x}|\psi)$. La distance entre les vecteurs des statistiques observées $t(\mathbf{s})$ et le vecteur des statistiques calculées à partir de la simulation $t(\mathbf{x})$, est calculée. Si cette distance est plus petite qu'une valeur seuil ϵ, alors le candidat ψ est accepté comme valeur du paramètre.

Cet algorithme est illustré plus bas :

Algorithme ABC classique : Fixer le seuil $\epsilon \in \mathbb{R}_+$.

1) faire
2) générer ψ selon la loi a priori $p(\psi)$
2) simuler $\mathbf{x}$ selon $p(\mathbf{x}|\psi)$
3) tant que $d\left(t(\mathbf{s}), t(\mathbf{x})\right) > \epsilon$
3) accepter $\theta = \psi$
4) si plusieurs valeurs de θ sont nécessaires aller en 1) autant des fois que nécessaire

Les points clefs de cette approche sont le choix du vecteur de statistiques, de la distance et du seuil ϵ. Si $p(\theta)$ est trop différent de $p(\theta|\mathbf{s})$ alors le taux de rejet de l'algorithme est vraiment très important. Les effets du choix du seuil sont discutés dans [Blum, 2010] et [Biau et al., 2015].

La méthode ABC Shadow suit un principe différent [Stoica et al., 2017, Stoica et al., 2021]. L'idée est de construire une chaîne de Markov capable d'approcher le comportement la chaîne de Markov "idéale" qui échantillonne la loi a posteriori d'intérêt. La chaîne "idéale" ne peut pas être construite afin que l'on puisse simuler. La construction de la chaîne approchée doit la suivre comme une "ombre" afin d'obtenir un vrai contrôle théorique de la solution proposée.

Supposant la chaîne "idéale" dans l'état θ, un nouvel état $\psi \in \Theta$ est généré à partir de l'état initial selon la densité $q(\theta \longrightarrow \psi)$. Cet état est accepté avec la probabilité

$$\alpha(\theta \longrightarrow \psi) = \min\left\{1, \frac{p(\psi|\mathbf{s})q(\psi \longrightarrow \theta)}{p(\theta|\mathbf{s})q(\theta \longrightarrow \psi)}\right\}. \tag{4.18}$$

Le nouvel état ψ est généré dans la boule de centre ψ et de rayon $\Delta/2 \in \mathbb{R}_+$, notée $b(\theta, \Delta/2)$. Cela se traduit par

$$q(\theta \longrightarrow \psi) = \frac{Z(\psi)^{-1}\exp(-U(\mathbf{x}|\psi))}{\int_{b(\theta,\Delta/2)} Z(\phi)^{-1}\exp(-U(\mathbf{x}|\phi))d\phi}\mathbb{1}_{b(\theta,\Delta/2)}(\psi). \tag{4.19}$$

Cette chaîne de Markov converge vers la loi *a posteriori*. Pour cette raison, elle porte le nom de chaîne idéale ("ideal chain" en anglais). La présence de la constante de normalisation dans (4.18) rend difficile la simulation de cette chaîne via cette approche.

Il est possible de donner une approximation du quotient des lois d'instrumenta-

tion nécessaire au calcul de (4.18) :

$$\frac{q(\psi \longrightarrow \theta)}{q(\theta \longrightarrow \psi)} \simeq \frac{Z(\psi)\exp(-U(\mathbf{x}|\theta))\mathbb{1}_{b(\psi,\Delta/2)}(\theta)}{Z(\theta)\exp(-U(\mathbf{x}|\psi))\mathbb{1}_{b(\theta,\Delta/2)}(\psi)}\}. \tag{4.20}$$

Le nouvel état ψ est alors accepté avec la probabilité

$$\alpha(\theta \longrightarrow \psi) = \min\left\{1, \frac{p(\psi|\mathbf{y})}{p(\theta|\mathbf{y})}\frac{Z(\psi)\exp(-U(\mathbf{x}|\theta))\mathbb{1}_{b(\psi,\Delta/2)}(\theta)}{Z(\theta)\exp(-U(\mathbf{x}|\psi))\mathbb{1}_{b(\theta,\Delta/2)}(\psi)}\right\}. \tag{4.21}$$

Lorsque Δ tend vers 0, cette nouvelle chaîne de Markov approche comme une ombre la chaîne idéale désirée, pendant un nombre fixé d'étapes.

Voilà le schéma de cet algorithme :

Algorithme ABC Shadow : Soient $\mathbf{s}$ l'observation, $\theta^{(0)} \in \Theta$ la valeur initiale du paramètre, $\Delta \in \mathbb{R}_+$ le paramètre de la loi d'instrumentation et le nombre d'itérations pour générer une valeur du paramètre $N_{ABC} \in \mathbb{N}$.

1) simuler $\mathbf{x}$ selon $p(\mathbf{x}|\theta^{(0)})$
2) Pour $k = 1, \ldots, N_{ABC}$
   a) générer ψ uniformément sur $b(\theta_k, \Delta/2)$
   b) poser $\theta^{(k)} = \psi$ avec probabilité $\alpha(\theta^{(k-1)} \longrightarrow \psi)$ et $\theta^{(k)} = \theta^{(k-1)}$ sinon rester dans le même état
3) renvoyer $\theta_1 = \theta^{(N_{ABC})}$
4) pour obtenir plusieurs échantillons aller en 1) en mettant $\theta^{(0)} = \theta_1$ autant des fois que nécessaire

4.3.3 Analyse des performances de l'algorithme ABC Shadow

Comme indiqué dans le Paragraphe 3.3, les différents algorithmes présentés dans ce chapitre ont été codés en C++, les résultats analysés par des scripts R.

Pour les mêmes raisons que dans le Paragraphe 3.3, les effets de l'initialisation de l'algorithme ABC Shadow sont observés sur des processus ponctuels déjà abordés précédemment : les processus ponctuels de Poisson et de Strauss.

Pendant les simulations, le vrai vecteur de paramètres θ^* est connu. Pour rappel, les algorithmes nécessaires utilisent uniquement les statistiques des réalisations

et le vecteur de paramètres. Le vecteur de statistiques de l'observation $t(\mathbf{s})$ est obtenu en moyennant les statistiques de réalisations obtenues avec θ^*, ici nous considérons 10^3 réalisations.

Nous établissons à partir des Exemples 4.3.1 et 4.3.2 les observations suivantes. Différentes valeurs du paramètre de la loi d'instrumentation Δ et du nombre d'itérations pour générer une valeur du paramètre N_{ABC} (le nombre de mises à jour du paramètre) sont testées pour comparer leurs effets sur la qualité de l'estimation. L'algorithme doit visiter le domaine d'étude suffisamment précisément pour trouver l'optimum de θ tout en étant suffisamment grand pour visiter tout le domaine durant l'étude. Le paramètre de la loi d'instrumentation Δ doit donc être adapté au domaine d'étude et au nombre d'itérations de l'algorithme. Le nombre de mises à jour doit rester suffisamment faible. En effet, la réalisation $\mathbf{x}$, nécessaire dans l'algorithme, est simulée selon $p(\mathbf{x}|\theta^{(0)})$ avec $\theta^{(0)}$ une valeur initiale. S'il y a trop de mise à jour, le nouvel état ψ risque d'être très différent de $\theta^{(0)}$ et donc $p(\mathbf{x}|\psi)$ et $p(\mathbf{x}|\theta^{(0)})$ seront aussi très différentes : même si $\psi = \theta^*$, cet état peut être rejeté si $\mathbf{x}$ est trop éloigné de $\mathbf{s}$.

C'est par ces observations que le choix de l'initialisation de l'algorithme a été fait. De manière empirique, le rayon de changement du vecteur θ est fixé à $\Delta = 0.01$ et le nombre d'itérations à $N_{ABC} = 200$.

Exemple 4.3.1 (Processus ponctuels de Poisson (suite))**.** *L'algorithme de l'ABC Shadow est utilisé pour estimer le paramètre* $\theta^* = -\log(20)$ *d'un processus ponctuel de Poisson, défini dans l'Exemple 2.4.1 et simulé au moyen des dynamiques étudiées dans l'Exemple 3.3.1.*

L'état initial du paramètre est fixé à $\theta^{(0)} = 0$*. Les nouveaux états sont générés dans* $[-7; 7]$*. L'algorithme ABC Shadow est utilisé pour générer* 10^5 *valeurs de* θ *avec seulement* 1000 *valeurs sauvegardées espacées de* 100 *itérations de l'algorithme. Afin de tester l'effet de* N_{ABC}*, l'algorithme est appliqué en fixant* $\Delta = 0.01$ *et* $N_{ABC} = 10$*,* $N_{ABC} = 200$ *et* $N_{ABC} = 1000$*. Pour tester l'effet de* Δ *l'algorithme est appliqué en fixant* $\Delta = 0.1$ *et* $N_{ABC} = 200$*. Dans les Figures 4.1 et 4.2 et le Tableau 4.1, ces simulations sont représentées respectivement par les indices (a), (b), (c) et (d).*

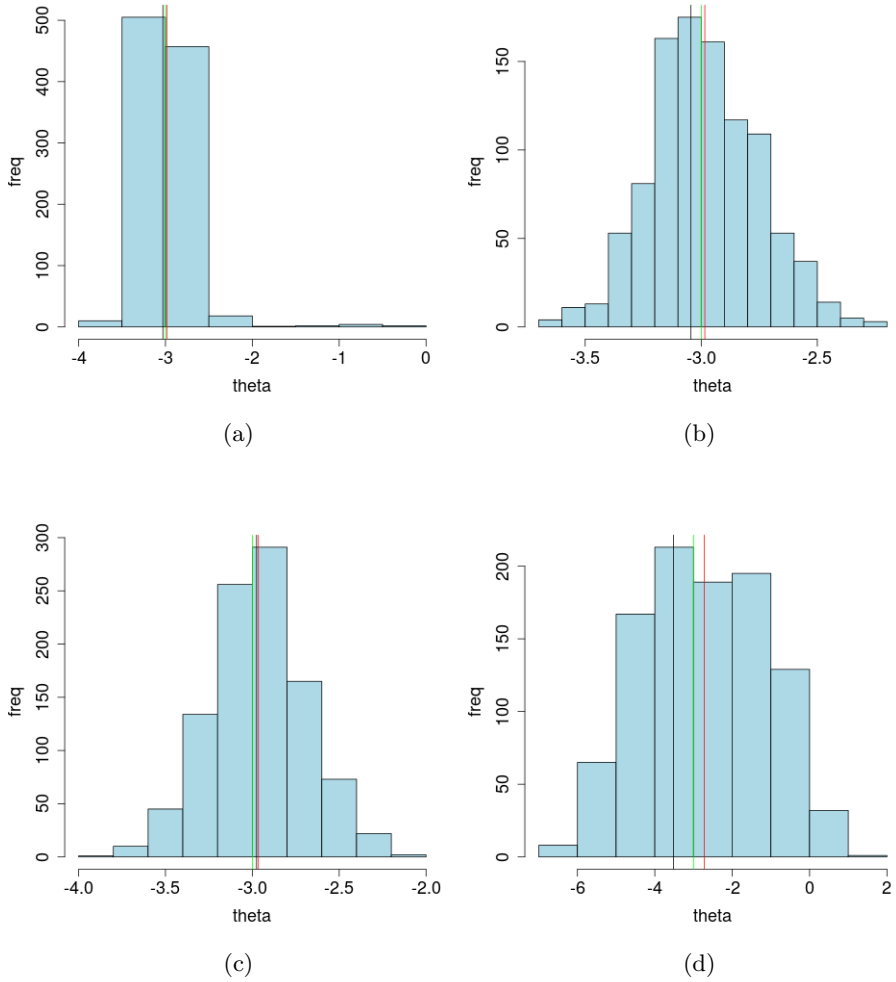


FIGURE 4.1 – Histogrammes des valeurs de θ obtenues par application de l'ABC Shadow sur des processus ponctuels de Poisson de paramètres $\theta = -\log(20)$ avec $\Delta = 0.01$ et $N_{ABC} = 10$ (a), $N_{ABC} = 200$ (b) et $N_{ABC} = 1000$ (c) et $\Delta = 0.1$ et $N_{ABC} = 200$ (d). La valeur théorique du paramètre est représentée en vert, la moyenne empirique du paramètre est représentée en rouge et son mode en noir.

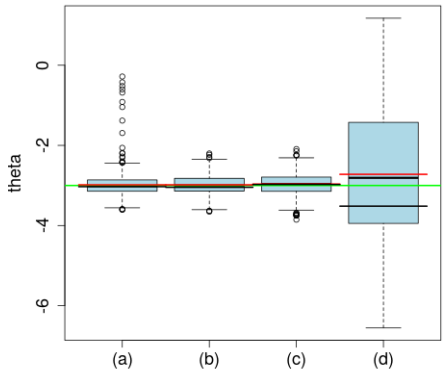


FIGURE 4.2 – Boite à moustaches des valeurs de θ obtenues par application de l'ABC Shadow sur des processus ponctuels de Poisson de paramètres $\theta = -\log(20)$ avec $\Delta = 0.01$ et $N_{ABC} = 10$ (a), $N_{ABC} = 200$ (b) et $N_{ABC} = 1000$ (c) et $\Delta = 0.1$ et $N_{ABC} = 200$ (d). La valeur théorique du paramètre est représentée en vert, la moyenne empirique du paramètre est représentée en rouge et son mode en noir.

	Q25	Q50	Q75	moyenne	mode
(a)	-3.14	-3.01	-2.86	-2.98	-3.03
(b)	-3.14	-3	-2.82	-2.98	-3.05
(c)	-3.15	-2.97	-2.79	-2.97	-2.98
(d)	-3.95	-2.81	-1.42	-2.72	-3.52

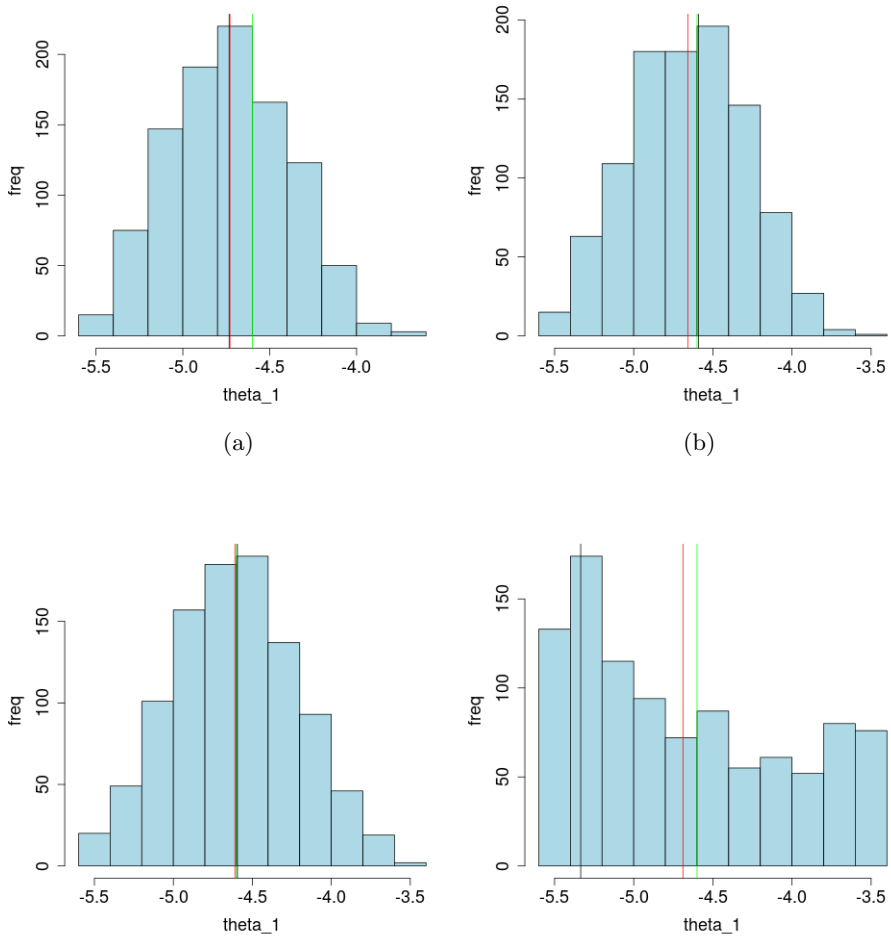
Tableau 4.1 – Valeurs des quantiles, de la moyenne et du mode des estimations de θ obtenu sur des processus ponctuels de Poisson de paramètres $\theta = -\log(20)$ avec $\Delta = 0.01$ et $N_{ABC} = 10$ (a), $N_{ABC} = 200$ (b) et $N_{ABC} = 1000$ (c) et $\Delta = 0.1$ et $N_{ABC} = 200$ (d).

Plus la valeur de N_{ABC} est faible et plus les θ générés sont concentrés autour de la vraie valeur en vert $\theta^* = -\log(20)$. Dans les cas (a), (b) et (c), les valeurs de la moyenne, représentées en rouge, et les valeurs du mode, représentées en noir, sont relativement proches. Il faut donc prendre une valeur de N_{ABC} suffisamment petite pour s'approcher de la vraie valeur θ^* mais assez grande pour balayer suffisamment la plage des valeurs possibles. C'est pour cette raison que le nombre d'itérations de l'algorithme ABC est fixé à $N_{ABC} = 200$.

L'effet de Δ est observée en comparant les résultats (b) et (d). Les θ générés ont un écart type plus grand lorsque Δ est grand. Pour cette raison, le rayon de changement de θ est fixé à $\Delta = 0.01$.

Exemple 4.3.2 (Processus ponctuels de Strauss (suite))**.** *L'algorithme de l'ABC Shadow est utilisé pour estimer le paramètre* $\theta^* = (-\log(100), 1.6)$ *d'un processus ponctuel de Strauss, défini dans l'Exemple 2.4.2 et simulé comme dans l'Exemple 3.3.2.*

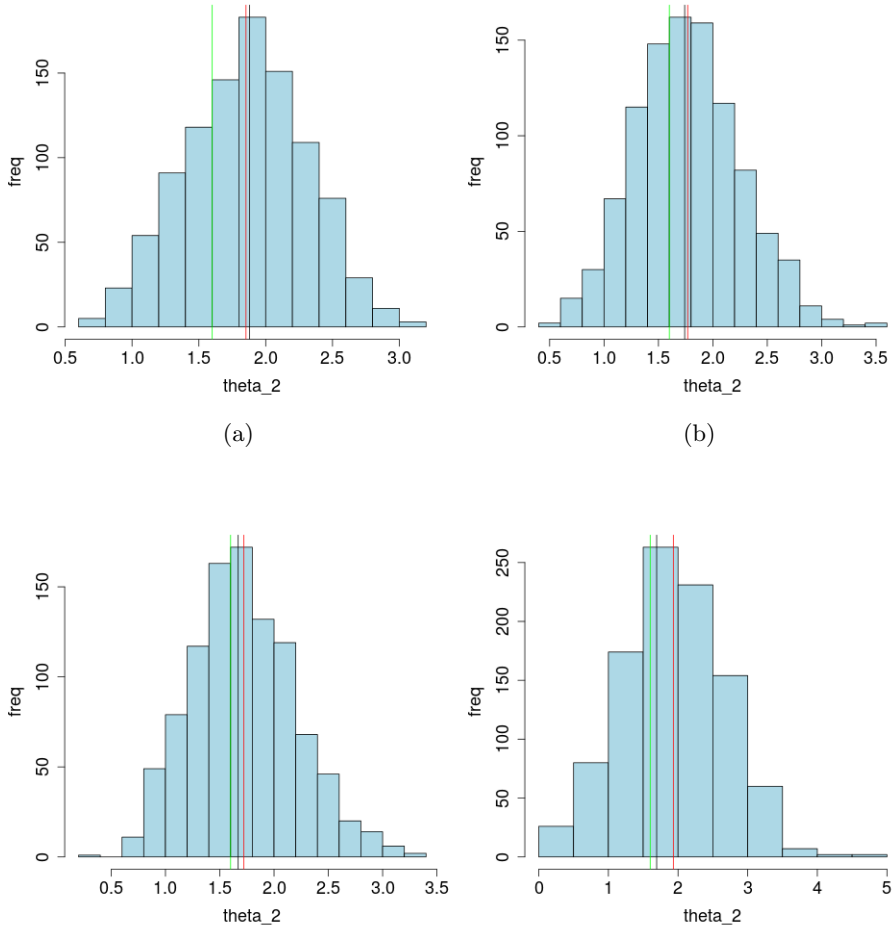
L'état initial du paramètre est fixé à $\theta^{(0)} = (-4, 2.5)$*. Les nouveaux états sont générés dans* $[-5.5; -3.5] \times [O; 5]$*. L'algorithme ABC Shadow est utilisé pour générer* 10^5 *valeurs de* θ *avec seulement* 1000 *valeurs sauvegardées espacées de* 100 *itérations de l'algorithme. Afin de tester l'effet de* N_{ABC}*, l'algorithme est appliqué en fixant* $\Delta = 0.01$ *et* $N_{ABC} = 10$*,* $N_{ABC} = 200$ *et* $N_{ABC} = 1000$*. Pour tester l'effet de* Δ *l'algorithme est appliqué en fixant* $\Delta = 0.1$ *et* $N_{ABC} = 200$*. Dans les Figures 4.3, 4.4 et 4.5 et les Tableaux 4.2 et 4.3, ces simulations sont représentées respectivement par les indices (a), (b), (c) et (d).*



(a) (b)

(c) (d)

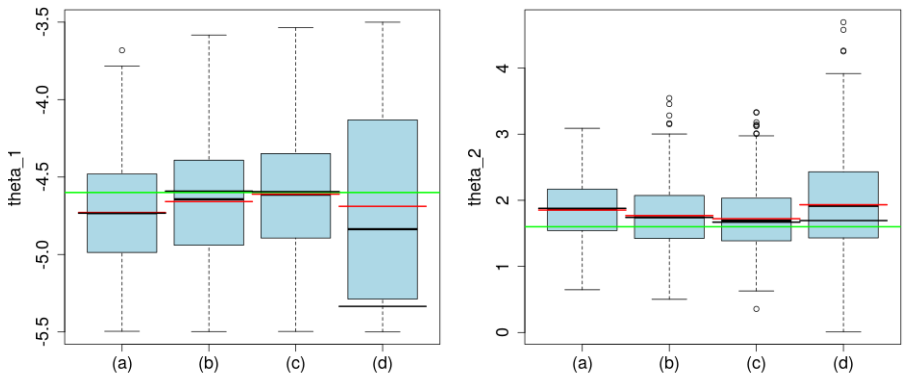
FIGURE 4.3 – Histogrammes des valeurs de θ_1 obtenues par application de l'ABC Shadow sur des processus ponctuels de Strauss de paramètres $\theta = (-\log(100), 1.6)$ et avec $\Delta = 0.01$ et $N_{ABC} = 10$ (a), $N_{ABC} = 200$ (b) et $N_{ABC} = 1000$ (c) et $\Delta = 0.1$ et $N_{ABC} = 200$ (d). La valeur théorique du paramètre est représentée en vert, la moyenne empirique du paramètre est représentée en rouge et son mode en noir.



(a) (b)

(c) (d)

FIGURE 4.4 – Histogrammes des valeurs de θ_1 obtenues par application de l'ABC Shadow sur des processus ponctuels de Strauss de paramètres $\theta = (-\log(100), 1.6)$ et avec $\Delta = 0.01$ et $N_{ABC} = 10$ (a), $N_{ABC} = 200$ (b) et $N_{ABC} = 1000$ (c) et $\Delta = 0.1$ et $N_{ABC} = 200$ (d). La valeur théorique du paramètre est représentée en vert, la moyenne empirique du paramètre est représentée en rouge et son mode en noir.



(a) (b)

FIGURE 4.5 – Boite à moustaches des valeurs de θ_1 (a) et de θ_2 (b) obtenues par application de l'ABC Shadow sur des processus ponctuels de Strauss de paramètres $\theta = (-\log(100), 1.6)$ et avec $\Delta = 0.01$ et $N_{ABC} = 10$ (a), $N_{ABC} = 200$ (b) et $N_{ABC} = 1000$ (c) et $\Delta = 0.1$ et $N_{ABC} = 200$ (d). La valeur théorique du paramètre est représentée en vert, la moyenne empirique du paramètre est représentée en rouge et son mode en noir.

	Q25	Q50	Q75	moyenne	mode
(a)	-4.99	-4.73	-4.48	-4.73	-4.73
(b)	-4.94	-4.64	-4.39	-4.66	-4.59
(c)	-4.89	-4.61	-4.35	-4.61	-4.59
(d)	-5.29	-4.84	-4.13	-4.69	-5.34

Tableau 4.2 – Valeurs de la moyenne, du mode et de la variance estimés pour θ obtenues par application de l'ABC Shadow sur des processus ponctuels de Strauss de paramètres $\theta = (-\log(100), 1.6)$ et avec $\Delta = 0.01$ et $N_{ABC} = 10$ (a), $N_{ABC} = 200$ (b) et $N_{ABC} = 1000$ (c) et $\Delta = 0.1$ et $N_{ABC} = 200$ (d).

	Q25	Q50	Q75	moyenne	mode
(a)	1.54	1.86	2.17	1.85	1.88
(b)	1.42	1.74	2.07	1.77	1.74
(c)	1.39	1.7	2.03	1.72	1.67
(d)	1.43	1.92	2.43	1.93	1.69

Tableau 4.3 – Valeurs de la moyenne, du mode et de la variance estimés pour θ obtenues par application de l'ABC Shadow sur des processus ponctuels de Strauss de paramètres $\theta = (-\log(100), 1.6)$ et avec $\Delta = 0.01$ et $N_{ABC} = 10$ (a), $N_{ABC} = 200$ (b) et $N_{ABC} = 1000$ (c) et $\Delta = 0.1$ et $N_{ABC} = 200$ (d).

Plus la valeur de N_{ABC} est élevée et plus les moyennes, en rouge, et les modes, en noir, des θ générés sont proches de la vraie valeur en vert $\theta^* = -\log(20)$. Il faut donc une valeur suffisamment grande de N_{ABC}, ainsi cette valeur est fixée à $N_{ABC} = 200$.

L'effet de Δ est observée en comparant les résultats (b) et (d). Les θ générés ont un écart type plus grand lorsque Δ est grand. Pour cette raison, le rayon de changement de θ est fixé à $\Delta = 0.01$.

Chapitre 5

HUG : un processus ponctuel de Gibbs pour la modélisation de la distribution des sources dans un mélange hydrochimique

Les sources à l'origine du système de mélange sont vues comme un ensemble de points dans l'espace des données. La position de ces sources est estimée à partir des données.

Ce chapitre est consacré à notre modèle probabiliste pour les mélanges : le modèle HUG. Ce modèle considère les données comme le résultat d'un système de mélange linéaire et conservatif. Le modèle HUG est un processus ponctuel de Gibbs qui distribue des configurations de points dans l'espace des données. La densité de probabilité de ce processus prend en considération les hypothèses sur les mélanges du paragraphe 1.4. La détection de sources consiste à trouver la configuration de points qui maximise la densité de probabilité décrivant le modèle.

Le nom de ce modèle vient du fait que les sources qui sont détectées entourent les données, un peu comme si elles leur faisaient un câlin ("hug" en anglais).

Les hypothèses utilisées pour développer le modèle sont présentées dans le paragraphe 5.1. Le paragraphe 5.2 présente la fonction énergie du modèle HUG appliqué à des données bidimensionnelles. Le paragraphe 5.3 généralise son application à des données multidimensionnelles. Le paragraphe 5.4 présente l'algorithme de simulation du modèle.

5.1 Hypothèses du modèle

Les données considérées sont des mesures de $K \in \mathbb{N}$ paramètres hydrochimiques (concentrations, ratios isotopiques,...) sur $m \in \mathbb{N}$ prélèvements. Les mesures du prélèvement numéro j sont rassemblées dans le vecteur $d_j = (d_{(j);1}, \dots, d_{(j);K})$, ce prélèvement est représenté par un point dans l'espace des paramètres hydrochimiques $W \subset \mathbb{R}^K$ avec W un espace borné. Le jeu de données s'écrit alors :

$$\mathbf{d} = \begin{pmatrix} d_{1;1} & \dots & d_{1;k} & \dots & d_{1;K} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ d_{j;1} & \dots & d_{j;k} & \dots & d_{j;K} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ d_{m;1} & \dots & d_{m;k} & \dots & d_{m;K} \end{pmatrix}.$$

Les données sont supposées issues d'un système de mélange linéaire conservatif, c'est-à-dire sans réaction chimique au cours du mélange. De plus, les données sont supposées être représentatives du mélange des sources : les sources doivent pouvoir être détectées à partir des données.

Les sources sont vues comme une configuration de points, notée $\mathbf{s} = (s_i, i = 1, \dots, n) \subset W$, qui satisfait les critères sur les sources et qui satisfait l'œil de l'expert. La composition de la source numéro i est $s_i = (s_{(i);1}, \dots, s_{(i);K})$. L'ensemble des sources s'écrit :

$$\mathbf{s} = \begin{pmatrix} s_{1;1} & \dots & s_{1;k} & \dots & s_{1;K} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ s_{i;1} & \dots & s_{i;k} & \dots & s_{i;K} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ s_{n;1} & \dots & s_{n;k} & \dots & s_{n;K} \end{pmatrix}.$$

Les critères que doivent satisfaire les sources sont des hypothèses souvent utilisées pour étudier les mélanges. Elles sont au nombre de quatre :

(a) les données ont une composition proche de celles des sources,

(b) les données sont contenues dans l'enveloppe convexe des sources,

(c) le nombre de sources est minimisé,

(d) la composition des sources est très différente d'une source à l'autre.

L'hypothèse (a) induit que les sources sont adaptées aux données. L'hypothèse (b) est la condition physique sur les mélanges. En effet, pour être expliqué par les sources, un point doit être un barycentre de ces dernières. Le poids de chaque source, aussi appelé contribution, est compris entre 0 et 1. Les données doivent donc être, par définition, des points du simplexe formé par les sources. Les hypothèses (c) et (d) sont des contraintes sur le milieu d'étude. Ces hypothèses traduisent les connaissances des experts sur les mélanges.

Le modèle que l'on propose, un processus ponctuel de Gibbs, intègre ces quatre hypothèses dans la densité de probabilité et donc dans sa fonction d'énergie. Cette construction se fait afin que les configurations de points qui satisfont le mieux ces hypothèses aient une probabilité de densité plus élevée. Plus précisément, la fonction énergie est une combinaison linéaire de $I \in \mathbb{N}$ statistiques traduisant les hypothèses précédentes faites sur les mélanges. La densité de probabilité du processus est ainsi connue à un vecteur de paramètres $\theta \in \Theta \subseteq \mathbb{R}^I$ près.

La densité de probabilité de ce processus pour une configuration de sources $\mathbf{s}$ de W est

$$p(\mathbf{s}|\theta, \mathbf{d}) = \frac{\exp\left[-U(\mathbf{s}|\theta, \mathbf{d})\right]}{Z(\theta)} = \frac{\exp\left[-U_{\mathbf{d}}(\mathbf{s}|\theta, \mathbf{d}) - U_{\mathbf{i}}(\mathbf{s}|\theta)\right]}{Z(\theta)} \tag{5.1}$$

avec $Z(\theta)$ la constante de normalisation, U la fonction énergie composée d'un terme d'attache aux données $U_{\mathbf{d}}$ et d'un terme d'interaction $U_{\mathbf{i}}$. Le terme d'attache aux données $U_{\mathbf{d}}^{(l)}(\mathbf{s}, \theta|\mathbf{d})$ tend à positionner les sources en fonction de leur relation par rapport aux données observées. Le terme d'interaction $U_{\mathbf{i}}^{(l)}(\mathbf{s}, \theta)$ contrôle ce positionnement en considérant uniquement les interactions entre sources.

Les hypothèses font intervenir la notion d'enveloppe convexe. Le calcul de cette dernière est simple en deux dimensions et devient plus difficile quand le nombre de dimensions est très grand. Ainsi, le modèle HUG est défini dans un premier temps sur des données bidimensionnelles ($K = 2$). La généralisation en dimensions plus grandes est faite par la suite.

Le modèle HUG est échantillonné en utilisant des algorithmes de simulation des chaînes de Markov de type Monte-Carlo (chaînes de Markov Monte-Carlo - MCMC). Un algorithme hybride contenant une procédure de type Metropolis-Hastings dans une dynamique de type Gibbs est mise en œuvre. Sur la base de cette dynamique, un algorithme de recuit simulé est proposé pour pouvoir maximiser la densité de probabilité du modèle proposé. Ceci considère un algorithme d'optimisation globale. Le vecteur de paramètres du modèle est estimé par l'algorithme ABC Shadow.

5.2 La fonction énergie du modèle HUG $K = 2$.

La considération des données plan par plan est souvent adoptée dans l'analyse des données hydrochimiques. En effet, il est plus simple de visualiser et d'interpréter les données et les sources en deux dimensions que de les considérer dans un espace en K-dimensions. Cette considération implique que les données doivent être analysées dans les $L = K(K-1)/2$ plans de projection. La projection des données dans le plan numéro l se fait en ne considérant que les deux dimensions du plan : k_{l_1} et k_{l_2}. La projection du jeu de données s'écrit :

$$\mathbf{d}_{\{l\}} = \begin{pmatrix} d_{1;k_{l_1}} & d_{1;k_{l_2}} \\ \vdots & \vdots \\ d_{j;k_{l_1}} & d_{j;k_{l_2}} \\ \vdots & \vdots \\ d_{m;k_{l_1}} & d_{m;k_{l_2}} \end{pmatrix}.$$

En pratique les dimensions du plan l sont le plus petit k tel que $l - \sum_{i=0}^{k-1}(K-i) \leq 0$ et $k_{l_2} = l - \sum_{i=0}^{k_{l_1}-1}(K-i)$).

Le modèle HUG doit détecter la projection sur le plan numéro l de l'ensemble des sources, noté $\mathbf{s}_{\{l\}}$. Pour simplifier les notations, la projection de l'ensemble des sources est notée $\mathbf{s}$.

Le terme d'attache aux données considère les hypothèses (a) et (b). Sur le plan numéro l, le terme d'attache aux données est notée $U^{(l)}(\mathbf{s}|\theta, \mathbf{d})$. Ce terme est composé de deux éléments. Le premier élément est un ratio qui considère les

aires des enveloppes convexes des sources et des données.

$$g^{(l)}(\mathbf{s}, \mathbf{d}) = \left| \frac{g^{(l)}(\mathbf{s})}{g^{(l)}(\mathbf{d})} - 1 \right|. \tag{5.2}$$

La fonction $g^{(l)} : W \to [0, +\infty)$ calcule l'aire de l'enveloppe convexe d'une configuration de points dans le plan l. Si les données sont représentatives du mélange, alors l'enveloppe convexe des données et l'enveloppe convexe des vraies sources devraient se confondre. La fonction valeur absolue $|\cdot|$ rend cette statistique plus flexible. En effet, cette statistique sera la même pour des configurations de points dont l'aire est plus petite que l'aire des données et des configurations de points donc l'aire est plus grande que l'aire des données. L'enveloppe convexe est obtenue par l'algorithme d'Andrew de la chaîne monotone de l'enveloppe convexe [Andrew, 1979].

Le second élément est la proportion de données non expliquées par $\mathbf{s}$, donc à l'extérieur de l'enveloppe convexe de $\mathbf{s}$. Cette statistique s'écrit :

$$n_e^{(l)}(\mathbf{s}, \mathbf{d}) = 1 - \frac{n_{expl}(\mathbf{s}, \mathbf{d})}{m}, \tag{5.3}$$

avec $n_{expl}(\mathbf{s}, \mathbf{d})$ le nombre de données expliquées par $\mathbf{s}$. Lorsque $\mathbf{s}$ explique toutes les données, $n_e^{(l)}(\mathbf{s}, \mathbf{d}) = 0$.

Ainsi, le terme d'attache aux données sur le plan numéro l est défini par la somme de (5.2) et de (5.3), pondérée par les paramètres $\theta_1, \theta_2 \in \mathbb{R}$

$$U_{\mathbf{d}}^{(l)}(\mathbf{s}|\theta, \mathbf{d}) = \theta_1 g^{(l)}(\mathbf{s}, \mathbf{d}) + \theta_2 n_{expl}(\mathbf{s}, \mathbf{d}). \tag{5.4}$$

Le terme d'interaction considère les hypothèses (c) et (d). Sur le plan numéro l, le terme d'interaction est noté $U_{\mathbf{i}}^{(l)}(\mathbf{s}|\theta)$. Ce terme est composé de deux éléments. Le premier élément est le nombre de points dans la configuration $\mathbf{s}$, noté $n(\mathbf{s})$.

Le second élément est le nombre de paires de points dans $\mathbf{s}$ à une distance inférieure à une constante fixée par l'utilisateur $r \in \mathbb{R}_+$ appelé rayon d'interaction. Cette statistique est notée $n_r^{(l)}(\mathbf{s})$. La distance choisie est la distance euclidienne.

Ainsi, le terme d'interaction dans le plan numéro l est défini par la somme des statistiques précédentes pondérée par les paramètres $\theta_3, \theta_4 \in \mathbb{R}$:

$$U_{\mathbf{i}}^{(l)}(\mathbf{s}|\theta) = \theta_3 n(\mathbf{s}) + \theta_4 n_r^{(l)}(\mathbf{s}). \tag{5.5}$$

La fonction énergie sur le plan numéro l s'écrit à partir des équations (5.4) et (5.5) :

$$U^{(l)}(\mathbf{s}|\theta,\mathbf{d}) = \theta_1 g^{(l)}(\mathbf{s},\mathbf{d}) + \theta_2 n_e^{(l)}(\mathbf{s},\mathbf{d}) + \theta_3 n(\mathbf{s}) + \theta_4 n_r^{(l)}(\mathbf{s}). \tag{5.6}$$

La répartition, la structure d'une configuration des sources et leur nombre sont contrôlées à l'aide de ces paramètres. Par exemple, si $\theta_3 > 0$, le modèle va pénaliser les configurations de points contenant beaucoup de points, et donc favoriser les configurations de points contenant peu de points. À l'inverse, si $\theta_3 < 0$ le modèle va favoriser les configurations de points contenant beaucoup de points.

Les paramètres $\theta_1, \theta_2, \theta_3$ et θ_4 sont choisis positifs afin de pénaliser les configurations avec des sources qui ont une grande différence d'aire avec les données, qui expliquent peu les données, qui contiennent beaucoup de sources et qui possèdent beaucoup de paires de sources trop proches.

Le choix de θ se fait à partir de connaissances sur les données. Cette connaissance est décrite par la loi *a priori* $p(\theta)$, proposée directement par l'utilisateur.

L'ensemble de sources est estimé par la configuration de points qui minimise la fonction énergie (5.6) et, de manière équivalente, maximise la densité de probabilité du modèle :

$$\widehat{(\mathbf{s},\theta)} = \arg\max_{\Omega\times\Theta} p(\mathbf{s},\theta|\mathbf{d}) = \arg\max_{\Omega\times\Theta} p(\mathbf{s}|\theta,\mathbf{d})p(\theta) \tag{5.7}$$

avec Ω l'espace des configurations des points dans W et $\Theta \subseteq \mathbb{R}_+^4$ l'espace des paramètres.

5.3 Généralisation en dimension $K > 2$

La généralisation du modèle HUG à $K > 2$ dimensions pourrait se faire à première vue, directement, en écrivant l'équivalent des statistiques du modèle en dimensions supérieures à deux.

Les statistiques qui forment la fonction énergie associée aux interactions sont aisément généralisables. La distance euclidienne en K dimensions lors des considérations des interactions pourrait être utilisée. Il existe des algorithmes ca-

pables de calculer des enveloppes convexes multidimensionnels [Preparata and Shamos, 1985, Hoff III et al., 1999, Tang et al., 2012]. Cependant, le modèle de HUG doit être simulée par des méthodes de type MCMC s'appuyant sur des calculs de variations locales des propriétés des configurations : comportement des distances inter, intra points pour les interactions, évolution du volume de l'enveloppe convexe du polytope formé par la position des sources. En grande dimension et en grand nombre, ces opérations induisent un coût de calcul beaucoup trop important.

Le choix adopté est une solution de compromis. Le modèle sera construit à partir des fonctions d'énergies définies sur des plans de projections (donc en dimension deux), qui seront ensuite regroupées dans une densité de probabilité définie en dimension K.

Cette généralisation s'inspire de la méthode de détection graphique utilisée par certains experts. Dans ces situations, les sources sont recherchées dans des plans de projections et reconstruites dans l'espace complet. Il est important de préciser que notre solution ne perd jamais le "contact" avec le nuage des données en dimension K. En plus, cette approche permet une simulation du modèle à travers un échantillonneur de Gibbs qui intègre un algorithme de type Metropolis-Hastings. Ces points sont précisés dans ce qui suit.

Une variable supplémentaire v qui permet de sélectionner un plan de projection est ajoutée dans le modèle. Cette variable est à valeur dans $V = \{1, \ldots, L\}$ et est caractérisée par une loi *a priori* $p(v)$. Pour le plan numéro l, la densité de probabilité conditionnelle du modèle HUG s'écrit :

$$p(\mathbf{s}|\theta, v = l, \mathbf{d}) = \frac{\exp\left[-\theta_1 g^{(l)}(\mathbf{s}, \mathbf{d}) - \theta_2 n_e^{(l)}(\mathbf{s}, \mathbf{d}) - \theta_3 n(\mathbf{s}) - \theta_4 n_r^{(l)}(\mathbf{s})\right]}{Z^{(l)}(\theta)}. \quad (5.8)$$

avec $\theta = (\theta_1, \theta_2, \theta_3, \theta_4)$ et

$$Z^{(l)}(\theta) = \int_W \exp\left[-\theta_1 g^{(l)}(\mathbf{y}, \mathbf{d}) - \theta_2 n_e^{(l)}(\mathbf{y}, \mathbf{d}) - \theta_3 n(\mathbf{y}) - \theta_4 n_r^{(l)}(\mathbf{y})\right] d\mu(\mathbf{y})$$

la constante de normalisation associée.

Le modèle HUG généralisé est alors

$$p(\mathbf{s}, \theta, v|\mathbf{d}) = p(\mathbf{s}, \theta|v, \mathbf{d})p(v) = p(\mathbf{s}|\theta, v, \mathbf{d})p(\theta)p(v). \quad (5.9)$$

La solution que l'on propose est donnée par la solution au problème d'optimisation suivant :

$$\widehat{(\mathbf{s},\theta,v)} = \arg\max_{\Omega\times\Theta\times V} p_{\mathbf{d}}(\mathbf{s},\theta,v) = \arg\max_{\Omega\times\Theta\times V} p_{\mathbf{d}}(\mathbf{s}|\theta,v)p(\theta)p(v). \tag{5.10}$$

5.4 Simulation et optimisation

Pour la suite, nous allons considérer que la surface de l'enveloppe convexe des données est positive, c'est-à-dire

$$g^{(l)}(\mathbf{d}) > 0, \forall l \in L. \tag{5.11}$$

Cette hypothèse n'est pas restrictive du point de vue pratique. Elle impose que les termes de mélange ne soient pas concentrés sur un point (uniquement une source), une ligne droite (mélange de deux sources) ou courbe parfaite (mélange non conservatif). En effet, cette hypothèse est garantie par le contexte géologique de notre problème et impose un nombre minimal de trois sources à détecter.

Les résultats suivants montrent que le modèle HUG pour $K = 2$ est localement stable, ce qui implique que le modèle soit intégrable [Geyer, 1999, Møller and Waagepetersen, 2004].

Proposition 5.4.1. *Sous l'hypothèse* (5.11)*, pour* $\theta_1, \theta_2 > 0$ *le terme d'attache aux données du modèle du HUG (K=2) est borné.*

Démonstration. Par définition $0 \leq g^{(l)}(\mathbf{s}) \leq \nu(W)$.

L'hypothèse $g^{(l)}(\mathbf{d}) \neq 0$ permet d'écrire :

$$\min_{t\in[0,\nu(W)]} \left\{ \left| \frac{t}{g^{(l)}(\mathbf{d})} - 1 \right| \right\} \leq g^{(l)}(\mathbf{s},\mathbf{d}) \leq \max_{t\in[0,\nu(W)]} \left\{ \left| \frac{t}{g^{(l)}(\mathbf{d})} - 1 \right| \right\} < \infty. \tag{5.12}$$

De plus $0 \leq n_e^{(l)}(\mathbf{s},\mathbf{d}) \leq 1$. En conclusion, si $\theta_1, \theta_2 > 0$ alors

$$\theta_1 \min_{t\in[0,\nu(W)]} \left\{ \left| \frac{t}{g^{(l)}(\mathbf{d})} - 1 \right| \right\} \leq U_{\mathbf{d}}^{(l)}(\mathbf{s},\theta|\mathbf{d}) \leq \theta_1 \max_{t\in[0,\nu(W)]} \left\{ \left| \frac{t}{g^{(l)}(\mathbf{d})} - 1 \right| \right\} + \theta_2. \tag{5.13}$$

□

Théorème 5.4.2. *Sous l'hypothèse* (5.11)*, le modèle de HUG (K=2) est intégrable pour tout* $\theta_1, \theta_2, \theta_3, \theta_4 > 0$.

Démonstration. Nous allons commencer par montrer que le modèle est localement stable.

Soit une configuration $\mathbf{s} \in \Omega$. D'après la Proposition 5.4.1, et en posant

$$M_1 = \theta_1 \min_{t \in [0, \nu(W)]} \left\{ \left| \frac{t}{g^{(l)}(\mathbf{d})} - 1 \right| \right\} \tag{5.14}$$

l'on obtient :

$$U^{(l)}(\mathbf{s}|\theta, \mathbf{d}) \geq M_1 + \theta_3 n(\mathbf{s}) + \theta_4 n_r^{(l)}(\mathbf{s}) \geq \theta_3 n(\mathbf{s}) + \theta_4 n_r^{(l)}(\mathbf{s}). \tag{5.15}$$

La fonction énergie du modèle HUG est plus grande que la fonction énergie d'un processus de type Strauss (2.7).

Le processus de Strauss est localement stable si $\theta_4 > 0$ [Geyer, 1999, M. N. M. van Lieshout, 2000, Møller and Waagepetersen, 2004]. En effet, pour toute configuration $\mathbf{s} \in \Omega$ et $\xi \in W$ tels que $s_i \neq \xi, \forall s_i \in \mathbf{s}$, l'intensité conditionnelle de ce processus de Strauss est majorée par :

$$\lambda^*(\mathbf{s}, \xi) = \exp\left[-\theta_3\right] \exp\left[-\theta_4\right]^{n_r(\mathbf{s} \cup \xi) - n_r(\mathbf{s})} \leq \exp\left[-\theta_3\right]. \tag{5.16}$$

Ainsi la stabilité locale de ce processus de Strauss implique celle du modèle HUG. En se rapportant au Chapitre 2, nous pouvons conclure que la stabilité locale du modèle HUG implique son intégrabilité . □

Le modèle de HUG ($K = 2$) avec paramètres fixés peut être simulé par un algorithme de Metropolis-Hasting qui est $\phi-$irréductible, Harris récurrent et géométrique ergodique [Geyer, 1999, M. N. M. van Lieshout, 2000, Møller and Waagepetersen, 2004]. Quand $K > 2$ ou quand les paramètres du modèle sont décrits par une loi a priori connue $p(\theta)$, un échantillonneur de Gibbs peut être construit. Cet algorithme demande une loi a priori $p(v)$ pour sélectionner un plan de projection. Puis, conditionnellement aux paramètres et au plan de projection choisis, l'algorithme de Metropolis-Hastings mentionné auparavant, peut y être intégré.

La recherche du maximum se fait à l'aide d'un algorithme de type recuit simulé *(SA - simulated annealing)* s'appuyant sur la dynamique de simulation qui vient d'être décrite. Pour les algorithmes de type recuit simulé basé sur des dynamiques

de type processus naissance-mort ou Metropolis-Hastings, des schémas optimaux de refroidissement ont été prouvés [van Lieshout, 1994, Stoica et al., 2005]. Dans notre situation, pour des raisons liées au temps de calcul, nous avons opté pour un schéma de refroidissement polynomial

$$T_{k+1} = cT_k, \tag{5.17}$$

avec $0 < c < 1$ le coefficient de refroidissement et $T_0 \in \mathbb{R}_+$ la température initiale.

Le schéma de l'algorithme de recuit simulé utilisé est présenté plus bas. Il faut noter que si la température est gardée constante, alors l'on obtient un échantillonneur du modèle avec paramètres θ/T.

Algorithme SA : Fixer $p(\theta)$ et $p(v)$. Choisir la configuration initiale $\mathbf{s}_0$, la température initiale T_0, le coefficient c, le nombre total d'itérations N, le nombre d'applications de la dynamique de Gibbs G et N_{MH} le nombre d'applications de l'algorithme MH.

1. Pour $k = 0, \ldots, N-1$ faire
   - $\theta^{(k+1)} \sim [p(\theta)]^{1/T_k}$
   - Pour $g = 1, \ldots, G$ faire
     (a) $v_{k+1}^{(g)} \sim [p(v)]^{1/T_k}$
     (b) $\mathbf{s}_{k+1}^{(g)} \sim [p(\mathbf{s}|\theta^{(k+1)}, v_{k+1}^{(g)})]^{1/T_k}$ en appliquant l'algorithme MH sur $\mathbf{s}_{k+1}^{(g-1)}$ successivement N_{MH} fois, avec $\mathbf{s}_{k+1}^{(0)} = \mathbf{s}_{k-1}$.
   - poser $\mathbf{s}_{k+1} = \mathbf{s}_{k+1}^{(G)}$ et $v_{k+1} = v_{k+1}^{(G)}$
   - $T_{k+1} = cT_k$
2. Renvoyer $(\mathbf{s}_N, \theta^{(N)}, v_N)$.

Ici l'on a présenté, la détection des sources effectuée en considérant $p(\theta)$ et $p(v)$ définies par l'utilisateur. Dans le chapitre suivant, une méthode pour choisir $p(\theta)$ est présentée. À partir d'une configuration des sources $\mathbf{s}$, les paramètres θ et la loi *a posteriori* $p(\theta|\mathbf{s})$ peuvent être estimés par l'algorithme ABC Shadow. Ainsi, nous allons y proposer une détection en trois étapes. La première consiste dans la détection utilisant des lois *a priori* construites ad-hoc par l'utilisateur. Ensuite, conditionnellement à la configuration des sources obtenues, l'inférence sur les paramètres θ permet construire des nouveaux *a priori*. Finalement, la procédure de détection des sources est relancée en utilisant ces nouveaux paramètres. Nous montrons que la qualité des résultats ainsi obtenus est améliorée.

Chapitre 6

Détection et caractérisation des sources dans un mélange hydrochimique. Analyse de la méthode proposée.

Ce chapitre est structuré en trois parties. La première partie met en place tous les éléments présentés jusqu'à ce moment pour illustrer les résultats de notre méthodologie capable de détecter les sources dans un mélange hydrochimique. Ensuite, l'estimation des paramètres est abordée avec un double but. Premièrement, l'on cherche à améliorer la construction des lois a priori pour les paramètres. Deuxièmement, les paramètres caractérisent la distribution spatiale des sources. Finalement, la troisième partie de ce chapitre développe une analyse détaillée des différents éléments qui forment la solution proposée.

6.1 Détection des sources en utilisant le modèle de HUG

Cette partie illustre les performances du modèle de HUG pour la détection des sources. Le vecteur de paramètres θ est décrit par une loi *a priori* $p(\theta)$. La méthode est appliquée à des données synthétiques et ensuite à des données réelles.

6.1.1 Création de données synthétiques et construction de la loi *a priori* sur θ

Les données synthétiques sont construites par étapes.

La première étape est d'établir le nombre de sources $n \in \mathbb{N}$ et le nombre de dimensions de l'espace de données $K \in \mathbb{N}$.

La seconde étape est de définir l'espace des données $W \subset \mathbb{R}^K$, qui est un espace borné. La composition des sources est alors fixée soit par l'utilisateur, soit de manière aléatoire. Il est important de noter que le modèle ne peut détecter une source que si elle est l'un des sommets de l'enveloppe convexe des sources sur au moins un des $L = K(K-1)/2$ plans d'études. La fonction énergie du modèle ne considère en effet que l'aire de l'enveloppe convexe et le nombre de données expliquées : ces valeurs ne sont pas modifiées par une source qui ne fait pas partie de l'enveloppe convexe des sources.

La troisième étape consiste à créer les $m \in \mathbb{N}$ données. Pour cela, nous avons choisi de générer leur vecteur de contribution des sources γ_j (avec $j \in [\![1; m]\!]$), utilisé dans l'équation (1.1) au moyen d'une loi de Dirichlet de paramètre $1_n = (1, \ldots, 1) \in \mathbb{R}^n$. Cette loi a la particularité de générer des vecteurs dont la somme des coordonnées soit égal à 1. La loi de Dirichlet de paramètre $1_n = (1, \ldots, 1) \in \mathbb{R}^n$ est équivalente à la distribution uniforme dans un simplexe à n dimensions. Cette dernière particularité, justifie le choix de cette loi pour construire la plus "simple" répartition de termes de mélange.

Les données sont ensuite normalisées : l'espace de données W est transformé en l'hypercube unité au moyen d'une transformation affine. Sur chaque axe, les données subissent une translation suivie d'une homothétie. Pour simplifier, les sources peuvent directement être générées dans l'hypercube unité, permettant d'éviter l'étape de normalisation. Cette opération permet de traiter d'une manière homogène chaque composante du fluide.

La loi $p(\theta)$ est fixée d'une manière empirique, par des essais successifs. La stratégie adoptée s'apparente aux principes des méthodes ABC classiques. Pour ce faire, des jeux de données synthétiques similaires au jeu de données étudié sont créés, c'est-à-dire avec un même nombre de données et une répartition spatiale proche, mais dans ce cas avec le nombre et la position des sources connues. Pour chacun de ses jeux de données, le modèle HUG est appliqué avec différentes va-

leurs constantes de θ. Lorsque l'utilisateur est satisfait des résultats sur un jeu de données, il conserve le θ correspondant. À la fin de ces simulations, l'utilisateur obtient autant de valeurs de θ que de jeux de données.

La loi *a priori* choisie pour $p(\theta)$ est une loi normale de paramètres obtenus construite avec les moyennes et les variances empiriques calculées utilisant les valeurs sélectionnées à l'étape précédente.

6.1.2 Application à des données synthétiques

La loi *a priori* $p(\theta)$ choisie est une loi gaussienne paramétrée selon le Tableau 6.1.

	θ_1	θ_2	θ_3	θ_4
μ_θ (moyenne)	11.25	250.0	0.25	1.0
σ^2_θ (variance)	1.0	10.0	0.01	0.01

Tableau 6.1 – Valeurs de la moyenne et de la variance pour la loi a priori Gaussienne de θ.

La méthode de détection des sources utilisant le modèle HUG est paramétrée selon le Tableau 6.2.

Variable	Description	Valeur
L	nombre de plans	$K(K-1)/2$
r	rayon d'interaction	0.01
N	nombre d'applications de l'algorithme SA	$3.5 * 10^6$
G	nombre d'applications de la dynamique de Gibbs	L
N_{MH}	nombre d'applications de l'algorithme MH	200
T_0	température initiale	10^4
c	coefficient de refroidissement	0.99999
$p_b; p_d; p_c$	probabilité de "naissance" ; "mort" ; "changement"	$0.2; 0.2; 0.6$
r_c	rayon de la boule utilisée dans l'évènement "changement"	0.3

Tableau 6.2 – Paramètres de la méthode de détection des sources utilisant le modèle de HUG.

Une fois l'algorithme SA lancé, l'on sauvegarde toutes les 1000 itérations, la valeur de θ, la configuration estimée de sources et les statistiques suffisantes. Cela

permet de suivre les simulations tout en minimisant la sauvegarde de résultats intermédiaires.

Le premier jeu de données synthétiques considéré est formé par $m = 200$ résultats d'un système de mélange de $n = 4$ sources dans un espace à $K = 3$ dimensions. Il faut donc considérer $L = 3$ plans d'études. La composition de la configuration des vraies sources $\mathbf{s}^*$ est donnée dans le Tableau 6.3.

indice	solute1	solute2	solute3
1	0.2	0.2	0.2
2	0.3	0.78	0.8
3	0.7	0.7	0.1
4	0.8	0.13	0.8

Tableau 6.3 – Coordonnées des sources du premier jeu de données synthétiques.

Les statistiques des vraies sources sont indiquées dans le Tableau 6.4.

l	$g^{(l)}(\mathbf{s}, \mathbf{d})$	$n_e^{(l)}(\mathbf{s}, \mathbf{d})$	$n(\mathbf{s})$	$n_r^{(l)}(\mathbf{s})$
1	0.358501	0	4	0
2	0.294945	0	4	0
3	0.299012	0	4	0

Tableau 6.4 – Statistiques des vraies sources du premier jeu de données synthétique sur chaque plan.

L'utilisation de l'algorithme de recuit simulé, demande de laisser la température refroidir. Une analyse robuste demande de s'assurer que tous les plans de projections aient été suffisamment parcourus. Dans le contexte des données traitées, les 500 dernières réalisations du modèle sont utilisées pour interpréter les résultats.

Les mises à jour de la configuration de sources se font plan par plan. Ainsi, pour augmenter la robustesse du modèle, ce n'est pas l'évolution des statistiques qui est présentée, mais l'évolution de la moyenne cumulée de chaque statistique. Cette évolution est représentée dans la Figure 6.1. L'inspection visuelle de ces courbes indique une possible convergence de ces valeurs.

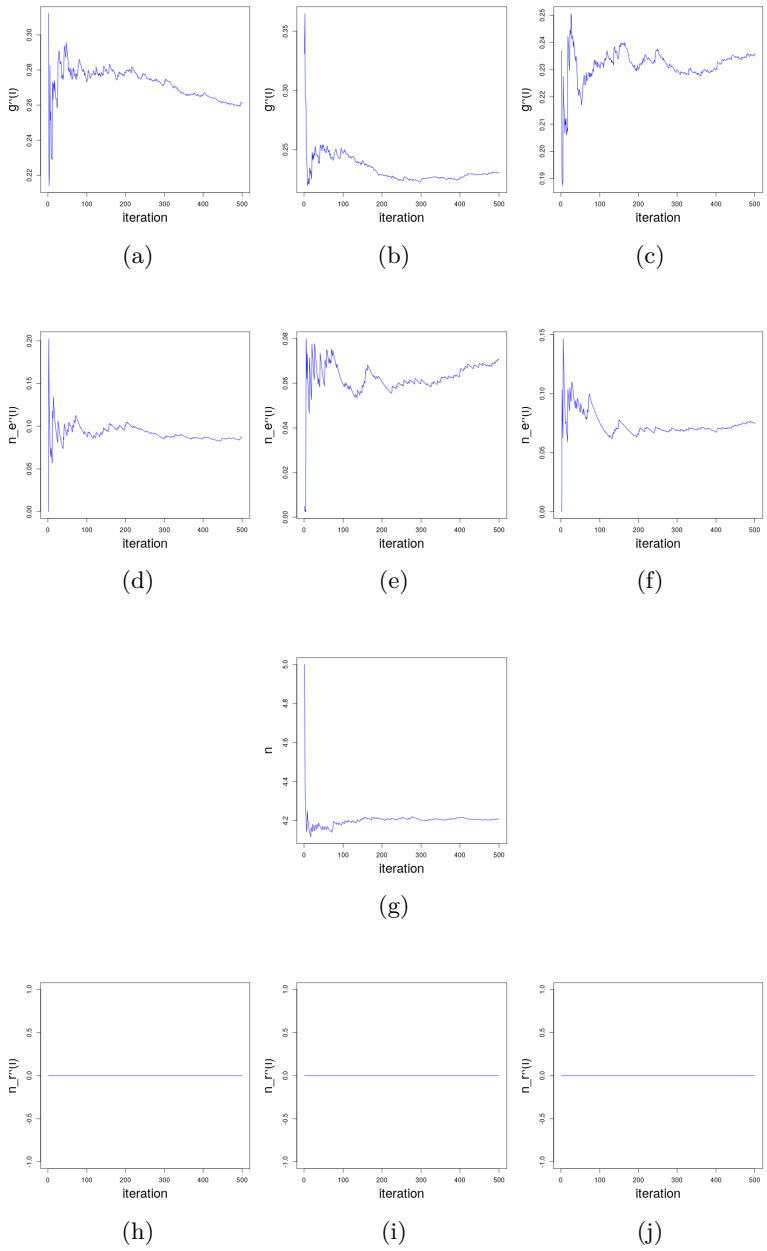
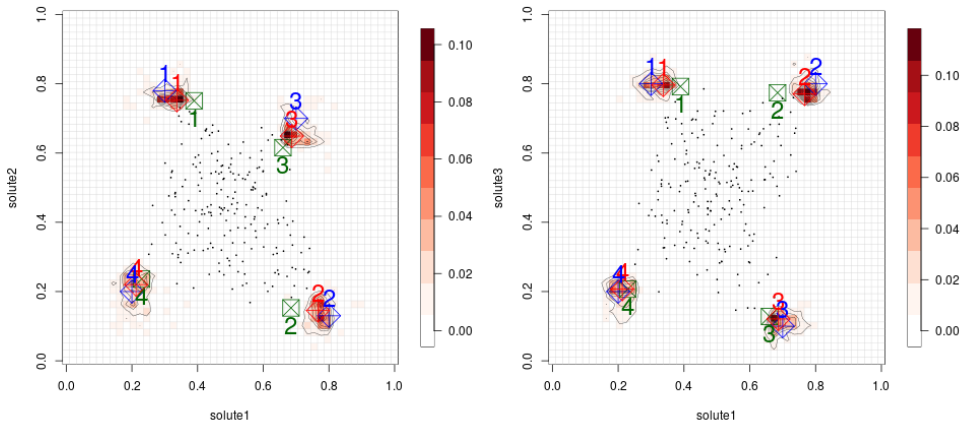


FIGURE 6.1 – Évolution des moyennes cumulées des statistiques suffisantes pour le premier jeu de données synthétique dans le plan un (a,d,g,h), deux (b,e,g,i) et trois (c,f,g,j), chaque ligne représente une statistique et chaque colonne un plan.

Les 500 configurations de points obtenues sont projetées dans les trois plans d'études (Figure 6.2). Chaque plan est subdivisé par une grille régulière composée de cellules carrées de côté 0.02. La probabilité qu'une cellule contienne une source simulée est calculée en divisant le nombre de sources simulées dans cette cellule

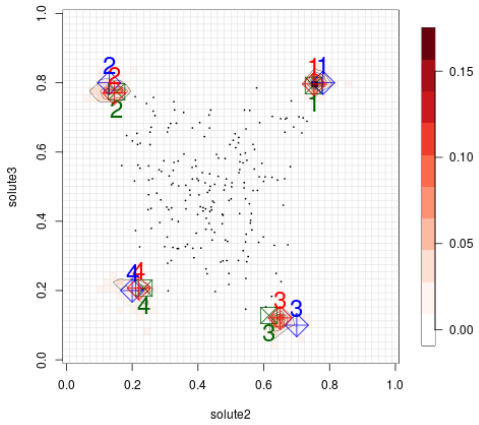
par le nombre de configurations considérées, ici 500.

Une première estimation des sources est obtenue en appliquant un algorithme de classification sur les 500 configurations. Le nombre de classes considéré est la valeur finale de la moyenne cumulée du nombre de points, arrondie à l'entier le plus proche, ici quatre. Les centres de ces quatre classes, obtenus en cherchant le point moyen de chaque classe, forment ainsi cette première estimation des sources. La moyenne est sensible aux valeurs extrêmes, d'où la nécessité d'utiliser d'autres estimations des sources à partir des résultats du modèle HUG. Une seconde estimation des sources est calculée en considérant le point médian de chaque classe. Ce point médian est obtenu en cherchant sur chaque dimension la valeur médiane : l'effet des valeurs extrêmes est ainsi réduit. Dans la Figure 6.2, les vraies sources sont représentées en bleu, les centres des classes en vert et les points médians en rouge. La première chose à remarquer est que le nombre de sources estimé est le même que le nombre de vraies sources. De plus, les sources simulées sur chaque plan sont principalement réparties dans quatre zones proches des vraies sources. Il faut aussi noter que les points médians sont plus proches des centres des zones, et donc des vraies sources, que les centres des classes.



(a)

(b)



(c)

FIGURE 6.2 – Ensembles de niveaux obtenus pour le premier jeu de données synthétiques dans le premier plan (a), le deuxième plan (b) et le troisième plan (c). Les vraies sources sont en bleu. Les centres obtenus en appliquant un algorithme des k-moyennes à quatre classes sont en vert. Les points médians de chaque classe sont en rouge.

Les sources proposées par le modèle (centres des classes et points médians) sont relativement proches des vraies sources. Les erreurs moyennes entre une vraie source et son estimation (le point médian) sur chaque dimension sont calculées en pourcentage par la formule

$$\frac{s_{(i);k} - s^*_{(i);k}}{s^*_{(i);k}} \times 100. \tag{6.1}$$

Les erreurs moyennes sont présentées dans le Tableau 6.5.

source	solute1	solute2	solute3	erreur moyenne par source
1	13.3	-3.8	0.00	3.2
2	-3.8	15.4	-3.8	2.6
3	-1.4	-7.1	20.0	3.8
4	5.00	10.0	5.0	6.7
erreur moyenne par dimension	3.3	3.6	5.3	4.1

Tableau 6.5 – Erreur moyenne relative en pourcentage entre les vraies sources et les points médians.

La méthode proposée a donc permis de détecter le nombre et la composition des sources dans l'espace des données multidimensionnelles à partir de la projection des données sur chacun des plans d'études. La qualité des résultats repose sur plusieurs critères :

— la construction du modèle,

— la paramétrisation des algorithmes,

— la qualité de l'échantillonnage du système de mélange,

— le choix des lois *a priori* de θ et de v.

6.1.3 Reconstruction des sources proposées

Le jeu de données considéré précédemment est relativement simple. En effet, les vraies sources sont "visibles" sur tous les plans, c'est-à-dire que les vraies sources sont des sommets de l'enveloppe convexe des sources sur chacun des plans d'études. La reconstruction des sources ne rencontre aucun problème majeur.

Cependant, si toutes les sources ne sont pas tout le temps visibles, la reconstruction des sources devient problématique. Le nombre de sources total ne peut plus être déterminé simplement en considérant les plans comme dans la Figure 1.4.

Le deuxième jeu de données synthétiques considéré est un mélange de quatre sources. Sur chaque plan, seulement trois sources sont visibles : la dernière source est confondue avec une autre source. La composition des sources est décrite dans le Tableau 6.6.

indice	solute1	solute2	solute3
1	0.29	0.32	0.33
2	0.67	0.32	0.33
3	0.67	0.67	0.33
4	0.67	0.67	0.76

Tableau 6.6 – Composition des sources du second jeu de données synthétiques.

Le modèle HUG est appliqué sur ce jeu de données. L'évolution des moyennes cumulées des statistiques est donnée dans la Figure 6.3. L'évolution de la moyenne cumulée du nombre de points indique une possible convergence vers une valeur. En prenant l'entier le plus proche, la méthode détecte trois sources sur chaque plan.

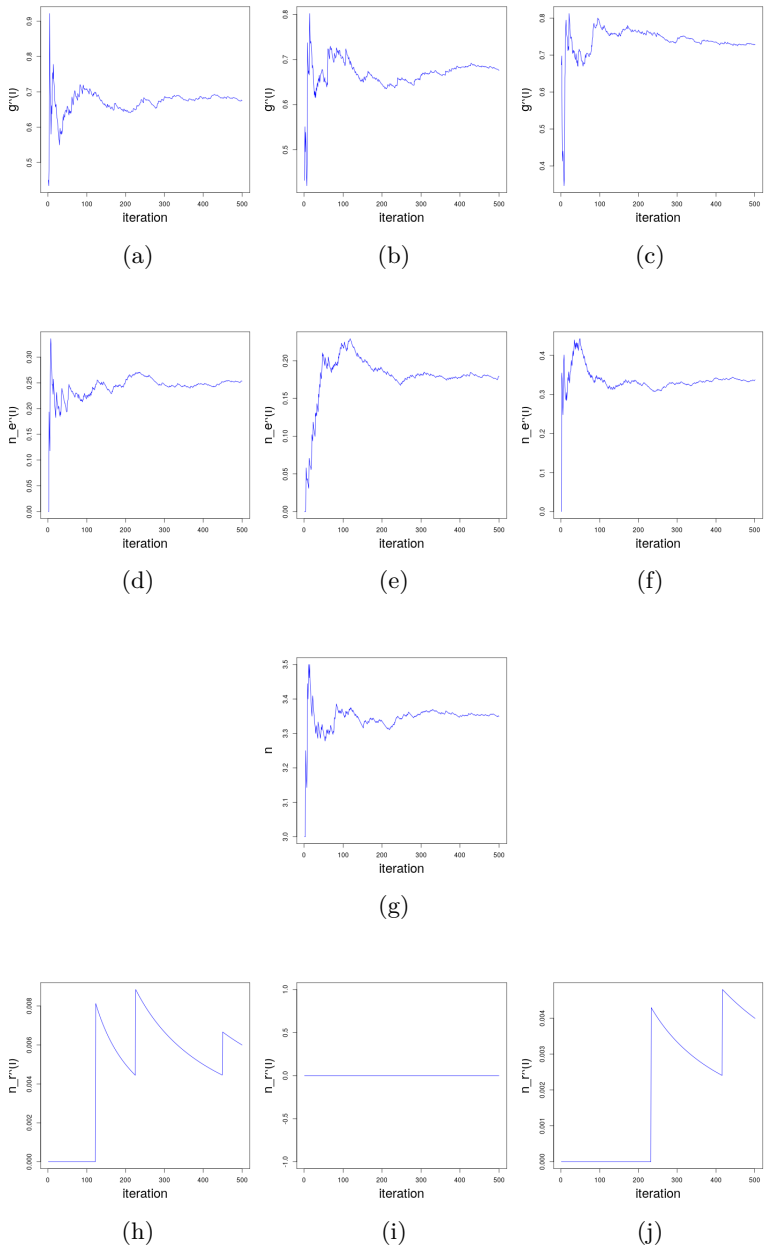
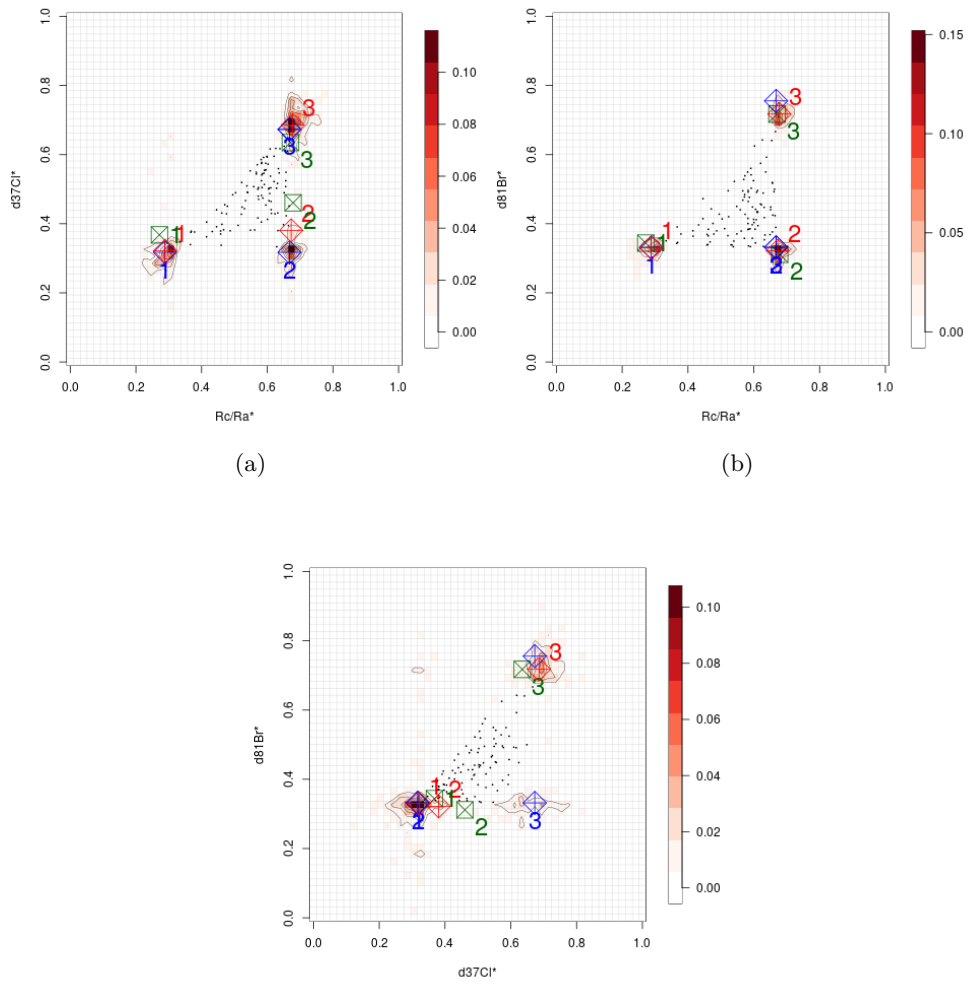


FIGURE 6.3 – Évolution des moyennes cumulées des statistiques suffisantes pour le deuxième jeu de données synthétique dans le plan un (a,d,g,h), deux (b,e,g,i) et trois (c,f,g,j), chaque ligne représente une statistique et chaque colonne un plan.

Les 500 configurations de points obtenues sont projetées sur les plans d'études dans la Figure 6.4. Le modèle détecte sur chaque plan trois zones relativement proches des vraies sources. Cependant, la classification en trois classes ne donne

pas de sources satisfaisantes.



(a) (b)

(c)

FIGURE 6.4 – Ensembles de niveaux obtenus pour le deuxième jeu de données synthétiques dans le premier plan (a), le deuxième plan (b) et le troisième plan (c). Les quatre vraies sources sont en bleu. Les centres obtenus en appliquant un algorithme des k-moyennes à trois classes sont en vert. Les points médians de chaque classe sont en rouge.

En appliquant l'algorithme des k-moyennes séquentiel, quatre sources sont reconstruites. Pour vérifier ce nombre, une classification hiérarchique est effectuée en utilisant les 500 configurations sauvegardées. Le résultat obtenu est montré dans la Figure 6.5.

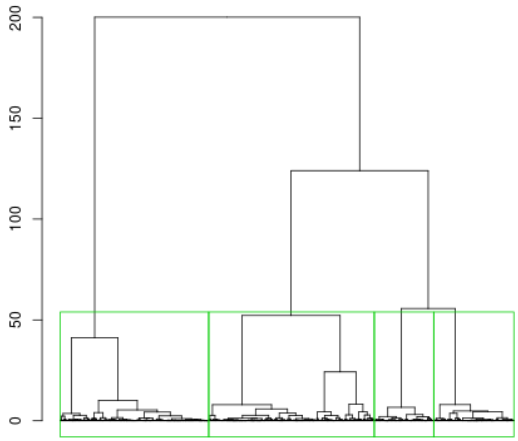


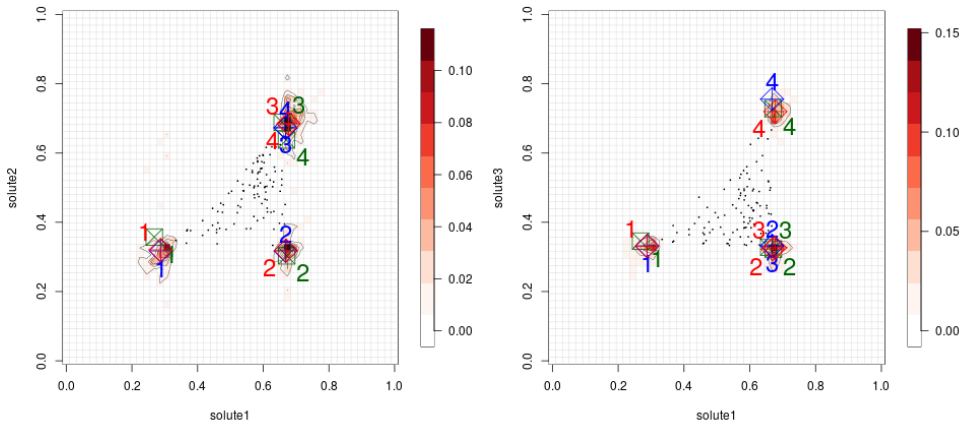
FIGURE 6.5 – Dendrogramme obtenu par un regroupement hiérarchique minimisant la variance intra classe pour le second jeu de données synthétiques. Les quatre classes sont délimitées par des rectangles verts.

Une autre vérification concernant le nombre de sources consiste en l'application d'une classification avec un nombre supérieur de classes et le calcul du pourcentage de points attribué à chaque classe. Les pourcentages de points contenus dans les plus grandes quatre classes, les quatre classes qui contiennent le plus de sources, sont présentés dans le Tableau 6.7. Les résultats montrent que les quatre principales classes obtenues dans chaque classification contiennent plus de 75% des sources. Le nombre de sources détectées est donc quatre.

Nombre de classes	5	6	7	8	9
Pourcentage	92	84	82	75	77

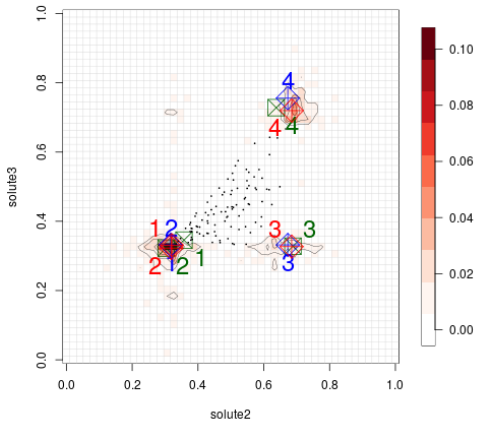
Tableau 6.7 – Pourcentage de la répartition des sources dans les quatre principales classes obtenues par un algorithme des k-moyennes à cinq, six, sept, huit et neuf classes.

Suite à cette analyse, une classification à quatre classes est appliquée sur les réalisations, les 500 configurations des sources. Les résultats sont représentés dans la Figure 6.6. Les centres des classes et les points médians forment deux configurations de points très proches des vraies sources.



(a)

(b)



(c)

FIGURE 6.6 – Ensembles de niveaux obtenus pour le premier jeu de données synthétiques dans le premier plan (a), le deuxième plan (b) et le troisième plan (c). Les vraies sources sont en bleu. Les centres obtenus en appliquant un algorithme des k-moyennes à quatre classes sont en vert. Les points médians de chaque classe sont en rouge.

Les erreurs relatives entre les points médians et les vraies sources en pourcentage sont décrites dans le Tableau 6.8.

indice	solute1	solute2	solute3	erreur moyenne par source
1	0.0	0.0	0.0	0.0
2	0.0	0.0	-3.0	-1.0
3	1.5	1.5	-0.0	1.0
4	1.5	1.5	-5.3	-0.8
erreur moyenne par dimension	0.8	0.8	-2.1	-0.2

Tableau 6.8 – Erreur moyenne relative en pourcentage entre les vraies sources et les points médians.

Le modèle, associé à l'algorithme des k-moyennes séquentiel, a donc permis de détecter le nombre et la composition des sources sur l'espace des données multidimensionnelles à partir de la projection des données sur chacun des plans d'études, malgré le fait que toutes les sources ne soient pas "visibles" sur tous les plans d'études. L'algorithme des k-moyennes séquentiel est un outil qui montre aussi ses capacités quant à la reconstruction des sources en multidimensions.

6.1.4 Application à des données réelles

Le modèle HUG est désormais appliqué à deux jeux de données réelles. La particularité des jeux de données réelles est que le nombre et la composition des sources ne sont pas connues. En effet, bien qu'il soit possible d'estimer le nombre et la composition des sources, dans ne nombreux cas, les prélèvements représentent des mélanges entre plusieurs sources et non les sources elles-mêmes.

Le premier jeu de données réelles étudié est présenté dans l'article [Pinti et al., 2020]. Dans cet article, une méthode d'estimation des sources est par ailleurs décrite. Les données sont considérées bidimensionnelles et résultantes d'un système de mélange de trois sources. Les trois sources identifiées sont l'une issue du manteau terrestre "manteau", une issue d'une zone de subductions "subduction" et la dernière issue de la croûte "croute". Dans l'espace des données, les sources sont les sommets du plus petit triangle en terme d'aire qui contient les données.

Les données considérées sont les compositions en isotopes stables du chlore $\delta^{34}Cl$ notées $d37Cl$, compositions en isotopes stables du brome $\delta^{81}Br$ noté $d81Br$ et le ratio $^3He/^4He$ notées Rc/Ra de $m = 75$ échantillons prélevés dans des puits géothermaux du Mexique. Aucune réaction impliquant, le chlore, le brome et

l'hélium n'est attendue au cours du mélange ([Pinti et al., 2020]). Les plans d'études sont au nombre de deux : un premier plan formé par Rc/Ra et $d37Cl$ et un second plan formé par Rc/Ra et $d81Br$. Aucun effet de courbure n'est attendu sur ces plans ([Pinti et al., 2020]). Le mélange n'est pas linéaire sur le dernier plan possible formé par $d37Cl$ et $d81Br$ en raison d'un effet de courbure ([Pinti et al., 2020]). Le modèle n'est donc pas appliqué sur ce plan. La loi *a priori* sur le choix du plan est donc ici 50% du temps le modèle choisit le plan Rc/Ra et $d37Cl$ et le reste du temps le modèle choisit le plan Rc/Ra et $d81Br$.

Les sources détectées dans l'article sont montrées dans le Tableau 6.9.

sources	Rc/Ra	delta37Cl en ‰	Rc/Ra	delta81Br en ‰
1 ("manteau")	7.76	0.88	8.26	0.75
2 ("subduction")	6.45	-0.43	7.17	-1.03
3 ("croute")	1.68	0.11	1.89	0.26

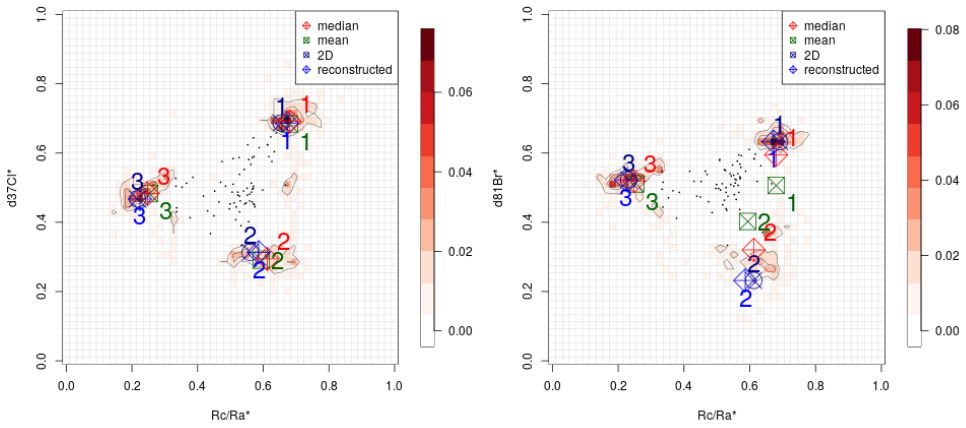
Tableau 6.9 – Sources estimées dans [Pinti et al., 2020] en considérant les données comme bidimensionnelles et résultantes d'un système de mélange à trois sources. Les sources estimées sont les sommets du plus petit triangle qui contient les données.

Pour reconstruire les sources en trois dimensions, le nombre de sources considéré est trois. Les sources sont fusionnées sur la dimension Rc/Ra. La valeur en Rc/Ra de chaque source reconstruite est le barycentre entre les deux sources les plus proches. Le Tableau 6.10 présente les sources reconstruites :

sources	Rc/Ra	delta37Cl en ‰	delta81Br en ‰
1 ("manteau")	8.01	0.88	0.75
2 ("subduction")	6.81	-0.43	-1.03
3 ("croute")	1.78	0.11	0.26

Tableau 6.10 – Sources reconstruites.

Comme précédemment, les 500 dernières configurations de points sauvegardées sont projetées dans les plans d'études (Figure 6.7). Sur chaque plan, il y a trois zones ayant une forte probabilité de contenir une source simulée. Ces sources sont relativement proches des sources détectées par l'autre méthode.



(a) (b)

FIGURE 6.7 – Ensembles de niveaux obtenus pour le premier jeu de données réelles dans le premier plan d'étude (a) et dans le second plan d'étude (b). Les sources détectées par l'autre méthode et les sources reconstruites sont en bleu. Les centres obtenus en appliquant un algorithme des k-moyennes à quatre classes sont en vert. Les points médians de chaque classe sont en rouge.

Pour être intéressantes pour les experts, les sources proposées par le modèle HUG, les points médians des classes, subissent la transformation inverse à la normalisation (Tableau 6.11).

sources	Rc/Ra	delta37Cl en ‰	delta81Br en ‰
1 (“manteau”)	8.12	0.90	0.58
2 (“subduction”)	7.17	-0.50	-0.64
3 (“croute”)	2.12	0.17	0.24

Tableau 6.11 – Sources proposées par le modèle HUG après la transformation inverse à la normalisation.

Les erreurs relatives entre les sources proposées par le modèle HUG est les sources présentées dans [Pinti et al., 2020] sont décrites dans le Tableau 6.12. Les sources proposées par le modèle HUG sont très proches des sources présentées dans [Pinti et al., 2020] : les deux méthodes ont ici détecté les mêmes sources.

indice	Rc/Ra	delta37Cl	delta81Br	erreur moyenne par source
1	1.4	2.3	-22.7	-6.3
2	5.3	16.3	-37.9	-5.4
3	19.1	54.5	-7.7	22.0
erreur moyenne par dimension	8.6	24.4	-22.8	3.4

Tableau 6.12 – Erreur moyenne relative en pourcentage entre les sources présentées et les sources proposées par le modèle HUG.

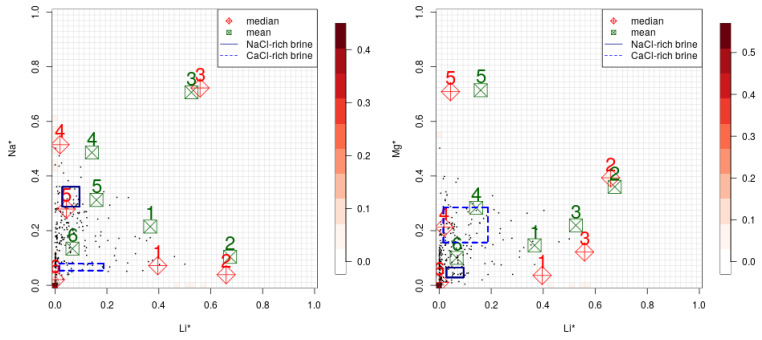
Le second jeu de données réelles est le jeu de données Athabasca présenté dans [Richard et al., 2010], [Richard et al., 2016] et [Martz et al., 2019]. Les données sont issues des prélèvements d'inclusions fluides de gisement d'uranium dans le bassin Athabasca au Canada. Ce jeu de données contient la concentration en parties par million (ppm) de 19 éléments chimiques pour 275 prélèvements. Les concentrations ont été mesurées par spectrométrie à plasma à couplage inductif par ablation au laser (LA-ICP-MS). Parmi ces 19 éléments, huit sont mesurés pour presque tous les prélèvements. Nous ne considérons ici, pour des raisons de temps de calculs, cinq éléments chimiques : lithium (Li), sodium (Na), magnésium (Mg), potassium (K) et calcium (Ca). Il y a donc $L = 5 * 4/2 = 10$ plans d'études.

Dans la littérature, deux sources ont été identifiées : une saumure dominée par le chlorure de sodium notée "saumure NaCl" et une saumure dominée par le chlorure de calcium notée "saumure $CaCl_2$". Les données sont séparées en deux classes, une première si la concentration en sodium est supérieure à 80000 ppm et une seconde si la concentration en sodium est inférieure à 30000 ppm. Pour chaque classe, les quantiles à 25 et 75 pourcent sont représentés dans le Tableau 6.13.

Sources	"saumure NaCl"		"saumure $CaCl_2$"	
	Q25	Q75	Q25	Q75
[Li] en ppm	900	3000	520	6000
[Na] en ppm	80000	100000	15000	22000
[Mg] en ppm	4000	9000	22000	40000
[K] en ppm	1700	5200	8000	17000
[Ca] en ppm	11000	32000	27000	60000

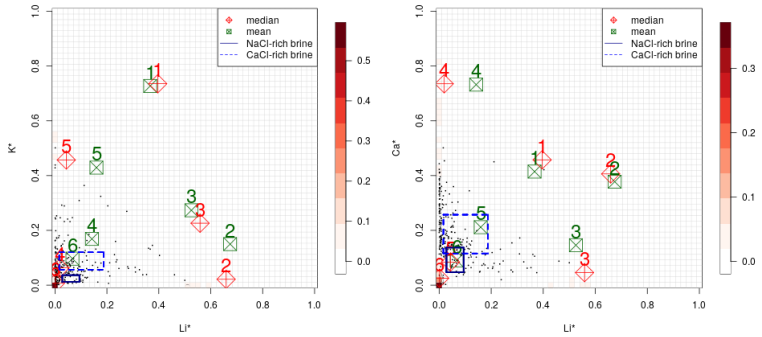
Tableau 6.13 – Quantile à 25% (Q25) et à 75% (Q75) des données dont $[Na] > 80000$ ("saumure NaCl") et des données dont $[Na] < 30000$ ("saumure $CaCl_2$") [Richard et al., 2016].

Les résultats de la méthode HUG sont présentés dans les Figures 6.8 et 6.9. Sur chaque plan d'étude, il y a trois zones qui ont une forte probabilité de contenir une source simulée. Les sources sont ici reconstruites par l'algorithme des k-moyennes séquentiel en considérant trois classes dans chaque plan. L'algorithme reconstruit six sources.

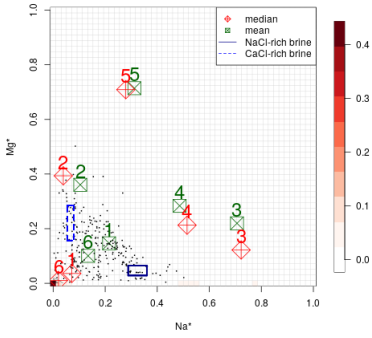


(a)

(b)



(c) (d)



(e)

FIGURE 6.8 – Ensembles de niveaux obtenus pour le second jeu de données réelles dans le premier plan d'étude (a) jusqu'au cinquième plan (e). Les centres obtenus en appliquant un algorithme des k-moyennes à quatre classes sont en vert. Les points médians de chaque classe sont en rouge. Les rectangles bleus symbolisent les quantiles définis dans le Tableau 6.13 pour "saumure NaCl" en traits continus et pour "saumure $CaCl_2$" en traits discontinus.

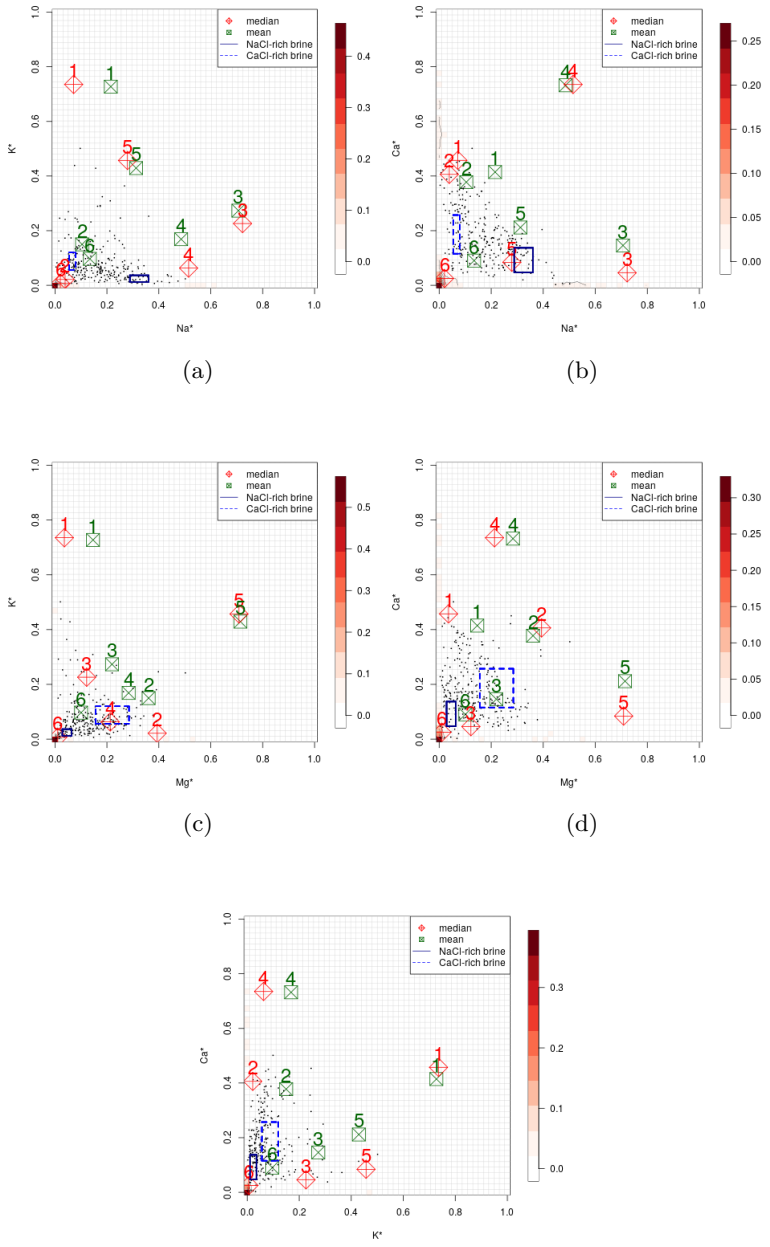


FIGURE 6.9 – Ensembles de niveaux obtenus pour le second jeu de données réelles dans le sixième plan d'étude (a) jusqu'au dixième plan (e). Les centres obtenus en appliquant un algorithme des k-moyennes à quatre classes sont en vert. Les points médians de chaque classe sont en rouge. Les rectangles bleus symbolisent les quantiles définis dans le Tableau 6.13 pour "saumure NaCl" en traits continus et pour "saumure $CaCl_2$" en traits discontinus.

Comme pour la reconstruction dans le cadre du deuxième jeu de données synthétiques, un regroupement hiérarchique est appliqué sur les configurations sauvegardées (Figure 6.10) ainsi que le calcul de la proportion de points dans les six principales classes obtenues en appliquant un algorithme des k-moyennes à sept, huit et neuf classes (Tableau 6.14).

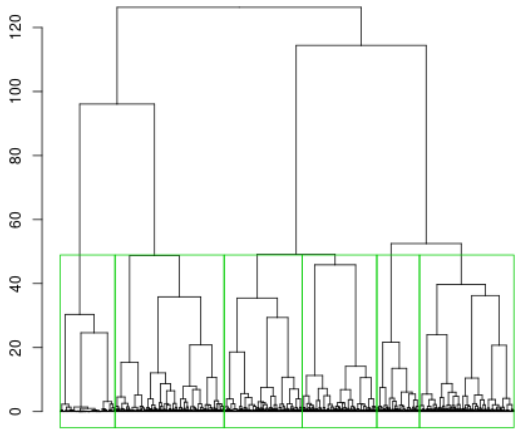


FIGURE 6.10 – Dendrogramme obtenu par un regroupement hiérarchique minimisant la variance intra classe pour le second jeu de données synthétiques. Les six classes sont délimitées par des rectangles verts.

Nombre de classes	7	8	9
Pourcentage	90	79	72

Tableau 6.14 – Pourcentage de sources simulées dans les six principales classes obtenues par un algorithme des k-moyennes à sept, huit et neuf classes.

Comme pour toute application sur des données réelles, il est nécessaire d'appliquer la transformation inverse à la normalisation sur les sources proposées. Les sources proposées sont alors représentées dans l'espace des données non normalisé (Figures 6.11 et 6.12).

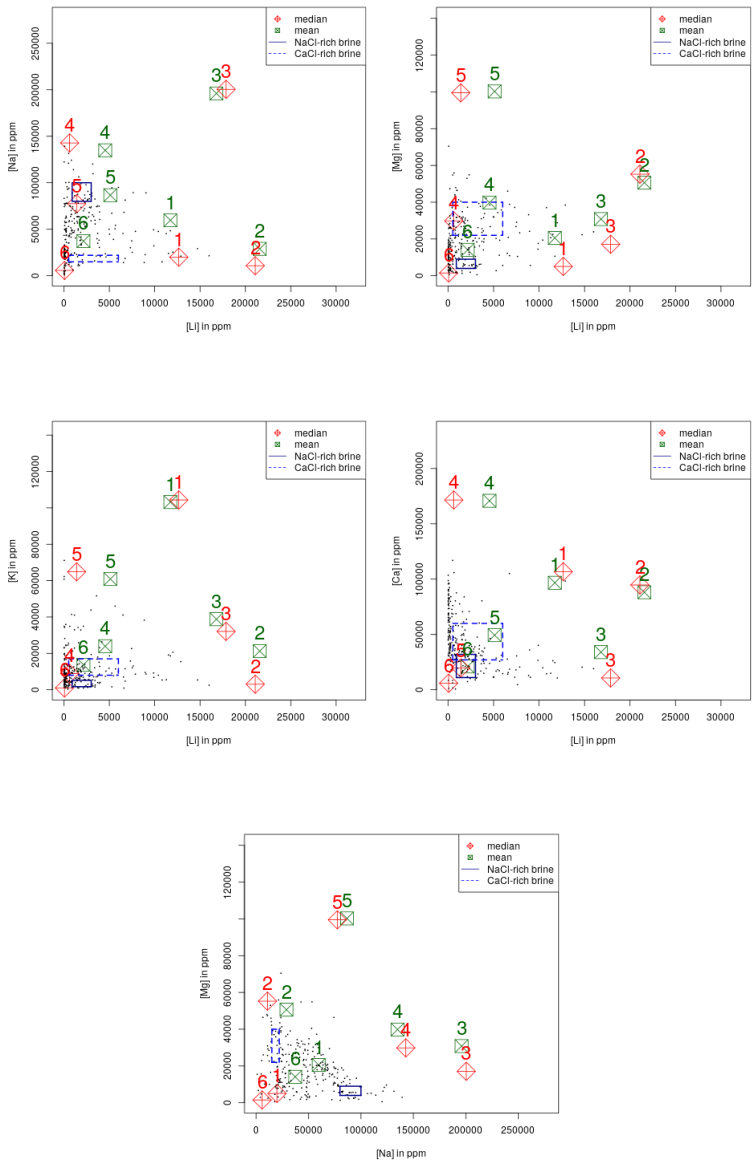


FIGURE 6.11 – Projection des sources proposées par le modèle HUG après la transformation inverse à la normalisation sur les cinq premiers plans. Les rectangles bleus symbolisent les quantiles définis dans le Tableau 6.13 pour "saumure NaCl" en traits continus et pour "saumure $CaCl_2$" en traits discontinus.

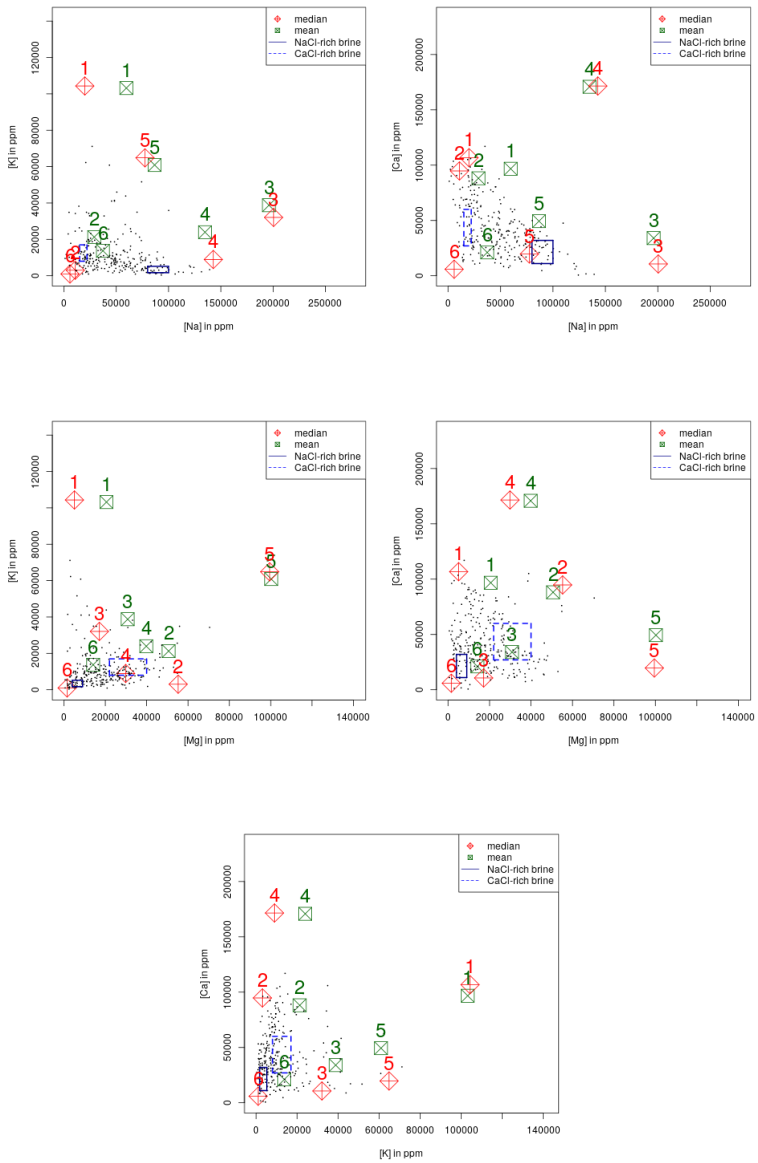


FIGURE 6.12 – Projection des sources proposées par le modèle HUG après la transformation inverse à la normalisation sur les cinq premiers plans. Les rectangles bleus symbolisent les quantiles définis dans le Tableau 6.13 pour “saumure NaCl” en traits continus et pour “saumure $CaCl_2$” en traits discontinus.

La composition des sources proposées est décrite dans le Tableau 6.15.

Sources	[Li] en ppm	[Na] en ppm	[Mg] en ppm	[K] en ppm	[Ca] en ppm
1	12658	19934	5070	104325	106728
2	21084	10813	55239	3051	94757
3	17870	200441	17131	32081	10663
4	624	142790	29854	8944	171537
5	1393	77471	99486	64833	19675
6	57	5744	1524	992	5900

Tableau 6.15 – Sources proposées par le modèle HUG après la transformation inverse à la normalisation.

HUG identifie six sources potentielles alors que deux seulement étaient identifiées précédemment. Il conviendra de vérifier tout d'abord si ces sources potentielles ne sont pas possiblement liées à des biais d'échantillonnage (fluides isolés n'interagissant pas avec les autres sources, échantillonnage partiel de mélanges de fluides dont la composition est trop éloignée de celles d'une des sources,...), des analyses entachées d'erreurs ou encore des réactions chimiques pendant le mélange. S'il s'avère que ces sources potentielles sont plausibles, alors les résultats de HUG invitent à reconsidérer le nombre de sources réellement impliquées et la composition des sources dites "saumure NaCl" et "saumure $CaCl_2$". La discussion sur l'implication géologique de ces résultats dépasse le cadre de cette thèse.

6.2 Estimation des paramètres du modèle HUG : méthode ABC Shadow

La qualité des estimations du modèle dépend du choix du paramètre θ. Ce paragraphe a pour objectif l'estimation des paramètres afin d'améliorer les performances du modèle.

L'algorithme ABC Shadow nécessite l'observation des sources pour estimer le θ. Lorsque les vraies sources sont connues, cette observation est donnée par la vraie configuration de sources $\mathbf{s}^*$. L'estimation de θ peut aussi se faire en considérant que les données observées sont représentées par la configuration de sources détectée.

Nous commençons par appliquer l'algorithme ABC Shadow sur le premier jeu

de données synthétiques : en considérant uniquement les données sur le premier plan d'étude, puis en considérant l'ensemble des dimensions. La détection des sources est par la suite faite en utilisant ces estimations de θ.

L'algorithme est lancé pour générer 10^6 échantillons du vecteur de paramètres θ. À chaque itération, le nombre de mises à jour du vecteur de paramètres est fixé à $N_{ABC} = 200$. La mise à jour de $\theta^{(k-1)}$ consiste à la proposition d'un nouveau vecteur ϕ généré uniformément dans la boule $B(\theta^{(k-1)}, \Delta)$ avec $\Delta = 0.1$. L'état initial du vecteur de paramètres est $\theta^{(0)} = (1, 1, 1, 1)$. Pour que les estimations soient suffisamment décorrélées les unes des autres, les estimations sont sauvegardées toutes les 10^3 itérations de l'algorithme : il en résulte 10^3 échantillons de θ.

La variable auxiliaire $\mathbf{x}$ simulée dans l'étape 1 de l'algorithme est obtenue en appliquant le modèle HUG paramétrisé selon le Tableau 6.16. A chaque itération de l'ABC Shadow, la nouvelle réalisation $\mathbf{x}$ est obtenue à partir du $\mathbf{x}$ précédant.

Variable	Description	Valeur
r	rayon d'interaction	0.01
N	nombre d'applications de l'algorithme SA	1
G	nombre d'applications de la dynamique de Gibbs	L
N_{MH}	nombre d'applications de l'algorithme MH	200
T_0	température initiale	1
c	coefficient de refroidissement	1
$p_b; p_d; p_c$	probabilité de "naissance" ;"mort" ; "changement"	$0.2; 0.2; 0.6$
r_c	rayon de la boule utilisée dans l'évènement "changement"	0.3

Tableau 6.16 – Paramétrisation du modèle HUG pour l'estimation de paramètres.

On considérera que les données sont bidimensionnelles.

La variable auxiliaire $\mathbf{x}$ a un comportement lié aux observations, c'est-à-dire les vraies sources [Caimo and Friel, 2011, Murray et al., 2012, Laporte-Chabasse et al., 2022]. Une première visualisation des résultats peut être obtenue en affichant ses états successifs (Figure 6.13). Les vraies sources sont détectées : quatre zones à forte probabilité, correspondant aux vraies sources, sont visibles sur la Figure 6.13.

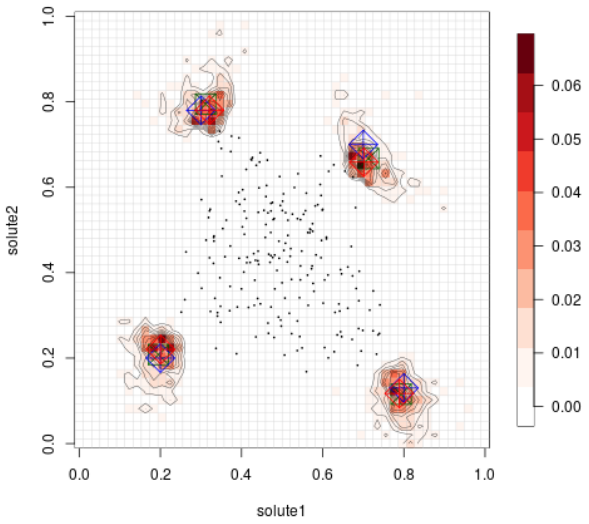


FIGURE 6.13 – Ensembles de niveaux obtenus pour les états successifs $\mathbf{x}$ simulés lors de l'estimation des paramètres pour le jeu de données synthétiques bidimensionnelles. Les vraies sources sont en bleu. Les centres obtenus en appliquant un algorithme des k-moyennes à quatre classes sont en vert. Les points médians de chaque classe sont en rouge.

La Figure 6.14 montre les histogrammes des valeurs de θ.

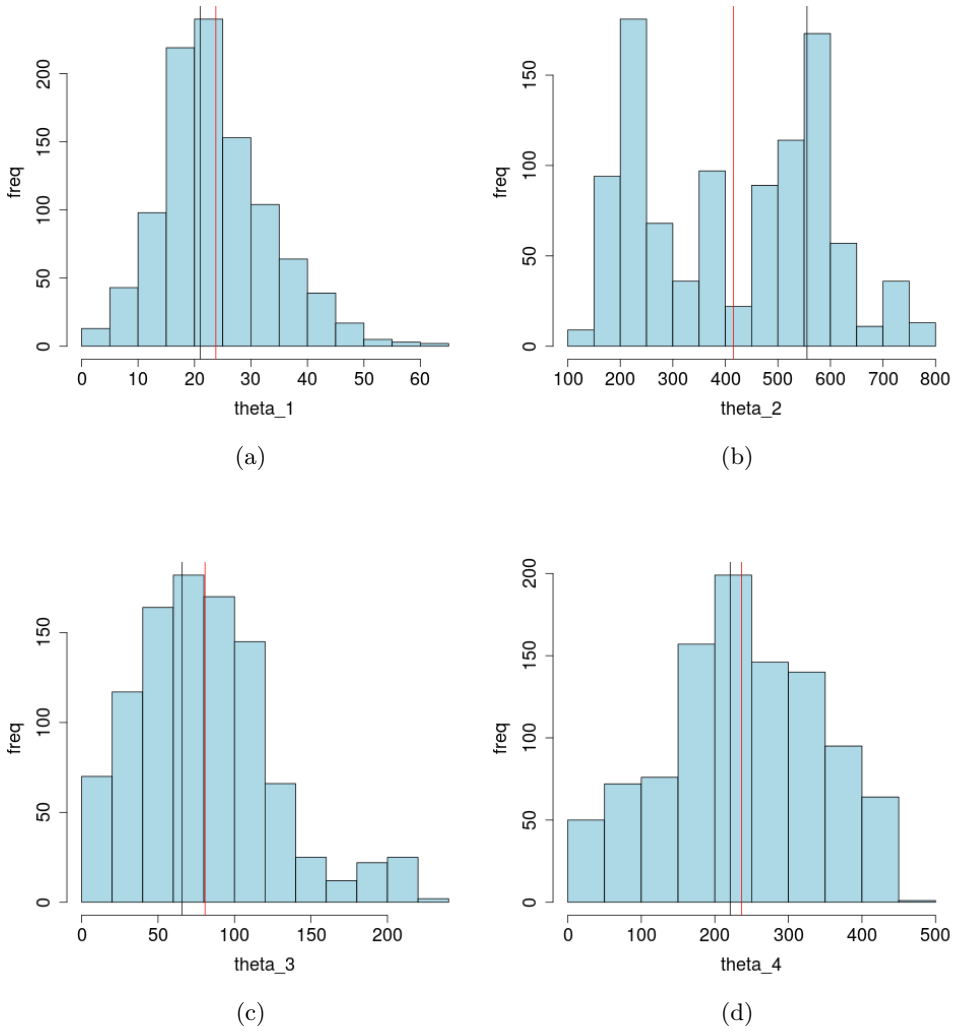
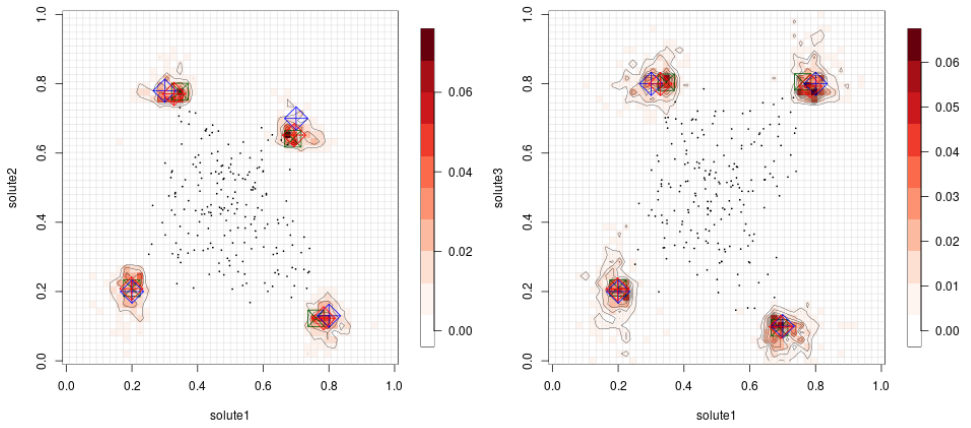


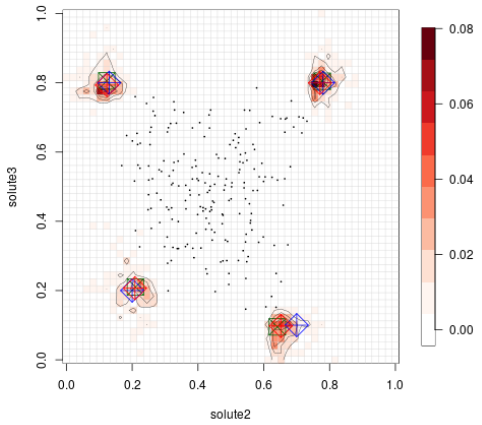
FIGURE 6.14 – Histogramme des valeurs de θ_1 (a), θ_2 (b), θ_3 (c) et θ_4 (d) obtenues par ABC Shadow pour le jeu de données synthétiques bidimensionnelles. La valeur moyenne est représentée en rouge, la valeur modale en noir.

L'estimation de paramètres se fait désormais sur le jeu de données en trois dimensions. Là encore, une première visualisation des résultats peut être obtenue en affichant les états de la variable auxiliaire $\mathbf{x}$ (Figure 6.15). Les vraies sources sont détectées : quatre zones à forte probabilité, correspondant aux vraies sources, sont visibles sur la Figure 6.15.



(a)

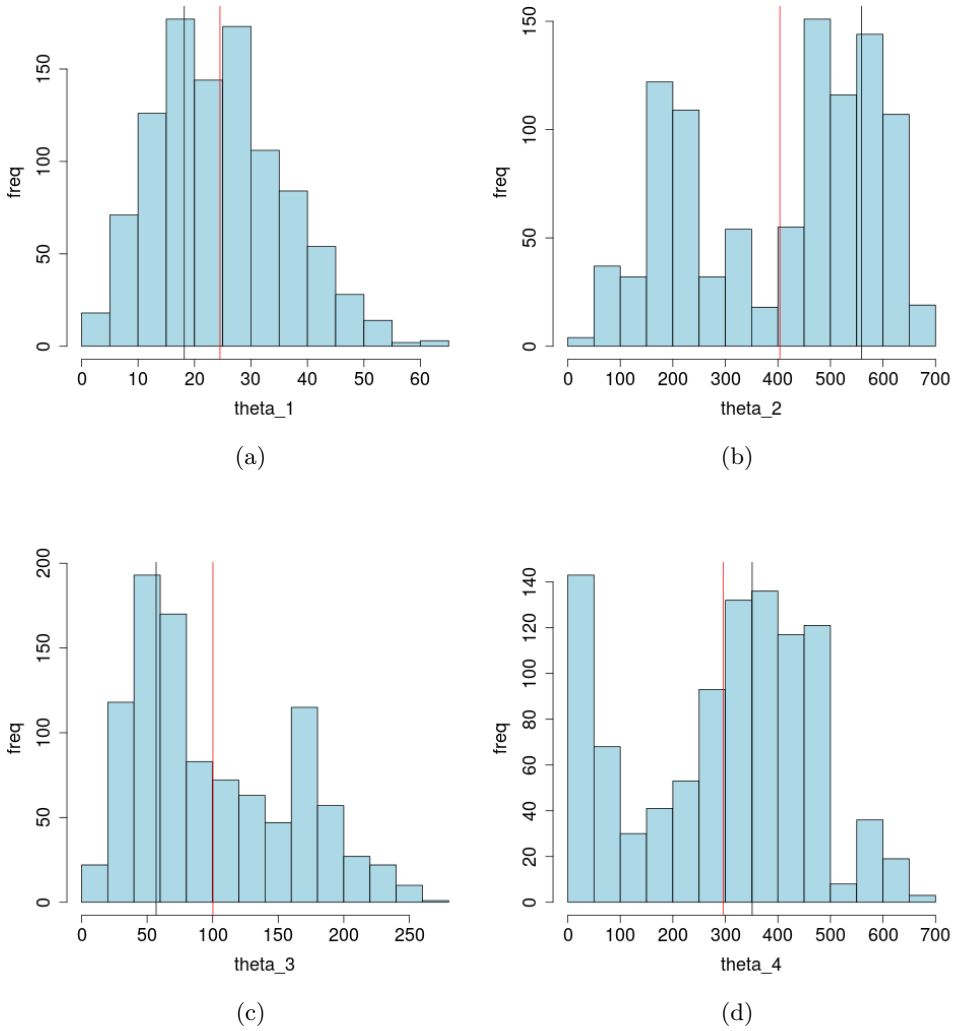
(b)



(c)

FIGURE 6.15 – Ensembles de niveaux obtenus pour les états successifs $\mathbf{x}$ simulés lors de l'estimation des paramètres pour le jeu de données synthétiques tridimensionnelles. Les vraies sources sont en bleu. Les centres obtenus en appliquant un algorithme des k-moyennes à quatre classes sont en vert. Les points médians de chaque classe sont en rouge.

Les histogrammes des valeurs de θ sont représentés sur la Figure 6.16.



(a) (b) (c) (d)

FIGURE 6.16 – Histogramme des valeurs de θ_1 (a), θ_2 (b), θ_3 (c) et θ_4 (d) obtenues par ABC Shadow pour le jeu de données synthétiques tridimensionnelles. La valeur moyenne est représentée en rouge, la valeur modale en noir.

Les valeurs des quantiles, de la moyenne, du mode et de l'écart type des estimations de θ en considérant uniquement le premier plan d'étude (i) et l'ensemble de l'espace des données (ii) sont décrites dans le Tableau 6.17. Les estimations de θ obtenues en considérant les données bidimensionnelles et tridimensionnelles sont similaires.

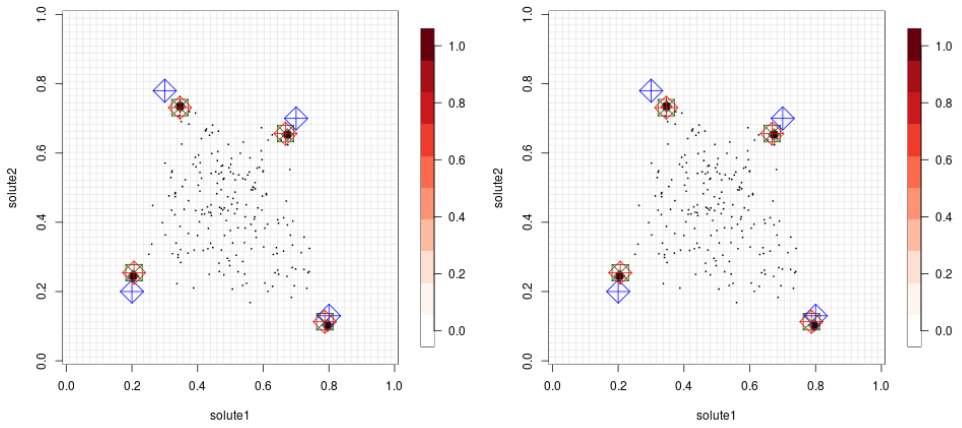
	indice	Q25	Q50	Q75	moyenne	mode	écart type
θ_1	(i)	17.52	22.78	29.45	23.78	21.02	9.59
	(ii)	16.07	23.81	31.74	24.51	18.19	11.32
θ_2	(i)	240.26	444.7	557.87	415.2	555.39	169.6
	(ii)	214.3	466.63	561.57	403.74	559.1	176.74
θ_3	(i)	48.75	75.6	105.75	80.72	65.71	44.95
	(ii)	53.18	79.81	154.36	100.2	56.87	59.39
θ_4	(i)	166.99	233.03	317.68	236.43	220.86	105.55
	(ii)	158.76	321.74	425.6	295.83	351.02	167.88

Tableau 6.17 – Valeurs des quantiles, de la moyenne, du mode et de l'écart type des estimations de θ obtenues par ABC Shadow en considérant uniquement le premier plan d'étude (i) et l'ensemble de l'espace des données (ii).

Nous allons tester quatre candidats pour la loi *a priori* sur θ : un premier avec la moyenne et l'écart type de θ obtenue pour les données bidimensionnelles, le deuxième avec le mode et l'écart type de θ obtenue pour les données bidimensionnelles, le troisième avec la moyenne et l'écart type de θ obtenue pour les données tridimensionnelles, le quatrième avec le mode et l'écart type de θ obtenue pour les données tridimensionnelles. Nous n'utilisons ici le mode, donc la valeur la plus fréquente des coordonnées de θ, afin d'avoir une estimation moins sensible aux valeurs extrêmes, mais aussi plus sensée : cette valeur est en effet celle qui a le plus souvent maximiser la probabilité de densité de la variable auxiliaire $\mathbf{x}$.

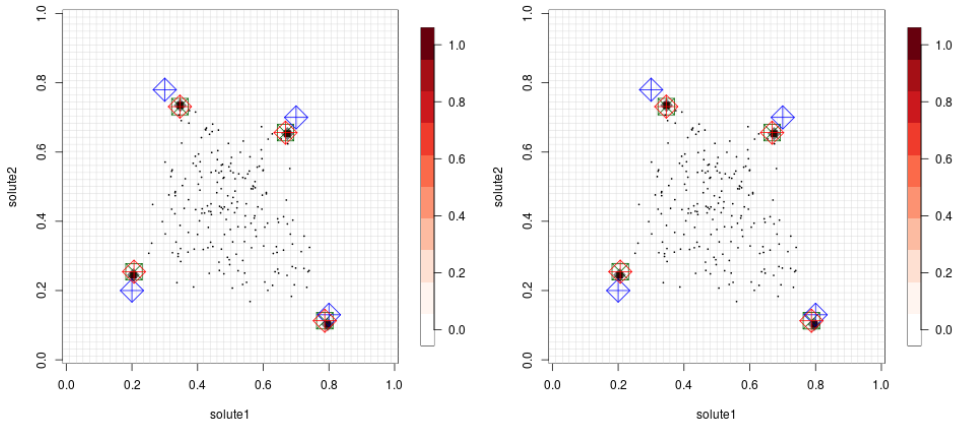
Le modèle HUG est appliqué sur le jeu de données avec ces quatre lois *a priori*.

Lorsque les données sont bidimensionnelles, la détection des sources est la même pour toutes les lois *a priori* (Figure 6.17).



(a)

(b)

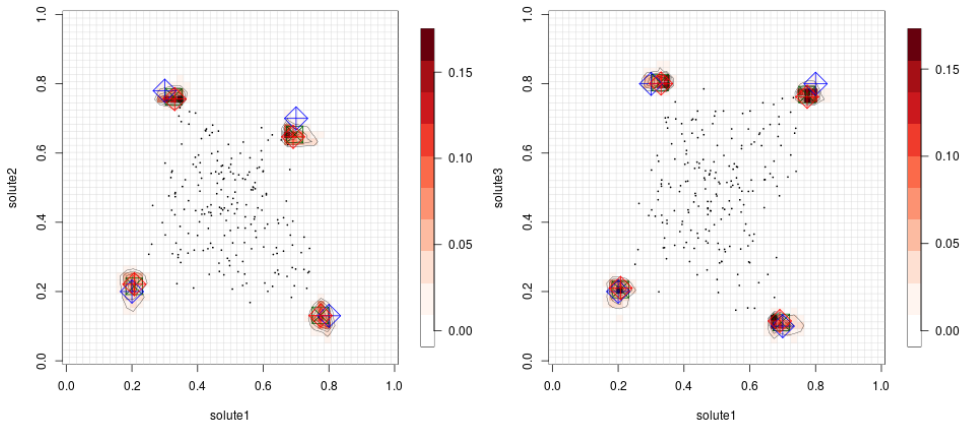


(c)

(d)

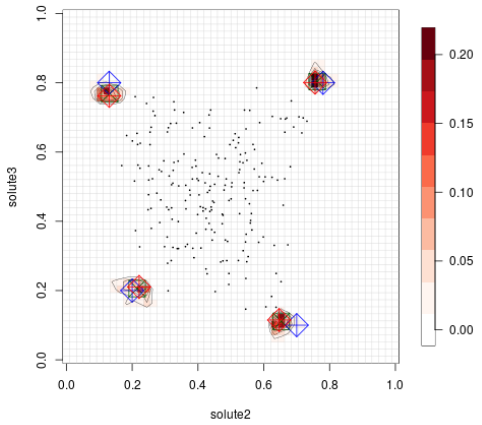
FIGURE 6.17 – Ensembles de niveaux obtenus pour les données bidimensionnelles avec $p(\theta)$ une gaussienne avec la moyenne et l'écart type de θ obtenue pour les données bidimensionnelles (a), le mode et l'écart type de θ obtenue pour les données bidimensionnelles (b), la moyenne et l'écart type de θ obtenue pour les données tridimensionnelles (c) et le mode et l'écart type de θ obtenue pour les données tridimensionnelles (d). Les vraies sources sont en bleu. Les centres obtenus en appliquant un algorithme des k-moyennes à quatre classes sont en vert. Les points médians de chaque classe sont en rouge.

Lorsque les données sont tridimensionnelles, la détection des sources est similaire pour chacune des lois (Figure 6.18-6.21).



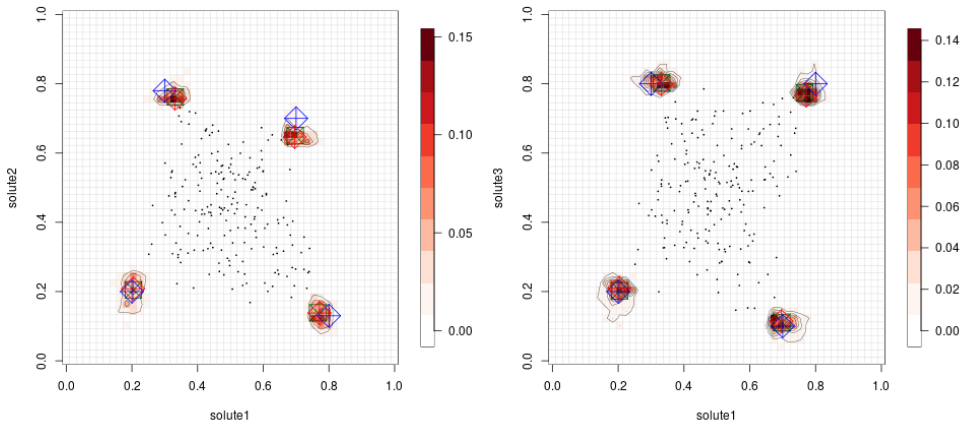
(a)

(b)



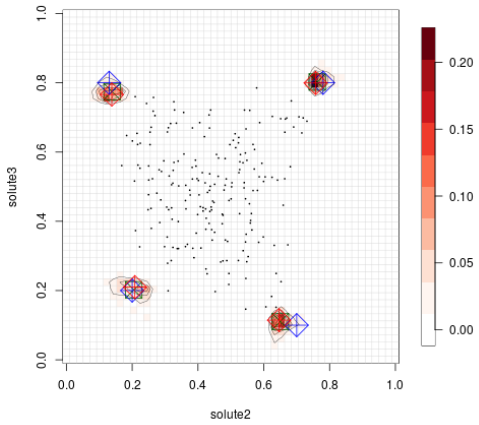
(c)

FIGURE 6.18 – Ensembles de niveaux obtenus pour les données tridimensionnelles avec $p(\theta)$ une gaussienne avec la moyenne et l'écart type de θ obtenue pour les données bidimensionnelles dans le plan un (a), deux (b) et trois (c). Les vraies sources sont en bleu. Les centres obtenus en appliquant un algorithme des k-moyennes à quatre classes sont en vert. Les points médians de chaque classe sont en rouge.



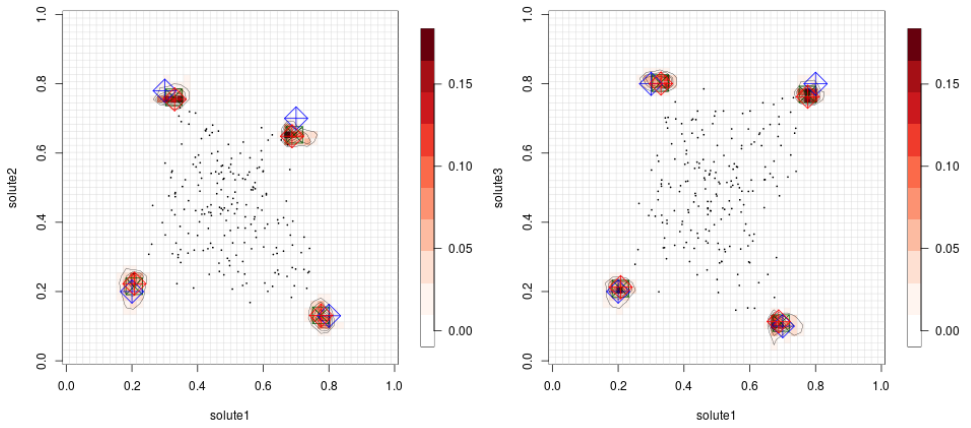
(a)

(b)



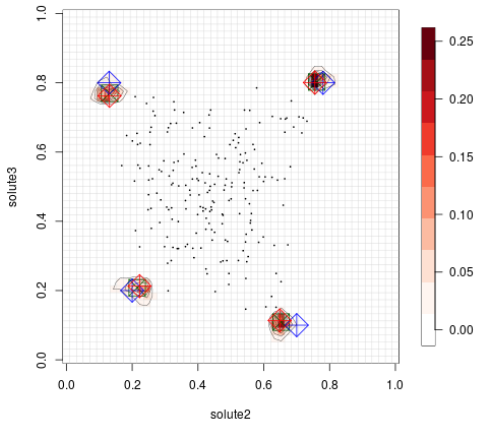
(c)

FIGURE 6.19 – Ensembles de niveaux obtenus pour les données tridimensionnelles avec $p(\theta)$ une gaussienne avec le mode et l'écart type de θ obtenue pour les données bidimensionnelles dans le plan un (a), deux (b) et trois (c). Les vraies sources sont en bleu. Les centres obtenus en appliquant un algorithme des k-moyennes à quatre classes sont en vert. Les points médians de chaque classe sont en rouge.



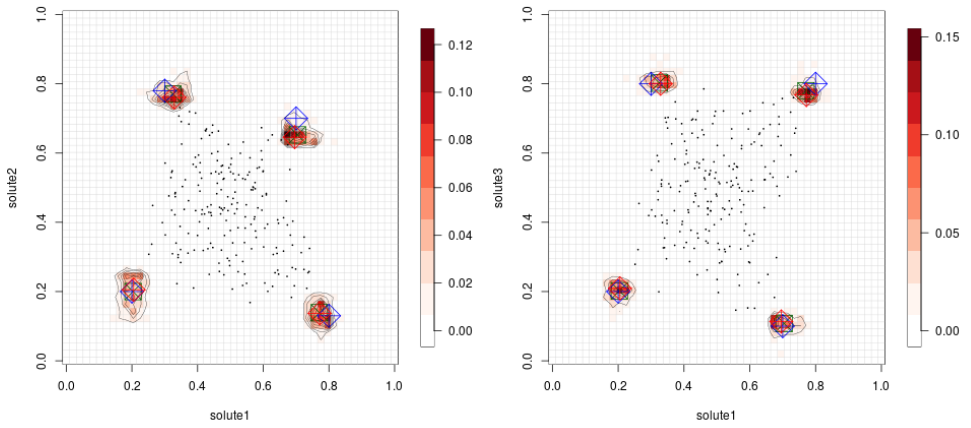
(a)

(b)



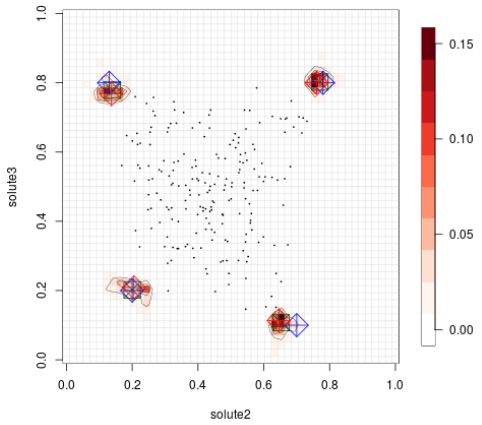
(c)

FIGURE 6.20 – Ensembles de niveaux obtenus pour les données tridimensionnelles avec $p(\theta)$ une gaussienne avec la moyenne et l'écart type de θ obtenue pour les données tridimensionnelles dans le plan un (a), deux (b) et trois (c). Les vraies sources sont en bleu. Les centres obtenus en appliquant un algorithme des k-moyennes à quatre classes sont en vert. Les points médians de chaque classe sont en rouge.



(a)

(b)



(c)

FIGURE 6.21 – Ensembles de niveaux obtenus pour les données tridimensionnelles avec $p(\theta)$ une gaussienne avec le mode et l'écart type de θ obtenue pour les données tridimensionnelles dans le plan un (a), deux (b) et trois (c). Les vraies sources sont en bleu. Les centres obtenus en appliquant un algorithme des k-moyennes à quatre classes sont en vert. Les points médians de chaque classe sont en rouge.

En conclusion, les résultats obtenus en choisissant la moyenne et le mode sont similaires. Nous pouvons donc choisir la valeur moyenne, plus simple à calculer, pour définir la loi *a priori* sur θ. De plus, les résultats obtenus par les lois *a priori* définies à partir des données bidimensionnelles et des données tridimensionnelles sont similaires. Dans cette situation, nous pouvons donc estimer les

paramètres seulement sur les données bidimensionnelles pour avoir une première approximation de la loi *a priori* sur les données multidimensionnelles. Une estimation jointe des sources et des paramètres peut être envisagée en se basant sur les travaux effectués sur les champs de Gibbs [Younes, 1989, Winkler, 2001].

6.3 Analyse des performances de la méthode

Si les sections précédentes montrent comment mettre en œuvre la méthode, comment obtenir des bons résultats et comment choisir ses différents paramètres et réglages en vue de cet objectif, dans cette partie du travail, on souhaite pousser notre méthode à toucher ses limites.

Le modèle HUG et les méthodes de calcul utilisés pour mettre en place une méthodologie d'inférence appropriée ont les propriétés mathématiques nécessaires : stabilité locale et intégrabilité du modèle, convergence des algorithmes construits. Cependant, ces propriétés sont garanties dans un cadre théorique strict, parfois difficilement vérifiables en pratique. C'est pour ces raisons que dans cette partie, l'on va effectuer une analyse par simulation des différentes composantes qui pourraient influencer les performances de la méthode proposée.

6.3.1 Sensibilité par rapport aux données

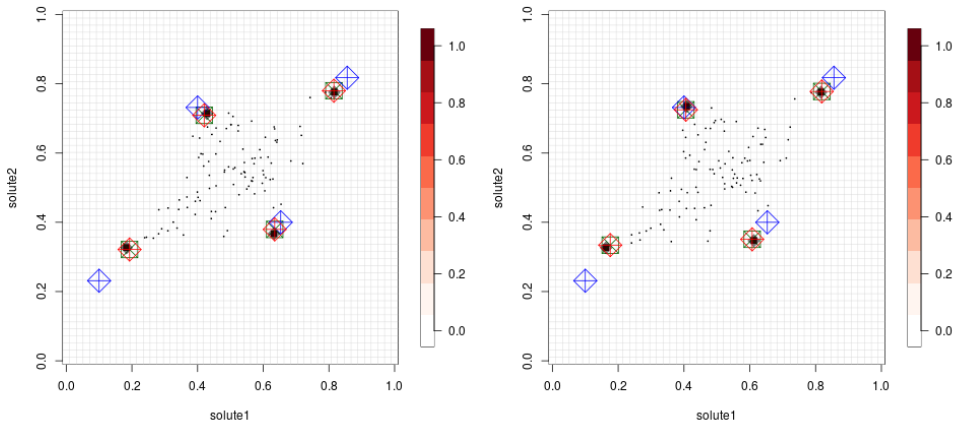
La qualité des données, induite par le prélèvement, les mesures et l'échantillonnage, a un impact direct sur les résultats de la détection. Si le nuage de points ne dessine pas l'enveloppe convexe des sources, alors la composition des sources ou pire le nombre de sources peut être mal estimé. La sensibilité du modèle aux données est examinée sur un nouveau jeu de données synthétiques modélisant le mélange de quatre sources dans un espace bi-dimensionnel. Pour ce faire le nuage de points subit différentes déformations. Les données sont perturbées pour former trois jeux de données supplémentaires : le premier est obtenu en perturbant les données selon une loi $\mathcal{N}(0, 0.01)$, le deuxième selon une loi $\mathcal{N}(0, 0.05)$ et le troisième selon une loi $\mathcal{N}(0, 0.1)$.

Le modèle HUG est appliqué avec comme loi *a priori* pour θ une loi normale définie dans le Tableau 6.18.

	θ_1	θ_2	θ_3	θ_4
moyenne	20	400	100	300
variance	5	10	10	10

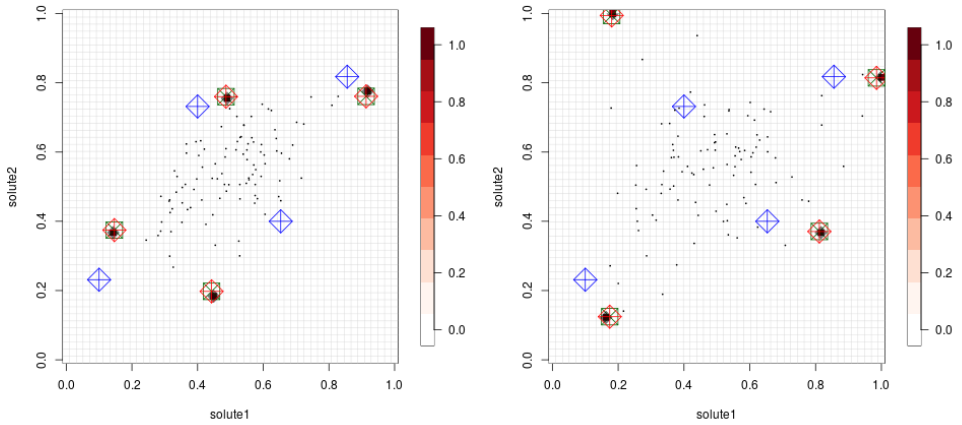
Tableau 6.18 – Paramètres de la loi a priori $p(\theta)$: moyenne et variance d'une loi Gaussienne

Les résultats sur chacun des jeux de données sont représentés dans la Figure 6.22. Bien que les données brutes ne dessinent pas parfaitement l'enveloppe convexe des vraies sources (a), le nombre est la composition des vraies sources sont bien détectées. Lorsque la perturbation ne modifie pas trop la forme des données (b), les sources proposées par le modèle HUG sont relativement les mêmes. Lorsque la perturbation est telle que le nuage de points ne dessine plus l'enveloppe convexe des sources (c) et (d), alors les sources proposées ne permettent pas de détecter les vraies sources.



(a)

(b)



(c)

(d)

FIGURE 6.22 – Ensembles de niveaux obtenus pour le jeu de données brutes (a), le jeu de données perturbé selon une loi $\mathcal{N}(0, 0.01)$ (b), le jeu de données perturbé selon une loi $\mathcal{N}(0, 0.05)$ (c) et le jeu de données perturbé selon une loi $\mathcal{N}(0, 0.1)$ (d). Les vraies sources sont en bleu. Les centres obtenus en appliquant un algorithme k-moyennes à quatre classes sont en vert. Les points médians de chaque classe sont en rouge.

En conclusion, lorsque les nuages de points dessinent l'enveloppe convexe, la détection des sources est la configuration de sources détectée est satisfaisante pour l'œil de l'expert. Ainsi, si l'incertitude sur les données ne déforme pas trop le nuage de points, l'hypothèse sur le caractère ponctuel des données peut être maintenue. Il est aussi nécessaire de supposer que les données sont représentatives

du système de mélange, c'est-à-dire que l'enveloppe convexe des vraies sources soit bien dessinée par les données.

6.3.2 Sensibilité par rapport aux paramètres du modèle

Le choix des paramètres et de la loi a priori étaient faits suivant une stratégie à deux niveaux. En première étape, nous avons simulé des données "semblables" aux observations, en connaissant les sources. Nous avons lancé la méthode de détection pour différentes valeurs des paramètres, et nous avons construit une loi a priori qui prend en compte les caractéristiques de ces paramètres qui donnent globalement des bons résultats.

La deuxième étape a consisté à estimer les paramètres du modèle à partir de la meilleure détection trouvée avec les paramètres ad-hoc, et relancé toute la machine avec ces nouveaux paramètres.

Ce paragraphe a pour but de voir les effets du choix des paramètres sur la qualité de la détection. Pour cela, nous allons considérer un même jeu de données, ici le troisième jeu de données synthétiques utilisé dans le Paragraphe 6.3.1. Pour générer différents paramètres, nous recourons à plusieurs configurations de sources. Pour chacune de ces configurations, les paramètres sont estimés puis l'algorithme SA est appliqué avec ces paramètres.

Les données sont issues d'un mélange de quatre sources. Plusieurs cas de détection sont possibles : le modèle détecte moins de sources que les vraies sources, le modèle détecte autant de sources et le modèle détecte plus de sources. Des configurations de sources à trois, quatre et huit sources sont étudiées.

Pour traiter le premier cas, trois configurations de sources à trois sources sont générées : une configuration expliquant 77% des données, une configuration expliquant 52% des données et une configuration expliquant 0% des données. Les statistiques de ces configurations de sources sont décrites dans le Tableau 6.19.

	$g(\mathbf{s}^*, \mathbf{d})$	$n_e(\mathbf{s}^*, \mathbf{d})$	$n(\mathbf{s}^*)$	$n_r(\mathbf{s}^*)$
(a)	0.14	0.23	3	0
(b)	0.02	0.48	3	0
(c)	0.54	1.00	3	0

Tableau 6.19 – Statistiques des configurations de sources à trois sources.

Les paramètres sont estimés par l'algorithme ABC Shadow. Les histogrammes des valeurs de θ sont représentés sur la Figure 6.23, chaque colonne représente les résultats pour une configuration de sources et chaque ligne les estimations d'une coordonnée de θ.

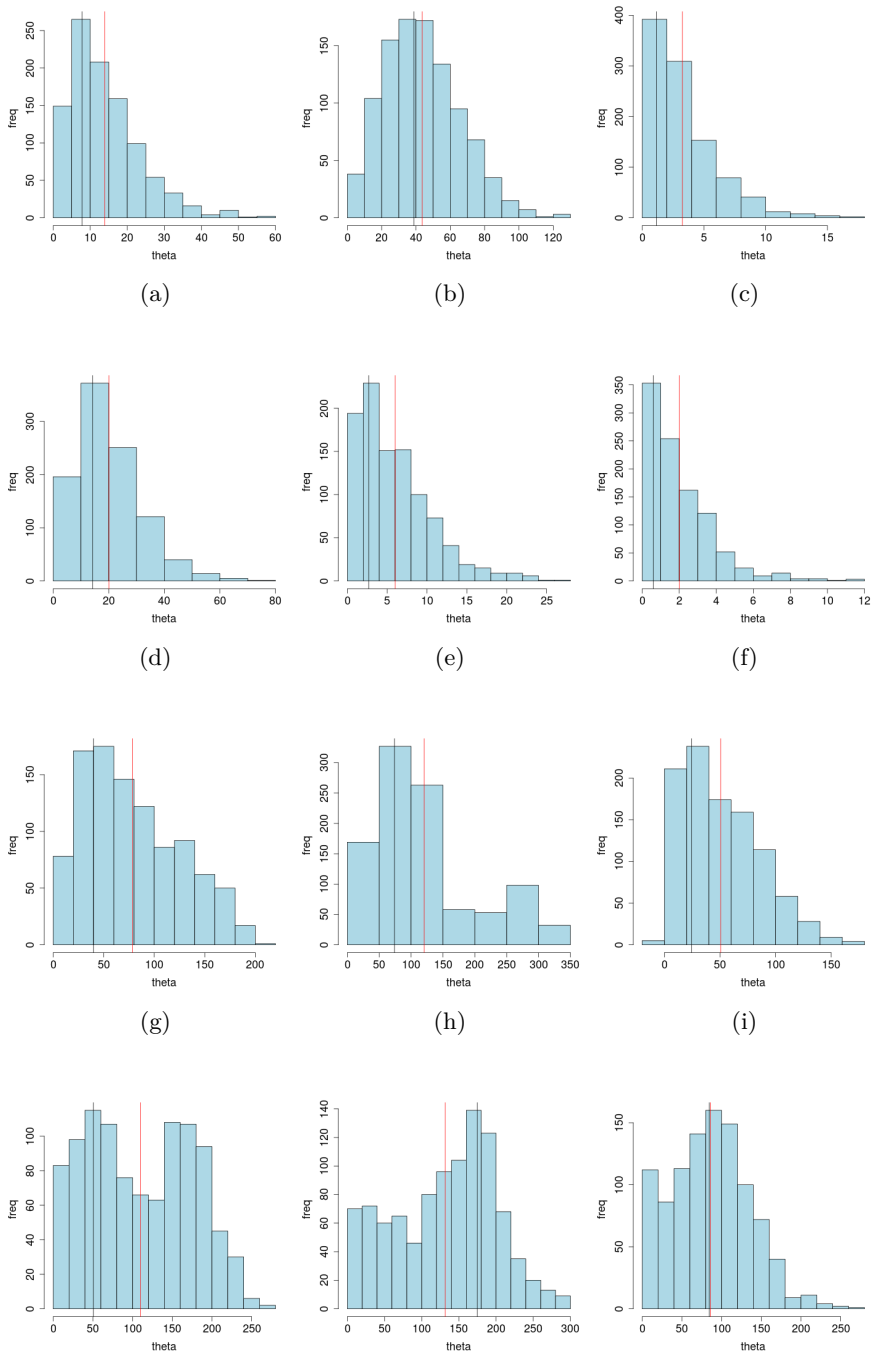


FIGURE 6.23 – Histogramme des valeurs de θ obtenues par ABC Shadow pour une configuration de trois sources qui explique 77% (a,d,g,j), 52% (b,e,h,k) et 0% (c,f,i,l) des données. Chaque colonne représente une configuration de sources et chaque ligne une coordonnée de θ. La valeur moyenne est représentée en rouge, la valeur modale en noir.

Pour traiter le deuxième cas, trois configurations de sources à quatre sources sont générées en plus de la configuration des vraies sources (a) : la première selon une loi $\mathcal{N}(0, 0.01)$ (b), la deuxième selon une loi $\mathcal{N}(0, 0.05)$ (c) et la troisième selon une loi $\mathcal{N}(0, 0.1)$ (d). Les statistiques de ces configurations de sources sont décrites dans le Tableau 6.20.

	$g(\mathbf{s}^*, \mathbf{d})$	$n_e(\mathbf{s}^*, \mathbf{d})$	$n(\mathbf{s}^*)$	$n_r(\mathbf{s}^*)$
(a)	0.38	0	4	0
(b)	0.38	0	4	0
(c)	0.08	0.14	4	0
(d)	0.75	0.04	4	0

Tableau 6.20 – Statistiques des configurations de sources à quatre sources.

Les histogrammes des valeurs de θ sont représentés sur les Figures 6.24, les lignes correspondent aux coordonnées de θ et les colonnes aux configurations de sources.

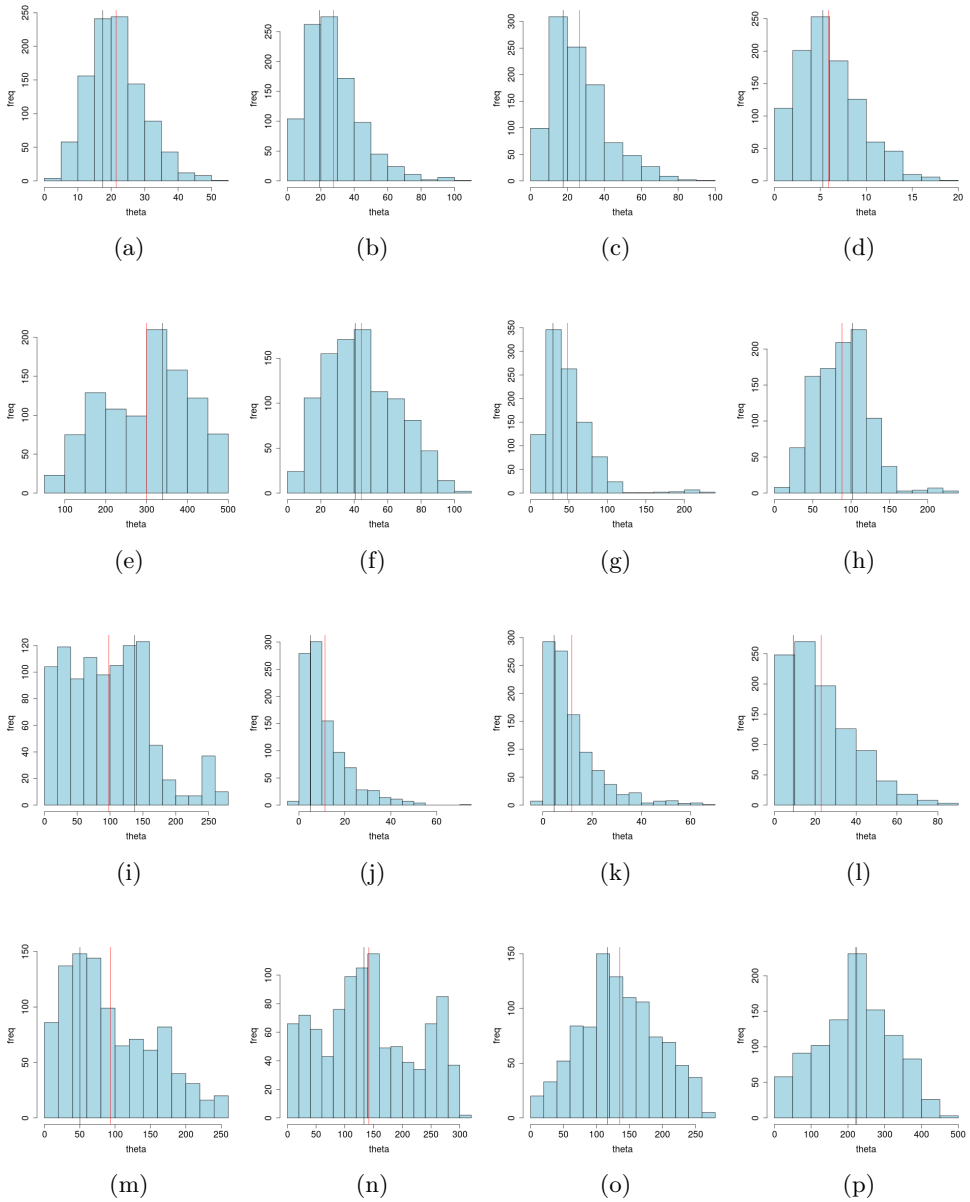


FIGURE 6.24 – Histogramme des valeurs de θ obtenues par ABC Shadow pour les vraies sources (a,e,i,m), les vraies sources perturbées selon une loi $\mathcal{N}(0, 0.01)$ (b,f,j,n), les vraies sources perturbées selon une loi $\mathcal{N}(0, 0.05)$ (c,g,k,o) et les vraies sources perturbées selon une loi $\mathcal{N}(0, 0.1)$ (d,h,l,p). La valeur moyenne est représentée en rouge, la valeur modale en noir.

Pour traiter le troisième cas, deux configurations de sources à huit sources sont générées : la première composée des vraies sources et de quatre sources superflues, la seconde composée d'une vraie source et de sept points aléatoirement répartis.

Les statistiques de ces configurations sont décrites dans le Tableau 6.21.

	$g(\mathbf{s}^*, \mathbf{d})$	$n_e(\mathbf{s}^*, \mathbf{d})$	$n(\mathbf{s}^*)$	$n_r(\mathbf{s}^*)$
(a)	1.96	0	8	0
(b)	1.73	0.01	8	0

Tableau 6.21 – Statistiques des configurations de sources à huit sources.

Les histogrammes des valeurs de θ sont représentés sur les Figures 6.25, les lignes correspondent aux coordonnées de θ et les colonnes aux configurations de sources.

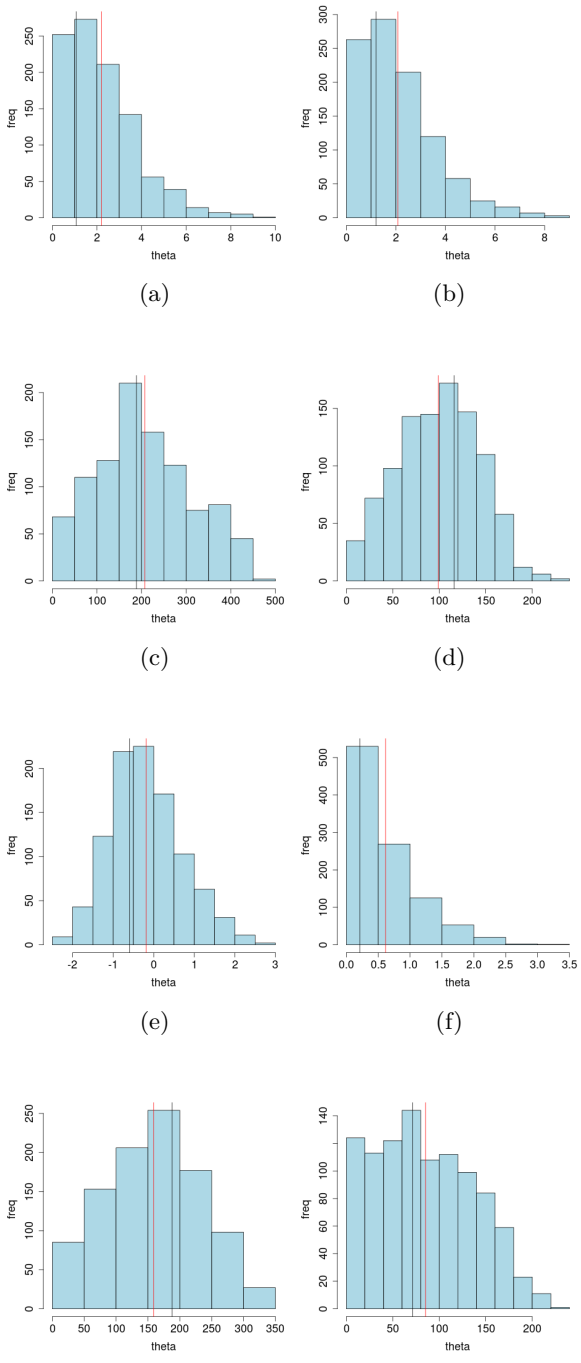


FIGURE 6.25 – Histogramme des valeurs de θ obtenues par ABC Shadow pour une configuration contenant les vraies sources (a,c,e,g) et une configuration contenant uniquement des points aléatoires (b,d,f,h). La valeur moyenne est représentée en rouge, la valeur modale en noir.

Pour faciliter la comparaison des différentes estimations de θ, les valeurs des quantiles, de la moyenne, du mode et de l'écart type des estimations de θ pour toutes les configurations considérées sont décrites dans les Tableaux 6.22 à 6.25.

nombre de sources	indice	Q25	Q50	Q75	moyenne	mode	écart type
3	(a)	7.04	12.01	19.07	13.94	7.81	9.42
	(b)	27.78	41.76	57.81	43.66	38.77	21.84
	(c)	1.24	2.59	4.43	3.26	1.14	2.7
4	(a)	15.56	20.96	26.26	21.49	17.43	8.21
	(b)	16.34	24.84	35.8	27.71	19.19	16.13
	(c)	15.64	23.81	34.29	26.5	17.63	15.08
	(d)	3.38	5.5	7.99	5.91	5.27	3.36
8	(a)	0.73	1.7	2.8	1.96	0.63	1.5
	(b)	0.94	1.83	2.85	2.08	1.2	1.53

Tableau 6.22 – Valeurs des quantiles, de la moyenne, du mode et de l'écart type des estimations de θ_1 obtenues par ABC Shadow pour les différentes configurations de sources considérées.

nombre de sources	indice	Q25	Q50	Q75	moyenne	mode	écart type
3	(a)	11.51	18.41	26.32	20.04	14.2	11.62
	(b)	2.53	4.95	8.32	6	2.68	4.62
	(c)	0.65	1.56	2.88	2.01	0.6	1.79
4	(a)	209.72	317.64	376.45	300.2	338.97	106.67
	(b)	27.55	42.17	59.77	44.32	40.64	21.11
	(c)	26.79	41.8	61.69	47.66	29.13	30
	(d)	62.34	89.26	110.44	88.37	102.26	34.74
8	(a)	137.94	210.33	297.45	213.29	171.32	98.29
	(b)	67.68	100.79	129.17	98.83	115.87	43.52

Tableau 6.23 – Valeurs des quantiles, de la moyenne, du mode et de l'écart type des estimations de θ_2 obtenues par ABC Shadow pour les différentes configurations de sources considérées.

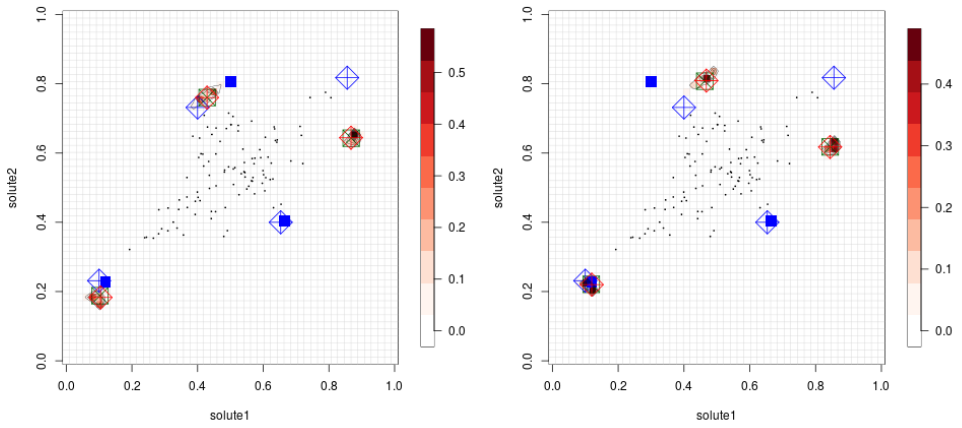
nombre de sources	indice	Q25	Q50	Q75	moyenne	mode	écart type
3	(a)	40.2	71.15	112.31	78.44	39.97	47.06
	(b)	64.47	100.61	145.74	120.53	73.98	80.11
	(c)	22.93	43.78	74.7	50.64	24.33	34.93
4	(a)	44.84	94.73	139.5	97.9	137.14	62.16
	(b)	4.49	8.31	15.65	11.4	5.08	9.83
	(c)	4.21	8.36	15.5	11.73	4.62	10.86
	(d)	10.08	19.26	31.78	22.88	9.35	16.1
8	(a)	0.23	0.56	1	0.71	0.21	0.61
	(b)	0.21	0.47	0.86	0.62	0.21	0.53

Tableau 6.24 – Valeurs des quantiles, de la moyenne, du mode et de l'écart type des estimations de θ_3 obtenues par ABC Shadow pour les différentes configurations de sources considérées.

nombre de sources	indice	Q25	Q50	Q75	moyenne	mode	écart type
3	(a)	51.41	105.99	165.03	109.82	50.7	65.66
	(b)	75.56	141.98	182.33	131.55	174.71	68.65
	(c)	49.1	85.43	118.77	85.9	84.3	49.08
4	(a)	43.54	77.62	140.03	93.48	50.33	62.17
	(b)	81.53	136.35	206.32	141.8	133.18	81.85
	(c)	94.35	131.81	175.91	135.16	116.89	58.65
	(d)	149.79	225.41	294.86	220.76	224.17	100.65
8	(a)	89.02	172.21	274.2	181.26	113.28	110.88
	(b)	41.49	79.47	125.63	85.18	71.16	52.45

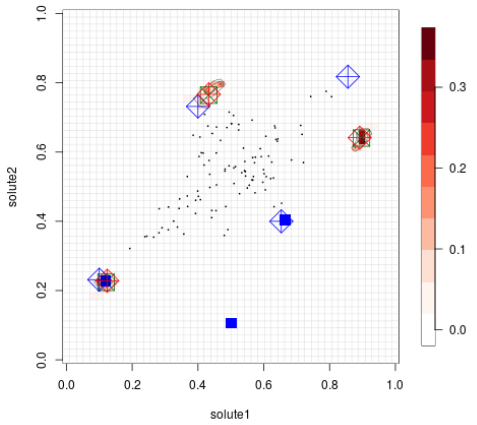
Tableau 6.25 – Valeurs des quantiles, de la moyenne, du mode et de l'écart type des estimations de θ_4 obtenues par ABC Shadow pour les différentes configurations de sources considérées.

Le modèle HUG est maintenant appliqué avec les différentes lois *a priori* gaussiennes sur θ obtenues en prenant pour moyenne la valeur moyenne des estimations et pour écart type l'écart type des estimations. Les résultats sont représentés dans les Figures 6.26 à 6.28. Les sources connues sont en bleu : les sources utilisées pour l'estimation des paramètres par les carrés, les vraies sources par l'autre symbole. À l'instar des figures précédentes, les centres des classes obtenus par un algorithme des k-moyennes à trois classes sont en vert. Les points médians de chaque classe sont en rouge.



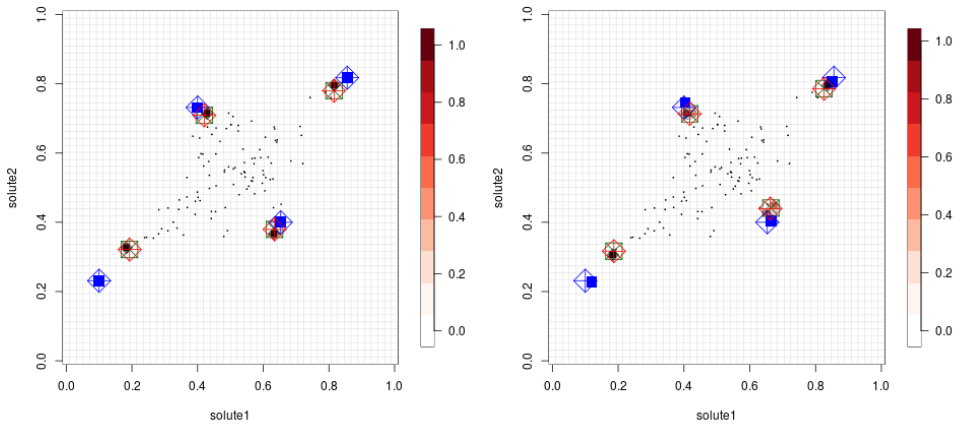
(a)

(b)



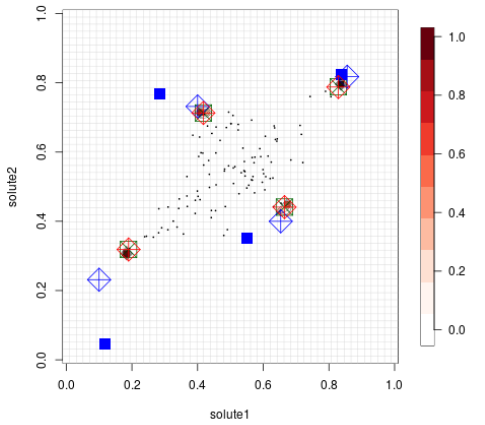
(c)

FIGURE 6.26 – Ensembles de niveaux obtenus avec la loi *a priori* sur θ défini lors de l'estimation des paramètres pour une configuration de trois sources qui explique 77% (a), 52% (b) et 0% (c) des données. Les configurations de sources sont en bleu et les premières sources "détectées" par les carrés bleus. Les centres obtenus en appliquant un algorithme des k-moyennes à trois classes sont en vert. Les points médians de chaque classe sont en rouge.

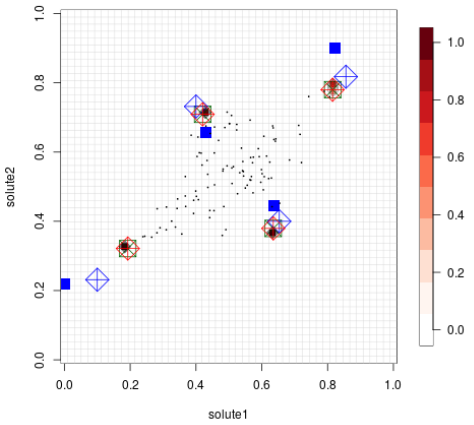


(a)

(b)

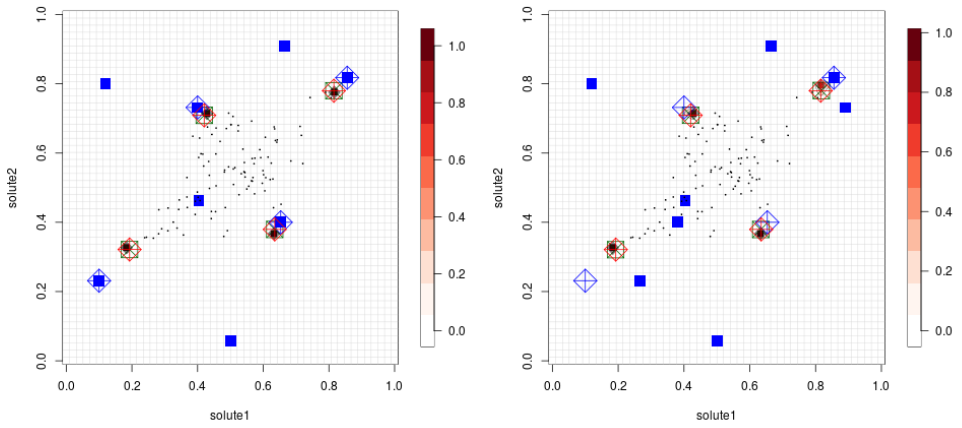


(c)



(d)

FIGURE 6.27 – Ensembles de niveaux obtenus avec la loi *a priori* sur θ défini lors de l'estimation des paramètres pour les vraies sources (a), les vraies sources perturbées selon une loi $\mathcal{N}(0, 0.01)$ (b), les vraies sources perturbées selon une loi $\mathcal{N}(0, 0.05)$ (c) et les vraies sources perturbées selon une loi $\mathcal{N}(0, 0.1)$ (d). Les configurations de points sont en bleu et les premières sources "détectées" par les carrés bleus. Les centres obtenus en appliquant un algorithme des k-moyennes à quatre classes sont en vert. Les points médians de chaque classe sont en rouge.



(a) (b)

FIGURE 6.28 – Ensembles de niveaux obtenus avec la loi *a priori* sur θ défini lors de l'estimation des paramètres pour une configuration de huit sources contenant les vraies sources (a) et une configuration contenant uniquement une vraie source (b). Les configurations de sources sont en bleu et les premières sources "détectées" par les carrés bleus. Les centres obtenus en appliquant un algorithme des k-moyennes à 8 classes sont en vert. Les points médians de chaque classe sont en rouge.

Dans le cas des paramètres estimés par les configurations à quatre et huit sources, le modèle détecte une même configuration : une configuration à quatre sources proches des vraies sources. Dans le cas des paramètres estimés par les configurations à trois sources, le modèle détecte une autre configuration : une configuration de trois sources qui explique 93% des données et dont l'aire de l'enveloppe convexe des données est la même que l'aire de l'enveloppe convexe des sources. Le couple formé par la configuration de sources détectées et le paramètre estimé est un maximum local de la loi jointe $p(\mathbf{s}, \theta|\mathbf{d})$. Cela est dû au fait que les sources utilisées dans l'estimation des paramètres n'expliquent pas suffisamment les données et donc la pénalisation de cette statistique est sous-estimée par rapport à la pénalisation des autres statistiques. Nous pouvons conclure que si la valeur de θ_2 est sous-estimée par rapport aux autres paramètres, alors le modèle aura tendance à détecter des configurations de sources qui contiennent moins de sources que les vraies sources. Néanmoins, même si θ_2 est sous-estimé, le modèle aura toujours tendance à détecter des sources qui expliquent une grande partie des données.

6.3.3 Sensibilité par rapport aux paramètres des différents algorithmes implémentés

La qualité des résultats est fortement liée aux algorithmes utilisés et notamment à leur initialisation.

La sensibilité du modèle HUG par rapport aux algorithmes consiste en trois points : la sensibilité des algorithmes de type Metropolis-Hastings et Gibbs avec Metropolis-Hastings, la sensibilité de la procédure de maximisation de la densité de probabilité et la sensibilité de la procédure d'estimation des paramètres.

Algorithmes de type Metropolis-Hastings et Gibbs avec Metropolis-Hastings

La sensibilité de l'algorithme de type Metropolis-Hastings a été étudiée pour la simulation des processus ponctuels de Poisson et de Strauss dans les exemples 3.3.1 et 3.3.2, respectivement. Différentes valeurs pour le nombre d'itérations de l'algorithme MH et pour les probabilités des évènements ont été utilisées pour en tester les effets.

Pour réduire la corrélation entre les itérations, il faut un nombre important d'itérations de MH noté N_{MH}. Cependant, modifier N_{MH} dans ce sens, augmente également le temps de calcul.

Le choix des probabilités des trois évènements ("naissance", "mort" et "changement") est fait de sorte à favoriser l'évènement "changement". Ceci s'argumente par le fait que le nombre de sources n'est pas très important. À l'équilibre, le nombre de points dans une configuration ne change presque plus : la recherche de la configuration qui maximise la densité de probabilité se fait donc principalement par l'évènement "changement".

6.3.4 Algorithme de type recuit simulé

Le schéma de refroidissement est primordial dans l'optimisation. Le schéma logarithmique qui garantit la convergence de certaines dynamiques est lent et nécessite donc un temps de calcul important. Le schéma polynomial peut être une bonne alternative. La vitesse de convergence dépend du coefficient de re-

froidissement. Un coefficient proche de 1 implique une vitesse de refroidissement lente, tandis qu'un coefficient proche de 0 implique une vitesse refroidissement rapide.

Les effets du choix de la température initiale et du coefficient de refroidissement sont testés dans ce paragraphe.

Dans la suite, l'algorithme SA est appliqué sur le quatrième jeu de données synthétique : la projection du premier jeu de données synthétiques dans le premier plan d'étude. Cet algorithme est initialisé selon le Tableau 6.26 et $p(\theta)$ est une loi gaussienne décrite dans le Tableau 6.27.

Variable	Description	Valeur
L	nombre de plans	1
r	rayon d'interaction	0.01
N	nombre d'applications de l'algorithme SA	$2 * 10^6$
G	nombre d'applications de la dynamique de Gibbs	1
N_{MH}	nombre d'applications de l'algorithme MH	200
$p_b; p_d; p_c$	probabilité de "naissance" ;"mort" ; "changement"	$0.2; 0.2; 0.6$
r_c	rayon de la boule utilisée dans l'évènement "changement"	0.3

Tableau 6.26 – Paramètres de l'algorithme SA pour tester les effets du choix de la température initiale et du coefficient de refroidissement.

Le modèle HUG est appliqué avec comme loi *a priori* pour θ une loi normale définie dans le Tableau 6.18.

	θ_1	θ_2	θ_3	θ_4
moyenne	20	400	100	300
variance	5	10	10	10

Tableau 6.27 – Paramètres de la loi a priori $p(\theta)$: moyenne et variance d'une loi Gaussienne

Lorsque $T_0 = 1$ et $c = 1$, l'algorithme SA génère des configurations de points distribuées selon la densité de probabilité du modèle HUG. Sur la Figure 6.29, quatre zones ont des probabilités plus élevées de contenir des sources, ces zones correspondent avec les vraies sources représentées en bleu. La densité de probabilité du modèle HUG est importante dans les zones proches des vraies sources. Ces

zones sont assez étendues, il faut gagner en précision, d'où la nécessité d'utiliser un algorithme d'optimisation, pour réduire la taille de ces zones.

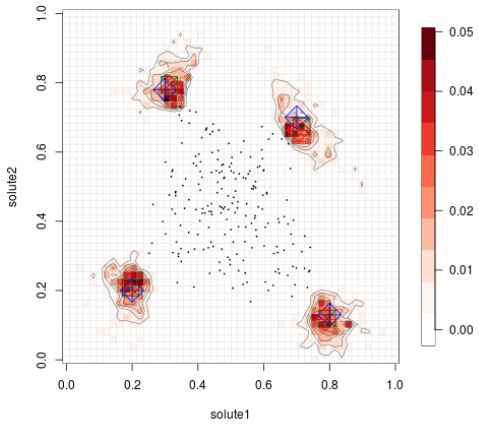


FIGURE 6.29 – Ensembles de niveaux obtenus sur le quatrième jeu de données synthétiques lorsque $T_0 = 1$ et $c = 1$. Les vraies sources sont en bleu. Les centres obtenus en appliquant un algorithme des k-moyennes à quatre classes sont en vert. Les points médians de chaque classe sont en rouge.

Le coefficient de refroidissement est fixé à $c = 0.99999$ et la température initiale est fixée à $T_0 = 1$ (a), $T_0 = 10$ (b), $T_0 = 100$ (c), $T_0 = 1000$ (d), $T_0 = 2 * 10^4$ (e) et $T_0 = 10^5$ (f). L'évolution de la température au cours des simulations est donnée par la Figure 6.30. Pour rappel, les itérations sauvegardées sont espacées de 1000 itérations.

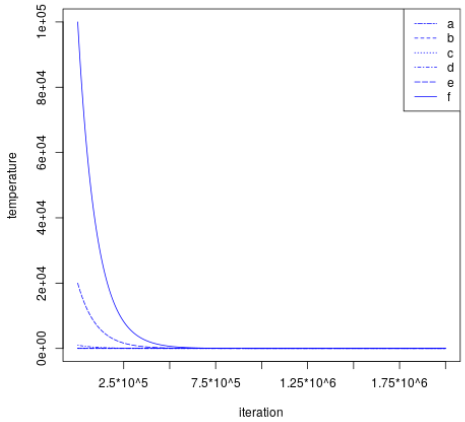


FIGURE 6.30 – Évolution de la température au cours des simulations lorsque $c = 0.99999$ et $T_0 = 1$ (a), $T_0 = 10$ (b), $T_0 = 100$ (c), $T_0 = 1000$ (d), $T_0 = 2 * 10^4$ (e) et $T_0 = 10^5$ (f).

Les résultats des 500 itérations sauvegardées entre les itérations $5 * 10^5$ et 10^6 sont utilisées pour calculer les ensembles de niveaux de la Figure 6.31. Le modèle a convergé vers une configuration de sources pour les températures initiales plus petites que $T_0 = 100$. Pour des températures initiales plus grandes, le modèle n'a pas encore convergé après $5 * 10^5$ itérations. Après 10^6 itérations, la température est à environs $4.5 * 10^{-5}$ (a), $4.5 * 10^{-4}$ (b), $4.5 * 10^{-3}$ (c), $4.5 * 10^{-2}$ (d), $9.1 * 10^{-1}$ (e) et $4.5 * 10$ (f).

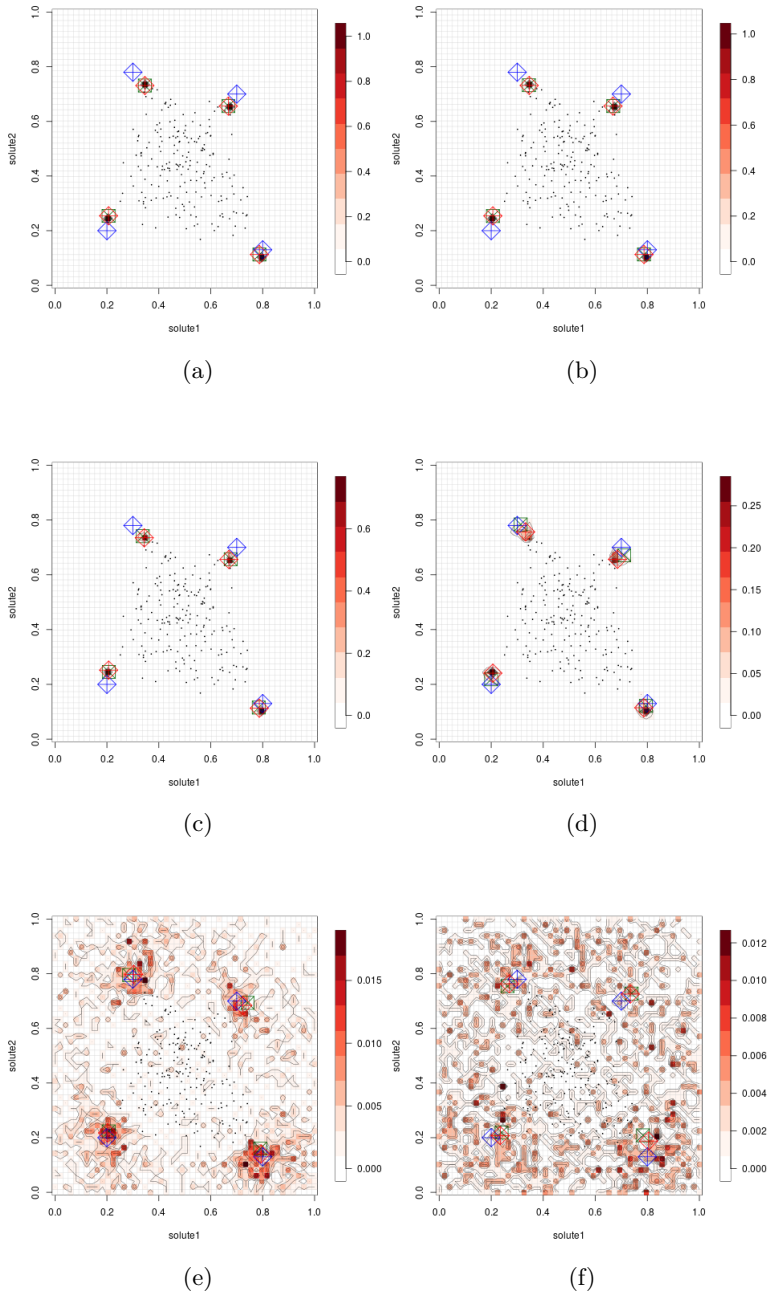
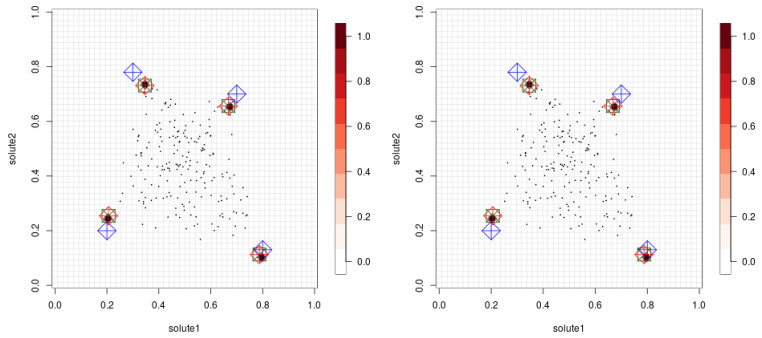


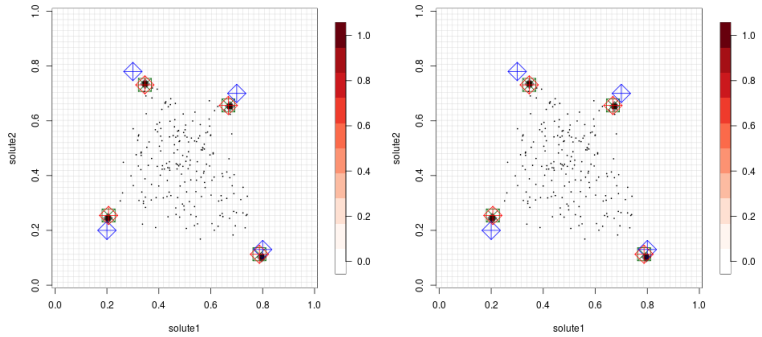
FIGURE 6.31 – Ensembles de niveaux, des 500 itérations sauvegardées entre les itérations $5 * 10^5$ et 10^6, obtenus sur le quatrième jeu de données synthétiques lorsque $c = 0.99999$ et $T_0 = 1$ (a), $T_0 = 10$ (b), $T_0 = 100$ (c), $T_0 = 1000$ (d), $T_0 = 2 * 10^4$ (e) et $T_0 = 10^5$ (f). Les vraies sources sont en bleu. Les centres obtenus en appliquant un algorithme des k-moyennes à quatre classes sont en vert. Les points médians de chaque classe sont en rouge.

Les ensembles de niveaux des 500 dernières itérations sauvegardées sont représentés sur la Figure 6.32. Les résultats obtenus sont les mêmes pour toutes les températures initiales : le modèle a convergé pour toutes les températures initiales. Après $2 * 10^6$ itérations, la température est à environs $2.1 * 10^{-9}$ (a),$2.1 * 10^{-8}$ (b), $2.1 * 10^{-7}$ (c), $2.1 * 10^{-6}$ (d), $4.1 * 10^{-5}$ (e) et $2.1 * 10^{-4}$ (f).



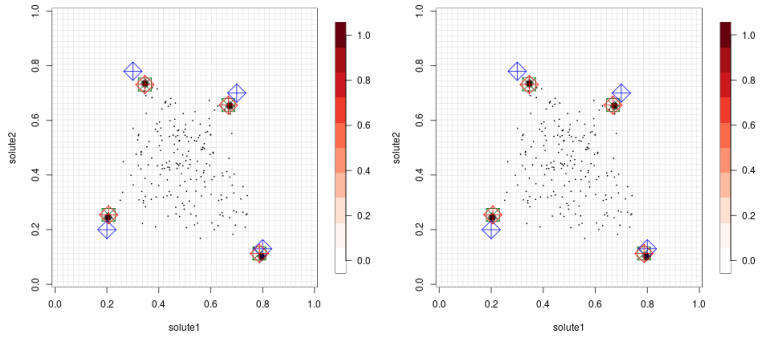
(a)

(b)



(c)

(d)



(e)

(f)

FIGURE 6.32 – Ensembles de niveaux, des 500 itérations sauvegardées entre les itérations $1.5{*}10^6$ et $2{*}10^6$, obtenus sur le quatrième jeu de données synthétiques lorsque $c = 0.99999$ et $T_0 = 1$ (a), $T_0 = 10$ (b), $T_0 = 100$ (c), $T_0 = 1000$ (d), $T_0 = 2 * 10^4$ (e) et $T_0 = 10^5$ (f). Les vraies sources sont en bleu. Les centres obtenus en appliquant un algorithme des k-moyennes à quatre classes sont en vert. Les points médians de chaque classe sont en rouge.

La température initiale est fixée à $T_0 = 1000$ et le coefficient de refroidissement est fixé à $c = 0.99$ (a), $c = 0.999$ (b), $c = 0.9999$ (c) et $c = 0.999995$ (d). L'évolution de la température au cours des simulations est donnée par la Figure 6.33.

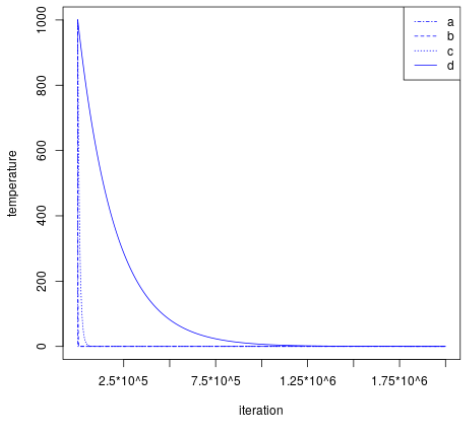


FIGURE 6.33 – Évolution de la température au cours des simulations lorsque $T_0 = 1000$ et $c = 0.99$ (a), $c = 0.999$ (b), $c = 0.9999$ (c) et $c = 0.999995$ (d).

Les résultats des 500 itérations sauvegardées entre les itérations $5*10^5$ et 10^6 sont utilisées pour calculer les ensembles de niveaux de la Figure 6.34. Le modèle a convergé vers une configuration de sources pour les coefficients de refroidissement plus petits que $c = 0.999$. En effet, après 10^6 itérations, la température est à environs $2.4 * 10^{-322}$ (a), $2.5 * 10^{-321}$ (b), $3.7 * 10^{-41}$ (c) et 6.7 (d).

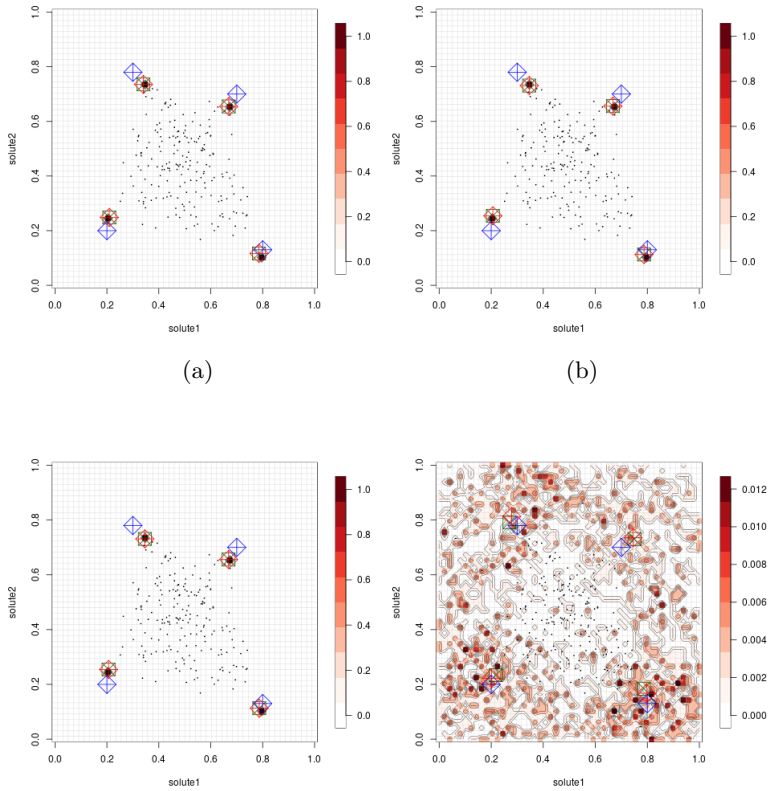


FIGURE 6.34 – Ensembles de niveaux, des 500 itérations sauvegardées entre les itérations $5 * 10^5$ et 10^6, obtenus sur le quatrième jeu de données synthétiques lorsque $T_0 = 1000$ et $c = 0.99$ (a), $c = 0.999$ (b), $c = 0.9999$ (c) et $c = 0.999995$ (d). Les vraies sources sont en bleu. Les centres obtenus en appliquant un algorithme des k-moyennes à quatre classes sont en vert. Les points médians de chaque classe sont en rouge.

Les ensembles de niveaux des 500 dernières itérations sauvegardées sont représentés sur la Figure 6.35. Le modèle converge pour $c = 0.9999$, la température après $2 * 10^6$ est de $4.5 * 10^{-2}$. Dans tous les cas considérés, le modèle a convergé vers la même configuration de sources.

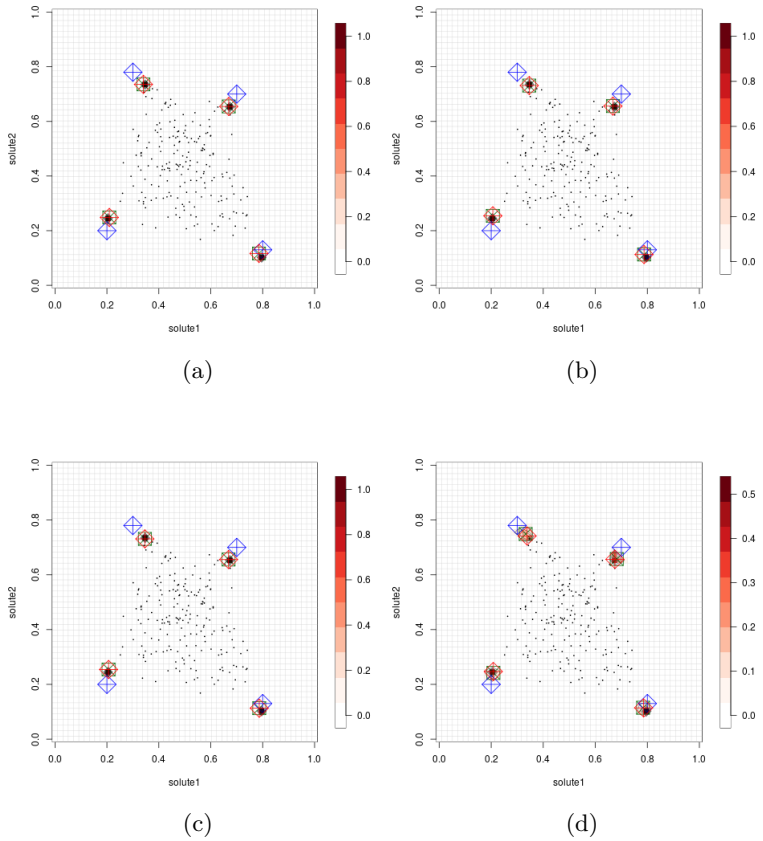


FIGURE 6.35 – Ensembles de niveaux, des 500 itérations sauvegardées entre les itérations $1.5{*}10^6$ et $2{*}10^6$, obtenus sur le quatrième jeu de données synthétiques lorsque $T_0 = 1000$ et $c = 0.99$ (a), $c = 0.999$ (b), $c = 0.9999$ (c) et $c = 0.999995$ (d). Les vraies sources sont en bleu. Les centres obtenus en appliquant un algorithme des k-moyennes à quatre classes sont en vert. Les points médians de chaque classe sont en rouge.

La température initiale et le coefficient de refroidissement ont une influence directe sur le temps de calcul nécessaire pour le modèle converge vers une solution. Dans le cas de ce jeu de données, le modèle converge vers les mêmes solutions pour toutes les températures initiales et tous les coefficients de refroidissement. Cependant, lorsque les jeux de données sont en plus de deux dimensions, le modèle ne converge pas vers une configuration de sources unique, mais vers un ensemble de configurations de sources concentrées autour des sources proposées. Cela vient du fait que la mise à jour de la configuration de sources se fait plan par plan. Et aussi du fait de la nature de la convergence de l'algorithme SA. Ainsi, contrôler la vitesse de refroidissement permet de s'assurer que le modèle converge

bien vers l'ensemble des configurations qui maximisent de manière globale la probabilité de densité du modèle HUG et pas vers un ensemble des configurations qui maximisent de manière locale.

Algorithme ABC Shadow

La sensibilité de l'algorithme ABC Shadow a été étudiée sur des processus ponctuels de Poisson et de Strauss dans les exemples 4.3.1 et 4.3.2. Le rayon de changement de θ, noté Δ et aussi appelé pas, doit être adapté à la plage de valeur possible que peut prendre θ mais aussi au nombre d'itérations. En effet, il faut que toute la plage de valeur puisse être visitée, mais si le pas est trop grand alors l'algorithme risque de ne pas trouver le bon θ. Si le nombre d'applications de l'ABC Shadow, noté N_{ABC}, est trop grand, alors la chaîne de Markov Shadow sera trop éloignée à la chaîne de Markov idéale, conformément au résultat présenté dans [Stoica et al., 2017].

Ces exemples sont complétés par les applications présentées dans [Stoica et al., 2017]. Dans cet article, l'ABC Shadow est utilisé pour estimer les paramètres d'un modèle Gaussien, d'un processus de Strauss, mais aussi du modèle Candy.

Les auteurs de [Stoica et al., 2021] associent l'ABC Shadow à un algorithme de recuit simulé et un résultat de convergence faible est obtenu.

Chapitre 7

Conclusions et perspectives

Ce manuscrit présente une méthode d'analyse de données hydrochimiques inédite dans le domaine via l'utilisation des processus ponctuels.

Les processus ponctuels sont utilisés pour modéliser des phénomènes de répartition d'objets dans l'espace et/ou dans le temps. Leur utilisation est de plus en plus répandue, notamment dans des domaines tels que l'astronomie, pour détecter des filaments cosmiques, la foresterie, pour étudier la variabilité des forêts, ou encore la sismologie.

L'idée clef de ce travail de thèse consiste à aborder les systèmes de mélanges comme des processus qui distribuent des points dans un espace. À partir de cette idée, les apports de cette thèse ont été les suivants :

— la création du modèle HUG, un nouveau processus ponctuel qui intègre des connaissances géologiques pour décrire la distribution des sources dans l'espace induit par les paramètres géologiques,

— la construction d'une dynamique MCMC adaptée pour simuler ce modèle ;

— la construction des outils d'inférence adaptés, capables de mettre en évidence les sources "cachées" dans les données,

— la construction du modèle, les dynamiques de simulation mises en œuvre, les procédures d'inférence ont toutes des propriétés théoriques aidant à garantir la qualité des résultats obtenus,

— une étude par simulation a été conduite pour vérifier les différentes composantes de la méthodologie,

— la validation des résultats sur des données simulées et réelles a été effectuée

en comparant les résultats obtenus par la méthode proposée soit avec la vérité terrain (données simulées) soit avec les résultats présentés dans la littérature (données réelles).

À notre meilleure connaissance, cette solution est la toute première qui peut gérer le fait que le nombre de sources soit inconnu. Est également une première, par rapport à la manière proposée pour gérer l'aspect multidimensionnel, permet de travailler en prenant en compte la totalité des données.

Le modèle proposé intègre sous la forme d'une densité de processus ponctuel avec interaction des caractéristiques des sources généralement admises par les géologues. L'hypothèse concernant l'enveloppe convexe des sources et des prélèvements signifie que le mélange se fait sans réaction chimique. Une perspective envisageable serait d'étudier l'adaptabilité du modèle à ce nouveau contexte.

La dynamique de simulation est un échantillonneur de Gibbs dans lequel un algorithme de type Metropolis-Hastings y est intégré. Idéalement, un algorithme de simulation exacte devrait y être préféré. Cependant, la vitesse de convergence des algorithmes, comme la CFTP ou le clan des ancêtres, dépend fortement des paramètres du modèle en augmentant le temps de calcul d'une manière prohibitive [van Lieshout and Stoica, 2006]. Une perspective plus abordable serait de considérer des stratégies de type "simulated tempering" afin d'obtenir des algorithmes de simulation exacte et des meilleures propriétés de mélange de la chaîne de Markov simulée [Döge et al., 2004].

Le modèle HUG recherche les sources en mettant à jour une configuration de points plan par plan. Du fait des effets de projection, des sources peuvent disparaître d'un plan à l'autre. Cette disparition peut survenir soit parce qu'une source n'est pas détectable par le modèle sur un plan, soit parce que des sources ont fusionné en raison de la proximité de leur projection sur un plan. Une méthode pour reconstruire les sources à partir de leurs projections a été développée Utilisant les résultats de la détection dans chaque plan d'étude, cette méthode applique des techniques de classification de type k-moyennes d'une manière séquentielle pour pouvoir reconstruire la composition des sources à partir de ses projections.

L'algorithme des k-moyennes dépend des conditions initiales. L'ordre des choix des plans de projections influence aussi le résultat final. L'étude de la robustesse de cette méthode est un problème que nous considérons actuellement. Quel que soit le résultat de cette étude, une classification hiérarchique directe de toutes

les configurations des sources devrait fournir les informations nécessaires quand le nombre de classes est proche de celui indiqué par la méthode séquentielle proposée.

La détection de sources utilisant le modèle HUG dépendent le plus fortement de ses paramètres d'interaction. Les paramètres ont été réglés en utilisant une procédure de type essai-erreur. En première étape, des données similaires aux observations ont été simulées. Dans ce cas, les sources sont connues. La méthode de détection y a été lancée. Si les sources détectées correspondaient aux sources connues, ce jeu de paramètres était préservé. Plusieurs jeux de paramètres ont été ainsi sélectionnés. Cette forme de connaissance a été intégrée à la méthode sous la forme d'un prior bayésien. Les résultats obtenus sur les données réelles et simulées ont été obtenus en utilisant cette loi *a priori*. Une fois les sources détectées, une estimation des paramètres a été effectué, et un nouveau prior pour les paramètres a été construit. La détection obtenue de cette manière était meilleure que la précédente, elle montrait des sources moins dispersées. Une étude a été menée pour étudier la robustesse de la détection par rapport aux choix de la loi *a priori*. Plusieurs perspectives méritent être mentionnées. Les résultats de cette étude pourraient intégrer les travaux de [Stoica et al., 2021] qui proposent un algorithme d'estimation de paramètres de type ABC évoluant sur une dynamique de type recuit simulé. Dans cette démarche, il faudrait considérer comme point de départ les travaux classiques de L. Younes [Younes, 1989] et G. Winkler [Winkler, 2001] qui ont abordé le problème d'estimation et détection jointes pour des champs de Markov. Il faudrait sans doute étudier des travaux plus récents, par exemple les auteurs dans [Liu et al., 2021] s'intéressent à la restauration d'un signal et l'estimation des hyper-paramètres, toujours d'une manière simultanée.

Les données réelles sont obtenues par des méthodes de mesure spécifiques. En général, l'incertitude des données est supposée par les géologues suivre une loi normale, qui dépend de la méthode de mesure. Notre étude par rapport à ce problème, nous a rassuré sur la possibilité d'intégration de ces incertitudes dans la méthode existante à l'aide d'une loi *a priori*. Il s'agit là d'une perspective immédiate qui devrait être abordée le plus rapidement possible.

Les incertitudes sur les données, quand elles sont disponibles, peuvent fournir des indications quant à la construction d'un modèle pour les données. Si un tel modèle est disponible, alors on pourrait envisager la construction des procédures de validation des résultats de notre méthode ou de choix de modèles, en utilisant

des tests d'enveloppe à la manière de [Myllymäki et al., 2017, Mrkvička et al., 2017, Myllymäki and Mrkvička, 2019].

Les méthodes construites durant cette thèse ont été développées en langage de programmation C++. Durant la période cette thèse, j'ai aidé au développement du DRLib une librairie C++ en libre accès construite à l'IECL [Gemmerlé et al., 2022]. Une autre perspective immédiate à mentionner est l'intégration des programmes développés pendant la thèse dans cette librairie.

Comme perspective à court terme, nous voulons rédiger un article qui présente l'estimation des paramètres et l'intégration des incertitudes sur les données intitulé "HUG model for source detection in geological data : data uncertainities integration and parameter estimation".

Comme perspective à long terme, il faudrait étudier dans quelle mesure le cadre de travail et de développement proposé par cette thèse pourrait permettre aborder des problèmes d'un grand impact, comme la traçabilité des minéraux.

Bibliographie

[Andrew, 1979] Andrew, A. M. (1979). Another efficient algorithm for convex hulls in two dimensions. *Information Processing Letters*, 9(5) :216–219.

[Arendt et al., 2015] Arendt, C., Aciego, S., and Hetland, E. (2015). An open source Bayesian Monte Carlo isotope mixing model with applications in Earth surface processes. *Geochemistry, Geophysics, Geosystems*, 16(5) :1274–1292.

[Biau et al., 2015] Biau, G., Cérou, F., and Guyader, A. (2015). New insights into Approximate bayesian Computation. *Annales de l'I.H.P. Probabilités et statistiques*, 51(1) :376–403.

[Blake et al., 2018] Blake, W. H., Boeckx, P., Stock, B. C., Smith, H. G., Bodé, S., Upadhayay, H. R., Gaspar, L., Goddard, R., Lennard, A. T., Lizaga, I., et al. (2018). A deconvolutional bayesian mixing model approach for river basin sediment source apportionment. *Scientific reports*, 8(1) :1–12.

[Blum, 2010] Blum, M. G. (2010). Approximate bayesian computation : a nonparametric perspective. *Journal of the American Statistical Association*, 105(491) :1178–1187.

[Caimo and Friel, 2011] Caimo, A. and Friel, N. (2011). Bayesian inference for exponential random graph models. *Social networks*, 33(1) :41–55.

[Christophersen and Hooper, 1992] Christophersen, N. and Hooper, R. P. (1992). Multivariate analysis of stream water chemical data : The use of principal components analysis for the end-member mixing problem. *Water Resources Research*, 28(1) :99–107.

[Christophersen et al., 1990] Christophersen, N., Neal, C., Hooper, R. P., Vogt, R. D., and Andersen, S. (1990). Modelling streamwater chemistry as a mixture of soilwater end-members—a step towards second-generation acidification models. *Journal of Hydrology*, 116(1-4) :307–320.

[Descombes X. (ed.), 2012] Descombes X. (ed.) (2012). *Stochastic geometry for image analysis*. John Wiley & Sons.

[Diggle et al., 2005] Diggle, P., Rowlingson, B., and Su, T.-l. (2005). Point process methodology for on-line spatio-temporal disease surveillance. *Environmetrics : The official journal of the International Environmetrics Society*, 16(5) :423–434.

[Döge et al., 2004] Döge, G., Mecke, K., Møller, J., Stoyan, D., and Waagepetersen, R. P. (2004). Grand canonical simulations of hard-disk systems by simulated tempering. *International Journal of Modern Physics C*, 15(01) :129–147.

[Faure, 1997] Faure, G. (1997). *Principles and applications of geochemistry*, volume 625. Prentice Hall New Jersey, United States,.

[Geman et al., 1990] Geman, D., Geman, S., Graffigne, C., and Dong, P. (1990). Boundary detection by constrained optimization. *IEEE Transactions on pattern analysis and machine intelligence*, 12(7) :609–628.

[Geman and Geman, 1984] Geman, S. and Geman, D. (1984). Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-6(6) :721–741.

[Gemmerlé et al., 2022] Gemmerlé, D., Stoica, R. S., and Reype, C. (2022). DRlib. https ://gitlab.univ-lorraine.fr/labos/iecl/DRlib.

[Geyer, 1999] Geyer, C. J. (1999). Likelihood inference for spatial point processes. In Barndorff-Nielsen, O., Kendall, W., and van Lieshout, M., editors, *Stochastic Geometry, Likelihood and Computation*. CRC Press/Chapman and Hall, Boca Raton.

[Gholami et al., 2019] Gholami, H., Kordestani, M. D., Li, J., Telfer, M. W., and Fathabadi, A. (2019). Diverse sources of aeolian sediment revealed in an arid landscape in southeastern iran using a modified bayesian un-mixing model. *Aeolian Research*, 41 :100547.

[Heinrich et al., 2012] Heinrich, P., Stoica, R. S., and Tran, V. C. (2012). Level sets estimation and Vorob'ev expectation of random compact sets. *Spatial Statistics*, 2 :47–61.

[Hoff III et al., 1999] Hoff III, K. E., Keyser, J., Lin, M., Manocha, D., and Culver, T. (1999). Fast computation of generalized voronoi diagrams using graphics hardware. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, pages 277–286.

[Holley and Phillips, 2022] Holley, E. and Phillips, D. (2022). Solving geological mixing problems with bayesian tracer models : A demonstration of the method applied to carlin-type pyrite. *Journal of Geochemical Exploration*, page 107091.

[Langmuir et al., 1978] Langmuir, C. H., Vocke Jr, R. D., Hanson, G. N., and Hart, S. R. (1978). A general mixing equation with applications to icelandic basalts. *Earth and Planetary Science Letters*, 37(3) :380–392.

[Laporte-Chabasse et al., 2022] Laporte-Chabasse, Q., Stoica, R. S., Clausel, M., Charoy, F., and Oster, G. (2022). Morpho-statistical description of networks through graph modelling and bayesian inference. *IEEE Transactions on Network Science and Engineering*, 9(4) :2123–2138.

[Liu et al., 2021] Liu, Z., Perrodin, M., Chambrion, T., and Stoica, R. (2021). Windowed total variation denoising and noise variance monitoring. *arXiv preprint arXiv :2101.11850*.

[Longman et al., 2018] Longman, J., Veres, D., Ersek, V., Phillips, D. L., Chauvel, C., and Tamas, C. G. (2018). Quantitative assessment of Pb sources in isotopic mixtures using a Bayesian mixing model. *Scientific reports*, 8(1) :6154.

[M. N. M. van Lieshout, 2000] M. N. M. van Lieshout (2000). *Markov Point Processes and their Applications*. Imperial College Press, London.

[Martz et al., 2019] Martz, P., Mercadier, J., Cathelineau, M., Boiron, M.-C., Quirt, D., Doney, A., Gerbeaud, O., De Wally, E., and Ledru, P. (2019). Formation of U-rich mineralizing fluids through basinal brine migration within basement-hosted shear zones : A large-scale study of the fluid chemistry around the unconformity-related Cigar Lake U deposit (Saskatchewan, Canada). *Chemical Geology*, 508 :116–143.

[Meyn and Tweedie, 2009] Meyn, S. P. and Tweedie, R. L. (2009). Markov chains and stochastic stability, cambridge university press.

[Mohler et al., 2011] Mohler, G. O., Short, M. B., Brantingham, P. J., Schoenberg, F. P., and Tita, G. E. (2011). Self-exciting point process modeling of crime. *Journal of the American Statistical Association*, 106(493) :100–108.

[Møller and Nicholls, 1999] Møller, J. and Nicholls, G. K. (1999). *Perfect simulation for sample-based inference*. University of Aarhus. Centre for Mathematical Physics and Stochastics ....

[Møller and Waagepetersen, 2004] Møller, J. and Waagepetersen, R. P. (2004). *Statistical inference and simulation for spatial point processes*. Chapman and Hall/CRC.

[Monfort, 1997] Monfort, A. (1997). *Cours de statistique mathématique*. Economica.

[Moore and Semmens, 2008] Moore, J. W. and Semmens, B. X. (2008). Incorporating uncertainty and prior information into stable isotope mixing models. *Ecology letters*, 11(5) :470–480.

[Mrkvička et al., 2017] Mrkvička, T., Myllymäki, M., and Hahn, U. (2017). Multiple monte carlo testing, with applications in spatial point processes. *Statistics and Computing*, 27(5) :1239–1255.

[Murray et al., 2012] Murray, I., Ghahramani, Z., and MacKay, D. (2012). Mcmc for doubly-intractable distributions. *arXiv preprint arXiv :1206.6848*.

[Myllymäki and Mrkvička, 2019] Myllymäki, M. and Mrkvička, T. (2019). Get : Global envelopes in r. *arXiv preprint arXiv :1911.06583*.

[Myllymäki et al., 2017] Myllymäki, M., Mrkvička, T., Grabarnik, P., Seijo, H., and Hahn, U. (2017). Global envelope tests for spatial processes. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, 79(2) :381–404.

[Ogata, 1998] Ogata, Y. (1998). Space-time point-process models for earthquake occurrences. *Annals of the Institute of Statistical Mathematics*, 50(2) :379–402.

[Papangelou, 1974] Papangelou, F. (1974). The conditional intensity of general point processes and an application to line processes. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 28(3) :207–226.

[Parnell et al., 2010] Parnell, A. C., Inger, R., Bearhop, S., and Jackson, A. L. (2010). Source partitioning using stable isotopes : coping with too much variation. *PloS one*, 5(3) :e9672.

[Parnell et al., 2013] Parnell, A. C., Phillips, D. L., Bearhop, S., Semmens, B. X., Ward, E. J., Moore, J. W., Jackson, A. L., Grey, J., Kelly, D. J., and Inger, R. (2013). Bayesian stable isotope mixing models. *Environmetrics*, 24(6) :387–399.

[Phillips and Gregg, 2001] Phillips, D. L. and Gregg, J. W. (2001). Uncertainty in source partitioning using stable isotopes. *Oecologia*, 127(2) :171–179.

[Pinti et al., 2020] Pinti, D. L., Shouakar-Stash, O., Castro, M. C., Lopez-Hernández, A., Hall, C. M., Rocher, O., Shibata, T., and Ramírez-Montes, M. (2020). The bromine and chlorine isotopic composition of the mantle as revealed by deep geothermal fluids. *Geochimica et Cosmochimica Acta*.

[Pirajno, 2009] Pirajno, F. (2009). *Hydrothermal Processes and Wall Rock Alteration*, pages 73–164. Springer Netherlands, Dordrecht.

[Preparata and Shamos, 1985] Preparata, F. P. and Shamos, M. I. (1985). Convex hulls : Basic algorithms. In *Computational geometry*, pages 95–149. Springer.

[Richard et al., 2016] Richard, A., Cathelineau, M., Boiron, M.-C., Mercadier, J., Banks, D. A., and Cuney, M. (2016). Metal-rich fluid inclusions provide new

insights into unconformity-related U deposits (Athabasca basin and basement, Canada). *Mineralium Deposita*, 51(2) :249–270.

[Richard et al., 2010] Richard, A., Pettke, T., Cathelineau, M., Boiron, M.-C., Mercadier, J., Cuney, M., and Derome, D. (2010). Brine–rock interaction in the Athabasca basement (McArthur river U deposit, Canada) : consequences for fluid chemistry and uranium uptake. *Terra Nova*, 22(4) :303–308.

[Robb, 2005] Robb, R. (2005). Introduction to ore-forming processes, book.

[Rodriguez-Iturbe et al., 1987] Rodriguez-Iturbe, I., De Power, B. F., and Valdes, J. (1987). Rectangular pulses point process models for rainfall : analysis of empirical data. *Journal of Geophysical Research : Atmospheres*, 92(D8) :9645–9656.

[Skuce et al., 2015] Skuce, M., Longstaffe, F., Carter, T., and Potter, J. (2015). Isotopic fingerprinting of groundwaters in southwestern Ontario : Applications to abandoned well remediation. *Applied Geochemistry*, 58 :1–13.

[Stock et al., 2018] Stock, B. C., Jackson, A. L., Ward, E. J., Parnell, A. C., Phillips, D. L., and Semmens, B. X. (2018). Analyzing mixing systems using a new generation of bayesian tracer mixing models. *PeerJ*, 6 :e5096.

[Stoica, 2014] Stoica, R. (2014). Modélisation probabiliste et inférence statistique pour l'analyse des données spatialisées. *Research Habilitation Thesis, Université Lille*, 1.

[Stoica et al., 2021] Stoica, R., Deaconu, M., Philippe, A., and Hurtado-Gil, L. (2021). Shadow simulated annealing : A new algorithm for approximate Bayesian inference of Gibbs point processes. *Spatial Statistics*, page 100505.

[Stoica et al., 2004] Stoica, R., Descombes, X., and Zerubia, J. (2004). A Gibbs point process for road extraction from remotely sensed images. *International Journal of Computer Vision*, 57(2) :121–136.

[Stoica et al., 2007a] Stoica, R., Gay, E., and Kretzschmar, A. (2007a). Cluster pattern detection in spatial data based on Monte Carlo inference. *Biometrical Journal : Journal of Mathematical Methods in Biosciences*, 49(4) :505–519.

[Stoica et al., 2005] Stoica, R., Gregori, P., and J. Mateu, J. (2005). Simulated annealing and object point processes : tools for analysis of spatial patterns. *Stochastic Processes and their Applications*, 115 :1860–1882.

[Stoica et al., 2007b] Stoica, R., Martínez, V. J., and Saar, E. (2007b). A three-dimensional object point process for detection of cosmic filaments. *Journal of the Royal Statistical Society : Series C (Applied Statistics)*, 56(4) :459–477.

[Stoica et al., 2017] Stoica, R. S., Philippe, A., Gregori, P., and Mateu, J. (2017). Abc shadow algorithm : a tool for statistical analysis of spatial patterns. *Statistics and computing*, 27(5) :1225–1238.

[Stoyan et al., 2013] Stoyan, D., Kendall, W. S., Chiu, S. N., and Mecke, J. (2013). *Stochastic geometry and its applications*. John Wiley & Sons.

[Stoyan and Penttinen, 2000] Stoyan, D. and Penttinen, A. (2000). Recent applications of point process methods in forestry statistics. *Statistical science*, pages 61–78.

[Szu and Hartley, 1987] Szu, H. and Hartley, R. (1987). Fast simulated annealing. *Physics letters A*, 122(3-4) :157–162.

[Tang et al., 2012] Tang, M., Zhao, J.-y., Tong, R.-f., and Manocha, D. (2012). Gpu accelerated convex hull computation. *Computers & Graphics*, 36(5) :498–506.

[van Lieshout, 1994] van Lieshout, M. (1994). Stochastic annealing for nearest-neighbour point processes with application to object recognition. *Advances in Applied Probability*, pages 281–300.

[van Lieshout and Stoica, 2006] van Lieshout, M. and Stoica, R. S. (2006). Perfect simulation for marked point processes. *Computational statistics & data analysis*, 51(2) :679–698.

[van Lieshout and van Zwet, 2001] van Lieshout, M. and van Zwet, E. (2001). Exact sampling from conditional boolean models with applications to maximum likelihood inference. *Advances in Applied Probability*, 33(2) :339–353.

[Weltje, 1997] Weltje, G. J. (1997). End-member modeling of compositional data : Numerical-statistical algorithms for solving the explicit mixing problem. *Mathematical Geology*, 29(4) :503–549.

[Winkler, 2001] Winkler, G. (2001). A stochastic algorithm for maximum likelihood estimation in imaging. *Statistics & Risk Modeling*, 19(2) :101–120.

[Winkler, 2003] Winkler, G. (2003). *Image analysis, random fields and Markov chain Monte Carlo methods : a mathematical introduction*, volume 27. Springer Science & Business Media.

[Yardley and Bodnar, 2014] Yardley, B. W. and Bodnar, R. J. (2014). Fluids in the continental crust. *Geochemical Perspectives*, 3(1) :1–2.

[Younes, 1989] Younes, L. (1989). Parametric inference for imperfectly observed Gibbsian fields. *Probability theory and related fields*, 82(4) :625–645.

Résumé

L'analyse de données hydrogéochimiques a pour objectif d'améliorer la compréhension des échanges de matières entre sol et du sous-sol. Ce travail se concentre sur l'étude des interactions fluides-fluides au travers des systèmes de mélange de fluides et plus particulièrement de la détection des compositions des sources du mélange. La détection se fait au moyen d'un processus ponctuel : le modèle proposé est non supervisée et applicable à des données multidimensionnelles.

Les connaissances physiques sur les mélanges et géologiques sur les données sont directement intégrés dans la densité de probabilité d'un processus ponctuel de Gibbs, qui distribue des configurations de points dans l'espace des données, appelé le modèle HUG. Les sources détectées forment la configuration de points qui maximise la densité de probabilité du modèle HUG. Cette densité de probabilité est connue à une constante de normalisation près. La connaissance sur les paramètres du modèle, qu'elle soit acquise d'une manière expérimentale ou bien en utilisant des méthodes d'inférence, y est intégrée sous forme des lois *a priori*. La configuration des sources est obtenue par un algorithme de type recuit simulé et des méthodes de type Monte-Carlo par Chaînes de Markov (MCMC). Les paramètres du modèle sont estimés par une méthode de calcul bayésien approximatif (ABC).

Dans un premier temps, le modèle est appliqué sur des données synthétiques, et après sur des données réelles. Les paramètres du modèle sont estimés ensuite pour un jeu de données synthétiques avec les sources connues. Enfin, la sensibilité du modèle aux données, aux paramètres et aux algorithmes est étudiée.

Mots-clés: processus ponctuels, processus ponctuel de Gibbs, analyse bayésienne, détection de sources, modèle de mélange, modèle non supervisé de détection de sources, modèle "HUG", analyse de données hydrogéochimiques, estimation paramétrique, dynamique MCMC, recuit simulé, k-moyennes séquentiel.

Abstract

The analysis of hydrogeochemical data aims to improve the understanding of mass transfer in the sub-surface and the Earth's crust. This work focuses on the study of fluid-fluid interactions through fluid mixing systems, and more particularly on the detection of the compositions of the mixing sources. The detection is done by means of a point process : the proposed model is unsupervised and applicable to multidimensional data.

Physical knowledge of the mixtures and geological knowledge of the data are directly integrated into the probability density of a Gibbs point process, which distributes point patterns in the data space, called the HUG model. The detected sources form the point pattern that maximises the probability density of the HUG model. This probability density is known up to the normalisation constant. The knowledge related to the parameters of the model, either acquired experimentally or by using inference methods, is integrated in the method under the form of prior distributions. The configuration of the sources is obtained by a simulated annealing algorithm and Markov Chain Monte Carlo (MCMC) methods. The parameters of the model are estimated by an approximate Bayesian computation method (ABC).

First, the model is applied to synthetic data, and then to real data. The parameters of the model are then estimated for a synthetic data set with known sources. Finally, the sensitivity of the model to data uncertainties, to parameters choices and to algorithms set-up is studied.

Keywords: point process, Gibbs point process, bayesian analysis, source detection, mixing model, unsupervised source detection model, HUG model, parameter estimation, MCMC dynamics, simulated annealing, sequential k-means.