N° d'ordre : 42125

UNIVERSITÉ LILLE 1 - SCIENCES ET TECHNOLOGIES
ECOLE DOCTORALE DES SCIENCES POUR L'INGÉNIEUR

THÈSE

présentée en vue d'obtenir le grade de :

DOCTEUR

Spécialité : Automatique, Génie Informatique,
Traitement du Signal et de l'Image

Par :

M. Harizo RAJAONA

Doctorat délivré par l'Université Lille 1 - Sciences et Technologies

Inférence bayésienne adaptative pour la reconstruction de source en dispersion atmosphérique

Soutenue le 21 novembre 2016 devant le jury d'examen :

M. Patrick Bas	Directeur de Recherche CNRS, Ecole Centrale de Lille	*Président du Jury*
M. Thierry Chonavel	Professeur, Télécom Bretagne	*Rapporteur*
M. Hichem Snoussi	Professeur, Université de Technologie de Troyes	*Rapporteur*
M. Lionel Soulhac	Maître de Conférences, Ecole Centrale de Lyon	*Examinateur*
M. Patrick Armand	Expert senior, CEA	*Examinateur*
M. François Septier	Maître de Conférences, Télécom Lille / CRIStAL	*Directeur de thèse*
M. Armand Albergel	Directeur Général Délégue, ARIA Technologies	*Invité*

Préparée au Centre de Recherche en Informatique, Signal et Automatique de Lille
CRIStAL UMR CNRS 9189 et Télécom Lille
Ecole Doctorale SPI 072
PRES Université Lille - Nord de France

Adaptive Bayesian inference for source term estimation in atmospheric dispersion

Abstract : In atmospheric physics, reconstructing a pollution source is a challenging but important question : it provides better input parameters to dispersion models, and gives useful information to first-responder teams in case of an accidental toxic release. Various methods already exist, but using them requires an important amount of computational resources, especially as the accuracy of the dispersion model increases. A minimal degree of precision for these models remains necessary, particularly in urban scenarios where the presence of obstacles and the unstationary meteorology have to be taken into account. One has also to account for all factors of uncertainty, from the observations and for the estimation. The topic of this thesis is the construction of a source term estimation method based on adaptive Bayesian inference and Monte Carlo methods. First, we describe the context of the problem and the existing methods. Next, we go into more details on the Bayesian formulation, focusing on adaptive importance sampling methods, especially on the AMIS algorithm. The third chapter presents an application of the AMIS to an experimental case study, and illustrates the mechanisms behind the estimation process that provides the source parameters' posterior density. Finally, the fourth chapter underlines an improvement of how the dispersion computations can be processed, thus allowing a considerable gain in computation time, and giving room for using a more complex dispersion model on both rural and urban use cases.

Keywords : Bayesian inference, Monte-Carlo methods, Adaptive methods, Atmospheric physics, Source term estimation.

Inférence bayésienne adaptative pour la reconstruction de source en dispersion atmosphérique

Résumé :En physique de l'atmosphère, la reconstruction d'une source polluante à partir des mesures de capteurs est une question importante. Elle permet en effet d'affiner les paramètres des modèles de dispersion servant à prévoir la propagation d'un panache de polluant, et donne aussi des informations aux primo-intervenants chargés d'assurer la sécurité des populations. Plusieurs méthodes existent pour estimer les paramètres de la source, mais leur application est coûteuse à cause de la complexité des modèles de dispersion. Toutefois, cette complexité est souvent nécessaire, surtout lorsqu'il s'agit de traiter des cas urbains où la présence d'obstacles et la météorologie instationnaire imposent un niveau de précision important. Il est aussi vital de tenir compte des différents facteurs d'incertitude, sur les observations et les estimations. Les travaux menés dans le cadre de cette thèse ont pour objectif de développer une méthodologie basée sur l'inférence bayésienne adaptative couplée aux méthodes de Monte Carlo pour résoudre le problème d'estimation du terme source. Pour cela, nous exposons d'abord le contexte scientifique du problème et établissons un état de l'art. Nous détaillons ensuite les formulations utilisées dans le cadre bayésien, plus particulièrement pour les algorithmes d'échantillonnage d'importance adaptatifs. Le troisième chapitre présente une application de l'algorithme AMIS dans un cadre expérimental, afin d'exposer la chaîne de calcul utilisée pour l'estimation de la source. Enfin, le quatrième chapitre se concentre sur une amélioration du traitement des calculs de dispersion, entraînant un gain important de temps de calcul à la fois en milieu rural et urbain.

Mots-clés : Inférence bayésienne, Méthodes de Monte-Carlo, Techniques adaptatives, Physique de l'atmosphère, Estimation du terme source.

Remerciements

Mes premiers remerciements vont au professeur Yves Delignon, qui fût le directeur originel de cette thèse et qui n'est malheureusement plus parmi nous pour en voir l'aboutissement. Ses conseils, sa patience, et son enseignement de la rigueur et de la persévérance ont été indispensables pour achever ce travail, et pour cela je lui en suis plus que reconnaissant.

Un immense merci à François Septier, qui m'a guidé tout au long de ces années, et qui a assuré la direction de ma thèse sur les derniers mois. Sa patience, son attention et sa disponibilité malgré la distance ont été indispensables au bon déroulement de cette thèse. Il a su trouver les mots pour répondre clairement à chacune de mes questions, mais aussi et surtout, pour m'encourager à franchir les étapes difficiles qui jalonnent la vie d'un doctorant.

J'adresse également mes sincères remerciements aux membres du jury pour l'attention qu'ils ont porté à l'évaluation de mon travail. Merci aux professeurs Thierry Chonavel et Hichem Snoussi pour avoir accepté d'endosser le rôle de rapporteur, pour leur relecture minutieuse de mon manuscrit ainsi que pour leurs intéressantes remarques sur son contenu. Merci aussi à MM. Patrick Bas et Lionel Soulhac d'avoir accepté de se déplacer pour assister à la soutenance.

Je remercie Patrick Armand pour m'avoir accueilli au sein de son laboratoire au CEA et co-encadré cette thèse : sa vision globale du sujet mais aussi son souci du détail ont toujours permis de faire avancer les choses dans le bon sens. Merci également à Christophe M. et Jean A., qui ont toujours été à l'écoute des doctorants du service, ainsi qu'à Karine P. et Emilie D. pour leur assistance dans les (nombreuses !) démarches administratives du CEA.

Merci évidemment à tous les membres du LIRC pour tous les bons moments passés ensemble, à parler boulot mais aussi foot ou encore robotique. Merci à Marguerite M., experte en titre et grande voyageuse du labo, et à Maud L.-W. Merci à Christophe D., expert en questions informatiques mais aussi imbattable sur l'actualité sportive. Merci à Luc P., en qui j'ai trouvé un hobbyiste tout aussi *geek* que moi. Merci à Sylvia S., Pascal A. et Thomas A., qui avaient toujours de quoi rire ou plaisanter pour garder la bonne humeur. Merci à Christelle C., voisine du sud qui a sû ne pas perdre l'accent. Merci à Jean-Baptiste S. et Thomas L., co-explorateurs (on sait ce qu'il y a au dernier étage !) et jardiniers improvisés du fameux "loft", dont je n'oublie pas non plus tous les occupants que j'ai eu le plaisir de connaître.

J'ai une pensée spéciale pour les collègues chimistes qui ont accepté de me faire une place dans leurs locaux, et avec qui j'ai eu le plaisir de discuter, plaisanter, et apprendre au quotidien : sans votre gentillesse et votre bonne humeur, cette thèse n'aurait pas été la même. Merci à Françoise Z. pour tous les cookies ramenés d'Angleterre. Merci à Françoise L. pour ses récits épiques de voyage au bout du monde. Merci à Frédéric P., qui a encore des efforts à faire en informatique :-) mais qui compense très largement par sa connaissance sans faille des musées de Paris. Merci à Maxime B., dont les histoire de vie aux USA m'ont convaincu de mettre la visite de ce pays en tête de ma *to-do list*.

Impossible également d'oublier tous ceux avec qui j'ai partagé le fameux bureau 274, et qui ont transformé cet espace de travail en un lieu de partage et d'échanges. Merci à Yasmine B., doctorante pionnière du laboratoire, pour toutes les discussions qu'on a partagées, parfois très sérieuses, parfois un peu moins, mais toujours avec quelque chose à apprendre à la clé. Merci à Sébastien S., chimiste de la pollution de l'air et comme moi, grand amateur des oeuvres de Nobuo Uematsu. Merci à Sébastien V. qui a sû mettre de la couleur (de l'orange en particulier) et de la musique dans la vie du bureau. Merci à Christophe J., le vétéran du bureau, pour avoir partagé son expérience de chercheur et pour ses conseils dont je saurai me rappeler à l'avenir. Merci à Robin L. pour ses explications sur l'assimilation de données et son lien avec mon sujet de recherche. Merci à Adrien N. et Daniel L., qui ont pris la route de la thèse

après leurs stages : je vous souhaite tout le meilleur pour la suite.

J'ai eu la chance de travailler sur un sujet de recherche appliquée, et donc de passer un certain temps au sein de la PME ARIA Technologies pour implémenter une grande partie de mes travaux. Je souhaite donc remercier ses dirigeants et co-fondateurs Jacques Moussafir et Armand Albergel pour la confiance qu'ils m'ont accordé, ainsi que pour leur présence et leurs avis éclairés lors des comités de thèse, malgré leurs agendas surchargés respectifs. Un grand merci à Christophe Olry, dont l'aide s'est révélée indispensable lorsqu'il a fallu "brancher" les algorithmes d'estimation sur PMSS, merci également à Laurent Makké qui a pris sa suite, et dont l'implication a permis d'obtenir les derniers résultats du cas urbain. Merci aussi à Jérôme, Maxime, Cyrille et Félix pour leurs conseils. Plus généralement, je remercie tous les "Ariotes" pour leur accueil chaleureux durant les nombreux jours passés à Boulogne-Billancourt, vous avez chacun contribué à votre façon à la partie "pratique" de cette thèse.

Terminer une thèse, c'est aussi la fin du chapitre "scolaire" d'une vie, et mon parcours n'aurait pas été ce qu'il est sans ceux qui m'ont appris et inspiré durant toutes ces années. Merci à Caroline Paulus et à George Malliaras, mes premiers mentors dans la Recherche, pour m'avoir mis le pied à l'étrier et initié à cet univers que je connais à présent un peu mieux. Merci à mes professeurs de l'ENSEA, en particulier Sophie Olijnyk et Thomas Tang, ainsi qu'au légendaire professeur Kasbari, dont le cours de physique des composants restera ma meilleure expérience pédagogique. Merci à Christophe Coste pour m'avoir donné la chance que je n'attendais plus. Enfin, merci à Eugénie Gaudin et Alain Tertzaguian, avec qui tout a commencé.

Merci du fond du coeur à mes amis dont la présence et le soutien ont beaucoup aidé, avec une pensée particulière pour les camarades de promo ENSEArques, ainsi que pour les incontournables HDP.

Pour terminer, je souhaiterais dire merci à ma famille, aux oncles, tantes et cousins à qui je vais enfin pouvoir dire "les études, c'est fini". Les derniers mots seront pour mes parents et ma soeur : c'est parce que vous n'avez jamais cessé de croire en moi que je suis allé aussi loin. Cette réussite, c'est aussi et surtout la vôtre. Vous êtes ma fierté, et ce manuscrit vous est dédié.

« Mais je connais la solitude. Trois années de désert m'en ont enseigné le goût. On ne s'y effraie point d'une jeunesse qui s'use dans un paysage minéral, mais il y apparait que, loin de soi, c'est le monde entier qui vieillit. Les arbres ont formé leurs fruits, les terres ont sorti leur blé, les femmes déjà sont belles... Mais la saison avance et l'on est retenu au loin... Et les biens de la terre glissent entre les doigts comme le sable fin des dunes. »

Terre des Hommes, A. de Saint-Exupéry

Sommaire

Table des figures

Chapitre 1

Introduction

1.1 Pourquoi chercher à reconstruire des sources de pollution ?

La menace de rejets de substances *Nucléaires*, *Radiologiques*, *Biologiques* ou *Chimiques* (NRBC) dans l'atmosphère suscite un fort intérêt, du fait des enjeux humains et environnementaux qu'ils affectent. De tels incidents peuvent être d'origine accidentelle, causés par des rejets ayant eu lieu dans des sites industriels stockant ou exploitant des matières dangereuses. Plusieurs événements historiques témoignent de l'impact de tels accidents :

- Seveso (Italie) en 1976 : un rejet accidentel de dioxine provenant d'une usine chimique engendre un nuage toxique contaminant une surface de près de 2.8 km^2, touchant plus de 400 personnes victimes de lésions cutanées, et nécessitant l'abattage de près de 80 000 bêtes contaminées dans les domaines agricoles atteints [BARPI-5620, 2008].
- Bhopal (Inde) en 1984 : une explosion dans une usine de pesticides entraîne un important rejet de substances chimiques toxiques (isocyanathe de méthyle, cyanure hydrogéné) touchant directement la population vivant aux alentours. Le bilan à long terme est d'au moins 16 000 morts et d'environ 500 000 intoxiqués [BARPI-7022, 2014].
- Tchernobyl (Ukraine) en 1986 : suite à des erreurs humaines lors d'opérations sur un réacteur de la centrale nucléaire locale, le cœur est entré en fusion : il s'en est suivi une explosion et la libération d'importantes quantités d'éléments radioactifs dans l'atmosphère. Le nuage formé par ces polluants s'est répandu à l'échelle continentale sur une grande partie de l'Europe [Repussard, 2006].
- Algésiras (Espagne) en 1998 : une usine espagnole incinère accidentellement une source radioactive dans ses hauts-fourneaux. Ce n'est que près de trois semaines plus tard que l'origine de la fuite est établie. Sur la base des mesures effectuées et de la reconstitution des courants atmosphériques, la quantité totale de césium 137 libérée a été évaluée à 1850 GBq [Estevan, 2003].
- Fukushima (Japon) en 2011 : suite au déclenchement d'un séisme de magnitude 9.0 au large des côtes japonaises, les dégâts causés par le tsunami induit ont entraîné la fusion

d'au moins deux réacteurs de la centrale de Fukushima, causant ainsi d'importants rejets radioactifs sur le territoire japonais, et plus largement, sur une large portion de l'océan Pacifique [IRSN, 2012].

— Igualada (Espagne) en 2015 : une citerne explose en effectuant une livraison dans une usine chimique, causant la propagation d'un nuage orange d'acide nitrique. L'incident a entraîné des mesures de confinement des populations dans le voisinage immédiat de l'usine.

— Los Angeles (Etats-Unis) en 2015 : une fuite localisée dans un puits de stockage de gaz de ville entraîne d'importants rejets de méthane dans l'atmosphère. L'incident a été déclaré le 23 octobre 2015, et annoncé être "sous contrôle" le 11 février 2016 par l'entreprise gestionnaire du puits. Durant la période de rejet, entre 30 et 50 tonnes par heure de méthane furent rejetés dans l'air. Le méthane étant un gaz à effet de serre, les conséquences environnementales à moyen et long terme sont déjà considérées comme graves.

Les incidents NRBC peuvent aussi être issus d'actes malveillants relevant du terrorisme. Ce fut le cas à Tokyo (Japon) en 1995, où des membres d'une secte ont percé des poches contenant du gaz sarin (un puissant neurotoxique) dans des rames de métro. Le bilan final fut de 12 morts et plus de 5500 blessés. Plus récemment, le risque d'attentats NRBC a également été mis en évidence en France suite aux attentats de l'année 2015, et amplifié en raison du contexte géopolitique actuel.

Dans tous les cas, il est vital de disposer de techniques rapides pour détecter et évaluer le risque, afin d'assurer au mieux la sécurité des personnes et de coordonner les manoeuvres des équipes de premier secours. De telles techniques reposent sur des méthodes de modélisation des phénomènes physiques régissant l'atmosphère, ainsi que sur un système d'instrumentation permettant, entre autres, de caractériser et quantifier la présence de substances toxiques dans l'air.

Cependant, pour que ces outils de modélisation puissent fonctionner, il est indispensable de disposer d'un certain nombre de paramètres, dont les caractéristiques de la source à l'origine du rejet. C'est ce point particulier que nous illustrerons dans la suite de ce chapitre, après un bref exposé introductif sur la physique de la dispersion atmosphérique.

1.2 Les principes de la dispersion atmosphérique

Nous décrivons ici les grandes lignes des règles physiques qui régissent la propagation d'un nuage de polluant dans l'atmosphère. Pour un exposé plus exhaustif, le lecteur peut se référer à [Sportisse, 2008].

1.2.1 L'équation d'advection-diffusion

Une fois qu'un polluant est émis dans l'atmosphère, son comportement est régi par plusieurs processus distincts :

1. le **transport**, ou **advection** qui se fait sous l'influence des circulations d'air dans l'atmosphère,
2. la **diffusion**, résultant de la nature turbulente des écoulements dans la partie basse de l'atmosphère (couche limite),
3. les processus de **pertes par dépôt sec ou humide**, diminuant la quantité de polluant transportée,
4. les éventuelles **transformations physico-chimiques** pouvant altérer l'état du polluant lors de son séjour dans l'atmosphère : la filiation radioactive (s'il s'agit d'un radionucléide), ou les diverses réactions chimiques pouvant avoir lieu avec les autres composants de l'air.

Nous supposerons en première approximation que les processus (3) et (4) énumérés précédemment ne sont pas pris en compte. Si on considère le transport d'un polluant unique dont la concentration peut être décrite au point $\boldsymbol{x} = (x, y, z) \in \mathbb{R}^3$ et à l'instant t par une fonction $C(\boldsymbol{x}, t)$. On peut alors écrire la loi de conservation de la masse pour C sous la forme suivante :

$$\frac{\partial C}{\partial t} + \nabla \cdot \vec{J}(\boldsymbol{x}, t) = \varsigma \tag{1.1}$$

où ς est le *terme source*, $\vec{J}(\boldsymbol{x}, t)$ représente le flux de masse du polluant, et ∇ désigne l'opérateur gradient. $\vec{J}$ est une fonction vectorielle qui regroupe la somme des phénomènes distincts d'advection et de diffusion :

$$\vec{J} = \vec{J}_A + \vec{J}_D \tag{1.2}$$

Le terme $\vec{J}_D$ est associé au phénomène de diffusion. Celui-ci est généralement considéré comme suivant la première loi de Fick, stipulant que le flux de diffusion $\vec{J}_D$ est proportionnel au gradient de concentration :

$$\vec{J}_D = -\boldsymbol{K} \nabla C \tag{1.3}$$

où $\boldsymbol{K}$ est une matrice contenant les coefficients de diffusion moléculaire, qui dépendent de l'espèce du polluant. Le terme $\vec{J}_A$ est associé au phénomène d'advection, et traduit une dépendance linéaire de la concentration par rapport au champ de vent $\vec{\boldsymbol{u}}$:

$$\vec{J}_A = C\vec{\boldsymbol{u}} \tag{1.4}$$

La combinaison des équations (1.2), (1.3) et (1.4) mène ainsi à la formulation de l'*équation d'advection-diffusion*, formulée sur le modèle de [Stockie, 2011] :

$$\frac{\partial C}{\partial t} + \nabla \cdot (C\vec{\boldsymbol{u}}) = \nabla \cdot (\boldsymbol{K}\nabla C) + \varsigma \tag{1.5}$$

Cette équation constitue la base de départ pour les différents modèles numériques cherchant à simuler les phénomènes de dispersion atmosphérique.

1.2.2 Les différents modèles de dispersion

Suivant les situations, la taille du domaine sur lequel les calculs sont effectués peut grandement différer. On distingue habituellement :

— l'*échelle locale*, aussi appelée *micro-échelle* (jusqu'à 1 km) : à ce niveau, des phénomènes spécifiques tels que la présence de bâtiments doivent être pris en compte dans le calcul des champs de vent. C'est à cette échelle qu'il convient de se placer pour traiter les études d'impact en milieu urbain (par exemple à l'échelle d'un quartier), ou sur un site industriel.

— la *méso-échelle*, ou *échelle régionale* (jusqu'à 1000 km) : elle est utilisée pour travailler sur des phénomènes plus larges, par exemple pour étudier l'impact de la pollution à l'ozone ou aux particules fines sur le territoire d'un pays tel que la France.

— l'*échelle planétaire*, ou *échelle synoptique* ($>$ 1000 km) : on s'intéresse ici à l'impact d'événements de très grande ampleur, tels que les accidents de Tchernobyl ou Fukushima.

On peut également noter la corrélation entre les échelles spatiales et temporelles dans le cadre accidentel : au niveau local, on s'intéresse typiquement à des phénomènes n'excédant pas quelques heures, alors qu'à l'échelle du globe on se place sur un intervalle de plusieurs semaines, voire plusieurs mois.

Pour couvrir efficacement ces différentes tailles de zones, plusieurs types de modèles de dispersion existent. Ceux-ci se divisent en quatre catégories principales :

— **Les modèles gaussiens** : Sous certaines hypothèses simplificatrices, l'équation (1.5) peut donner des solutions analytiques pour déterminer la concentration de polluant au sein de *panaches* ou de *bouffées* selon la configuration choisie. [Stockie, 2011] résume les principes régissant les modèles gaussiens, et une présentation plus détaillée de leur fonctionnement sera introduite au Chapitre 3 du présent manuscrit.

— **Les modèles eulériens** : Le domaine de simulation sur lequel est résolue l'équation (1.5) est discrétisé en un maillage de calcul par des méthodes numériques. On retrouve ce type d'approche dans des études de cas à l'échelle continentale [Saunier et al., 2013] ou globale.

— **Les modèles lagrangiens particulaires** : Dans le cadre lagrangien, le rejet est modélisé comme un ensemble de particules numériques porteuses d'une masse élémentaire. Le modèle suit la trajectoire de chaque particule, dont le mouvement moyen est constitué d'une composante régie par le champ de vent, et d'une composante stochastique traduisant la variabilité causée par la turbulence. La concentration mesurée sur un volume élémentaire du domaine à un temps donné est ainsi égale à la somme des masses élémentaires portées par chaque particule contenue dans ce volume à cet instant. Ce type de modèle est présenté plus en détails au Chapitre 4.

— **Les modèles de mécanique des fluides** : Aussi appelés *Computational Fluid Dynamics* (CFD), ces modèles sont utilisés à petite échelle, et impliquent une résolution des équations de Navier-Stokes sur un maillage relativement fin. Il en résulte une très bonne précision des calculs, en particulier en cas de présence d'obstacles menant à des écoulements complexes.

1.3 Les caractéristiques du terme source

Les modèles de dispersion présentés dans le paragraphe précédent nécessitent plusieurs types de données d'entrée :

— un certain nombre de **paramètres météorologiques** : dans le cas le plus simple, on considère la direction et la vitesse du vent mesurés par une station au sol dans le domaine d'étude ou à sa proximité, ainsi que la stratification atmosphérique. Ces données sont alors supposées homogènes. Dans des cas plus complexes, plusieurs instruments météorologiques peuvent être pris en compte, au sol comme en altitude. Enfin, il est également possible d'exploiter les résultats d'un système de prévision météorologique : les grandeurs considérées sont alors des champs de vitesse et direction de vent, température, humidité, nébulosité et flux de rayonnement.

— les **paramètres de la source** à l'origine des émissions à modéliser ;

C'est sur ce second point que nous allons plus particulièrement nous pencher dans la suite de ce paragraphe. En effet, la terminologie employée pour décrire la source d'un rejet polluant peut parfois prêter à confusion, car il existe plusieurs façons d'en décrire les caractéristiques. Nous fixons ici les termes utilisés pour distinguer les différents types de source rencontrés.

Une source est dite *localisée* si sa position peut être réduite à un point de l'espace, on peut également parler de *source ponctuelle*. Par opposition, une *source étendue* est caractérisée par une surface d'émission, il peut, par exemple, s'agir d'un bac de décantation dans une usine de retraitement des déchets, si on se place à une échelle suffisamment petite.

Une source est également définie par le profil temporel de son rejet, autrement dit la variation de la quantité de polluant rejetée par unité de temps. Comme nous travaillons dans un cadre de modélisation numérique d'un phénomène physique, il est nécessaire de définir une discrétisation temporelle sur T_s instants d'émission, sous la forme d'un vecteur $\boldsymbol{t'_s} = (t'_1, \ldots, t'_{T_s})$. On peut alors

construire le vecteur du profil d'émission comme étant la liste des quantités rejetées par pas de temps d'émission (figure 1.1) :

$$\boldsymbol{q} = \big(q(t'_1), q(t'_2), \ldots, q(t'_{T_s})\big)$$

Si le rejet est suffisamment bref pour n'être effectif que sur un unique pas de temps, alors $\boldsymbol{t'_s}$ est réduit à un scalaire t'_s, et $\boldsymbol{q}$ à une masse totale émise q_s : dans ce cas, on parle d'une source dite *instantanée*. Dans le cas inverse, et lorsque le débit n'est pas constant, on parle d'une source *non-instantanée*. Si $\forall i \in \{1, \ldots, T_s\},\ q(t'_i)$ est constant, la source est dite *continue*.

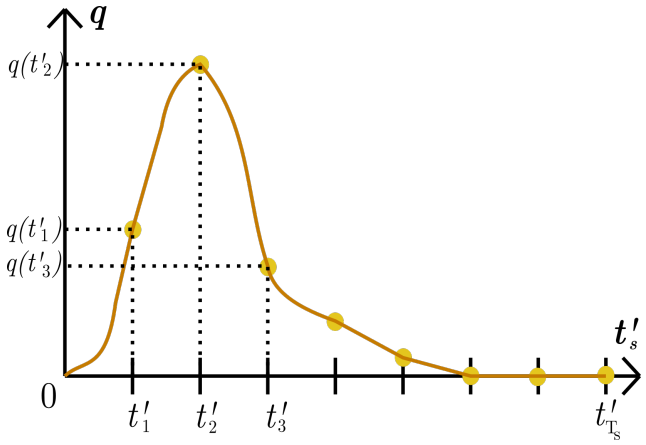


FIGURE 1.1 – Exemple de discrétisation d'une source non-instantanée.

Enfin, un cas d'étude peut comporter une *source unique*, ou bien avoir une configuration *multi-sources*.

1.4 Estimation du terme source : un état de l'art

1.4.1 Une brève introduction aux problèmes inverses

Le domaine des problèmes inverses constitue un vaste champ de la littérature scientifique, et de nombreux travaux y ont été menés. En pratique, résoudre un problème inverse revient à reconstituer un signal, une image, ou plus généralement une donnée non-observable à partir de mesures existantes, appelées observations. Cette approche est duale à celle du problème direct où, à partir d'un signal ou d'un ensemble de paramètres initiaux, on cherche à en calculer les effets après une transformation donnée (figure 1.2). De nombreux domaines théoriques et pratiques font appel aux méthodes inverses. Le lecteur pourra trouver une revue complète de ces applications dans [Tarantola, 2004], on peut en citer quelques exemples tels que la géophysique [Backus and Gilbert, 1967], l'acoustique [Kirsch and Kress, 1988], l'imagerie satellite [Park et al., 2003] et médicale [Arridge, 1999], les transferts thermiques [McCormick, 1992] ou encore la finance quantitative [Dembo and Rosen, 1999].

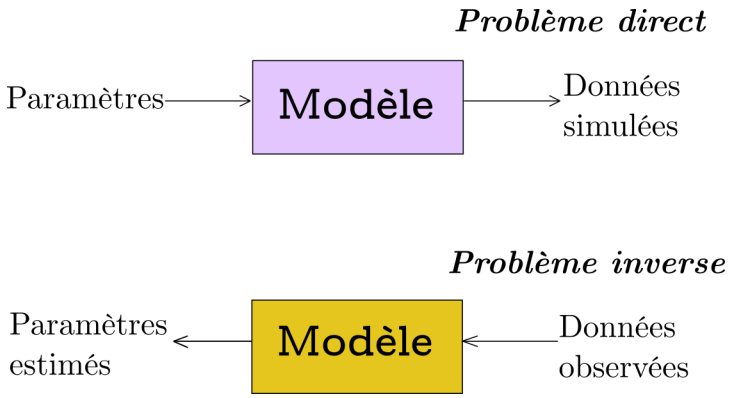


FIGURE 1.2 – Schéma de principe illustrant la dualité entre problèmes direct et inverse.

En dispersion atmosphérique, le problème direct peut ainsi se traduire par la donnée des paramètres de la source au modèle de dispersion, qui va produire le champ de concentration résultant. Cela revient bien à définir le problème inverse comme étant la reconstruction des paramètres de la source à partir des mesures de concentrations aux capteurs. Nous décrivons dans les paragraphes suivants les différentes approches existant dans la littérature qui permettent de résoudre ce problème. Des éléments complémentaires sur la question sont également disponibles dans [Rao, 2007]. Dans un souci de brièveté, par la suite nous qualifierons de "problèmes STE" (*source term estimation*) toutes les études de cas visant à estimer les paramètres d'une ou de plusieurs sources inconnues.

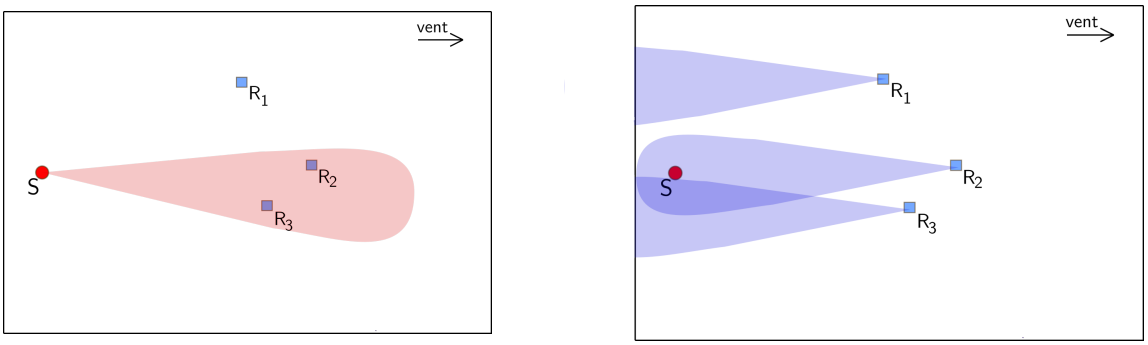
1.4.2 Rétro-transport et modèles rétrogrades

Une première possibilité consiste à étudier le problème dual associé à l'équation d'advection-diffusion (1.5), en inversant la flèche du temps, et par conséquent la direction du vent. Par le principe de symétrie présenté dans [Hourdin and Talagrand, 2006], on peut alors introduire la notion de *rétro-transport* ou *backtracking* en réécrivant l'équation d'advection-diffusion sous la forme :

$$\frac{-\partial C^*}{\partial t} - \nabla \cdot (C^* \vec{\boldsymbol{u}}) = \nabla \cdot (\boldsymbol{K} \nabla C^*) + \boldsymbol{\eta} \tag{1.6}$$

où C^* est un champ de concentrations conjuguées, ou *rétro-concentrations*, dont les valeurs sont obtenues par des rétro-rejets virtuels depuis les capteurs vers la source. Le modèle de dispersion associé à cette démarche duale est alors appelé *modèle de rétro-dispersion*, ou *modèle rétrograde*. Cette démarche revient ainsi à transformer le champ d'émission obtenu par l'équation (1.5) en un champ de rétro-émissions issues des capteurs, comme illustré sur les figures 1.3a et 1.3b, et ainsi fournir une estimation du terme source. Dans [Flesch et al., 1995], un modèle rétrograde lagrangien est construit pour estimer le profil de rejet d'une source surfacique. Cette méthode est également appliquée dans [Pudykiewicz, 1998], où la position et l'intensité du terme source de Tchernobyl sont reconstruits.

De même, dans [Hourdin et al., 2006] le principe du *backtracking* est appliqué à un modèle eulérien pour retrouver la source de la campagne ETEX [1].



(a) Approche orientée source

(b) Approche orientée récepteur

FIGURE 1.3 – Exemple illustrant la relation entre une source unique S et trois capteurs R_1, R_2, R_3 sur un problème en deux dimensions pour les approches en "direct" (1.3a) et en "adjoint" (1.3b).

1.4.3 Formulation linéaire du problème et optimisation

Le modèle direct de simulation des concentrations résultant d'un rejet de polluant peut être décrit de façon linéaire. Pour cela, on définit de façon arbitraire un maillage dans le temps et l'espace qui couvre respectivement la période d'observation et le domaine spatial considérés. En pratique, cela revient à discrétiser l'espace en $N_x N_y N_z$ points, chaque point étant associé à un profil d'émission de longueur T_s. Dans un tel contexte, le terme source est défini comme un vecteur $\boldsymbol{\varsigma} \in \mathbb{R}^{N_{\boldsymbol{\varsigma}}}$ (où $N_{\boldsymbol{\varsigma}} = N_x N_y N_z T_s$). Il est ainsi possible de lier les observations générées par le modèle et le vecteur source $\boldsymbol{\varsigma}$ par la formule suivante, appelée *relation source-récepteur* :

$$\boldsymbol{\eta} = \boldsymbol{H}\boldsymbol{\varsigma} + \boldsymbol{\varepsilon} \tag{1.7}$$

où $\boldsymbol{\eta} \in \mathbb{R}^m$ représente le vecteur des observations, $\boldsymbol{\varsigma} \in \mathbb{R}^{N_{\boldsymbol{\varsigma}}}$ la source discrétisée et $\boldsymbol{H} \in \mathbb{R}^{m \times N_{\boldsymbol{\varsigma}}}$ la *matrice de transfert*. Le rôle de $\boldsymbol{H}$ est d'assurer la projection depuis l'espace $\mathbb{R}^{N_{\boldsymbol{\varsigma}}}$ de la source vers l'espace $\mathbb{R}^m$ des observations grâce à l'utilisation d'un modèle de dispersion atmosphérique. $\boldsymbol{\varepsilon} \in \mathbb{R}^m$ est le vecteur des erreurs d'observation, et prend en compte les erreurs de mesure, de représentativité et de modèle.

Retrouver le terme $\boldsymbol{\varsigma}$ dans l'équation (1.7) revient alors à résoudre un problème inverse linéaire : il existe pour cela différentes méthodes que nous détaillons ci-après.

Définition de la fonction-coût à optimiser

L'équation (1.7) peut être réécrite sous la forme :

$$\boldsymbol{\eta} = \widehat{\boldsymbol{\eta}} + \boldsymbol{\varepsilon} \tag{1.8}$$

1. L'expérience *European Tracer EXperiment 1* (ETEX-1), menée en 1994, a consisté à mesurer et étudier l'impact à l'échelle européenne de rejets de gaz traceurs passifs émis depuis une source située dans la ville de Monterfil, en Bretagne.

où $\widehat{\boldsymbol{\eta}} = \boldsymbol{H}\boldsymbol{\varsigma}$ est l'approximation des observations $\boldsymbol{\eta}$ par le processus de modélisation. On voit bien que dans le cas parfait où ce dernier reproduit exactement les observations attendues, on a $\boldsymbol{\varepsilon} = 0$ et par suite $\widehat{\boldsymbol{\eta}} = \boldsymbol{\eta}$. Le fait de chercher la valeur $\boldsymbol{\varsigma}$ pour laquelle $\widehat{\boldsymbol{\eta}}$ se rapproche le plus de $\boldsymbol{\eta}$ permet ainsi de définir le problème STE sous la forme d'une minimisation de l'erreur $\boldsymbol{\varepsilon}$: on définit pour cela une fonction-coût qui s'écrit :

$$\mathcal{J}(\boldsymbol{\varsigma}) = ||\boldsymbol{\eta} - \boldsymbol{H}\boldsymbol{\varsigma}||_p = ||\boldsymbol{\eta} - \widehat{\boldsymbol{\eta}}||_p \tag{1.9}$$

où $||\cdot||_p$ désigne une norme L_p arbitrairement choisie. La résolution du problème STE se traduit alors par la minimisation de cette fonction-coût :

$$\widehat{\boldsymbol{\varsigma}} = \underset{\boldsymbol{\varsigma}}{\operatorname{argmin}}\, \mathcal{J}(\boldsymbol{\varsigma}) \tag{1.10}$$

où $\widehat{\boldsymbol{\varsigma}}$ est l'estimation du terme source recherché.

Minimisation d'une fonction-coût quadratique

Une des approches les plus courantes pour résoudre l'équation (1.10) consiste à choisir $p = 2$, autrement dit à minimiser l'écart quadratique entre les observations $\boldsymbol{\eta}$ et les données générées par le modèle direct. Concrètement, cela revient à optimiser la fonction-coût suivante :

$$\mathcal{J}_2(\boldsymbol{\varsigma}) = ||\boldsymbol{\eta} - \boldsymbol{H}\boldsymbol{\varsigma}||_2^2 \tag{1.11}$$

Le fait de minimiser $\mathcal{J}_2$ revient à appliquer la méthode des *moindres carrés* pour résoudre le problème inverse.

Le terme d'erreur que nous cherchons à minimiser est une combinaison des différentes sources d'incertitudes existantes (mesure, représentativité, modèle). Comme il est difficile de quantifier les contributions respectives de ces dernières, on peut invoquer le principe du maximum d'entropie pour définir la représentation statistique de l'ensemble des incertitudes comme étant un bruit gaussien centré et de matrice de covariance $\boldsymbol{R}$:

$$\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \boldsymbol{R}) \tag{1.12}$$

Dans le cadre de statistiques gaussiennes, la fonction-coût peut alors s'écrire sous la forme suivante [Winiarek et al., 2011] :

$$\mathcal{J}_2(\boldsymbol{\varsigma}) = \frac{1}{2}(\boldsymbol{\eta} - \boldsymbol{H}\boldsymbol{\varsigma})^T \boldsymbol{R}^{-1}(\boldsymbol{\eta} - \boldsymbol{H}\boldsymbol{\varsigma}) \tag{1.13}$$

Dans l'hypothèse où les éléments du vecteur de bruit sont indépendants et identiquement distribués, alors la matrice $\boldsymbol{B}$ est diagonale :

$$\boldsymbol{B} = \sigma_R^2 \boldsymbol{I} \tag{1.14}$$

où σ_R^2 est la variance d'erreur. Plusieurs travaux dans la littérature ont recours a cette formulation :

— Dans [Kathirgamanathan et al., 2002], un point source instantané est estimé à l'aide d'une méthode des moindres carrés, cette source étant caractérisée par sa masse totale émise, sa

position et son instant d'émission. L'étude qui y est menée met notamment l'accent sur l'influence du dimensionnement du réseau de capteurs sur la performance de la reconstruction : au moins trois points de mesure sont nécessaires, et la distance entre les différents capteurs influe sur la qualité de l'estimation.

— Dans [Ryall et al., 2001], ce sont des zones d'émission plus grandes qui sont estimées pour des rejets de gaz à effet de serre sur plusieurs années.

— Dans [Matthes et al., 2005], une approche en deux temps est formulée afin de résoudre un problème de complexité proportionnellement croissante au nombre de capteurs. Dans une première étape, les mesures individuelles de chaque capteur produisent des ensembles de positions probables pour la source. Dans un deuxième temps, ces ensembles sont comparés pour estimer la meilleure position en faisant varier l'intensité de la source à retrouver.

— Dans [Robertson and Langner, 1998], l'inversion est faite dans le cadre de l'assimilation de données : cette méthodologie permet de corriger de façon variationnelle les paramètres du modèle de dispersion selon une boucle de rétro-action basée sur des observations. Très utilisée en météorologie, l'assimilation de données agit en deux temps selon un principe de "prévision-correction", et itère sur le cycle suivant :

1. on calcule d'abord une ébauche de l'état à l'instant présent t, en appliquant un modèle de prévision en $t-1$,
2. on exploite ensuite les observations disponibles à l'instant t pour corriger les paramètres du modèle en les comparant avec l'ébauche.

— [Issartel, 2005] mentionne le principe d'*illumination*, qui est une mesure quantitative de la représentativité des mesures dans le temps et l'espace. Dans le cadre d'un modèle adjoint, l'illumination permet ainsi de caractériser des zones qui ne sont pas forcément couvertes par les rétro-émissions issues des capteurs, par exemple si la zone considérée est relativement éloignée de tout point de mesure. Cependant, si l'information d'illumination est utilisée pour inverser un terme source, alors elle aura tendance à fortement favoriser les solutions proches des capteurs. Pour compenser ce problème, une phase de *renormalisation* est introduite ([Issartel et al., 2007] et [Sharan et al., 2009]) pour équilibrer l'information apportée par chacun des capteurs. D'abord utilisée à grande échelle, cette méthodologie a récemment été validée à l'échelle locale dans [Singh and Rani, 2014], et dans un contexte urbain par [Kumar et al., 2015].

Régularisations

Bien souvent, le nombre d'observations à notre disposition est limité, à cause par exemple d'un faible nombre de capteurs ou d'une mauvaise représentativité temporelle dans les mesures fournies par ces derniers. Dans de telles situations, le problème inverse peut admettre une infinité de solutions, et ainsi devenir *mal-posé*. Rappelons qu'un problème *bien posé* au sens de Hadamard

[Hadamard, 1902] désigne un modèle mathématique dont la solution existe, est unique, et dépend de façon continue des données.

A l'inverse, un problème mal-posé déroge à au moins une des règles précédemment citées.

De nombreux problèmes physiques sont intrinsèquement mal-posés, et des solutions existent pour les *régulariser*, c'est-à-dire leur appliquer une transformation qui permet de revenir à un problème bien posé. Parmi ces méthodes, la *régularisation de Tikhonov* [Tikhonov, 1963] est la plus fréquemment employée : elle consiste à ajouter un terme $\lambda_B||\boldsymbol{\varsigma_b}||_2$ dit *régularisant* à l'expression de la fonction-coût $\mathcal{J}_2$, qui devient alors :

$$\mathcal{J}_{2_B}(\boldsymbol{\varsigma}) = ||\boldsymbol{\eta} - \boldsymbol{H\varsigma}||_2^2 + \lambda_B||\boldsymbol{\varsigma}||_2 \tag{1.15}$$

Le rôle du terme régularisant est de pénaliser les solutions indésirables, afin de garantir le critère d'unicité de la solution du problème inverse. Dans le cas des statistiques gaussiennes, l'erreur $\boldsymbol{\varsigma} - \boldsymbol{\varsigma_b}$ suit également une loi normale centrée, de matrice de covariance $\boldsymbol{B}$:

$$(\boldsymbol{\varsigma} - \boldsymbol{\varsigma_b}) \sim \mathcal{N}(0, \boldsymbol{B}) \tag{1.16}$$

L'équation (1.13) devient alors [Winiarek et al., 2011] :

$$\mathcal{J}_{2_B}(\boldsymbol{\varsigma}) = \frac{1}{2}(\boldsymbol{\eta} - \boldsymbol{H\varsigma})^T\boldsymbol{R}^{-1}(\boldsymbol{\eta} - \boldsymbol{H\varsigma}) + \frac{1}{2}(\boldsymbol{\varsigma} - \boldsymbol{\varsigma_b})^T\boldsymbol{B}^{-1}(\boldsymbol{\varsigma} - \boldsymbol{\varsigma}_b) \tag{1.17}$$

Dans la littérature des méthodes d'estimation du terme source, le terme régularisant est qualifié d'*ébauche*, de *first-guess*, ou encore de *background*. Les travaux présentés dans [Winiarek et al., 2012] introduisent une approche basée sur la résolution de l'équation (1.17) via la méthode du *Best Linear Unbiased Estimator* (BLUE), qui vise à annuler le gradient de la fonction-coût cible. Une approche similaire est proposée par [Saunier et al., 2013], élaborant une reconstruction du terme source de l'accident de Fukushima en deux temps. Tout d'abord, une première fonction-coût est minimisée pour estimer la période du rejet, puis tenant compte de cette contrainte, une seconde fonction-coût est optimisée pour retrouver la localisation de la source.

Il est intéressant de remarquer que l'équation (1.17) correspond à celle de la méthode d'assimilation variationnelle de données 3D-var [Courtier et al., 1998]. La vision de ce type d'estimation correspond ainsi à celle de l'assimilation des concentrations mesurées par les capteurs dans le but de reconstruire un état inconnu, qui est ici le champ d'émission de la source recherchée. Notons également que des études ont été menées dans le cadre de statistiques non-gaussiennes, en particulier dans les cas où on cherche à caractériser l'ébauche suivant certaines hypothèses spécifiques [Bocquet, 2005]. Ainsi, les auteurs de [Krysta and Bocquet, 2007] font l'hypothèse d'une source ponctuelle et instantanée pour définir une forme particulière d'ébauche, et utilisent un principe de maximum d'entropie sur la moyenne [Jaynes, 1957] pour réécrire la fonction-coût et en dériver un estimateur approprié pour la source qu'ils recherchent.

Il peut exister des situations où les observations prennent l'allure de *signaux parcimonieux* (ou *sparse signals*), autrement dit elles sont décrites par un très faible nombre de valeurs non-nulles sur l'intervalle temporel de mesure. Pour tenir compte de cela, l'équation (1.11) est modifiée, et

c'est cette fois un terme régularisant sous une norme L_1 qui est utilisé :

$$\mathcal{J}_{2_L} = ||\boldsymbol{\eta} - \boldsymbol{H}\boldsymbol{\varsigma}||_2^2 + \lambda_L||\boldsymbol{\varsigma}||_1 \tag{1.18}$$

Une application de cette approche à la reconstruction de sources multiples est présenté dans [Cheng and Singh, 2008], où la régularisation L_1 est justifiée par la nature parcimonieuse du vecteur d'état à estimer. Une autre étude, menée dans [Martinez-Camara et al., 2013] porte sur la modélisation des rejets de xénon durant l'incident de Fukushima sous la forme de brèves impulsions (ou *short bursts*), fournissant ainsi un cadre approprié à la régularisation L_1.

Une vision probabiliste du problème d'optimisation

Il est possible d'établir une analogie entre les méthodes d'optimisation précédemment mentionnées et les approches se basant sur des calculs de lois de probabilité.

En effet, dans un cadre probabiliste, l'optimisation sans terme régularisant est équivalente à la résolution d'un problème de *maximum de vraisemblance*. En pratique, le fait de minimiser la fonction-coût $\mathcal{J}_2$ revient ainsi à déterminer la grandeur suivante :

$$\widehat{\boldsymbol{\varsigma}_2} = \underset{\boldsymbol{\varsigma}}{\operatorname{argmax}}\ \log p(\boldsymbol{\eta}|\boldsymbol{\varsigma}) \tag{1.19}$$

où $p(\boldsymbol{\eta}|\boldsymbol{\varsigma})$ est la vraisemblance des observations pour un terme source donné.

De même, le fait de régulariser la fonction-coût par l'ajout d'un terme d'ébauche revient à adopter une démarche *bayésienne*[2], où le terme régularisant représente une information a priori $p(\boldsymbol{\varsigma})$, et l'estimateur recherché devient alors le maximum de la loi a posteriori de $\boldsymbol{\varsigma}$:

$$\begin{aligned}\widehat{\boldsymbol{\varsigma}_{2_B}} &= \underset{\boldsymbol{\varsigma}}{\operatorname{argmax}}\ \log p(\boldsymbol{\varsigma}|\boldsymbol{\eta}) \\ &= \underset{\boldsymbol{\varsigma}}{\operatorname{argmax}}\ (\log p(\boldsymbol{\eta}|\boldsymbol{\varsigma}) + \log p(\boldsymbol{\varsigma}))\end{aligned} \tag{1.20}$$

Dans le cas de la régularisation L_1 présenté à l'équation (1.18), la démarche est équivalente à l'introduction d'une information a priori de parcimonie sous la forme d'une loi laplacienne pour $p(\boldsymbol{\varsigma})$. Cette démarche de maximum a posteriori est alors équivalente à l'estimateur *Least Absolute Shrinkage and Selection Operator* (LASSO) [Tibshirani, 1996].

1.4.4 Formulation générale et méthodes de résolution

Le fait de présenter le problème inverse de la reconstruction du terme source sous une forme linéaire permet sa résolution directe grâce aux méthodes précédemment présentées. Cependant, le fait de dépendre d'un maillage sur le temps et l'espace peut poser problème, en particulier dans les cas où la finesse de ce maillage ramène l'estimation du terme source à un problème de grande dimension, qui devient alors difficile à résoudre.

2. Nous décrirons plus en détail les principes fondamentaux de la statistique bayésienne dans le prochain chapitre.

Il reste toutefois possible d'écrire le problème direct sous une forme plus générale, qui ne tient pas compte d'un maillage et qui reste valable pour le cas multi-source. Dans un tel contexte, la relation source-récepteur se traduit par l'équation suivante :

$$\boldsymbol{\eta} = \sum_{i=1}^{N_s} \boldsymbol{C}(\boldsymbol{x}_s{}^{(i)})\boldsymbol{q}^{(i)} + \boldsymbol{\varepsilon} \tag{1.21}$$

où :

— N_s est le nombre de sources,

— $\boldsymbol{x}_s{}^{(i)}$ et $\boldsymbol{q}^{(i)}$ sont respectivement la position et le profil temporel d'émission de la i-ème source,

— $\boldsymbol{C}(\boldsymbol{x}_s{}^{(i)})$ est la concentration résultant d'un rejet unitaire de la i-ème source.

Ici, la relation source-récepteur ne suppose désormais qu'une discrétisation temporelle sur les T_s dimensions du vecteur $\boldsymbol{q}$, contrairement à un cas maillé où la dimension du problème est de $N_x N_y N_z T_s$. Toutefois, il devient impossible d'obtenir directement les paramètres $\boldsymbol{x}_s$ et $\boldsymbol{q}$ de la source, contrairement à l'équation (1.7) où $\boldsymbol{\varsigma}$ peut se calculer analytiquement grâce à la nature linéaire du système à résoudre. En conséquence, d'autres méthodes plus appropriées sont nécessaires pour résoudre le problème inverse.

Algorithmes évolutionnaires

Le gain en complexité du problème d'optimisation à résoudre avec la formulation non-linéaire nécessite une approche suffisamment robuste afin de réussir à traiter les fonctions-coût résultantes. Cela est possible par le biais de *métaheuristiques* (méthodes d'optimisation généralement appliquées des problèmes à forte complexité combinatoire). Parmi celles-ci, les *approches évolutionnaires*, et en particulier les *algorithmes génétiques* ont été utilisés à plusieurs reprises pour résoudre des problèmes STE.
Les algorithmes évolutionnaires s'inspirent du principe de l'évolution darwinienne des populations biologiques, et leur construction s'appuie sur le fait que l'apparition d'espèces adaptées au milieu est la conséquence de la conjonction de deux phénomènes distincts :

— d'une part, la *sélection naturelle* imposée par le milieu (les individus les plus adaptés survivent et se reproduisent),

— d'autre part, les variations non-contrôlées du matériel génétique des espèces.

Pour un problème d'optimisation standard, la fonction-coût $\mathcal{J}$ à minimiser devient, dans le langage évolutionnaire, une *fonction d'adaptation*. Les points du domaine Ω des paramètres à explorer sont appelés *individus*, et l'ensemble d'individus est appelé *population*.
La structure d'un algorithme évolutionnaire se compose de plusieurs étapes. Dans un premier temps, on initialise une population Π_0 en tirant p individus dans Ω de façon uniformément aléatoire, et on l'évalue en calculant les valeurs de $\mathcal{J}$ pour chaque individu. Vient ensuite une boucle itérative qui implémente le processus de sélection naturelle :

1. on sélectionne les individus les plus performants (au sens de $\mathcal{J}$) de la population Π_i à l'itération courante i, que l'on appelle *parents*,

2. on applique des opérateurs de variation aux parents sélectionnés, pour générer de nouveaux individus : les *enfants*. On parle de mutation si l'opérateur est 1-aire (i.e. ne prend qu'un seul individu en argument) et de croisement si l'opérateur est n-aire (i.e. prend n individus en argument, avec $n \geq 2$). Notons que cette étape est purement stochastique, et que les opérateurs n'utilisent pas d'information sur la performance des précédentes générations : on parle alors d'opérateurs semi-aveugles.
3. on évalue les enfants avec $\mathcal{J}$,
4. on remplace Π_i par une nouvelle population Π_{i+1} créée à partir des enfants et des parents de Π_i au moyen d'une sélection darwinienne.

Dans notre contexte, on peut assimiler Ω à l'espace des paramètres du terme source. Historiquement, les premières méthodes d'estimation basées sur les algorithmes génétiques (voir par exemple [Haupt, 2005]) supposaient connues certaines informations a priori comme les emplacements potentiels de la source.

L'utilisation des algorithmes génétiques pour la résolution des problèmes STE peut être illustré à travers plusieurs exemples :

— Dans [Allen et al., 2007] et [Haupt et al., 2007], il est envisagé de se placer dans une situation plus réaliste où les paramètres à estimer comprennent non seulement la position de la source et la masse totale émise, mais également la vitesse et la direction du vent. [Cervone and Franzese, 2011] proposent une extension du processus de variation en couplant les opérateurs de croisement et de mutation avec un algorithme non-darwinien d'inférence statistique.

— L'approche génétique a également permis d'étudier le rapport entre la précision du terme source reconstruit et le nombre de capteurs utilisés : [Long et al., 2010] reprend ce problème de précision en testant un algorithme génétique sur différentes configurations de réseaux de capteurs plus ou moins denses, et en tenant compte du caractère bruité des données fournies par les détecteurs.

— Plus récemment, la méthodologie a été validée sur des données expérimentales de rejet multi-sources, comme présenté dans les travaux de [Cantelli et al., 2015] où jusqu'à trois sources distinctes ont pu être caractérisées.

Méthodes bayésiennes par simulation stochastique

On a vu qu'il est possible de voir l'estimation du terme source sous un angle bayésien comme étant la recherche du maximum de la loi a posteriori de ses paramètres. Au lieu de se limiter à une estimation ponctuelle, une meilleure alternative consiste à calculer l'ensemble de cette distribution a posteriori, afin de pouvoir bénéficier de l'ensemble de l'information statistique disponible sur les paramètres à estimer.

L'étude menée par [Sohn et al., 2002] se penche sur le cas particulier de l'estimation d'une source à l'intérieur d'un bâtiment. Pour cela, un calcul initial simule un nombre N fixé de scénarios $S_1, \ldots, S_N$ possibles à partir d'un échantillon de paramètres $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_N$ potentiels de la source. Chaque scénario S_k représente un jeu de mesures simulées issues d'une source de paramètres $\boldsymbol{\theta}_k$.

Une fois cette collection constituée, la loi a posteriori suivante est calculée via la règle de Bayes, pour chaque scénario S_k :

$$p(S_k|\boldsymbol{\eta}) = \frac{p(S_k)p(\boldsymbol{\eta}|S_k)}{p(\boldsymbol{\eta})} \tag{1.22}$$

Cette première approche permet de caractériser un rejet, et nuance les performances de l'algorithme d'estimation en fonction du mode d'acquisition des mesures de concentrations. La comparaison est faite entre des observations acquises simultanément depuis tous les capteurs, et des observations recueillies de façon séquentielle, capteur par capteur.

Toutefois, l'expression analytique de la loi a posteriori des paramètres de la source n'est pas accessible de façon directe : du fait de la formulation présentée à l'équation (1.21), elle nécessite un calcul du modèle de dispersion à chaque fois qu'un jeu de paramètres doit être évalué. L'exploration systématique de l'espace des paramètres peut alors se révéler très coûteuse en temps de calcul. Il devient alors nécessaire d'introduire des méthodes numériques afin d'explorer de façon optimale cet espace : c'est l'objet des techniques de *simulation stochastique*, aussi appelées *méthodes de Monte-Carlo*.

Parmi ces algorithmes, la catégorie la plus connue est certainement celle des méthodes *Markov Chain Monte-Carlo* (MCMC). Dans [Keats et al., 2007] et [Chow et al., 2008], la méthode MCMC est employée dans un contexte urbain, et couplée à des modèles de dispersion suffisamment performants pour une prise en compte du milieu bâti. [Senocak et al., 2008] propose l'utilisation d'un modèle de dispersion plus simple de type gaussien, amélioré par l'ajout de paramètres stochastiques relatifs à la diffusion turbulente, eux-mêmes estimés par l'algorithme MCMC en plus des paramètres de la source. [Yee, 2008] étend la méthodologie aux cas où le nombre de sources est inconnu et considéré comme un paramètre supplémentaire à estimer, ajoutant ainsi une étape de sélection de modèle dans la procédure d'estimation. Cela est possible grâce à une méthode MCMC à *sauts réversibles*, qui associe pour chaque nombre de sources possible un espace de paramètres différent à explorer. Enfin, [Yee et al., 2014] propose une extension des concepts de [Keats et al., 2007] à l'échelle globale, où un algorithme MCMC est utilisé pour reconstruire la position et le profil d'émission d'une usine d'isotopes médicaux à partir des mesures de xénon relevées par le réseau mondial des capteurs de l'OTICE [3].

1.5 Problématique de recherche

Le développement des méthodes STE dans le contexte de la dispersion atmosphérique est une branche scientifique relativement récente par rapport au cadre général de la physique de l'atmosphère. Il s'agit également d'un domaine de recherche largement pluridisciplinaire, mêlant des compétences en physique mais également sur des aspects mathématiques (optimisation, statistiques, simulation aléatoire...) et informatiques (algorithmique, calcul scientifique haute performance...) variés.

3. L'*Organisation du Traité d'Interdiction Complète des Essais Nucléaires* (OTICE, ou CTBTO en anglais : *Comprehensive Nuclear-Test Ban Treaty Organization)* a pour rôle de détecter, grâce à un réseau de capteurs déployés sur l'ensemble du globe, et de signaler à tous les pays signataires du traité toute explosion d'origine atomique, pour prendre des mesures afin d'empêcher les puissances nucléaires actuelles de poursuivre leurs essais, et les Etats ne disposant pas de l'arme atomique de s'en doter.

Même si de nombreux aspects sont couverts par les éléments présents dans l'état de l'art exposé en paragraphe §1.4, de nombreuses questions restent encore ouvertes. Dans le cas particulier de situations accidentelles, il est ainsi important pour les primo-intervenants de disposer d'une information fiable avec une certaine quantification de l'incertitude, à partir d'un nombre potentiellement faible de mesures, et dans un intervalle de temps raisonnablement court pour assurer l'efficacité de l'intervention. Les questions de vitesse de calcul sont alors, de fait, importantes : si elles ne posent pas de problèmes dans le cadre d'une étude *a posteriori*, il en va autrement en situation de crise, où le temps de réponse à un incident est un paramètre primordial. Dans l'inventaire des méthodes existantes, les approches basées sur l'optimisation d'une fonction-coût fournissent une estimation déterministe, dont l'interprétation en termes d'évaluation de l'incertitude est plus difficile qu'avec des méthodes probabilistes. Les méthodes bayésiennes stochastiques ne sont pas non plus exemptes de défauts, à l'image de l'application de l'algorithme MCMC dans les travaux de [Chow et al., 2008], où les temps de calcul sont bien trop importants pour une application en situation opérationnelle. Ce cas d'étude illustre ainsi le fait que l'approche bayésienne peut rapidement se révéler coûteuse si l'algorithme choisi n'est pas suffisamment élaboré pour tenir la charge de calcul nécessaire au processus d'estimation.

Un autre aspect important est celui de la nature de la source à retrouver. Pour le contexte accidentel, si on se place dans le cas d'un attentat à la bombe sale ou dans celui d'une fuite sur un complexe industriel, l'hypothèse d'une source localisée est raisonnable. Toutefois il peut être intéressant de se pencher plus en détail sur le profil du rejet, celui-ci n'étant pas forcément instantané. Plusieurs techniques existantes introduisent des hypothèses sur le type de source recherché, notamment par rapport à son profil d'émission : celui-ci est le plus souvent considéré comme constant. Dans les cas où le rejet est potentiellement plus complexe, on a souvent l'information sur la position réelle de la source ou du moins sur un ensemble d'emplacements potentiels. Il n'existe pas à ce jour de cadre méthodologique combinant une localisation de la source sans a priori sur sa position et une estimation de son profil d'émission pouvant varier dans le temps.

Les travaux présentés dans ce manuscrit et ayant fait l'objet du travail de thèse ont donc pour but :

- *de développer et valider une méthode basée sur l'inférence bayésienne, capable de caractériser un point source par sa localisation et son profil de rejet à partir de mesures issues d'un réseau de capteurs,*
- *de coupler cette méthode avec un modèle de dispersion atmosphérique au sein d'une chaîne de calcul pouvant, à terme, être utilisée en situation opérationnelle.*

Pour cela, nous aurons recours à une approche bayésienne stochastique différente des algorithmes MCMC de par sa philosophie, et basée sur un principe d'échantillonnage d'importance adaptatif. De telles méthodes permettent en effet une convergence suffisamment rapide, et pallient certaines difficultés rencontrées par les MCMC. Nous associons cette méthode adaptative, utilisée pour localiser la source, à un calcul analytique du profil d'émission, permis par le choix d'une hypothèse de gaussianité sur le vecteur $\boldsymbol{\varsigma}$, et accompagné par l'implémentation d'une étape de contrainte visant à assurer la positivité de $\boldsymbol{\varsigma}$. Le schéma d'implémentation que nous avons conçu

permet ainsi de produire une estimation des paramètres de position et d'émission d'une source unique via un schéma itératif de couplage avec un modèle de dispersion, la procédure d'estimation à proprement parler demeurant indépendante du modèle choisi.

Pour développer cette thèse, le présent manuscrit est organisé de la façon suivante. Après avoir présenté la problématique posée par le sujet et effectué un premier tour d'horizon de la thématique dans ce chapitre introductif, nous développons au Chapitre 2 les principes régissant l'inférence bayésienne, en nous concentrant plus particulièrement sur les applications des méthodes de Monte-Carlo en statistique bayésienne. Nous y détaillons le cheminement théorique permettant la construction de l'algorithme AMIS, qui constitue un élément central des travaux de cette thèse. Dans le Chapitre 3, nous présentons l'application de cet algorithme adaptatif à la question de la reconstruction du terme source en dispersion atmosphérique. Pour cela, nous exploitons un cas d'application pratique issu d'une campagne de mesures expérimentales dont nous expliquons les conditions de réalisation. Le Chapitre 4 détaille une variante de la méthodologie présentée au Chapitre 3, avec l'utilisation d'une approche orientée récepteur, et l'emploi d'un code de dispersion de type lagrangien dans des configurations simulées de milieu naturel et de milieu urbain. Enfin, la dernière partie de ce manuscrit présentera une conclusion générale sur les résultats présentés, et introduira les différentes perspectives que ceux-ci ont permis d'ouvrir.

Chapitre 2

Inférence bayésienne et méthodes de Monte-Carlo

Ce chapitre introduit la méthodologie et les objectifs de l'inférence bayésienne, puis présente les principes de base régissant les méthodes de Monte Carlo. Un exposé plus détaillé est développé autour des deux grandes familles des techniques Monte Carlo, à savoir les *Markov Chain Monte Carlo* et les algorithmes d'échantillonnage d'importance. Concernant ces derniers, l'aspect adaptatif sur la loi de proposition est introduit au travers des méthodes Population Monte Carlo, puis complété par la notion d'adaptativité multiple qui caractérise l'algorithme AMIS.

2.1 Eléments de statistique bayésienne

2.1.1 L'intérêt d'une approche statistique

Le cadre des problèmes STE offre plusieurs arguments en faveur d'une approche de type statistique.

La physique de l'atmosphère ainsi que les systèmes instrumentaux permettant d'en mesurer les variables sont soumis à des phénomènes de fluctuations aléatoires dont il est, par définition, impossible de quantifier la nature exacte. Le fait d'utiliser un cadre de travail probabiliste et de considérer la nature statistique des grandeurs d'intérêt prend donc tout son sens dans un tel contexte. Une approche statistique permet en effet de s'appuyer sur le socle théorique rigoureux des probabilités afin de chercher à estimer non plus des valeurs ponctuelles par des méthodes déterministes, mais plutôt des distributions de probabilité liées à ces valeurs. Il devient alors possible :

— de modéliser *l'incertitude relative aux données observées* en considérant qu'elles sont affectées par un certain type de *bruit*, qui est alors défini comme suivant une loi de probabilité spécifique,

— de quantifier *l'incertitude autour des valeurs estimées* pour la grandeur d'intérêt, en l'occurrence les paramètres d'un terme source dans notre cas. Comme on travaille désormais sur des distributions statistiques, on peut s'appuyer sur un *estimateur* afin de caractériser le

résultat recherché, ainsi que, par exemple, sur la *variance* qui lui est associée pour évaluer le degré de plausibilité du résultat d'estimation.

Erreurs d'observation

Toute l'*information disponible* pour traiter un problème STE est contenue dans les observations issues du réseau de capteurs. Or ces mesures ne représentent pas parfaitement la "vérité physique" illustrant le rejet caractérisé, celle-ci étant altérée par plusieurs facteurs :

— **la représentativité spatiale du réseau** : plus le nombre de capteurs est important, mieux le phénomène sera mesuré. Cependant, en pratique, le cas parfait où chaque point du domaine est instrumenté n'existe pas ;

— **la représentativité temporelle du capteur** : les observations sont obtenues suivant une certaine fréquence d'acquisition. Celle-ci peut être définie par les procédés physiques inhérents à la mesure, ceux-ci pouvant impliquer une chaîne de traitement plus ou moins complexe avant obtention d'une valeur de concentration. Dans d'autres cas l'acquisition de ces valeurs peut être quasi- instantanée, mais les contraintes de stockage et de transfert des données font qu'une opération de moyennage devient nécessaire afin de réduire la quantité des observations à un nombre compatible avec les capacités du dispositif instrumental considéré. Dans tous les cas, cela génère un biais sur les observations.

— les différences sources de bruit imputables à **l'électronique du capteur**.

Erreurs du modèle

Une autre source d'incertitude concerne l'*information générée* par le modèle de dispersion, elle-même confrontée à l'information disponible pour inférer les paramètres de la source. En effet, les approximations consécutives à la mise en équation des phénomènes de dispersion engendrent déjà un premier écart par rapport à la réalité. Ces approximations peuvent être de nature physique ou numérique (par exemple les dimensions du maillage pour un modèle eulérien, ou le nombre de particules numériques émises pour un modèle lagrangien). A cela viennent ensuite s'ajouter les incertitudes autour des données météorologiques, qu'elles soient obtenues grâce à une station de mesure ou par un modèle de prévision.

Dans la suite, nous décrivons un formalisme classique dès qu'il s'agit d'inférence statistique qui est celui du paradigme bayésien. L'analyse des incertitudes dans les méthodes STE constitue un volet de recherche important (voir par exemple [Rodriguez et al., 2011] et [Rodriguez et al., 2013]), mais n'est pas l'objet du travail de recherche présenté dans ce manuscrit.

2.1.2 Formulation du paradigme bayésien

Dans le cadre d'une approche statistique, on considère que les éléments $\eta^{(1)}, \eta^{(2)}, \ldots, \eta^{(m)}$ du vecteur d'observation $\boldsymbol{\eta}$ proviennent de lois de probabilités de paramètres respectifs $\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}, \ldots, \boldsymbol{\theta}^{(m)}$. Pour un problème STE, les observations sont *indépendantes*, et peuvent également être considérées

comme *identiquement distribuées*, c'est-à-dire suivant une unique loi de probabilité de paramètre $\boldsymbol{\theta} \in \Theta$, où Θ est un espace de dimension finie. Dans ce cas, cette loi peut alors s'écrire :

$$p(\boldsymbol{\eta}|\boldsymbol{\theta}) = \prod_{i=1}^{m} p(\eta^{(i)}|\boldsymbol{\theta}) \tag{2.1}$$

Dans la démarche bayésienne, l'objectif est d'exploiter au mieux l'information apportée par $\boldsymbol{\eta}$ sur le paramètre $\boldsymbol{\theta}$, pour ensuite créer des procédures d'inférence sur $\boldsymbol{\theta}$. Pour cela, l'idée de base consiste à définir $\boldsymbol{\theta}$ non plus comme un paramètre déterministe, mais désormais comme une variable aléatoire. D'après la définition des probabilités conditionnelles, on a alors :

$$p(\boldsymbol{\theta}|\boldsymbol{\eta}) = \frac{p(\boldsymbol{\theta}, \boldsymbol{\eta})}{p(\boldsymbol{\eta})} \tag{2.2}$$

où $p(\boldsymbol{\theta}, \boldsymbol{\eta})$ est la loi jointe de $\boldsymbol{\theta}$ et $\boldsymbol{\eta}$. Par cette même définition on a également :

$$p(\boldsymbol{\eta}|\boldsymbol{\theta}) = \frac{p(\boldsymbol{\theta}, \boldsymbol{\eta})}{p(\boldsymbol{\theta})} \tag{2.3}$$

En combinant (2.2) et (2.3) on obtient ainsi :

$$p(\boldsymbol{\theta}|\boldsymbol{\eta}) = \frac{p(\boldsymbol{\theta})p(\boldsymbol{\eta}|\boldsymbol{\theta})}{p(\boldsymbol{\eta})} \tag{2.4}$$

Il s'agit de la *règle de Bayes*, qui peut s'écrire sous sa forme complète grâce à la définition de la loi marginale $p(\boldsymbol{\eta})$:

$$p(\boldsymbol{\eta}) = \int_{\Theta} p(\boldsymbol{\theta}, \boldsymbol{\eta}) d\boldsymbol{\theta} = \int_{\Theta} p(\boldsymbol{\theta}) p(\boldsymbol{\eta}|\boldsymbol{\theta}) d\boldsymbol{\theta} \tag{2.5}$$

D'où :

$$p(\boldsymbol{\theta}|\boldsymbol{\eta}) = \frac{p(\boldsymbol{\theta})p(\boldsymbol{\eta}|\boldsymbol{\theta})}{\int_{\Theta} p(\boldsymbol{\theta})p(\boldsymbol{\eta}|\boldsymbol{\theta}) d\boldsymbol{\theta}} \tag{2.6}$$

Le terme $p(\boldsymbol{\eta}|\boldsymbol{\theta})$ contient l'information fournie par l'observation, et définit la *fonction de vraisemblance* de $\boldsymbol{\theta}$. L'incertitude sur $\boldsymbol{\theta}$ peut désormais être traduite par une loi de probabilité $p(\boldsymbol{\theta})$, ce qui signifie que $\boldsymbol{\theta}$ suit cette loi en l'absence d'information d'observation. Cette loi est également appelée loi *a priori*. Enfin, la loi de probabilité de $\boldsymbol{\theta}$ après acquisition des observations $\boldsymbol{\eta}$ et définie par $p(\boldsymbol{\theta}|\boldsymbol{\eta})$ est appelée loi *a posteriori*.

Dans l'équation (2.6), la loi marginale $p(\boldsymbol{\eta})$ sert de constante de normalisation pour que $p(\boldsymbol{\theta}|\boldsymbol{\eta})$ réponde bien à la définition d'une densité de probabilité. Comme cette constante ne dépend pas de $\boldsymbol{\theta}$, on utilise souvent la notation de proportionnalité suivante :

$$p(\boldsymbol{\theta}|\boldsymbol{\eta}) \propto p(\boldsymbol{\theta})p(\boldsymbol{\eta}|\boldsymbol{\theta}) \tag{2.7}$$

En une seule équation, le raisonnement bayésien permet ainsi de joindre l'information d'observation, l'information initialement disponible ou manquante sur le paramètre d'intérêt, et toutes les incertitudes associées grâce au cadre de la théorie des probabilités.

A propos du choix de l'a priori

La quantification de l'information disponible a priori est souvent perçue comme l'une des difficultés principales en statistique bayésienne, celle-ci étant faite de façon relativement arbitraire et ne faisant pas systématiquement de lien avec une distribution connue. Il existe néanmoins un ensemble de règles permettant de mieux appréhender cette étape.

— Il est d'abord nécessaire de définir le cas où aucune information préalable n'est disponible. Il reste alors possible d'utiliser la règle de Bayes grâce à la définition d'un type d'a priori non-informatif appelé "a priori de Jeffreys" et défini par la loi :

$$p(\boldsymbol{\theta}) \propto \sqrt{\mathcal{I}(\boldsymbol{\theta}|\boldsymbol{\eta})} \tag{2.8}$$

où $\mathcal{I}(\boldsymbol{\theta}|\boldsymbol{\eta})$ est l'information de Fisher de la vraisemblance :

$$\mathcal{I}(\boldsymbol{\theta}|\boldsymbol{\eta}) = -\mathbb{E}_{\boldsymbol{\theta}}\left[\frac{\partial^2 \log p(\boldsymbol{\eta}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^2}\right] \tag{2.9}$$

— Il existe également des cas où il n'y a pas de raison de favoriser initialement une ou plusieurs valeurs de $\boldsymbol{\theta}$. On peut alors choisir une loi uniforme pour $p(\boldsymbol{\theta})$, à condition que le support de cette distribution soit compact, i.e. que l'ensemble des valeurs pouvant être prises par $\boldsymbol{\theta}$ soit contenu dans un intervalle fermé. A la suite de l'équation (2.7), la loi a posteriori devient alors simplement proportionnelle à la vraisemblance. Notons que le fait de borner $\boldsymbol{\theta}$ introduit d'ores et déjà de l'information a priori, une loi uniforme n'est donc pas strictement non-informative.

— Dans un cadre plus général, la famille de la distribution a priori est souvent choisie de façon à faciliter le calcul de la loi a posteriori. Cela est possible grâce à l'utilisation des *a priori conjugués* qui, pour une vraisemblance donnée, retournent une distribution a posteriori de la même famille que l'a priori, et dont les paramètres peuvent être calculés de façon analytique.

2.1.3 Estimateurs ponctuels

Comme $\boldsymbol{\theta}$ est désormais une variable aléatoire, sa distribution a posteriori suffit pour prendre la décision sous-jacente à son interprétation. On peut donc caractériser cette loi par un estimateur ponctuel $\hat{\boldsymbol{\theta}}$, dont la construction est choisie selon la façon dont on interprète l'écart entre les valeurs estimées et les valeurs réelles du paramètre.

On peut ainsi définir une fonction-coût, ou *fonction de perte* $L(\varrho, \boldsymbol{\theta})$, où ϱ est l'approximation de $\boldsymbol{\theta}$ issue du processus d'estimation. L'*estimateur bayésien* (ou *estimateur optimal*) est par définition celui qui minimise le *risque bayésien*, dont la valeur est donnée par :

$$\mathcal{R}(\varrho) = \int_{\mathbb{R}^m} \int_{\Theta} L(\varrho, \boldsymbol{\theta}) p(\boldsymbol{\eta}|\boldsymbol{\theta}) p(\boldsymbol{\theta}) d\boldsymbol{\eta} d\boldsymbol{\theta} \tag{2.10}$$

D'après le paragraphe §1.3 de [Robert and Casella, 2004], cela revient de façon équivalente à

minimiser la *perte a posteriori* :

$$\mathbb{E}[L(\varrho, \boldsymbol{\theta})|\boldsymbol{\eta}] = \int_{\Theta} L(\varrho, \boldsymbol{\eta}) p(\boldsymbol{\theta}|\boldsymbol{\eta}) d\boldsymbol{\theta} \tag{2.11}$$

Pour cela, une approche courante consiste à écrire une fonction-coût de type quadratique :

$$L(\varrho, \boldsymbol{\theta}) = ||\varrho - \boldsymbol{\theta}||^2 \tag{2.12}$$

Dans ce cas, l'estimateur optimal est *l'espérance a posteriori*, aussi appelé MMSE (*Minimum Mean Square Error*), et s'écrit comme l'intégration de $\boldsymbol{\theta}$ sous la mesure $p(\boldsymbol{\theta}|\boldsymbol{\eta})$:

$$\varrho_{MMSE} = \hat{\boldsymbol{\theta}}_{MMSE} = \mathbb{E}_{\boldsymbol{\theta}|\boldsymbol{\eta}}[\boldsymbol{\theta}] = \int_{\Theta} \boldsymbol{\theta} p(\boldsymbol{\theta}|\boldsymbol{\eta}) d\boldsymbol{\theta} \tag{2.13}$$

Une autre possibilité consiste à choisir la fonction-coût 0-1, qui associe un coût nul à une estimation correcte (i.e. $\delta = \boldsymbol{\theta}$) et un coût unitaire à n'importe quelle autre valeur. On montre alors que l'estimateur optimal est le *maximum a posteriori* (MAP) et s'écrit :

$$\varrho_{MAP} = \hat{\boldsymbol{\eta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, p(\boldsymbol{\theta}|\boldsymbol{\eta}) \tag{2.14}$$

En pratique il s'agit simplement du mode de la loi a posteriori $p(\boldsymbol{\theta}|\boldsymbol{\eta})$.

2.2 Méthodes de Monte Carlo

2.2.1 Définitions et principes

Historiquement, le terme "Monte Carlo" a été inventé par N. Metropolis en 1947, en faisant allusion aux jeux de hasard pratiqués dans le casino du même nom, situé au coeur de la Principauté de Monaco. L'expression fût reprise en 1949 dans les travaux fondateurs de Von Neumann, Metropolis, et Ulam, publiés dans [Metropolis and Ulam, 1949].

Des différentes définitions données aux méthodes de Monte Carlo, celle énoncée dans [Halton, 1970] est sans doute la plus appropriée pour caractériser les travaux présentés dans ce manuscrit. Elle énonce les objectifs de ces méthodes comme étant :

- la représentation de la solution d'un problème par les paramètres d'une population statistique hypothétique,
- l'utilisation d'une suite aléatoire de valeurs pour construire un ensemble d'échantillons de cette population, dont les paramètres statistiques peuvent ensuite être obtenus et répondre au problème initial.

Dans un contexte général, cela revient à tirer depuis un espace Υ un vecteur $\boldsymbol{\upsilon}$ de N éléments (ou N-échantillon) $\upsilon^{(1)}, \upsilon^{(2)}, \ldots, \upsilon^{(N)}$ indépendants et identiquement distribués pour simuler la densité $p(\boldsymbol{\upsilon})$ définie comme étant la *loi cible*, i.e. celle dont les paramètres permettent de résoudre

le problème initial. Ce N-échantillon permet d'approximer $p(\boldsymbol{\upsilon})$ par une loi $p_N(\boldsymbol{\upsilon})$ qui peut s'écrire :

$$p_N(\boldsymbol{\upsilon}) = \frac{1}{N}\sum_{i=1}^{N}\delta_{\upsilon^{(i)}}(\boldsymbol{\upsilon}) \tag{2.15}$$

où $\delta_{\upsilon^{(i)}}(\boldsymbol{\upsilon})$ représente la distribution de Dirac centrée au point $\upsilon^{(i)}$.

On peut utiliser l'approximation de Monte Carlo pour approcher l'espérance de n'importe quelle fonction d'une variable aléatoire : il suffit pour cela de tirer des échantillons depuis la loi de cette variable, puis de calculer la moyenne arithmétique de la fonction en question appliquée aux échantillons. En d'autres termes, l'intégrale sur la variable aléatoire X suivante :

$$\mathbb{E}_p[\psi(X)] = \int \psi(x)p(x)dx \tag{2.16}$$

peut ainsi être approximée par :

$$\mathbb{E}_p[\psi(X)] \simeq \frac{1}{N}\sum_{i=1}^{N}\psi(x^{(i)}) \tag{2.17}$$

Une autre fonctionnalité des méthodes Monte Carlo, directement liée à l'inférence bayésienne appliquée aux problèmes inverses, concerne la simulation de distribution de probabilités complexes. Généralement, la valeur liant les observations au modèle de données et représentée par la vraisemblance $p(\boldsymbol{\eta}|\boldsymbol{\theta})$, présente un aspect fortement non-linéaire induit par la nature des phénomènes physiques mis en jeu. Par conséquent, il devient difficile d'échantillonner facilement depuis la loi a posteriori $p(\boldsymbol{\theta}|\boldsymbol{\eta})$, qui devient ici la loi cible.

La solution consiste alors à tirer depuis une loi alternative plus accessible appelée *loi de proposition*, et de choisir les paramètres de cette loi de façon à ce que les éléments qui en sont issus reflètent au mieux les propriétés statistiques de la loi cible. C'est dans cette optique que nous allons développer les différentes familles de méthodes Monte Carlo présentées dans la suite de ce chapitre.

2.2.2 Algorithmes Markov Chain Monte Carlo (MCMC)

Les méthodes *Markov Chain Monte Carlo* (MCMC) associent l'approche Monte Carlo classique avec un outil spécifique pour la simulation aléatoire appelé *chaîne de Markov*.

A propos des chaînes de Markov

Une chaîne de Markov est un processus aléatoire $(\boldsymbol{\theta}^{(0)}, \boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(N)})$ à temps discret tel que, $\forall i \in \{1, \ldots, N\}$ la loi de l'état i ne dépend uniquement que de celui qui le précède (donc l'état $i-1$). Autrement dit :

$$p(\boldsymbol{\theta}^{(i)}|\boldsymbol{\theta}^{(0)}, \boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(i-1)}) = p(\boldsymbol{\theta}^{(i)}|\boldsymbol{\theta}^{(i-1)}) \tag{2.18}$$

Toute chaîne de Markov peut être caractérisée par :

— une distribution marginale $p_0(\boldsymbol{\theta})$ permettant d'échantillonner le premier élément $\boldsymbol{\theta}^{(0)}$ du processus,

— un *noyau de transition* $\mathcal{K}(\boldsymbol{\theta}^{(i-1)}, \boldsymbol{\theta}^{(i)})$, aussi noté $\mathcal{K}(\boldsymbol{\theta}^{(i)}|\boldsymbol{\theta}^{(i-1)})$, qui est la distribution conditionnelle de $\boldsymbol{\theta}^{(i)}$ sachant $\boldsymbol{\theta}^{(i-1)}$ et qui définit la façon dont on passe d'un état à un autre.

On peut toutefois distinguer certains types particuliers de chaînes définis par des propriétés spécifiques. Ainsi, une chaîne de Markov est dite :

— *homogène* si son noyau de transition est invariant durant toute la durée de la simulation,

— *irréductible* si tout état du processus est accessible à partir de n'importe quel autre état, i.e. $\forall i, j \in \{0, \ldots, N\};\ \mathcal{K}(\boldsymbol{\theta}^{(i-1)}, \boldsymbol{\theta}^{(j)}) > 0$,

— *récurrente positive* si pour tout état du processus, la probabilité de retour à cet état est toujours non-nulle,

— *apériodique* si la chaîne ne se retrouve jamais bloquée dans la répétition infinie d'une suite donnée d'états [1].

La propriété qui nous intéresse concerne la *loi de probabilité stationnaire* : si, à partir d'un certain seuil, chacune des variables aléatoires de la chaîne suit la même loi, alors on dit que l'état stationnaire est atteint. La convergence vers une telle loi est garantie si la chaîne de Markov est dite *ergodique*, i.e. si elle respecte les propriétés d'homogénéité, d'irréductibilité, de récurrence positive et d'apériodicité. En d'autres termes, pour tout état j d'une chaîne de Markov ergodique :

$$\lim_{n \to +\infty} P(\boldsymbol{\theta}^{(n)} = j) = \pi(j) \tag{2.19}$$

où π est la distribution stationnaire de la chaîne.

La stratégie des méthodes MCMC consiste ainsi à définir une chaîne de Markov ergodique dont la distribution stationnaire est la loi cible recherchée, soit dans le cadre bayésien la loi a posteriori $p(\boldsymbol{\theta}|\boldsymbol{\eta})$, que nous noterons $\pi(\boldsymbol{\theta})$. Une fois l'état stationnaire atteint, les éléments tirés depuis la loi associée permettent d'approximer empiriquement la loi cible.

Algorithme de Metropolis-Hastings

C'est une méthode qui fut originalement présentée dans [Metropolis et al., 1953], puis plus largement développée dans [Hastings, 1970]. L'idée de base de l'algorithme Metropolis-Hastings (MH) est de proposer une transition de l'état courant $\boldsymbol{\theta}^{(i-1)}$ vers un nouvel état $\boldsymbol{\theta}^{(i)}$ en tirant un candidat $\boldsymbol{\theta}^*$ suivant la loi $\mathcal{K}(\boldsymbol{\theta}^{(i-1)}, \boldsymbol{\theta}^*)$ qui est ensuite soumis à une épreuve d'acceptation-rejet.

On choisit souvent $\mathcal{K}$ comme étant une distribution gaussienne symétrique et centrée sur l'état courant, i.e. $\mathcal{K}(\boldsymbol{\theta}^{(i-1)}, \boldsymbol{\theta}^*) = \mathcal{N}(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(i-1)}, \boldsymbol{\Sigma})$ [2] : on parle alors de *Random Walk Metropolis*.

1. La définition mathématique est ici un peu plus complexe. Considérons un état $i \in \{1, \ldots, N\}$: on peut alors définir la *période* $d(i)$ de i comme étant le plus grand diviseur commun (PGCD) de tous les entiers $k \geq 1$ pour lesquels on a $p(\boldsymbol{\theta}^{(k)} = i|\boldsymbol{\theta}^{(0)} = i)$. Si $d(i) = 1$, on dit que l'état i est apériodique. Par extension, si tous les états de la chaîne sont apériodiques, alors la chaîne elle-même est dite apériodique.

2. La notation $\mathcal{N}(x|\mu, \sigma^2)$ désigne l'évaluation au point x de la densité de probabilité d'une loi normale de moyenne μ et de variance σ^2.

Une autre possibilité consiste à utiliser un noyau de transition où le nouvel état ne dépend pas de l'état courant, autrement dit $\mathcal{K}(\boldsymbol{\theta}^{(i-1)}, \boldsymbol{\theta}^*) = \mathcal{K}(\boldsymbol{\theta}^*)$: cette approche est alors logiquement appelée *Independent Metropolis-Hastings*.

En pratique, une fois que $\boldsymbol{\theta}^*$ est obtenu, on cherche à déterminer si on accepte ou non ce candidat, en calculant une probabilité d'acceptation r définie par :

$$\begin{aligned} r &= \min\{1, \beta\} \\ \beta &= \frac{\pi(\boldsymbol{\theta}^*)\mathcal{K}(\boldsymbol{\theta}^*, \boldsymbol{\theta}^{(i-1)})}{\pi(\boldsymbol{\theta}^{(i-1)})\mathcal{K}(\boldsymbol{\theta}^{(i-1)}, \boldsymbol{\theta}^*)} \end{aligned} \tag{2.20}$$

Si $\mathcal{K}$ est symétrique, alors cette grandeur est donnée par :

$$r = \min\left\{1, \frac{\pi(\boldsymbol{\theta}^*)}{\pi(\boldsymbol{\theta}^{(i-1)})}\right\} \tag{2.21}$$

On voit facilement que si $\boldsymbol{\theta}^*$ est plus proche de la loi cible que $\boldsymbol{\theta}^{(i-1)}$, alors il devient avec certitude le nouvel état, car $\frac{\pi(\boldsymbol{\theta}^*)}{\pi(\boldsymbol{\theta}^{(i-1)})} > 1$. Toutefois, si ce n'est pas le cas, on tire une valeur u suivant une loi uniforme sur $[0, 1]$, et le passage de $\boldsymbol{\theta}^*$ au statut de nouvel état est alors conditionné par la valeur de r par rapport à celle de u. Une telle démarche permet ainsi de ne pas complètement discriminer les états candidats dont la probabilité est moins élevée.

La génération des différents états de la chaîne de Markov s'effectue ainsi suivant un schéma itératif décrit dans l'algorithme (1).

Algorithm 1. Metropolis-Hastings

Initialiser $\boldsymbol{\theta}^{(0)}$
for $i = 1, \ldots, N$ **do**
  Tirer $\boldsymbol{\theta}^*$ depuis $\mathcal{K}(\boldsymbol{\theta}^{(i-1)}, \boldsymbol{\theta}^*)$
  Calculer le ratio $\beta = \frac{\pi(\boldsymbol{\theta}^*)\mathcal{K}(\boldsymbol{\theta}^*, \boldsymbol{\theta}^{(i-1)})}{\pi(\boldsymbol{\theta}^{(i-1)})\mathcal{K}(\boldsymbol{\theta}^{(i-1)}, \boldsymbol{\theta}^*)}$
  Calculer la probabilité d'acceptation $r = \min\{1, \beta\}$
  Tirer ζ depuis $\mathcal{U}_{[0,1]}$
  if $\zeta < r$ **then**
    $\boldsymbol{\theta}^{(i)} = \boldsymbol{\theta}^*$
  else
    $\boldsymbol{\theta}^{(i)} = \boldsymbol{\theta}^{(i-1)}$
  end if
end for

Pour illustrer le fonctionnement de l'algorithme MH, nous pouvons prendre l'exemple d'une mixture de deux lois normales, dont la loi s'écrit :

$$\pi(\boldsymbol{\theta}) = \gamma\mathcal{N}(\boldsymbol{\theta}|\mu_1, \sigma_1^2) + (1-\gamma)\mathcal{N}(\boldsymbol{\theta}|\mu_2, \sigma_2^2) \tag{2.22}$$

Le paramètre γ permet de définir l'influence de chacune des composantes dans la mixture. Pour la suite, nous reprenons les paramètres utilisés dans le chapitre 24 de [Murphy, 2012], à

savoir $\gamma = 0.3$, $\mu_1 = -20$, $\sigma_1 = 10$, (en orange sur la figure 2.1), et $\mu_2 = 20$, $\sigma_2 = 10$ (en bleu sur la figure 2.1)

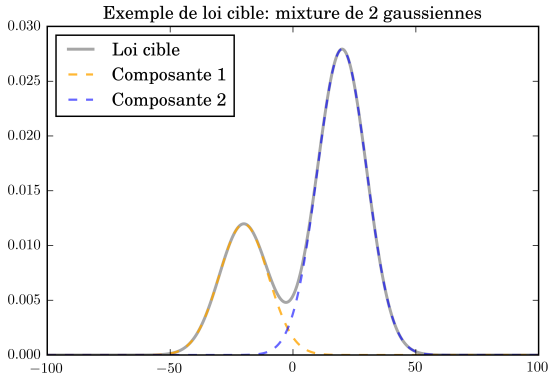


FIGURE 2.1 – Exemple de loi cible : mixture de deux gaussiennes

Supposons que l'on cherche à échantillonner depuis cette distribution : nous définissons un noyau de transition gaussien afin d'appliquer une méthodologie de type *random-walk Metropolis* :

$$\mathcal{K}(\boldsymbol{\theta}, \boldsymbol{\theta}^*) = \mathcal{N}(\boldsymbol{\theta}^* | \boldsymbol{\theta}, \sigma_K^2) \tag{2.23}$$

La figure 2.2 illustre ainsi l'approximation de π grâce aux histogrammes des échantillons obtenus à différentes itérations de l'algorithme.

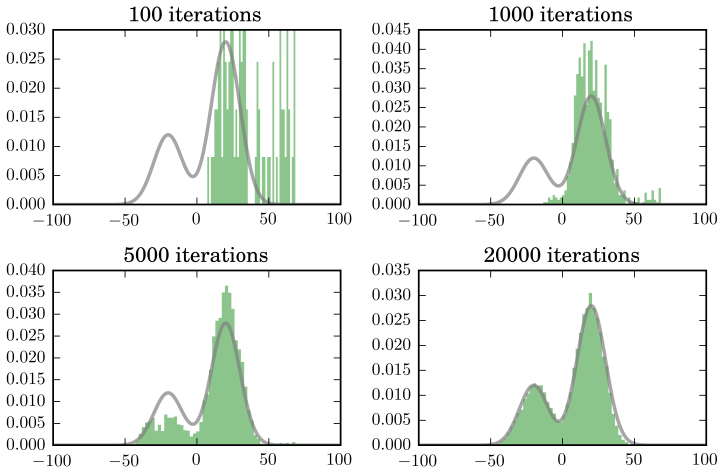


FIGURE 2.2 – Résultats de l'algorithme MH : histogrammes des éléments tirés (en vert) et comparaison avec la loi cible (en rouge)

On remarque qu'avant d'atteindre la distribution stationnaire, il existe un régime transitoire durant lequel les états de la chaîne ne permettent pas de représenter correctement la loi cible. Il est alors d'usage de ne pas tenir compte des éléments tirés durant cette période , appelée *temps de chauffe* ou *burn-in*.

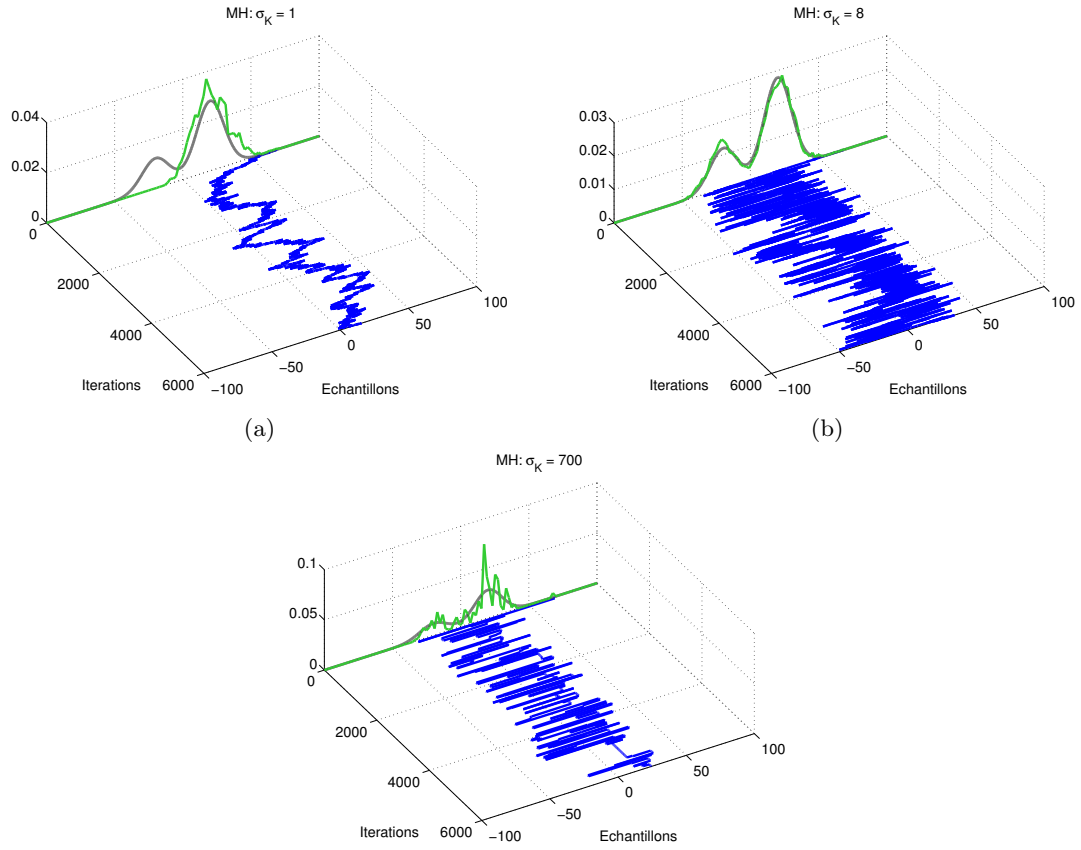


FIGURE 2.3 – Exemples de réalisations de l'algorithme MH avec 3 valeurs différentes pour la variance du noyau de transition

La vitesse de convergence et la qualité de l'estimation sont conditionnés par deux facteurs importants :

— **l'initialisation de la chaîne** : plus l'état de départ est proche d'une valeur ayant une probabilité élevée suivant la loi cible, plus vite l'algorithme convergera. S'il est impossible de favoriser a priori une ou plusieurs valeurs de l'espace d'état, on peut simplement tirer l'état initial avec une loi uniforme sur cet espace.

— **le choix du noyau de transition** : dans le cas d'un random-walk Metropolis, le paramètre σ_K reflète l'écart potentiel entre deux états consécutifs de la chaîne. Si celui-ci est trop important, les amplitudes des transitions proposées par l'algorithme deviennent trop importantes, causant un rejet fréquent des états candidats. Cela se traduit par la présence de plateaux sur la trajectoire de la chaîne, comme on le voit sur la figure 2.3c. A l'inverse, si σ_K est choisi trop petit, l'exploration de l'espace d'état ne se fait pas de façon optimale, car la chaîne évolue beaucoup plus lentement, avec le risque de ne pas atteindre certains états représentatifs de la loi cible du fait de leur "éloignement". On retrouve ce phénomène sur la figure 2.3a, où la chaîne reste bloquée sur l'un des modes, lui empêchant d'échantillonner depuis la composante de moyenne $\mu_1 = -20$. Il faut donc choisir une valeur offrant un bon compromis, et qui permet de bien retrouver la loi cible, comme c'est le cas sur la figure 2.3b.

Echantillonneur de Gibbs

L'échantillonneur de Gibbs a initialement été proposé dans [Geman and Geman, 1984] puis repris et développé par [Gelfand and Smith, 1990]. Il permet de passer du problème de l'échantillonnage d'un état $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)$ sur un espace pouvant être potentiellement de grande dimension, à une succession de sous-problèmes de dimension réduite, en générant successivement chacun des éléments composant $\boldsymbol{\theta}$ à l'aide de leurs dépendances par rapport aux autres. C'est ainsi la loi conditionnelle de chaque élément de $\boldsymbol{\theta}$ sous la loi cible qui fait office de loi de proposition.

Le parcours des composantes de $\boldsymbol{\theta}$ peut se faire de façon déterministe, par exemple en les traitant dans leur ordre naturel les unes après les autres : on parle alors de *balayage systématique*, ou *systematic scan*, comme illustré dans l'algorithme 2. Une autre possibilité consiste à choisir aléatoirement cet ordre d'exploration : il s'agit du *balayage aléatoire*, ou *random scan*.

Algorithm 2. Echantillonneur de Gibbs (balayage systématique)

Initialiser $\boldsymbol{\theta}^{(0)} = \left(\theta_1^{(0)}, \ldots, \theta_p^{(0)}\right)$
for $i = 1, \ldots, N$ **do**
  for $k = 1, \ldots, p$ **do**
    Tirer $\theta_k^{(i)}$ depuis $p(\theta_k^{(i)} | \theta_1^{(i)}, \ldots, \theta_{k-1}^{(i)}, \theta_{k+1}^{(i-1)}, \ldots, \theta_p^{(i-1)})$
  end for
end for

Pour illustrer le fonctionnement de ce type d'algorithme, nous allons chercher à échantillonner depuis une gaussienne bivariée à deux dimensions. Dans ce cas, on va donc noter $\boldsymbol{\theta} = (\theta_1, \theta_2)^T$, on a ainsi $\boldsymbol{\theta} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ avec :

$$\begin{aligned} \boldsymbol{\mu} &= (\mu_1, \mu_2)^T \\ \boldsymbol{\Sigma} &= \begin{pmatrix} \sigma_1^2 & \rho \\ \rho & \sigma_2^2 \end{pmatrix} \end{aligned} \tag{2.24}$$

où σ_1^2 et σ_2^2 sont les variances respectives de θ_1 et θ_2, et ρ est la corrélation entre θ_1 et θ_2 définie par :

$$\rho = \frac{\mathrm{Cov}(\theta_1, \theta_2)}{\sigma_1 \sigma_2}$$

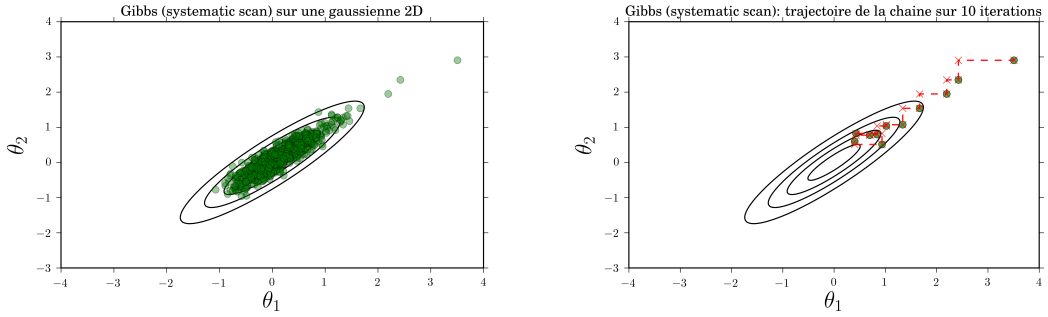
.

Le cas gaussien bivarié est relativement facile à traiter, car on sait calculer directement les expressions des lois conditionnelles, qui sont elle-mêmes gaussiennes, et qui s'écrivent sous la forme suivante :

$$\begin{aligned} p(\theta_1|\theta_2) &= \mathcal{N}\left(\theta_1 \middle| \mu_1 + \sigma_1 \rho \left(\frac{\theta_2 - \mu_2}{\sigma_2}\right), \sigma_1^2(1-\rho)\right) \\ p(\theta_2|\theta_1) &= \mathcal{N}\left(\theta_2 \middle| \mu_2 + \sigma_2 \rho \left(\frac{\theta_1 - \mu_1}{\sigma_1}\right), \sigma_2^2(1-\rho)\right) \end{aligned} \tag{2.25}$$

La figure 2.4b reflète bien le comportement de l'algorithme : la trajectoire de la chaîne (en pointillés gris) met en évidence le fait que les composantes de $\boldsymbol{\theta}$ sont évaluées une par une.

Cette famille d'algorithmes, contrairement à l'approche MH, n'effectue pas de sélection des états échantillonnés, et exploite ainsi toute l'information générée lors de la simulation. Cependant,



(a) (b)

FIGURE 2.4 – Illustration de l'échantillonneur de Gibbs pour une loi cible de type gaussienne bivariée (en noir) avec le résultat sur 800 itérations (a) et sur 10 itérations avec trajectoire (b)

elle nécessite de pouvoir échantillonner depuis les lois conditionnelles, ce qui n'est généralement pas facile. Aussi, dans ces cas-là, une variante consiste à avoir recours à une itération de type MH avec une loi de proposition choisie par l'utilisateur : cette méthode porte le nom de *Metropolis-Within-Gibbs* (MWG).

2.2.3 Echantillonnage d'importance classique (IS)

Les algorithmes MCMC se basent sur une procédure d'acceptation-rejet pour corriger le processus de tirage depuis la loi cible. Une autre approche de ce problème est proposée par les méthodes d'échantillonnage d'importance, ou *importance sampling* (IS), où la correction se fait cette fois via une procédure de pondération de chacun des échantillons tirés depuis la loi de proposition. Dans la suite de ce paragraphe, nous emploierons le terme de *particules* afin de désigner les éléments d'un échantillon.

L'idée de départ des méthodes IS est de chercher à approximer des intégrales de la forme :

$$I = \mathbb{E}_\pi[f(\boldsymbol{\theta})] = \int f(\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta} \tag{2.26}$$

Pour cela, on définit une loi de proposition φ permettant de réécrire l'équation (2.26) sous la forme :

$$\mathbb{E}_\pi[f(\boldsymbol{\theta})] = \int f(\boldsymbol{\theta})\frac{\pi(\boldsymbol{\theta})}{\varphi(\boldsymbol{\theta})}\varphi(\boldsymbol{\theta})d\boldsymbol{\theta} = \int f(\boldsymbol{\theta})w(\boldsymbol{\theta})\varphi(\boldsymbol{\theta})d\boldsymbol{\theta} \tag{2.27}$$

où $w(\boldsymbol{\theta}) = \dfrac{\pi(\boldsymbol{\theta})}{\varphi(\boldsymbol{\theta})}$ définit les *poids d'importance*, et φ est telle que $\varphi(\boldsymbol{\theta}) > 0$ pour tout $\boldsymbol{\theta}$ où $\pi(\boldsymbol{\theta}) > 0$, autrement dit le support de φ inclut celui de π. La forme de l'intégrale obtenue permet alors de dire que :

$$\mathbb{E}_\pi[f(\boldsymbol{\theta})] = \mathbb{E}_\varphi[f(\boldsymbol{\theta})w(\boldsymbol{\theta})] \tag{2.28}$$

D'après la loi forte des grands nombres, en notant $w^{(i)}(\boldsymbol{\theta}) = \dfrac{\pi^{(i)}(\boldsymbol{\theta})}{\varphi^{(i)}(\boldsymbol{\theta})}$, la quantité $\sum_{i=1}^{N} f(\boldsymbol{\theta}^{(i)})w^{(i)}$

converge presque sûrement vers $\mathbb{E}_\varphi[f(\boldsymbol{\theta})w(\boldsymbol{\theta})]$ pour $N \to +\infty$. En associant cela avec l'équation (2.28), on obtient ainsi :

$$\mathbb{E}_\pi[f(\boldsymbol{\theta})] \simeq I_N = \frac{1}{N}\sum_{i=1}^{N} f(\boldsymbol{\theta}^{(i)})w^{(i)} \tag{2.29}$$

Dans notre contexte, l'équation (2.7) nous rappelle que la loi cible n'est connue qu'à une constante près : pour en tenir compte, il est nécessaire de normaliser les poids d'importance. En écrivant :

$$\forall i \in \{1, \ldots, N\},\ \tilde{w}^{(i)} = \frac{w^{(i)}}{\sum_{i=1}^{N} w^{(i)}} \tag{2.30}$$

on garantit bien que toutes les valeurs de $\tilde{\boldsymbol{w}}$ sont comprises entre 0 et 1.

En reprenant l'équation (2.29), on peut alors obtenir une approximation de la loi cible, qui s'écrit :

$$\pi(\boldsymbol{\theta}) \simeq \sum_{i=1}^{N} \tilde{w}(\boldsymbol{\theta}^{(i)})\delta_{\boldsymbol{\theta}^{(i)}}(\boldsymbol{\theta}) \tag{2.31}$$

L'équation (2.31) permet ainsi d'obtenir des particules pondérées de la loi cible. Pour cela, il faut choisir une loi de proposition φ qui soit suffisamment proche de la loi cible, dont le support englobe celui de cette dernière, et à partir de laquelle on puisse échantillonner facilement.

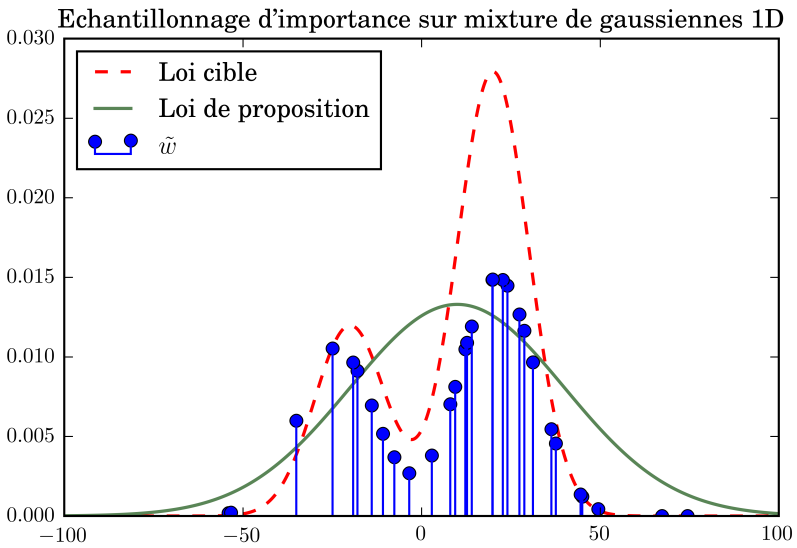


FIGURE 2.5 – Calcul des poids d'importance normalisés (en bleu) sur les 30 premiers éléments d'un vecteur de 1000 particules , avec la loi cible de la figure 2.1 (pointillés rouges) et une loi de proposition $\mathcal{N}(\mu = 10, \sigma = 30)$ (en vert).

En reprenant l'exemple de la mixture de deux gaussiennes dont la loi cible est présentée en équation (2.22), on observe sur la figure 2.5 que si la loi de proposition est bien choisie, alors on obtient une bonne approximation de la loi cible. Le choix de cette loi est en fait un point crucial des méthodes IS, et présente certaines difficultés s'il est impossible d'associer "intuitivement" une distribution φ à une loi cible π, par exemple lorsqu'on se place dans des grandes dimensions, ou plus généralement lorsque π est trop complexe pour être rattachée à une famille de lois connue.

Une loi de proposition mal choisie peut engendrer un ralentissement de la convergence de l'algorithme : par exemple, sur la figure 2.6, la nouvelle loi de proposition ne permet pas d'échantillonner correctement depuis le premier mode à -20.

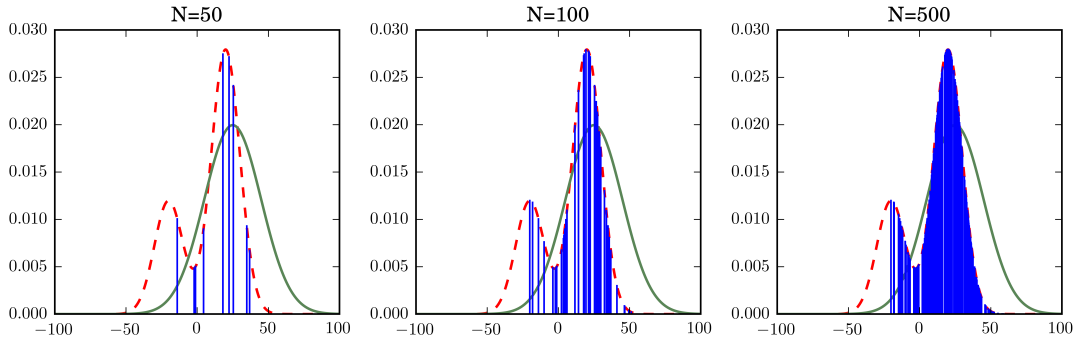


FIGURE 2.6 – IS pour une loi de proposition $\mathcal{N}(\mu = 25, \sigma = 20)$ à différentes étapes de l'échantillonnage.

On remarque également qu'un petit nombre de particules agrège les valeurs les plus importantes des poids, alors que la majorité restante détient des poids de faibles valeurs. Ce déséquilibre de répartition est plus généralement connu sous le nom de *dégénérescence des poids*. La métrique permettant de quantifier ce phénomène et de facto d'évaluer les performances de la loi de proposition, est appelée *Effective Sample Size* ($\mathbb{ESS}$), et se calcule selon la formule suivante :

$$\mathbb{ESS} = \frac{\left(\sum_{i=1}^{N} w^{(i)}\right)^2}{\sum_{i=1}^{N} \left(w^{(i)}\right)^2} \tag{2.32}$$

Dans le cas le plus extrême où une unique particule à un poids égal à 1 et toutes les autres ont un poids nul, on a $\mathbb{ESS} = 1$. Plus sa valeur croît, plus la loi de proposition est proche de la loi cible, le cas parfait étant $\mathbb{ESS} = N$ où φ et π sont identiques.

Une façon de contourner le problème de la dégénérescence consiste à ré-échantillonner sur les points où les poids d'importance sont élevés : cette méthode fût développée par [Rubin et al., 1988] sous le nom d'algorithme *Sampling Importance Resampling* (SIR). Une autre alternative plus intéressante revient à moduler la loi de proposition en fonction des valeurs de poids obtenues, ce qui est possible grâce aux méthodes d'*échantillonnage d'importance adaptatives* sur lesquelles nous allons à présent nous pencher.

2.2.4 Méthodes d'échantillonnage d'importance adaptatif

Le facteur-clé de la performance d'un algorithme de type IS réside dans l'utilisation d'une bonne loi de proposition : il paraît donc naturel d'envisager des méthodes permettant l'ajustement de cette loi pour qu'elle permette une bonne approximation de la loi cible.

Population Monte Carlo

Les travaux présentés dans [Cappé et al., 2004] illustrent une première approche, appelée *Population Monte Carlo* (PMC), qui peut être vue comme une version itérative de l'IS visant à adapter la loi de proposition afin de la rapprocher au mieux de la loi cible. Plus précisément, à chaque itération k :

1. on tire un échantillon $\boldsymbol{\theta}_k = (\theta_{1,k}, \ldots, \theta_{N_p,k})$ de particules obtenues par une perturbation stochastique qui revient à appliquer un noyau de transition de type "marche aléatoire" sur chacun des $\theta_{1,k-1}, \ldots, \theta_{N_p,k-1}$, à la façon d'une chaîne de Markov,
2. on calcule le vecteur des poids d'importance normalisés $\tilde{\boldsymbol{w}}_k$ de ces particules,
3. on génère un nouvel échantillon $(\tilde{\theta}_{1,k}, \ldots, \tilde{\theta}_{N_p,k})$ obtenu par *resampling* de $(\theta_{1,k}, \ldots, \theta_{N_p,k})$ en utilisant les poids $\tilde{\boldsymbol{w}}_k$.

Le point 3. donne le résultat de l'algorithme à l'itération k, à savoir un vecteur de particules ajustées pour être proche de la loi cible. A partir de là, on peut ré-itérer en appliquant de nouveau une perturbation sur ce vecteur, et ainsi de suite.

D-kernel PMC

[Douc et al., 2007] formalise le concept d'adaptation de la loi de proposition en la définissant comme étant un mélange de D noyaux fixes $\varphi_1, \ldots, \varphi_D$ défini par :

$$\varphi_\alpha(\boldsymbol{\theta}, \boldsymbol{\theta}^*) = \sum_{d=1}^{D} \alpha_d \varphi_d(\boldsymbol{\theta}, \boldsymbol{\theta}^*) \tag{2.33}$$

où les α_d définissent les *facteurs d'influence* des composantes respectives φ_d, et sont tels que

$$\sum_{d=1}^{D} \alpha_d = 1$$

La démarche proposée, portant le nom de *D-kernel PMC*, consiste à ajuster les valeurs des α_d à chaque itération d'un algorithme PMC, sur la base d'un critère de performance visant à minimiser l'écart statistique entre la distribution de proposition et la loi cible. Un tel écart peut être quantifié par la *divergence de Kullback-Leibler* (KL) $D(\pi||\varphi_\alpha)$, définie par :

$$D(\pi||q_\alpha) = \int \log\left(\frac{\pi(\boldsymbol{\theta})}{q_\alpha(\boldsymbol{\theta})}\right) \pi(\boldsymbol{\theta}) d\boldsymbol{\theta} \tag{2.34}$$

M-PMC

[Cappé et al., 2008] propose une extension de ces travaux en introduisant une méthode appelée *Mixture-PMC* (M-PMC) reposant sur le même type de critère de performance, mais visant cette

fois à étendre la phase d'adaptation à tous les paramètres de la mixture q_α, sans se limiter aux seuls α_d . La loi de proposition peut alors s'écrire sous la forme suivante :

$$\varphi_{(\alpha,\nu)}(\boldsymbol{\theta}) = \sum_{d=1}^{D} \alpha_d \varphi_d(\boldsymbol{\theta}|\nu_d) \tag{2.35}$$

où ν_d représente les paramètres de la d-ième composante de la loi de proposition.

La procédure d'adaptation qui en découle s'inspire de l'algorithme dit d'*Expectation-Maximization* (EM) pour construire les équations de mise à jour des éléments constituant ν_d : en effet, le fait de minimiser l'équation (2.34) équivaut ici à maximiser la grandeur :

$$\int \log\left(\sum_{d=1}^{D} \alpha_d \varphi_d(\boldsymbol{\theta}|\nu_d)\right) \pi(\boldsymbol{\theta}) d\boldsymbol{\theta} \tag{2.36}$$

On se ramène ainsi à un problème similaire à celui du calcul d'un estimateur du maximum de vraisemblance pour un modèle de mixture, d'où l'emploi d'une méthode de type EM. $\pi(\boldsymbol{\theta})$ est alors approchée par l'exécution itérative des formules d'approximation empirique présentées dans [Cappé et al., 2008]. Ce type de méthode est particulièrement approprié si on choisit la loi de proposition comme étant une mixture de gaussiennes multivariées, i.e. dont la densité en dimension p prend la forme suivante :

$$\forall d = 1, \ldots, D,\ \varphi_d(\boldsymbol{\theta}|\nu_d) = \frac{1}{\left((2\pi)^p \det \Sigma_d\right)^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{\theta} - \mu_d)^T \Sigma_d^{-1} (\boldsymbol{\theta} - \mu_d)\right) \tag{2.37}$$

où $\nu_d = (\mu_d, \Sigma_d)$ représente les paramètres classiques (moyenne et matrice de covariance) de la d-ième composante. En effet, dans ce cas-là, les équations de mise à jour de ν_d peuvent être obtenues de façon analytique. Pour cela, en se plaçant à l'itération k sur la d-ième composante et pour la i-ème particule, on définit tout d'abord la grandeur intermédiaire suivante :

$$\rho_d^k(\boldsymbol{\theta}_k^{(i)}|\alpha_d^k, \nu_d^k) = \frac{\alpha_d^k \varphi_d(\boldsymbol{\theta}_k^{(i)}|\nu_d^k)}{\sum_{m=1}^{D} \alpha_m^k \varphi_d(\boldsymbol{\theta}_k^{(i)}|\nu_m^k)}, \quad i \in \{1, \ldots, N\} \tag{2.38}$$

La mise à jour de α_d^k et des éléments μ_d^k et Σ_d^k de ν_d^k est alors donnée par :

$$\begin{aligned}
\alpha_d^{k+1} &= \sum_{i=1}^{N} \tilde{w}_k^{(i)} \rho_d^k(\boldsymbol{\theta}_k^{(i)}|\alpha_d^k, \nu_d^k) \\
\mu_d^{k+1} &= \frac{\sum_{i=1}^{N} \tilde{w}_k^{(i)} \rho_d^k(\boldsymbol{\theta}_k^{(i)}|\alpha_d^k, \nu_d^k) \boldsymbol{\theta}_k^{(i)}}{\alpha_d^{k+1}} \\
\Sigma_d^{k+1} &= \frac{\sum_{i=1}^{N} \tilde{w}_k^{(i)} \rho_d^k(\boldsymbol{\theta}_k^{(i)}|\alpha_d^k, \nu_d^k)(\boldsymbol{\theta}_k^{(i)} - \mu_d^{k+1})(\boldsymbol{\theta}_k^{(i)} - \mu_d^{k+1})^T}{\alpha_d^{k+1}}
\end{aligned} \tag{2.39}$$

Une démonstration de ces résultats est détaillée dans [Douc et al., 2007].

A chaque itération $k \in 1, \ldots, K$, l'approche M-PMC permet ainsi, à partir des particules échantillonnées depuis la loi de proposition courante et de leurs poids d'importance respectifs, d'obtenir la nouvelle proposition $\varphi_{\alpha^{k+1},\nu^{k+1}}$ qui sera réutilisée à l'itération suivante. Contrairement à la version originale du PMC, ici le tirage des nouvelles particules ne dépend plus directement des anciennes, qui ne sont pas explicitement considérées comme des paramètres de la loi de tirage.

Adaptive Multiple Importance Sampling (AMIS)

Il faut noter que malgré son aspect adaptatif, la méthode M-PMC n'exploite que les particules de l'itération courante pour l'ajustement des paramètres de la loi de proposition, ainsi que pour l'approximation empirique de la loi a posteriori. Une amélioration supplémentaire consiste à reprendre les idées développées dans [Veach and Guibas, 1995] et [Owen and Zhou, 2000], avec l'utilisation d'une méthode appelée *deterministic multiple mixture*, qui permet de réutiliser toutes les particules générées durant la simulation. En pratique,cela consiste à modifier la formule de calcul du poids d'importance $w_k^{(i)}$ associé à la particule $\theta_k^{(i)}$ à l'itération k. Le but est de prendre en compte la suite $(\varphi_0, \ldots, \varphi_k)$ des distributions de proposition, et non plus uniquement la loi courante φ_k. Pour cela, on définit la grandeur suivante :

$$\vartheta_k^{(i)} = N_p \varphi_{\alpha^0,\nu^0}(\theta_k^{(i)}) + \sum_{l=1}^{k} N_p \varphi_{\alpha^l,\nu^l}(\theta_k^{(i)}) \tag{2.40}$$

On peut alors écrire la nouvelle formule permettant de calculer le poids d'importance $w_k^{(i)}$:

$$w_k^{(i)} = \frac{\pi(\theta_k^{(i)})}{\left(\frac{1}{(k+1)N_p}\vartheta_k^{(i)}\right)} = (k+1)N_p \frac{\pi(\theta_k^{(i)})}{\vartheta_k^{(i)}} \tag{2.41}$$

Ce cheminement peut être appliqué à la fois pour le calcul du poids courant, mais également pour ajuster les poids calculés lors des itérations précédentes : on parle alors de *recyclage des poids*. Un tel processus de stabilisation permet ainsi d'atténuer automatiquement les valeurs des poids d'importance sujets à dégénérescence en les pénalisant par rapport aux poids obtenus aux itérations suivantes.

Le fait de coupler au sein d'un même algorithme :

— le recyclage optimal des poids d'importance générés entre le début de la simulation et l'itération courante,

— la mise à jour des paramètres de la loi de proposition en fonction des poids d'importance,

constitue ainsi le coeur des travaux présentés dans [Cornuet et al., 2012], qui ont mené à la création de l'algorithme d'*Adaptive Multiple Importance Sampling* (AMIS).

La caractéristique principale de l'AMIS est la combinaison d'une loi de proposition adaptive et d'une "mise en mémoire" progressive des populations de particules générées au fil des itérations.

Cela permet à l'algorithme de ne pas se limiter aux particules de l'itération courante, mais de travailler sur l'ensemble des populations de particules générées depuis la première itération. L'AMIS ré-exploite ainsi toute l'information disponible sur les poids d'importance et les particules de façon optimale. Son fonctionnement est détaillé dans l'algorithme 3 : les valeurs à fournir en entrée sont les paramètres α^0, ν^0 de la loi de proposition initiale φ_{α^0,ν^0}, et les résultats obtenus en sortie sont :

— la collection de toutes les particules $\theta_0^{(1)},\ldots,\theta_0^{(N_p)},\theta_1^{(1)},\ldots,\theta_K^{(1)},\ldots,\theta_K^{(N_p)}$,

— les poids recyclés associés $\tilde{w}_0^{(1)},\ldots,\tilde{w}_0^{(N_p)},\tilde{w}_1^{(1)},\ldots,\tilde{w}_K^{(1)},\ldots,\tilde{w}_K^{(N_p)}$.

L'aspect adaptatif de l'AMIS est ainsi renforcé par le recyclage de tous les poids générés depuis le début de la simulation. Cela lui permet, en pratique, de mieux exploiter l'information disponible, ce qui est illustré par la figure 2.7 , où l'exemple du modèle 2.22 est appliqué aux algorithmes basés sur l'IS.

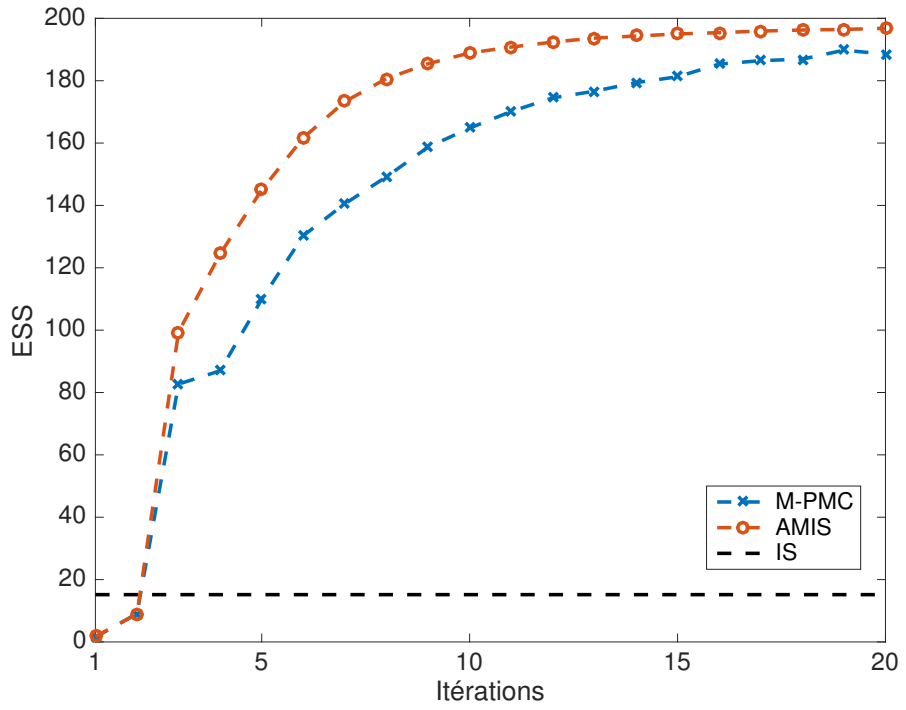


FIGURE 2.7 – Mixture de gaussiennes : évolution moyenne (sur 100 réplicats) de l'ESS sur l'itération courante pour différents algorithmes basés sur l'IS

Les différentes méthodes sont comparées à nombre total équivalent de particules, c'est-à-dire 4000 pour l'IS et 200 par itération (pour un total de 20 itérations) pour le M-PMC et l'AMIS. On peut dans un premier temps observer que les deux algorithmes adaptatifs surclassent rapidement l'IS classique, ce dernier étant pénalisé par la nature statique de sa loi de proposition. On note également que l'ESS pour l'AMIS augmente plus vite, ce qui distingue la caractéristique de recyclage des poids dans l'AMIS qu'on ne retrouve pas dans le M-PMC, celui-ci n'exploitant systématiquement que les particules et les poids de la dernière itération pour ajuster les paramètres de l'itération courante. L'effet de seuil à 200 (nombre de particules générées par itération) illustre

Algorithm 3. Adaptive Multiple Importance Sampling (AMIS)

Générer N_p particules $\theta_0^{(1)}, \ldots, \theta_0^{(N_p)}$ depuis φ_{α^0,ν^0}
for $i = 1, \ldots, N_p$ **do**
Calculer les poids d'importance initiaux :
$\vartheta_0^{(i)} = N_p \varphi_{\alpha^0,\nu^0}(\theta_0^{(i)})$
$w_0^{(i)} = \dfrac{\pi(\theta_0^{(i)})}{\varphi_{\alpha^0,\nu^0}(\theta_0^{(i)})}$
end for
Normaliser les poids initiaux :
for $i = 1, \ldots, N_p$ **do**
$\tilde{w}_0^{(i)} = \dfrac{w_0^{(i)}}{\sum_{j=1}^{N_p} w_0^{(j)}}$
end for
Mettre à jour la loi de proposition en calculant α^1 et ν^1 à partir de α^0, ν^0, $\tilde{w}_0$ et θ_0
for $k = 1, \ldots, K$ **do**
Générer N_p particules $\theta_k^{(1)}, \ldots, \theta_k^{(N_p)}$ depuis la loi de proposition courante φ_{α^k,ν^k} :
for $i = 1, \ldots, N_p$ **do**
Calculer les poids d'importance de l'échantillon courant :
$\vartheta_k^{(i)} = N_p \varphi_{\alpha^0,\nu^0}(\theta_k^{(i)}) + \sum_{l=1}^{k} N_p \varphi_{\alpha^l,\nu^l}(\theta_k^{(i)})$
$w_k^{(i)} = (k+1) N_p \dfrac{\pi(\theta_k^{(i)})}{\vartheta_k^{(i)}} = (k+1) N_p \dfrac{p(\boldsymbol{\eta}|\theta_k^{(i)}) p(\theta_k^{(i)})}{\vartheta_k^{(i)}}$
end for
Recycler les poids des particules obtenues par les précédentes lois de proposition :
for $l = 0, \ldots, k-1$ **do**
for $i = 1, \ldots, N_p$ **do**
$\vartheta_l^{(i)} = \vartheta_l^{(i)} + N_p \varphi_{\alpha^k,\nu^k}(\theta_l^{(i)})$
$w_l^{(i)} = (k+1) N_p \dfrac{\pi(\theta_l^{(i)})}{\vartheta_l^{(i)}}$
end for
end for
Normaliser les poids recyclés et courants :
for $l = 0, \ldots, k$ **do**
for $i = 1, \ldots, N_p$ **do**
$\tilde{w}_l^{(i)} = \dfrac{w_l^{(i)}}{\sum_{l'=1}^{k} \sum_{j=1}^{N_p} w_{l'}^{(j)}}$
end for
end for
Mettre à jour la loi de proposition en calculant α^{k+1} et ν^{k+1} à partir de α^k, ν^k , de la concaténation des poids recyclés normalisés $\tilde{w}_0, \ldots, \tilde{w}_k$ et des particules correspondantes $\theta_0, \cdots, \theta_K$
end for

le fait que les deux algorithmes atteignent une bonne qualité d'estimation de la loi cible.

L'AMIS permet également de fournir un estimateur dont la variance est moins élevée que celle du M-PMC, comme présenté en figure 2.8, ce qui lui confère une meilleure robustesse statistique.

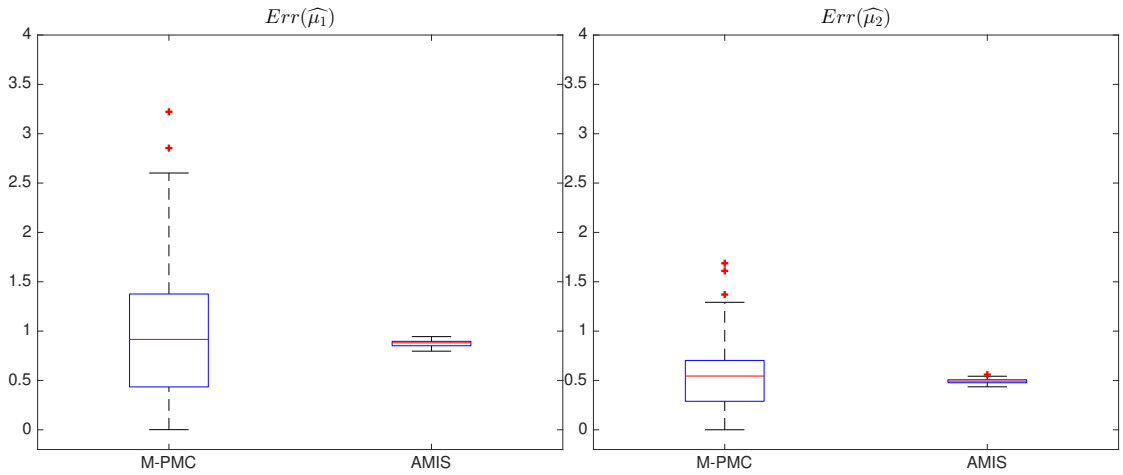


FIGURE 2.8 – Mixtures de gaussiennes : erreurs absolues (sur 100 réplicats) des estimateurs MMSE pour les algorithmes IS adaptatifs

Les avantages que présentent l'AMIS en font ainsi un choix privilégié pour la suite des travaux présentés dans ce manuscrit : les deux prochains chapitres introduisent son application à la problématique d'estimation du terme source dans différents cas pratiques.

Chapitre 3

Application de la méthodologie AMIS au cas expérimental FFT07

Dans ce chapitre, nous appliquons l'algorithme AMIS à un problème STE avec des données d'observations réelles issues d'une campagne expérimentale de mesures. Nous présentons dans un premier temps le contexte de la campagne, puis nous détaillons l'approche méthodologique ainsi que les résultats obtenus. Le contenu de ce chapitre reprend les travaux illustrés dans [Rajaona et al., 2015].

3.1 Contexte : l'expérience FFT07

La campagne expérimentale *FUSION Fields Trials 2007* (FFT07) fut conçue et menée en 2007 par la *Defense Threat Reduction Agency* (DTRA), une agence du Département de la Défense (DoD) des Etats-Unis dont la mission principale est axée autour de la prévention et de la protection vis-à-vis des risques NRBC.

Le but de FFT07 était de constituer une base de données météorologiques et de mesure de concentration issues d'une série de tests, ou *trials*, chacun d'entre eux consistant en un rejet de gaz traceur sur une zone fortement instrumentée du site militaire de *Dugway Proving Ground*, dans le désert de l'Utah.

Certains éléments de cette base de données ont été transmis à plusieurs équipes de recherche, chacune ayant pour tâche d'estimer au mieux les paramètres du terme source de chacun des *trials* fournis ([Platt and Deriggi, 2010] fournit un récapitulatif des différentes équipes et méthodes employées).

L'expérience FFT07 a été formatée pour étudier l'impact à courte portée des rejets de gaz traceur : le domaine considéré est un carré de 500 mètres de côté. Plusieurs configurations de rejet ont été utilisées, chacune caractérisant un *trial* distinct par :

— la période de la journée où le rejet a eu lieu,

— la vitesse du vent et sa direction,

— la classe de stabilité atmosphérique,

— le nombre de sources ayant simultanément émis un rejet (1, 2 ou 3),

— dans le cas d'une source unique, le type du rejet : continu ou instantané.

L'acquisition des données s'est faite via un réseau de 100 capteurs à photo-ionisation (*digital photoionization detectors*, ou digiPID), répartis en un maillage uniforme régulier sur l'ensemble du domaine. Ces capteurs sont situés à 50m les uns des autres, et à une hauteur de 2m du sol. La fréquence d'acquisition des concentrations est relativement élevée (50Hz) : dans notre étude, nous avons réduit la dimension du vecteur d'observation en effectuant un moyennage sur des fenêtres de 10s afin que les calculs puissent se faire dans des temps raisonnables sans pour autant déplorer une perte significative d'information (voir figure 3.1).

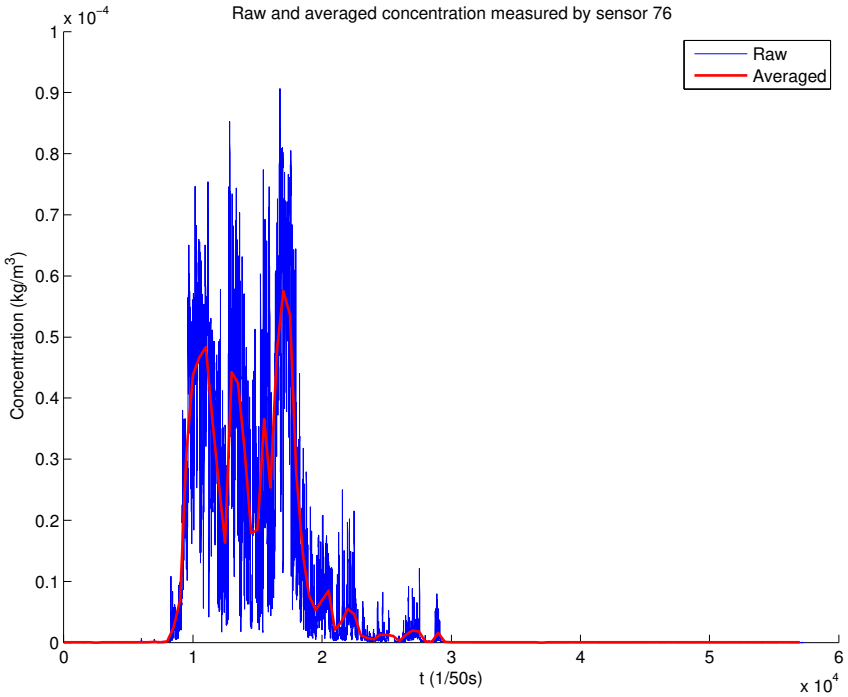


FIGURE 3.1 – Concentrations brutes (en bleu) et concentrations moyennées (en rouge) sur une fenêtre glissante de 10s.

L'expérience FFT07 étant un outil de validation, le domaine d'étude comprend beaucoup plus de capteurs que dans un cas standard, où le domaine est moins instrumenté. Afin d'envisager un cas réaliste tout en conservant une quantité suffisante de données d'observation à traiter, nous avons choisi de limiter le nombre de capteurs exploités à 25 (voir figure 3.2). Dans notre cas, nous travaillons avec les configurations issues des *trials* 7 et 30 dont les données nous sont disponibles :

— Dans un premier temps, nous simulons à l'aide d'un modèle de dispersion les concentrations mesurées aux capteurs pour chaque *trial*, afin d'exclure toute source d'erreur pouvant être causée par les différents facteurs instrumentaux.

— Nous confrontons ensuite l'algorithme de reconstruction du terme source aux mesures réelles des capteurs, afin de vérifier si l'estimation reste suffisamment précise par rapport au cas purement synthétique.

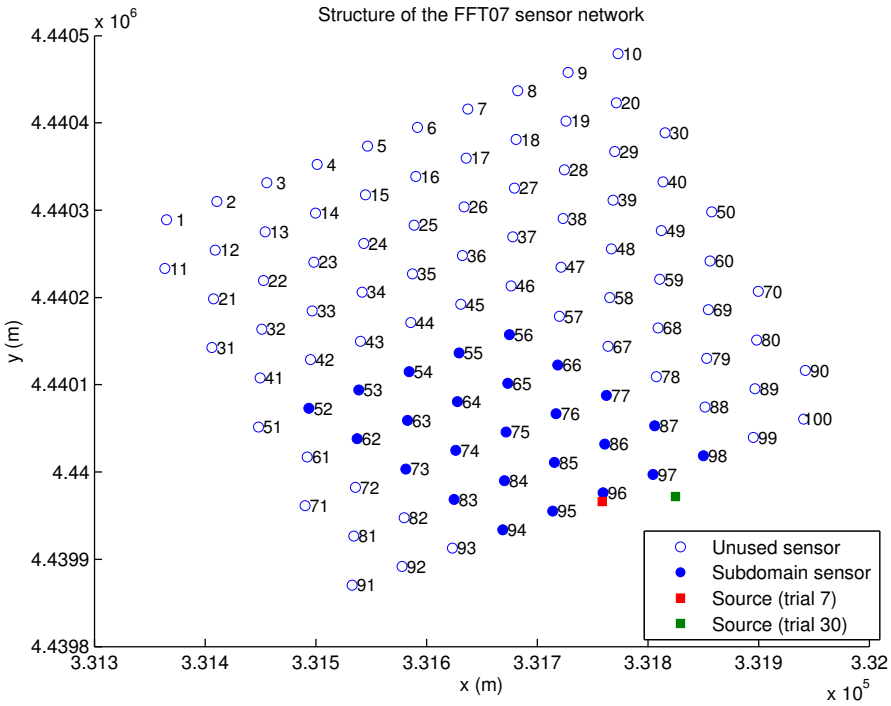


FIGURE 3.2 – Choix du sous-réseau de 25 capteurs utilisé dans notre étude.

3.2 Formulation bayésienne du problème STE

3.2.1 Modèle de données et vraisemblance

Nous considérons ici une source localisée en un point $\boldsymbol{x}_s = (x_s, y_s, z_s)$ de l'espace et caractérisée par un profil de rejet $\boldsymbol{q}$. Pour les besoins de la modélisation, ce dernier est discrétisé en T_s échéances d'émission $t'_1, \cdots, t'_{T_s}$, l'intervalle entre deux échéances consécutives demeurant strictement identique. On peut ainsi définir $\boldsymbol{q}$ comme une succession de paliers d'émissions constantes, le débit de la source ne variant pas entre deux instants d'émission consécutifs t'_n et t'_{n+1}.

On suppose que les observations sont définies par des mesures de concentration en un nombre fini de points ${\boldsymbol{x}_c}^{(1)}, \cdots, {\boldsymbol{x}_c}^{(N_c)}$ du domaine, qui constituent les positions d'un réseau de N_c capteurs. On considère que les capteurs et la source se situent à la même hauteur, ce qui permet de n'étudier la position de la source que sur deux dimensions, autrement dit écrire $\boldsymbol{x}_s = (x_s, y_s)$. Les mesures fournies par ces capteurs sont données suivant une discrétisation temporelle donnée, chaque capteur délivrant ainsi une valeur de concentration à chacun des instants d'observation $t_1, \cdots, t_{T_c}$.

La concentration $\eta_{i,j}$ fournie par le i-ème capteur à la position ${\boldsymbol{x}_c}^{(i)}$ et à l'échéance d'observation t_j est alors modélisée par l'équation suivante :

$$\eta_{i,j} = \sum_{n=1}^{T_s} q(t'_n) \boldsymbol{C}_{i,j}(\boldsymbol{x}_s, t'_n) + \varepsilon_{i,j} \tag{3.1}$$

Le premier terme de l'équation (3.1) représente à la concentration moyenne obtenue par la superposition des T_s rejets aux différents temps d'émission $\{t'_n\}_{1\leq n\leq T_s}$et pondérés par les quantités émises $\{q(t'_n)\}_{1\leq n\leq T_s}$ associées. Ainsi, $\boldsymbol{C}_{i,j}(\boldsymbol{x}_s, t'_n)$ désigne la concentration moyenne observée par le i-ème capteur à la position ${\boldsymbol{x}_c}^{(i)}$ et à l'échéance d'observation t_j pour un rejet unitaire émis à l'instant t'_n par une source située à la position $\boldsymbol{x}_s$. Enfin, $\varepsilon_{i,j}$ représente le terme regroupant toutes les sources d'erreur présentées au paragraphe §2.1.1.

Il est possible de réécrire l'équation (3.1) sous la forme matricielle suivante :

$$\boldsymbol{\eta} = \boldsymbol{C}(\boldsymbol{x}_s)\boldsymbol{q} + \boldsymbol{\varepsilon} \tag{3.2}$$

qui n'est autre que l'équivalent de l'équation (1.7). $\boldsymbol{\eta} \in \mathbb{R}^{N_c T_c}$ est un vecteur où toutes les observations de concentration sont concaténées sous la forme suivante :

$$\boldsymbol{\eta} = (\eta_{1,1}, \eta_{1,2}, \cdots, \eta_{1,T_c}, \eta_{2,1}, \cdots, \cdots, \eta_{N_c,T_c})^T \tag{3.3}$$

$\boldsymbol{\varepsilon} \in \mathbb{R}^{N_c T_c}$ est un vecteur d'erreur qui suit, comme présenté à l'équation (1.12), une loi normale centrée de matrice de covariance $\boldsymbol{R} \in \mathbb{R}^{N_c T_c \times N_c T_c}$. De plus, on considère que ce vecteur de bruit affecte les observations de façon indépendante et identiquement distribuée (i.i.d.) : par conséquent, la matrice $\boldsymbol{R}$ est diagonale :

$$\boldsymbol{R} = \sigma^2_{obs} \boldsymbol{I}_{N_c T_c} \tag{3.4}$$

Le terme $\boldsymbol{C}(\boldsymbol{x}_s) \in \mathbb{R}^{N_c T_c \times T_s}$ est la matrice source-récepteur suivante :

$$\boldsymbol{C}(\boldsymbol{x}_s) = \begin{pmatrix} \boldsymbol{C}_{1,1}(\boldsymbol{x}_s, t'_1) & \boldsymbol{C}_{1,1}(\boldsymbol{x}_s, t'_2) & \cdots & \boldsymbol{C}_{1,}(\boldsymbol{x}_s, t'_{T_s}) \\ \boldsymbol{C}_{1,2}(\boldsymbol{x}_s, t'_1) & \boldsymbol{C}_{1,2}(\boldsymbol{x}_s, t'_2) & \cdots & \boldsymbol{C}_{1,2}(\boldsymbol{x}_s, t'_{T_s}) \\ \vdots & \vdots & & \vdots \\ \boldsymbol{C}_{1,T_c}(\boldsymbol{x}_s, t'_1) & \boldsymbol{C}_{1,T_c}(\boldsymbol{x}_s, t'_2) & \cdots & \boldsymbol{C}_{1,T_c}(\boldsymbol{x}_s, t'_{T_s}) \\ \boldsymbol{C}_{2,1}(\boldsymbol{x}_s, t'_1) & \boldsymbol{C}_{2,1}(\boldsymbol{x}_s, t'_2) & \cdots & \boldsymbol{C}_{2,1}(\boldsymbol{x}_s, t'_{T_s}) \\ \vdots & \vdots & & \vdots \\ \vdots & \vdots & & \vdots \\ \boldsymbol{C}_{N_c,T_c}(\boldsymbol{x}_s, t'_1) & \boldsymbol{C}_{N_c,T_c}(\boldsymbol{x}_s, t'_2) & \cdots & \boldsymbol{C}_{N_c,T_c}(\boldsymbol{x}_s, t'_{T_s}) \end{pmatrix} \tag{3.5}$$

Si on note $\boldsymbol{\Theta} = (\boldsymbol{x}_s, \boldsymbol{q})$ le vecteur des paramètres caractérisant le terme source, la règle de Bayes permet de définir la loi a posteriori de $\boldsymbol{\Theta}$:

$$p(\boldsymbol{\Theta}|\boldsymbol{\eta}) = \frac{p(\boldsymbol{\eta}|\boldsymbol{\Theta})p(\boldsymbol{\Theta})}{p(\boldsymbol{\eta})} \tag{3.6}$$

Comme expliqué au paragraphe 2.1.2, on cherche à estimer cette loi a posteriori à un facteur multiplicatif près, l'équation (3.6) devient alors :

$$p(\boldsymbol{\Theta}|\boldsymbol{\eta}) \propto p(\boldsymbol{\eta}|\boldsymbol{\Theta})p(\boldsymbol{\Theta}) \tag{3.7}$$

Comme le vecteur d'erreur est supposé gaussien centré, on peut alors définir la vraisemblance des observations $\boldsymbol{\eta}$ sachant un terme source donné $\boldsymbol{\Theta}$:

$$p(\boldsymbol{\eta}|\boldsymbol{\Theta}) = \prod_{i=1}^{N_c} \prod_{j=1}^{T_c} \mathcal{N}\left(\eta_{i,j} \middle| \boldsymbol{C}_{i,j}(\boldsymbol{x}_s)\boldsymbol{q}, \sigma_{obs}^2\right) \tag{3.8}$$

3.2.2 Choix des lois a priori

Position de la source

On considère que la source est forcément contenue dans les limites du domaine spatial $\mathcal{D}$ considéré, mais qu'elle peut se situer en n'importe quel point de ce domaine. En termes probabilistes, cela se traduit par une loi a priori uniforme sur la position $\boldsymbol{x}_s$ de la source :

$$p(\boldsymbol{x}_s) = \mathcal{U}_{\mathcal{D}}(\boldsymbol{x}_s) \tag{3.9}$$

Profil d'émission

Comme expliqué dans [Winiarek et al., 2011], la pratique la plus courante consiste à choisir un a priori gaussien pour le vecteur $\boldsymbol{q}$, qui s'écrit alors :

$$p(\boldsymbol{q}) = \mathcal{N}\left(\boldsymbol{q}|\boldsymbol{\mu}_q, \boldsymbol{\Sigma}_q\right) \tag{3.10}$$

Dans [Bocquet, 2008], il est expliqué que l'hypothèse gaussienne sur $\boldsymbol{q}$ entraîne potentiellement des incohérences physiques telles que des valeurs d'émissions négatives. Cependant, une telle hypothèse demeure fréquemment utilisée dans la littérature, et conduit à des résultats satisfaisants (voir par exemple [Issartel and Baverel, 2003]) ainsi qu'une meilleure flexibilité quant à la quantification des connaissances a priori sur le type de rejet étudié. Par exemple, si on sait d'avance que le rejet se fait à un débit relativement faible, alors il est possible d'ajuster les valeurs de la diagonale de la matrice de covariance en y mettant des quantités faibles.

Il est toutefois possible d'atténuer les effets indésirables de l'hypothèse gaussienne sans avoir à changer la nature de la loi de probabilité a priori de $\boldsymbol{q}$, nous verrons comment cela est possible dans le prochain paragraphe.

3.3 Démarche de résolution du problème bayésien

Nous développons ici en détail le processus de résolution du problème tel qu'il a été défini dans le paragraphe précédent, à savoir *le calcul de la loi a posteriori* $p(\boldsymbol{\Theta}|\boldsymbol{\eta})$ *des paramètres de la source*. En pratique, il est impossible d'obtenir cette dernière de façon analytique, à cause de la complexité et de l'aspect fortement non-linéaire de la vraisemblance $p(\boldsymbol{\eta}|\boldsymbol{\Theta})$ présentée à l'équation (3.8). Afin de contourner ce problème, une solution consiste à décomposer la loi d'intérêt en plusieurs éléments distincts, suivant une opération de *marginalisation*.

La loi $p(\boldsymbol{\Theta}|\boldsymbol{\eta})$ est dite *jointe*, car elle regroupe tous les paramètres de la source autour d'une unique loi a posteriori. Par définition des probabilités conditionnelles, cette loi peut s'écrire de la façon suivante :

$$p(\boldsymbol{\Theta}|\boldsymbol{\eta}) = p(\boldsymbol{x}_s, \boldsymbol{q}|\boldsymbol{\eta}) = \frac{p(\boldsymbol{x}_s, \boldsymbol{q}, \boldsymbol{\eta})}{p(\boldsymbol{\eta})} \tag{3.11}$$

En appliquant la règle du conditionnement en chaîne, ou *chain rule*[1], sur le numérateur, on arrive à l'expression suivante :

$$p(\boldsymbol{x}_s, \boldsymbol{q}|\boldsymbol{\eta}) = p(\boldsymbol{q}|\boldsymbol{x}_s, \boldsymbol{\eta})p(\boldsymbol{x}_s|\boldsymbol{\eta}) \tag{3.12}$$

où :

— $p(\boldsymbol{q}|\boldsymbol{x}_s, \boldsymbol{\eta})$ est la loi a posteriori conditionnelle de $\boldsymbol{q}$,

— $p(\boldsymbol{x}_s|\boldsymbol{\eta})$ est la loi a posteriori marginale de $\boldsymbol{x}_s$.

3.3.1 Loi conditionnelle du profil d'émission

A priori gaussien et solution analytique

$$p(\boldsymbol{\eta}|\boldsymbol{x}_s) = \mathcal{N}(\boldsymbol{\eta}|\boldsymbol{C}(\boldsymbol{x}_s)\boldsymbol{\mu}_q, \boldsymbol{C}(\boldsymbol{x}_s)\boldsymbol{\Sigma}_q\boldsymbol{C}(\boldsymbol{x}_s)^T + \boldsymbol{R}) \tag{3.13}$$

La loi a priori de $\boldsymbol{q}$ ainsi que la vraisemblance de l'équation (3.13) suivant toutes deux des lois normales, alors la loi a posteriori conditionnelle de $\boldsymbol{q}$ suit également une loi normale de moyenne $\widetilde{\boldsymbol{\mu}_q}$ et de matrice de covariance $\widetilde{\boldsymbol{\Sigma}_q}$, dont les valeurs sont obtenues par les formules empiriques :

$$\begin{aligned} \widetilde{\boldsymbol{\mu}_q}(\boldsymbol{x_s}) &= \boldsymbol{\mu}_q + \boldsymbol{K}(\boldsymbol{\eta} - \boldsymbol{C}(\boldsymbol{x}_s)\boldsymbol{\mu}_q) \\ \widetilde{\boldsymbol{\Sigma}_q}(\boldsymbol{x_s}) &= \boldsymbol{\Sigma}_q - \boldsymbol{K}\boldsymbol{C}(\boldsymbol{x}_s)\boldsymbol{\Sigma}_q \end{aligned} \tag{3.14}$$

où $\boldsymbol{K}$ est définie par :

$$\boldsymbol{K} = \boldsymbol{\Sigma}_q\boldsymbol{C}(\boldsymbol{x}_s)^T(\boldsymbol{C}(\boldsymbol{x}_s)\boldsymbol{\Sigma}_q\boldsymbol{C}(\boldsymbol{x}_s)^T + \boldsymbol{R})^{-1} \tag{3.15}$$

Pour une valeur de $\boldsymbol{x_s}$ donnée, les paramètres $\widetilde{\boldsymbol{\mu}_q}(\boldsymbol{x_s})$ et $\widetilde{\boldsymbol{\Sigma}_q}(\boldsymbol{x_s})$ sont ainsi obtenus de façon analytique (i.e. sans approximation) par les équations (3.14). Dans un souci de lisibilité, on notera plus simplement $\widetilde{\boldsymbol{\mu}_q}$ et $\widetilde{\boldsymbol{\Sigma}_q}$ par la suite.

Contrainte de positivité

Dans le but d'assurer la positivité des valeurs d'émission de la source, une contrainte peut être appliquée sur les résultats de l'équation (3.14), inspirée par les travaux de [Simon and Simon, 2010]. Il s'agit d'utiliser une méthode permettant de restreindre les valeurs d'un vecteur d'état à un intervalle borné ou semi-borné prédéfini : pour cela, la densité de probabilité de ce vecteur d'état est tronquée suivant la contrainte que l'on cherche à appliquer.

1. La *chain rule* permet d'exprimer une loi jointe sous la forme d'un produit de lois conditionnelles. Si on considère les n variables aléatoires $X_1, \ldots, X_n$, alors on a $p(X_1, \ldots, X_n) = p(X_n|X_1, \ldots, X_{n-1})p(X_{n-1}|X_1, \ldots, X_{n-2}) \ldots p(X_2|X_1)p(X_1)$.

Le processus de troncature est ainsi appliqué de façon séquentielle sur chaque composante de $\boldsymbol{q}$: on travaille ainsi à tronquer T_s densités de loi univariées. Le détail de cette démarche est décrit par l'algorithme 4, qui permet d'approximer la loi a posteriori conditionnelle de $\boldsymbol{q}$ par :

$$p^c(\boldsymbol{q}|\boldsymbol{x}_s,\boldsymbol{\eta}) = \mathcal{N}(\boldsymbol{q}|\widetilde{\boldsymbol{\mu}_q}^c, \widetilde{\boldsymbol{\Sigma}_q}^c) \tag{3.16}$$

Algorithm 4. Contrainte de positivité sur $\boldsymbol{q}$ par troncature de la densité de $p(\boldsymbol{q}|\boldsymbol{x}_s,\boldsymbol{\eta})$

Entrées : $\widetilde{\boldsymbol{\mu}_q}$ et $\widetilde{\boldsymbol{\Sigma}_q}$
Initialisation : $\widetilde{\boldsymbol{\mu}_q}^c = \widetilde{\boldsymbol{\mu}_q}$ et $\widetilde{\boldsymbol{\Sigma}_q}^c = \widetilde{\boldsymbol{\Sigma}_q}$
for $i = 1 : T_s$ **do**
$\boldsymbol{\gamma}_i = \begin{bmatrix}\mathbf{0}_{1\times(i-1)} & 1 & \mathbf{0}_{1\times(T_s-i)}\end{bmatrix}^T$
Calculer $\boldsymbol{W}_i$ et $\boldsymbol{T}_i$ par la réduction de Jordan de $\widetilde{\boldsymbol{\Sigma}_q}^c$, i.e. $\boldsymbol{T}_i\boldsymbol{W}_i\boldsymbol{T}_i^T = \widetilde{\boldsymbol{\Sigma}_q}^c$
Calculer $\boldsymbol{S}_i$ par l'orthogonalisation de Gram-Schmidt pour obtenir la matrice orthogonale $\boldsymbol{S}_i$ telle que

$$\boldsymbol{S}_i\boldsymbol{W}_i^{1/2}\boldsymbol{T}_i^T\boldsymbol{\gamma}_i = \begin{bmatrix}(\boldsymbol{\gamma}_i^T\widetilde{\boldsymbol{\Sigma}_q}^c\boldsymbol{\gamma}_i)^{1/2} & 0 & \cdots & 0\end{bmatrix}$$

$c_i = -\dfrac{\boldsymbol{\gamma}_i^T\widetilde{\boldsymbol{\mu}_q}^c}{(\boldsymbol{\gamma}_i^T\widetilde{\boldsymbol{\Sigma}_q}^c\boldsymbol{\gamma}_i)^{1/2}}$
$\mu_i = \dfrac{\phi(c_i)}{1-\Phi(c_i)}$ avec $\phi(\cdot)$ la densité de la loi normale centrée réduite, et $\Phi(\cdot)$ la fonction de répartition de la loi normale centreé réduite.
$\sigma_i^2 = 1 - \mu_i(\mu_i - c_i)$
$\boldsymbol{z}_i = \begin{bmatrix}\mu_i & 0 & \cdots & 0\end{bmatrix}^T$
$\boldsymbol{D}_i = \mathrm{diag}(\sigma_i^2, 1, \ldots, 1)$
Calculer les paramètres de la densité tronquée :

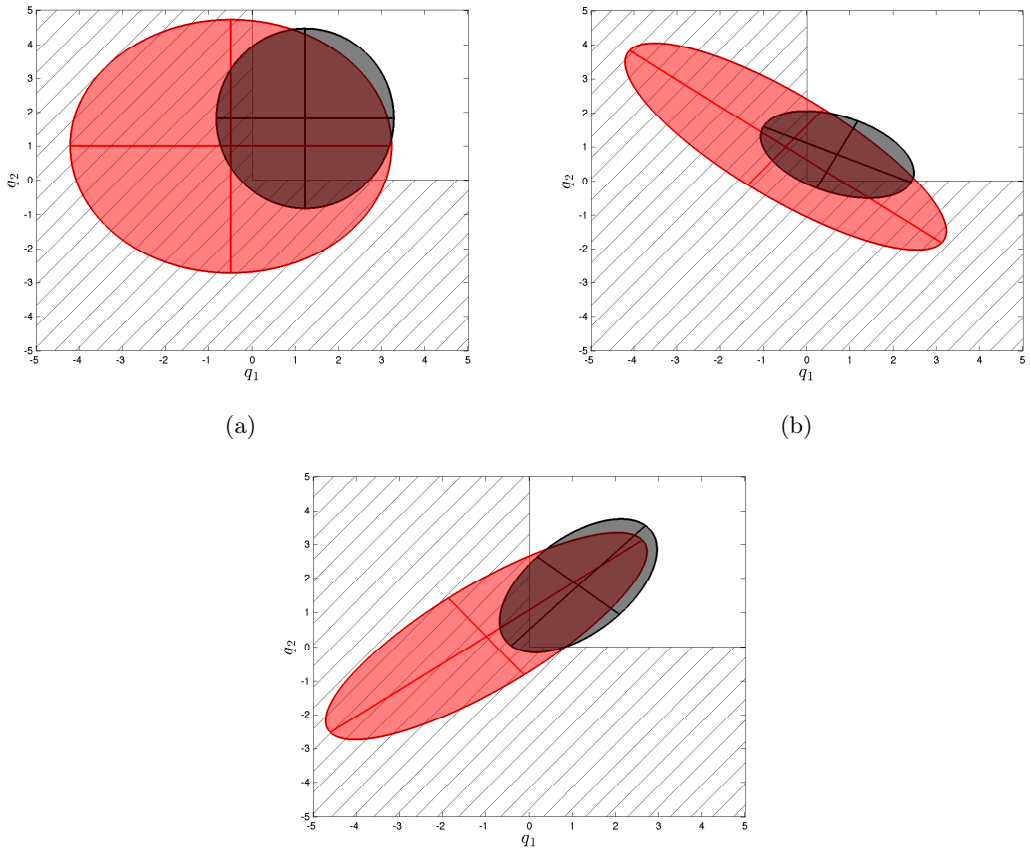
$$\widetilde{\boldsymbol{\mu}_q}^c = \boldsymbol{T}_i\boldsymbol{W}_i^{1/2}\boldsymbol{S}_i^T\boldsymbol{z}_i + \widetilde{\boldsymbol{\mu}_q}^c$$
$$\widetilde{\boldsymbol{\Sigma}_q}^c = \boldsymbol{T}_i\boldsymbol{W}_i^{1/2}\boldsymbol{S}_i^T\boldsymbol{D}_i\boldsymbol{S}_i\boldsymbol{W}_i^{1/2}\boldsymbol{T}_i^T$$

end for
Sorties : $\widetilde{\boldsymbol{\mu}_q}^c$ et $\widetilde{\boldsymbol{\Sigma}_q}^c$

Notons qu'une telle procédure permet également d'optimiser le calcul de la vraisemblance marginale de la localisation de la source $p(\boldsymbol{\eta}|\boldsymbol{x}_s)$: celle-ci peut alors se calculer par le même procédé que celui employé dans l'équation (3.12), et devient alors :

$$p^c(\boldsymbol{\eta}|\boldsymbol{x}_s) = \frac{p(\boldsymbol{\eta}|\boldsymbol{q},\boldsymbol{x}_s)p(\boldsymbol{q})}{p^c(\boldsymbol{q}|\boldsymbol{x}_s,\boldsymbol{\eta})} \tag{3.17}$$

La figure 3.3 résume bien le fonctionnement de l'algorithme 4 : la zone en noir représente la version tronquée aux valeurs positives de la distribution initiale (en rouge).



(a) (b)

(c)

FIGURE 3.3 – Illustration de l'application de la contrainte de positivité (en noir) sur les paramètres d'une distribution gaussienne bivariée (en rouge) dans trois cas distincts : sans corrélation (3.3a), avec corrélation négative (3.3b) et avec corrélation positive (3.3c).

3.3.2 Localisation de la source avec l'algorithme AMIS

Il s'agit ici d'étudier la loi a posteriori marginale $p(\boldsymbol{x}_s|\boldsymbol{\eta})$ de la position de la source. Contrairement au profil d'émission, il est impossible d'obtenir une solution analytique pour caractériser cette distribution. Par conséquent, l'idée est d'utiliser les méthodes de simulation stochastique afin d'approximer la distribution $p(\boldsymbol{x}_s|\boldsymbol{\eta})$.

Principe

Nous appliquons l'algorithme AMIS décrit dans le chapitre 2 afin de localiser la source. Pour cela, on va ainsi construire une procédure itérative basée sur l'algorithme 3 afin de créer sur K itérations et en générant N_p particules par itération un échantillon de KN_p particules $\left\{\boldsymbol{x}_{s_k}^{(i)}\right\}_{0\leq k\leq K}^{1\leq i\leq N_p}$ dont la distribution approxime celle de la loi marginale a posteriori $p(\boldsymbol{x}_s|\boldsymbol{\eta})$ de la position de la source.

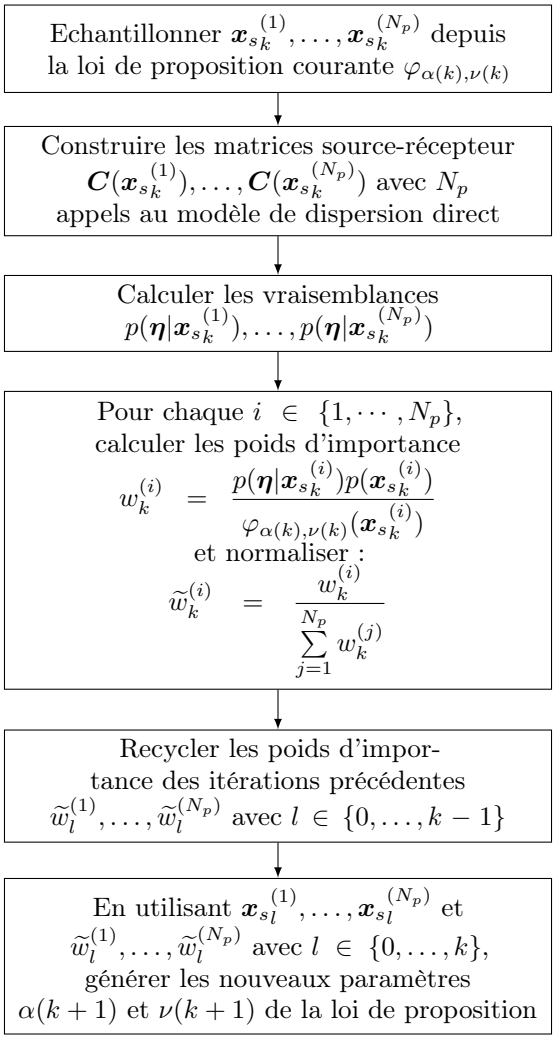


FIGURE 3.4 – Diagramme de fonctionnement de l'algorithme AMIS sur la k-ième itération pour la localisation de la source.

Dans un cadre bayésien, la loi cible recherchée peut alors s'écrire à une constante multiplicative près, suivant la relation de proportionnalité suivante :

$$p(\boldsymbol{x}_s|\boldsymbol{\eta}) \propto p(\boldsymbol{\eta}|\boldsymbol{x}_s)p(\boldsymbol{x}_s) \tag{3.18}$$

La fonction de vraisemblance $p(\boldsymbol{\eta}|\boldsymbol{x}_s)$ est définie par l'équation (3.8) (ou (3.17) si on choisit d'appliquer la contrainte de positivité), et la loi a priori $p(\boldsymbol{x}_s)$ par l'équation (3.9). L'équation (3.18) permet ainsi d'évaluer la loi cible toute particule $\boldsymbol{x}_s$ échantillonnée depuis la loi de proposition courante, afin de calculer les poids d'importance associés. La structure globale de la démarche est décrite par la figure 3.4.

Choix de la loi de proposition

L'algorithme AMIS requiert une loi de proposition paramétrique qui soit suffisamment flexible pour :

— pouvoir s'adapter de façon à être capable d'approximer correctement la loi cible une fois ses paramètres ajustés,

— effectuer simplement l'opération d'adaptation des paramètres. Sur ce point, le choix d'une mixture de gaussiennes telle que présentée dans [Cappé et al., 2008] est cohérent, car les formules de mise à jour sont directement disponibles via l'algorithme EM.

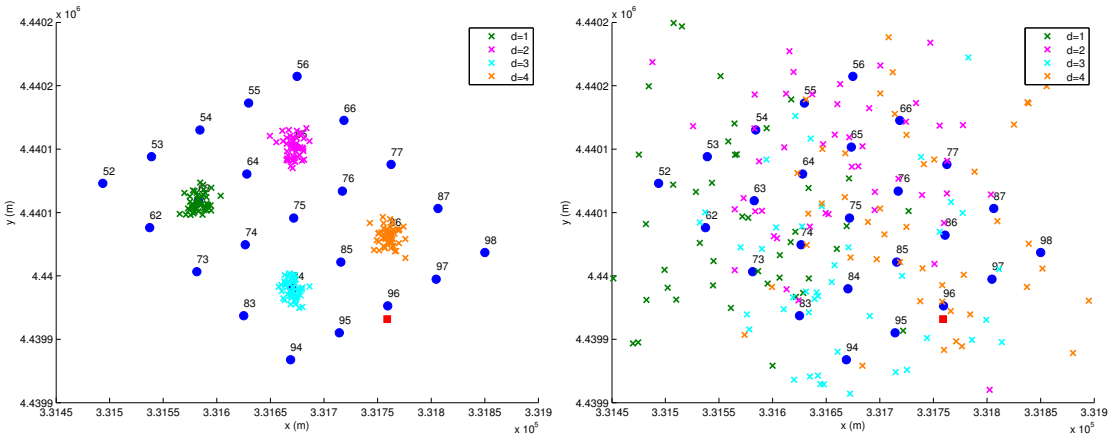
On définit donc la loi de proposition comme une mixture de D distribution gaussiennes bivariées :

$$\varphi_{(\alpha,\nu)}(\boldsymbol{x}_s) = \sum_{d=1}^{D} \alpha_d \varphi_d(\boldsymbol{x}_s|\nu_d) \tag{3.19}$$

où les α_d sont les facteurs d'influence de chacune des composantes de la mixture, et ν_d représente le vecteur des paramètres de la d-ième composante, telle que :

$$\varphi_d(\boldsymbol{x}_s|\nu_d) = \mathcal{N}(\boldsymbol{x}_s|\boldsymbol{\mu}_d, \boldsymbol{\Sigma}_d) \tag{3.20}$$

avec $\boldsymbol{\mu}_d \in \mathbb{R}^2$ et $\boldsymbol{\Sigma}_d \in \mathbb{R}^{2\times 2}$.



(a) $\boldsymbol{\Sigma}_d = \begin{pmatrix} 50 & 0 \\ 0 & 50 \end{pmatrix}$ (b) $\boldsymbol{\Sigma}_d = \begin{pmatrix} 5000 & 0 \\ 0 & 5000 \end{pmatrix}$

FIGURE 3.5 – Exemples d'initialisation de l'AMIS avec $D = 4$ composantes pour la loi de proposition : un bon choix des $\boldsymbol{\Sigma}_d$ permet un tirage homogène sur le domaine (3.5b) tandis qu'une covariance trop faible ne permettra pas d'explorer tout le domaine (3.5a).

Le choix des paramètres initiaux permet de régler la surface couverte par chacune des composantes afin d'obtenir une répartition relativement homogène des particules, illustrant une absence de connaissance a priori sur une localisation potentielle de la source (voir figure 3.5).

En pratique, nous avons :

1. divisé le domaine en quatre quadrants de surface égale,
2. placé les moyennes $\boldsymbol{\mu}_d$ sur le centre de chacun des quadrants,

3. réglé les matrices de covariance $\mathbf{\Sigma}_d$ de façon à ce que chacune des composantes de la mixture fournisse des échantillons couvrant toute la surface de son quadrant.

Si une information a priori est disponible (par exemple, des zones particulières du domaine à exclure ou à favoriser), les paramètres $\alpha_d, \boldsymbol{\mu}_d$ et $\mathbf{\Sigma}_d$ avec $d \in \{1, \ldots, D\}$ peuvent être préalablement ajustés afin d'orienter en conséquence l'échantillonnage des particules dès la première itération. Un exemple de ce type d'initialisation est présenté au Chapitre 4.

Une fois l'algorithme lancé, à chaque itération, ces paramètres sont mis à jour grâce aux formules des équations (2.39).

3.3.3 Loi a posteriori jointe des paramètres de la source

En suivant la procédure décrite par l'algorithme 3.4 sur un total de K itérations, la loi a posteriori marginale de la position de la source peut être approximée grâce à (2.30) de la façon suivante :

$$p(\boldsymbol{x_s}|\boldsymbol{\eta}) \simeq \sum_{k=1}^{K} \sum_{i=1}^{N_p} \widetilde{w}_k^{(i)} \delta_{\boldsymbol{x_s}_k^{(i)}}(\boldsymbol{x_s}) \tag{3.21}$$

En reprenant l'expression de la loi a posteriori jointe (3.12), on obtient finalement la solution suivante :

$$p(\boldsymbol{x_s}, \boldsymbol{q}|\boldsymbol{\eta}) \simeq \sum_{k=1}^{K} \sum_{i=1}^{N_p} \widetilde{w}_k^{(i)} p(\boldsymbol{q}|\boldsymbol{x_s}_k^{(i)}, \boldsymbol{\eta}) \delta_{\boldsymbol{x_s}_k^{(i)}}(\boldsymbol{x_s}) \tag{3.22}$$

Les estimations respectives (au sens du MMSE) de la position de la source et de son profil d'émission sont alors données par les relations suivantes :

$$\begin{aligned} \widehat{\boldsymbol{x_s}} &= \sum_{k=0}^{K} \sum_{i=1}^{N_p} \widetilde{w}_k^{(i)} \boldsymbol{x_s}_k^{(i)} \\ \widehat{\boldsymbol{q}} &= \sum_{k=1}^{K} \sum_{i=1}^{N_p} \widetilde{w}_k^{(i)} \widetilde{\boldsymbol{\mu}_q}(\boldsymbol{x_s}_k^{(i)}) \end{aligned} \tag{3.23}$$

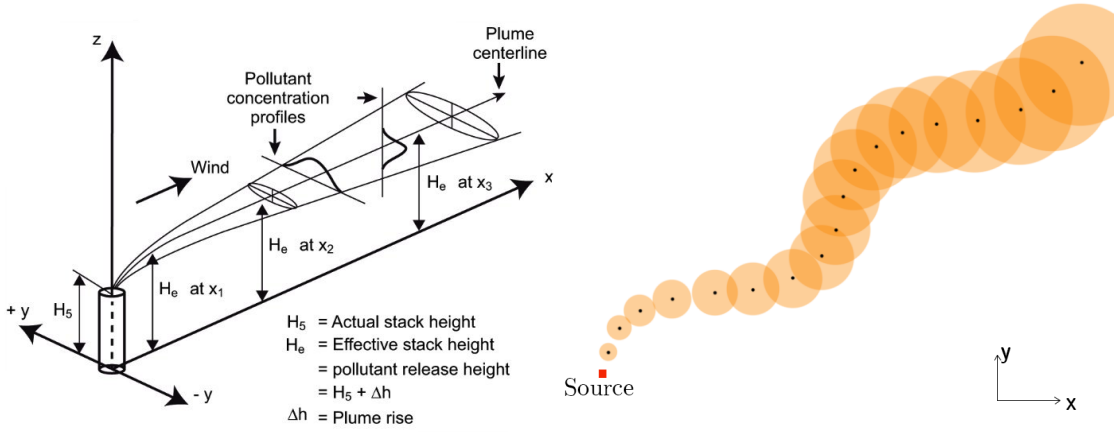
3.4 Description du modèle de dispersion à bouffées gaussiennes

Une fois la loi de proposition initialisée, on peut exécuter la boucle itérative de l'algorithme AMIS telle que décrite à l'algorithme 3. Dans cet algorithme, le calcul de la loi de vraisemblance telle que décrite à l'équation (3.8) requiert l'exécution d'un modèle de dispersion pour obtenir $\boldsymbol{C}(\boldsymbol{x}_s)$. Ici, nous choisissons d'utiliser un modèle *à boufféees gaussiennes*, ou *Gaussian Puff Model* (GPM).

Les modèles gaussiens permettent de représenter la dispersion à petite échelle autour d'une source. Pour cela, une formulation analytique de la concentration du polluant est calculée suivant plusieurs hypothèses permettant une mise en oeuvre simple et peu coûteuse en temps de calcul. Il existe deux types différents de modèles gaussiens :

— les modèles dits *de panache* (ou *Gaussian plume*), où la dispersion est modélisée dans des conditions météorologiques supposées uniformes et stationnaires. Le panache émis par la source est alors modélisé par une distribution de type gaussienne dans deux directions : celle orthogonale au vent, et la direction verticale (voir figure 3.6a).

— les modèles dits *à bouffées* (ou *Gaussian puff*), modélisant une série d'émissions instantanées représentées par des bouffées à distribution gaussienne dans les trois directions de l'espace, et dont les centres sont transportés par un champ de vent qui peut être non-uniforme et non-stationnaire. La globalité du rejet à un instant donné est ainsi illustrée par la somme des différentes bouffées émises par la source (voir figure 3.6b). C'est ce type de modèle que nous utilisons dans l'étude de cas sur FFT07.



(a) Modèle de panache gaussien [Schulze and Turner, 1996]

(b) Modèle de bouffées gaussiennes

FIGURE 3.6 – Schémas de principe des modèles de dispersion gaussiens à panache (3.6a) et à bouffées (3.6b)

Dans le cadre de FFT07, le choix d'un modèle à bouffées par rapport à un modèle à panache est justifié par le caractère instationnaire du profil d'émission que nous cherchons à estimer. La grandeur $\boldsymbol{q}$ est en effet un vecteur, et non une valeur constante, comme c'est le cas si on appliquait un modèle de type *Gaussian plume*. Cela permet également de prendre en compte les variations météorologiques telles que des changements de direction de vent.

La concentration mesurée au i-ème capteur ${\boldsymbol{x}_c}^{(i)} = (x_c^{(i)}, y_c^{(i)})$ au temps d'observation t_j résultant d'une source localisée en $\boldsymbol{x}_s$ émettant un rejet instantané à l'instant t'_n est donnée par :

$$\boldsymbol{C}_{i,j}(\boldsymbol{x}_s, t'_n) = \frac{2Q_u \Delta t_p}{(2\pi)^{3/2} s_x s_y s_z} \exp\left[-\frac{1}{2}(\Lambda_x^2 + \Lambda_y^2)\right] \tag{3.24}$$

où :

— Q_u est la quantité unitaire de polluant émise : comme les concentrations fournies par les capteurs sont en kg/m^3 on a $Q_u = 1$kg/s,

— Δt_p est l'écart temporel entre l'émission de deux bouffées consécutives dans le modèle de dispersion, ici $\Delta t_p = 1$s,

— s_x, s_y, s_z sont les coefficients de dispersion sur les trois axes de l'espace,

— les grandeurs Λ_x et Λ_y sont définies par :

$$\begin{aligned} \Lambda_x &= \frac{x_c^{(i)} - \left(\boldsymbol{x}_s + \sum_{t=t'_n}^{t_j} u_x(t)\delta_w\right)}{s_x} \\ \Lambda_y &= \frac{y_c^{(i)} - \left(\boldsymbol{x}_s + \sum_{t=t'_n}^{t_j} v_x(t)\delta_w\right)}{s_y} \end{aligned} \tag{3.25}$$

δ_w est l'intervalle de temps entre deux valeurs de mesures du vent, ce dernier étant représenté par ses deux composantes u_x et v_x.

On se place dans le cas d'une diffusion isotrope en x et y, on a ainsi $s_x = s_y$. Les valeurs de s_y et s_z sont obtenues par la formule suivante :

$$\begin{aligned} s_y &= a_y[d(\boldsymbol{x}_s, \boldsymbol{x}_p)]^{b_y} \\ s_z &= a_z[d(\boldsymbol{x}_s, \boldsymbol{x}_p)]^{b_z} \end{aligned} \tag{3.26}$$

où a_x, a_y, b_x, b_y sont des coefficients empiriques qui dépendent de la classe de stabilité atmosphérique [Pasquill and Smith, 1983] et $d(\boldsymbol{x}_s, \boldsymbol{x}_p)$ est la distance totale parcourue par la bouffée depuis son émission. On utilise ici la classe de Pasquill D, qui est la plus appropriée pour des conditions stables en topographie non-urbaine. Comme les valeurs de la composante verticale du vent qui ont été relevées sont relativement faibles, et étant donné que les capteurs et la source se trouvent assez près du sol (2m), on ne tient pas non plus compte du terme en z qui apparaît dans l'exponentielle dans la formulation générale du modèle à bouffées gaussiennes.

3.5 Présentation des résultats

3.5.1 Résultats obtenus avec des données simulées

Dans un premier temps, afin de valider la méthode, le choix est fait de se placer dans un cadre synthétique, à savoir reproduire les conditions de l'expérience FFT07 et simuler les concentrations mesurées aux capteurs grâce à l'exécution d'un modèle de dispersion, en l'occurence le modèle

gaussien défini au paragraphe précédent. Ainsi, les positions des capteurs et des sources sont exactement les mêmes que celles des *trials* 7 et 30 de FFT07. Les données météorologiques sont ajustées afin que les concentrations simulées soient les plus proches possibles des relevés expérimentaux, et le vent est considéré constant en direction et en vitesse.

L'algorithme AMIS a été exécuté sur $K = 10$ itérations, avec $N_p = 100$ particules échantillonnées par itération. La loi de proposition comporte $D = 4$ composantes. Les observations synthétiques ont été générées avec un ajout de bruit gaussien centré de variance 10^{-5} afin de simuler les différents éléments d'erreurs présents dans le cas expérimental. La variance d'observation est fixée à $\sigma^2_{obs} = 10^{-5}$. La loi a priori de $\boldsymbol{q}$ a été choisie telle que $\boldsymbol{\mu}_q = 0$ et $\boldsymbol{\Sigma}_q = \sigma^2_q \boldsymbol{I}$ avec $\sigma^2_q = 10^{-3}$. En pratique, diminuer σ^2_q revient à supposer a priori que la source est proche des capteurs ayant mesuré une concentration non-nulle, et à l'inverse une variance σ^2_q élevée implique une source potentiellement plus éloignée. Les valeurs choisies pour les différents paramètres énumérés ci-dessus sont celles ayant permis d'obtenir les résultats les plus plausibles, présentés ci-après.

Estimation de la position

L'algorithme a été testé avec deux positions de source différentes, correspondant respectivement aux *trials* 7 et 30, leurs coordonnées (en km) sont données par :

$$\begin{aligned} \boldsymbol{x}_s{}^{(7)} &= (331.759; 4439.960) \\ \boldsymbol{x}_s{}^{(30)} &= (331.825; 4439.972) \end{aligned} \tag{3.27}$$

On peut voir sur la figure 3.7 que l'AMIS fournit une bonne estimation de la position de la source dans chacun des *trials*.

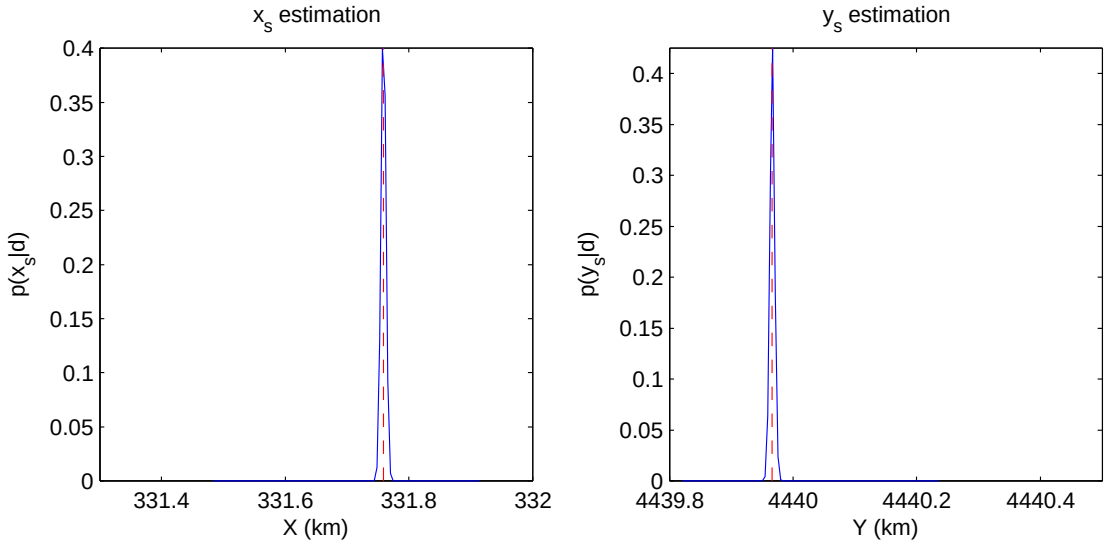
Le fait de visualiser une distribution de probabilité permet ainsi une vision plus explicite des aspects liés aux incertitudes autour de l'estimation, en comparaison avec des méthodes de type optimisationoù seule une estimation ponctuelle est donnée.

Pour obtenir les distributions présentées à la figure 3.7, un lissage par noyau a été effectué afin de construire une densité de probabilité à partir d'une population statistique, enl'occurrence les particules $\left\{\boldsymbol{\theta}_k^{(i)}\right\}_{0\leq k\leq K}^{1\leq i\leq N_p}$ associées à leurs poids d'importance respectifs $\left\{\widetilde{w}_k^{(i)}\right\}_{0\leq k\leq K}^{1\leq i\leq N_p}$.

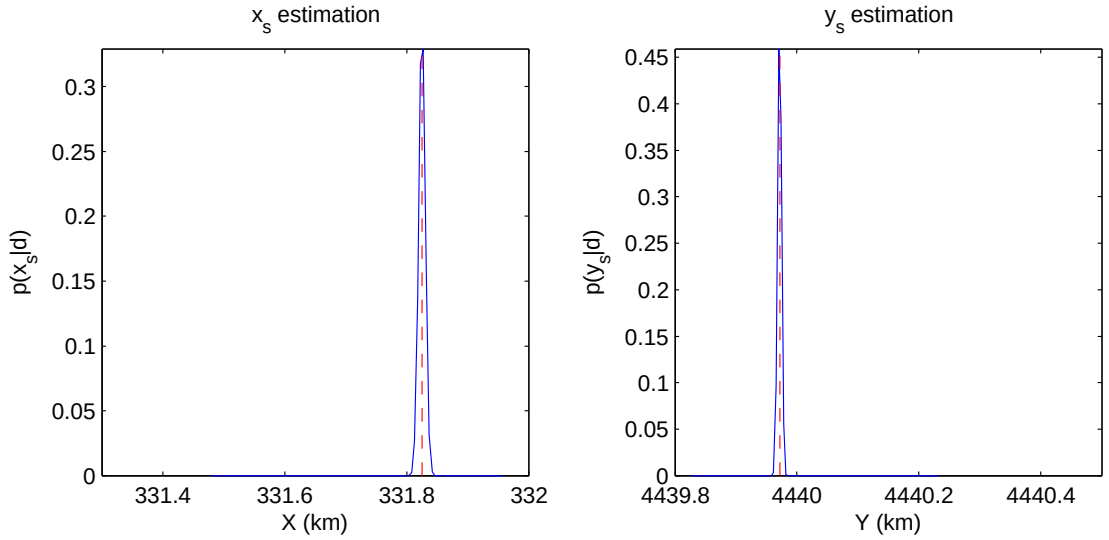
Néanmoins, il est toujours possible d'obtenir un estimateur ponctuel à partir des lois a posteriori, il suffit pour cela de se ramener à l'estimation MMSE présentée à l'équation (3.23). Par ce procédé, on arrive ainsi à un point estimé distant de moins d'un mètre par rapport à la position réelle de la source.

Estimation du profil de rejet

Afin de visualiser l'estimation du profil d'émission, on utilise également une approche basée sur le MMSE. Le vecteur $\widehat{\boldsymbol{q}}$ estimé est obtenu depuis l'équation (3.23), et la matrice de covariance



(a) Trial 7



(b) Trial 30

FIGURE 3.7 – Distributions a posteriori de la position de la source (en bleu) à partir de données d'observations synthétiques pour le trial 7 (3.7a) et le trial 30 (3.7b). La positionréelle de la source est en pointillés rouges.

estimée associée est donnée par :

$$\boldsymbol{\Sigma}_{\widehat{\boldsymbol{q}}} = \sum_{k=1}^{K} \sum_{i=1}^{N_p} \widetilde{w}_k^{(i)} \widetilde{\boldsymbol{\Sigma}_q}(\boldsymbol{x}_{\boldsymbol{s}_k}^{(i)}) \tag{3.28}$$

La figure 3.8 illustre l'impact de l'utilisation de la contrainte de positivité pour l'estimation du profil du rejet $\boldsymbol{q}$.

L'application de la contrainte de positivité améliore le résultat de l'estimation et permet de retrouver un profil d'émission suffisamment proche de celui recherché pour avoir une bonne estimation, tant en termes d'amplitude du débit que de détection des temps d'activation et d'arrêt

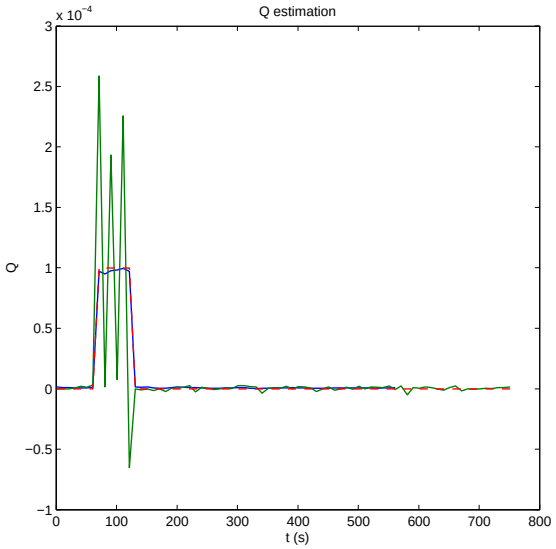


FIGURE 3.8 – Estimation de $\boldsymbol{q}$ sans (vert) et avec (bleu) application de la contrainte de positivité, comparaison avec le profil d'émission recherché (rouge).

d'émission de la source.

En général, du fait de sa nature potentiellement variable, le profil d'émission de la source peut prendre différents types d'allures. Nous avons ainsi modifié ce profil afin de représenter des cas non-triviaux, généré les observations correspondantes, et cherché à retrouver les caractéristiques de la source.

Comme présenté en figure 3.9, la méthode parvient à fournir une bonne estimation du profil de rejet dans les trois cas présentés. Sachant que l'estimation de la loi a posteriori de $\boldsymbol{q}$ dépend à la fois de $\widetilde{\boldsymbol{\mu}_q}$ et $\widetilde{\boldsymbol{\Sigma}_q}$, il est possible de représenter la marge d'incertitude associée à chaque valeur estimée : pour chacun des cas de la figure 3.9 la valeur réelle du débit est bien incluse dans l'intervalle à $\pm 2\widetilde{\sigma}_q$, où $\widetilde{\sigma}_q^2 = \text{diag}(\boldsymbol{\Sigma}_{\widetilde{\boldsymbol{q}}})$. Dans le cas de la figure 3.9c, on observe un effet de bord sur la fin du rejet estimé, phénomène qui est absent des autres configurations. Cela est dû au fait que l'algorithme ne dispose pas des mesures permettant de confirmer ou non les valeurs non-nulles prises par le débit estimé, ces observations étant après la limite supérieure de la fenêtre temporelle d'observation. Le problème ne se pose pas dans les configurations précédentes, car comme les concentrations mesurées sont nulles au dernier instant d'observation, l'algorithme laisse la valeur de débit estimé égale à sa moyenne a priori, à savoir 0.

Le pic à la limite de l'intervalle d'émission est dû au fait que le rejet est plus long que la période d'observation : en l'absence d'information, l'algorithme ne peut fournir une estimation correcte, mais demeure cohérent en augmentant fortement la taille de la marge d'incertitude, signifiant ainsi que l'information manque.

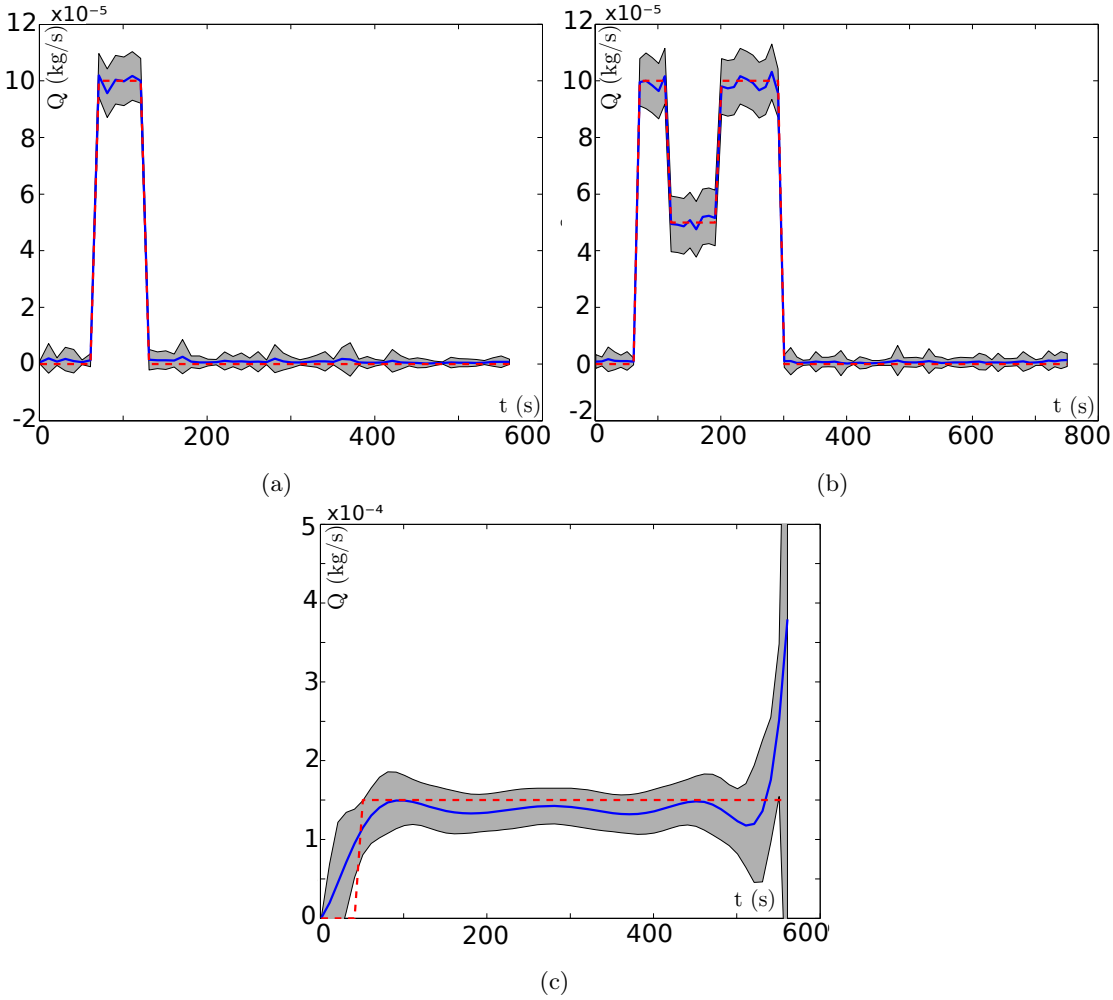


FIGURE 3.9 – Simulation et résultats d'estimation pour différents profils d'émission : rejet simple (3.9a), rejet variable (3.9b), rejet continu (3.9c). Le profil à retrouver (en rouge) est comparé au profil estimé (en bleu) et à l'intervalle de confiance à $\pm 2\widetilde{\sigma}_q$ (en gris).

Robustesse statistique de l'estimation

L'algorithme AMIS est une procédure basée sur l'échantillonnage d'importance, qui est par nature une démarche aléatoire. Pour tenir compte de cette caractéristique dans la validation de notre méthodologie, nous avons exécuté l'AMIS sur plusieurs *runs*, avec et sans application de la contrainte de positivité, et en perturbant les observations de concentrations fournies. Nous avons ensuite comparé les performances dans chacun des cas, en analysant les résultats présentés au tableau 3.1.

Contrainte de positivité	$\bar{d}(\widehat{\boldsymbol{x}_s}, \boldsymbol{x}_s)$ (m)	$\sigma_{\widehat{x_s}}$ (m)	$\sigma_{\widehat{y_s}}$ (m)
OUI	< 1	0.9	3.2
NON	10.44	0.4	1.0

TABLE 3.1 – Evaluation statistique des estimations sur 40 runs de l'AMIS avec et sans application de la contrainte de positivité. $\bar{d}(\widehat{\boldsymbol{x}_s}, \boldsymbol{x}_s)$ représente la distance moyenne entre la position de source estimée par MMSE et l'emplacement réel de la source. $\sigma_{\widehat{x_s}}$ et $\sigma_{\widehat{y_s}}$ sont respectivement les moyennes des écarts-types des distributions a posteriori marginales $p(x_s|\boldsymbol{\eta})$ et $p(y_s|\boldsymbol{\eta})$.

Ces résultats confirment que l'application de la contrainte de positivité apporte en moyenne une certaine amélioration au niveau de la localisation de la source. Si on cherche à quantifier la précision de l'estimation du profil de rejet, on peut calculer l'erreur quadratique moyenne, ou *mean square error* (MSE), entre la valeur estimée $\widehat{\boldsymbol{q}}$ et la valeur théorique $\boldsymbol{q}$ selon la formule suivante :

$$\mathrm{MSE}(\widehat{\boldsymbol{q}}, \boldsymbol{q}) = \frac{1}{T_s} \sum_{n=1}^{T_s} \left(\widehat{\boldsymbol{q}}(t'_n) - \boldsymbol{q}(t'_n)\right)^2 \tag{3.29}$$

Dans le cas où on applique la contrainte de positivité, on a une erreur MSE de 1.37×10^{-12}, qui est une meilleure valeur que pour le cas sans contrainte, où on a une erreur MSE de 1.26×10^{-9}.

Comparaison avec le MCMC

Nous avons ensuite cherché à comparer les performances de l'AMIS avec un autre algorithme d'inférence bayésienne couramment utilisé pour les problèmes STE, en l'occurence la méthode MCMC, présentée dans le Chapitre 2.

Pour cela, nous avons implémenté une algorithme de type *random walk Metropolis* avec un noyau de transition gaussien fixe, et l'intégration de la contrainte de positivité de l'algorithme 4 dans le calcul de la vraisemblance, sur 1000 itérations. Afin de tenir compte du phénomène de *burn-in*, les 100 premiers états sont ignorés.

Un test de robustesse statistique a ensuite été effectué, en changeant de façon aléatoire le point de départ de la chaîne de Markov à chaque *run* (uniformément sur le domaine). Pour la localisation de la source, l'estimateur MMSE moyen obtenu par MCMC a été comparé aux résultats de l'AMIS dans le tableau 3.2.

Algorithme	$\bar{d}(\widehat{\boldsymbol{x}_s}, \boldsymbol{x}_s)$ (m)	$\sigma_{\widehat{x_s}}$ (m)	$\sigma_{\widehat{y_s}}$ (m)
AMIS	< 1	0.9	3.2
MCMC	22.7	130	95

TABLE 3.2 – Evaluation statistique des estimations sur 40 *runs* de l'AMIS et du MCMC avec application de la contrainte de positivité.

Dans cette étude de cas, on remarque ainsi que les performances de l'AMIS sont meilleures en termes de précision pour localiser la source. De fait, l'approche MCMC est en partie pénalisée par l'étape d'initialisation. En effet, la convergence vers une position suffisamment proche de la vraie source est d'autant plus longue et incertaine que l'état initial de cette chaîne est éloigné de la vraie source. En ce sens, le fait de travailler sur une population de particules plutôt que sur la succession d'états séquentiels, ainsi que l'aspect adaptatif de l'AMIS lui permettent une meilleure flexibilité pour la mise à jour des paramètres de la loi de proposition, réduisant ainsi la dépendance de ses performances par rapport à l'état initial. Une alternative serait d'instancier plusieurs chaînes de Markov dont les états initiaux sont suffisamment éloignés les uns des autres pour bien couvrir le domaine, mais cela augmente la charge de calcul et les ressources requises pour l'exécution de l'algorithme.

En plus de la précision de chacun des algorithmes, nous nous sommes également penchés sur leur vitesse de convergence.

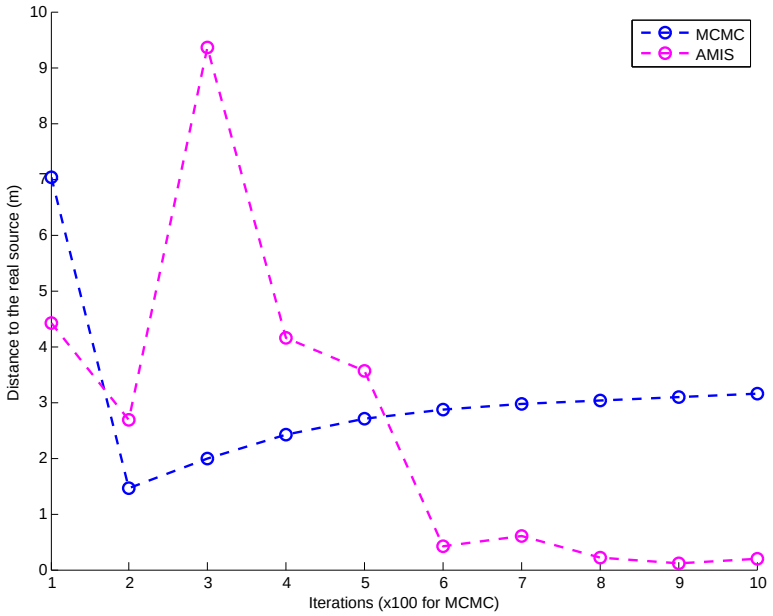


FIGURE 3.10 – Comparaison empirique de la vitesse de convergence entre MCMC (en bleu) et AMIS (en magenta)

Sur la figure 3.10, on utilise un critère qualitatif qui est la distance de l'estimation MMSE du lieu du rejet à une itération donnée. Les calculs ont été calibrés afin que chacun des algorithmes traite une charge équivalente de calcul. Ainsi, avec :

— $N_p = 100$ particules par itérations et $K = 10$ itérations pour l'AMIS,

— 1000 itérations pour le MCMC,

le nombre de calculs de la loi de vraisemblance, et par extension le nombre d'appels au modèle de dispersion, est strictement le même pour les deux algorithmes.

Les résultats obtenus démontrent l'efficacité de l'aspect adaptatif de l'AMIS, qui lui permet d'atteindre rapidement une estimation de bonne qualité pour la localisation de la source.

Encore une fois, dans la démarche MCMC, l'initialisation de la chaîne de Markov influe sur la vitesse de convergence, cette dernière étant d'autant plus élevée que l'état initial se situe près de la vraie source. Un autre paramètre du MCMC intervient également ici : la variance du noyau de transition. En effet, comme expliqué au Chapitre 2, si la valeur choisie pour cette variance est trop faible, même si l'initialisation place la chaîne dans une position relativement favorable, la convergence sera trop lente car le nombre d'états nécessaires pour atteindre une estimation correcte sera plus élevé. A l'inverse, il est risqué de trop augmenter la valeur de cette variance, car cela entraînerait des transitions d'amplitude trop élevée, pouvant potentiellement empêcher la convergence vers le point source recherché. L'AMIS est quant à lui plus souple dans sa démarche

de transition d'une itération à l'autre, cependant il peut être sujet à des problèmes de convergence si, par exemple, les particules échantillonnées par la loi de proposition initiale ne couvrent pas la zone où se situe la vraie source, d'où l'intérêt d'une initialisation homogène (figure 3.5b).

Effective Sample Size

Un critère spécifique aux méthodes basées sur l'échantillonnage d'importance est la représentativité de la distribution cible en fonction des particules ayant été échantillonnées afin d'approximer cette loi. Un outil permettant de quantifier cette grandeur est l'ESS, tel que présenté à l'équation (2.32), il permet ici en particulier de surveiller si l'AMIS est bloqué par un effet de dégénérescence des poids, ce qui se manifeste par une valeur d'ESS constamment faible.

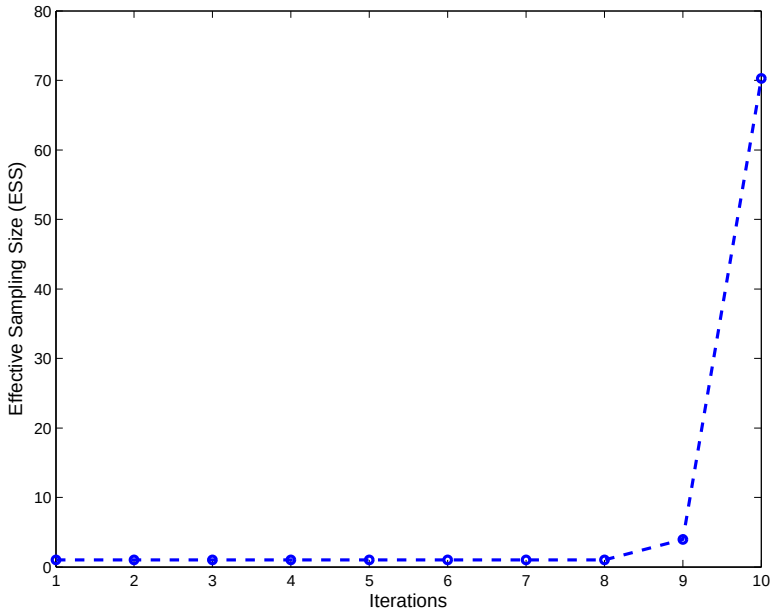


FIGURE 3.11 – Evolution de l'ESS par itération (avec $N_p = 100$ particules tirées par itération)

On observe sur la figure 3.11 que l'ESS prend un certain nombre d'itérations avant d'atteindre des valeurs suffisamment élevées, néanmoins l'estimation de la position de la source devient rapidement pertinente si on se réfère à la figure 3.10. En pratique, cela illustre le fait que l'AMIS a tendance à affecter un poids relativement élevé à un nombre restreint de particules (voire à une particule unique) proches de la vraie source.

3.5.2 Résultats obtenus avec des données expérimentales

Après un premier ensemble de tests sur données simulées, nous avons appliqué la même méthodologie en utilisant les mesures de concentrations expérimentales directement issues de l'expérience FFT07.

Traitement des données manquantes

Une première analyse des données d'observation a permis de mettre en évidence l'absence de mesures de concentrations sur certaines plages temporelles, voire sur la totalité de la fenêtre d'observation, pour plusieurs capteurs du réseau, dans les *trials* 7 et 30. Ces données manquantes sont le résultat de défaillances ponctuelles des dispositifs instrumentaux au moment des essais. Nous avons ainsi eu à gérer deux types de situations :

— Dans les cas où les données sont partiellement manquantes, typiquement s'il existe une quantité finie de points de mesure où les concentrations ne sont pas disponibles, alors une interpolation linéaire est faite pour remplir les valeurs manquantes à partir des concentrations disponibles.

— Dans les cas où aucune concentration n'est disponible sur un capteur donné, ce dernier est écarté du réseau et n'est pas intégré au vecteur d'observations $\boldsymbol{\eta}$. La même opération est faite si la proportion des points de mesures manquants est significativement supérieure à la quantité de mesures valables, l'opération d'interpolation étant dans ces cas-là susceptible d'introduire une information faussée.

Variabilité météorologique

Le fait d'utiliser des données expérimentales permet également de prendre en compte la présence de variations de vitesse et de direction du vent, et voir si l'algorithme d'estimation de la source arrive à gérer une telle situation. En particulier dans le *trial* 7, la direction du vent change sensiblement durant la période du rejet, comme le montre la figure 3.12.

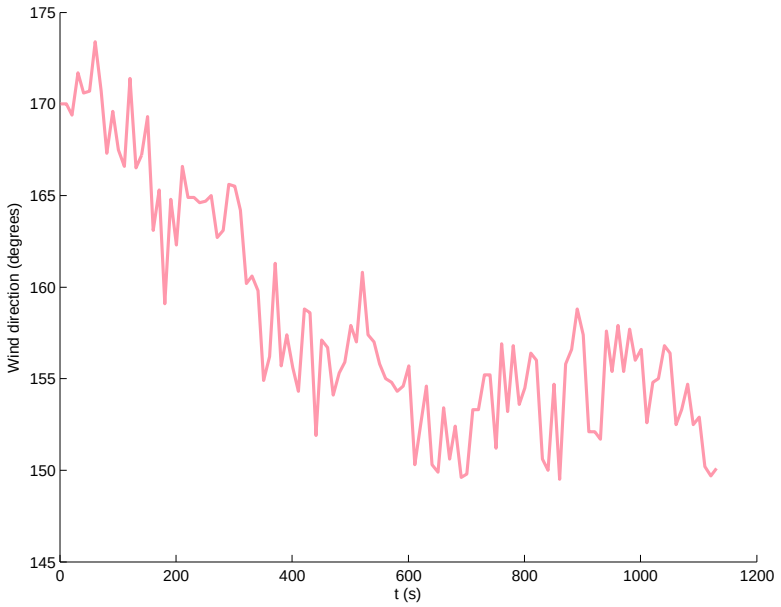


FIGURE 3.12 – Direction du vent mesurée par la station météorologique la plus proche de la source (*trial* 7)

Le choix d'un modèle de dispersion à bouffées gaussiennes permet ainsi d'intégrer ces variations de vent : pour le *trial* 7 nous avons donc recueilli et moyenné les mesures de vent issues des deux

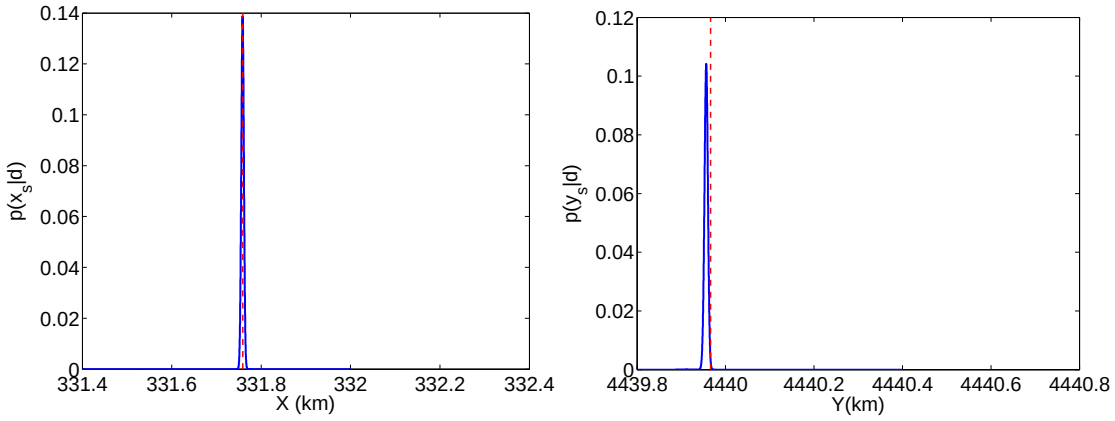
capteurs de vent soniques (anémomètres à ultrasons) présents sur le domaine, puis intégré les valeurs résultantes dans le modèle de dispersion. Pour le *trial* 30 les variations sont beaucoup moins importantes, nous avons donc gardé l'hypothèse d'un vent stationnaire uniforme.

Estimation de la position (*trials* 7 et 30)

De même que pour le cas synthétique, nous avons obtenu par KDE les distributions a posteriori de la position de la source sur chacun des axes, et représenté les résultats pour les *trials* 7 et 30 sur les figures 3.13 et 3.14. Les paramètres suivants ont été utilisés :

— *trial* 7 : $\sigma^2_{obs} = 10^{-10}$ et $\sigma^2_q = 5 \times 10^{-2}$,

— *trial* 30 : $\sigma^2_{obs} = 5 \times 10^{-8}$ et $\sigma^2_q = 5 \times 10^{-3}$.



(a) (b)

FIGURE 3.13 – Distribution a posteriori de la position de la source (en bleu) à partir des données d'observations expérimentales FFT07 pour le *trial* 7. La position réelle de la source est en pointillés rouges.

On observe que les résultats sont presque aussi bons que dans le cas synthétique. En termes de distance de l'estimée ponctuelle MMSE par rapport à la vraie source, on reste sur des valeurs satisfaisantes de 9.8m pour le *trial* 7 et 1.5m pour le *trial* 30.

Comparaison avec le MCMC (*trial* 7)

Une comparaison avec le même type d'algorithme MCMC que le paragraphe §3.5.1 a également été menée, en utilisant les données expérimentales du *trial* 7. On remarque sur la figure 3.15 que pour l'estimation de la position, les deux méthodes fournissent des résultats globalement satisfaisants.

La comparaison a aussi été faite avec les résultats d'estimation du profil de rejet, comme illustré en figure 3.16. Les résultats ont été moyennés sur des fenêtres d'1 minute afin de lisser l'allure du débit.

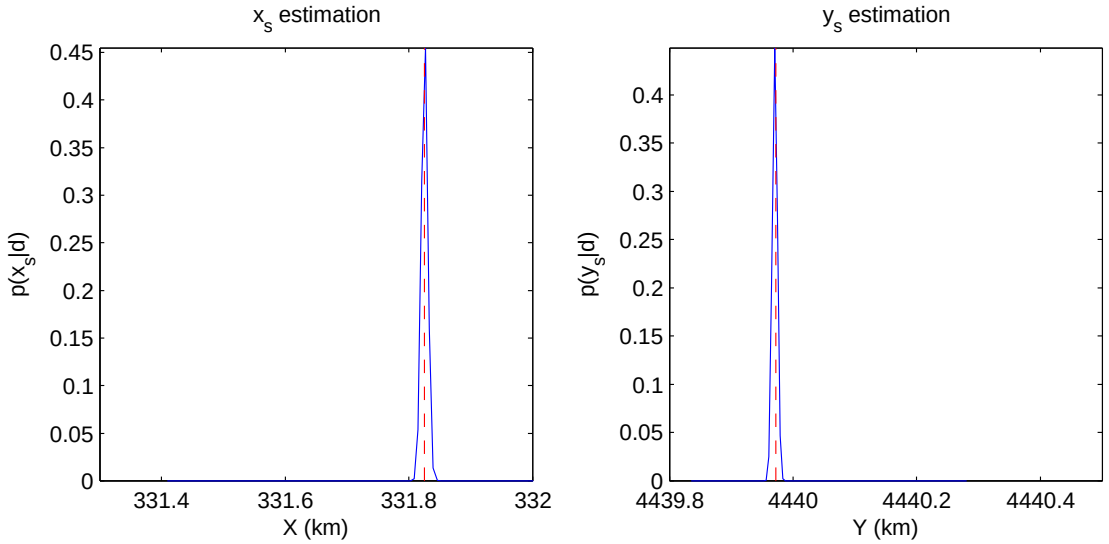
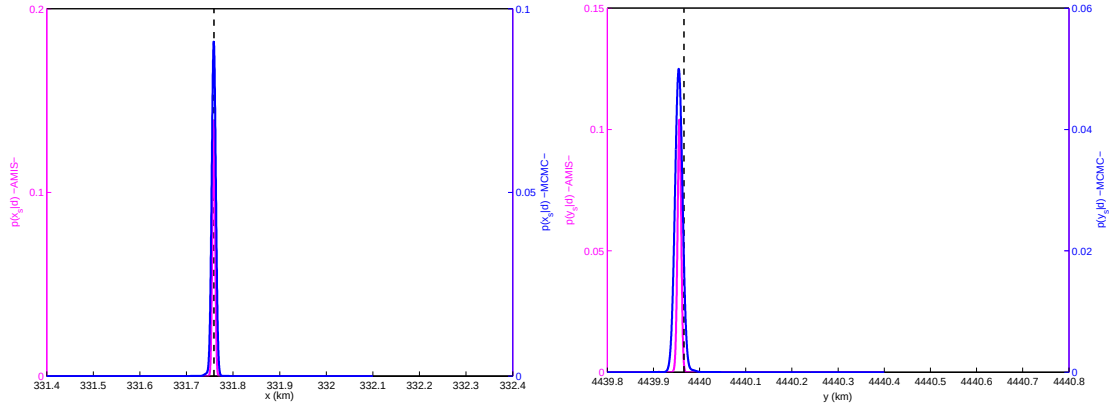


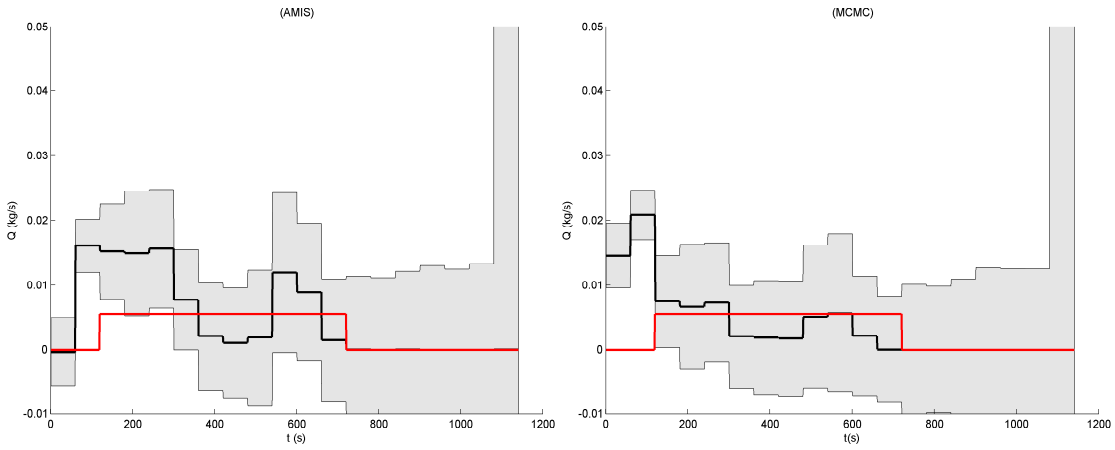
FIGURE 3.14 – Distribution a posteriori de la position de la source (en bleu) à partir des données d'observations expérimentales FFT07 pour le *trial* 30. La position réelle de la source est en pointillés rouges.



(a) (b)

FIGURE 3.15 – Comparaison de l'estimation de la position de la source du *trial* 7 avec des observations réelles en utilisant AMIS (en magenta) et MCMC (en bleu). La position réelle de la source est en pointillés noirs.

On note que les premières valeurs non-nulles du débit estimé apparaissent plus tôt qu'attendu, et avec des amplitudes plus importantes. Cela est en partie causé par le fait que l'estimée ponctuelle de la source a été située en amont de la position réelle, ce qui a forcé l'algorithme à produire un débit plus important pour ajuster en conséquence les concentrations résultantes au niveau des capteurs, et à avancer l'instant de "démarrage" du rejet. Ce phénomène est un peu plus marqué dans le cas du MCMC (3.16b).



(a) AMIS (b) MCMC

FIGURE 3.16 – Estimation du profil d'émission du *trial* 7 (en noir), intervalle de confiance à $\pm 2\widetilde{\sigma}_q$ (en gris) et comparaison avec les valeurs réelles (en rouge), avec AMIS (3.16a) et MCMC (3.16b).

3.6 Conclusions

L'étude de cas sur l'expérience FFT07 présentée dans ce chapitre a permis une première validation de la méthodologie axée sur l'algorithme AMIS. Celle-ci a en effet permis une bonne estimation de la position de la source dans les cas synthétiques et expérimentaux, en fournissant une distribution a posteriori des coordonnées de la source. Cet outil statistique permet ainsi d'effectuer une analyse autour des incertitudes associées à l'estimation, tout en permettant un accès à l'information ponctuelle par le biais du calcul de l'estimateur MMSE.

La démarche et a également permis de reconstruire un profil d'émission complet, fournissant une information temporelle sur les quantités émises par la source ainsi que ses temps de démarrage et d'arrêt. Cet aspect temporel présente un aspect innovant par rapport aux approches conventionnelles présentées dans la littérature, où l'hypothèse d'un débit d'émission constant est formulée, se traduisant dans la pratique par la recherche d'une unique valeur scalaire q.

Cette étude a démontré que la quasi-totalité de la charge de calcul est concentrée sur la construction des matrices source-récepteur, plus précisément sur les appels au modèle de dispersion. Dans le cas d'un modèle gaussien les temps de calcul demeurent raisonnables, mais ils dépendent directement du nombre de particules échantillonnées par itération de l'AMIS. Nous avons pu vérifier que le fait de tirer plus de 100 particules par itération dans cette étude de cas n'améliorait pas significativement les résultats d'estimation en termes de précision, mais dans le cadre général un nombre trop faible de particules pourrait dégrader la qualité de l'estimation, car l'exploration de l'espace des paramètres s'en trouverait restreinte. De plus, à nombre égal de particules, si le contexte exige l'utilisation d'un modèle de dispersion plus élaboré (par exemple dans un milieu urbain), la charge et le temps de calcul seront bien plus importants.
Il devient ainsi nécessaire d'optimiser à la fois le mode d'usage du modèle de dispersion ainsi que l'exploration des différentes valeurs possibles pour les particules. Ces deux points sont abordés dans le Chapitre 4, qui présente une nouvelle façon d'obtenir les résultats des calculs de dispersion, ainsi qu'une initialisation améliorée de la loi de proposition.

Chapitre 4

AMIS et modèle lagrangien rétrograde

La méthodologie utilisée dans le Chapitre 3 introduit une dépendance directe (linéaire) du nombre d'appels au modèle de dispersion avec la quantité de particules échantillonnées. Pour les cas les plus simples, il est possible de se limiter à des modèles de dispersion suffisamment rapides, mais la situation change dès qu'il s'agit de considérer un contexte plus élaboré (par exemple en présence d'obstacles). Il devient en effet nécessaire d'utiliser des outils de simulation permettant une meilleure représentation des phénomènes de dispersion. Cependant, le nombre et la complexité des calculs supplémentaires effectués au sein de ces outils rallongent le temps d'exécution du modèle de dispersion, et par conséquent réduisent la vitesse à laquelle l'opération de reconstruction de la source est accomplie.

Dans ce chapitre, nous proposons une modification de l'approche initiale, en remplaçant l'utilisation d'un modèle de dispersion direct par celle d'un modèle rétrograde. Cette substitution permet à la fois :

— d'optimiser le calcul des matrices source-récepteur, afin de permettre l'utilisation d'outils de simulation appropriés à la complexité de la situation, sans toutefois augmenter de façon excessive les temps de calcul ;
— d'exploiter des résultats de rétro-propagation pour construire une procédure d'initialisation améliorée de la loi de proposition de l'AMIS.

4.1 Le système de modélisation PMSS

4.1.1 L'approche lagrangienne de la dispersion atmosphérique

Pour les modèles dits eulériens, la résolution du problème de la dispersion d'un polluant dans l'atmosphère passe par la construction d'un maillage sur le domaine étudié, afin de pouvoir observer l'évolution des concentrations du polluant porté par les mouvements de l'air.

Le point de vue lagrangien est différent : il s'agit ici de résoudre un système d'équations dans un repère lié au déplacement de la masse d'air contenant le polluant. Pour cela, on représente le panache sous la forme d'un ensemble de *particules lagrangiennes*[1], chacune étant porteuse d'une masse élémentaire du polluant considéré. Le principe d'un modèle lagrangien consiste ainsi à étudier les trajectoires de ces éléments discrets dans le domaine au fil du temps.

Le fait de modéliser le panache par un ensemble de PL permet de tenir compte de la nature stochastique de leur déplacement, qui traduit la variabilité inhérente aux processus de turbulences auxquels est soumis le panache : on parle d'ailleurs plus précisément de *modèle lagrangien stochastique*. On va ainsi travailler sur une équation de transport portant sur la densité de probabilité associée à chaque trajectoire. Plus formellement, d'après [Flesch et al., 1995], la formulation classique régissant un modèle de dispersion lagrangien se présente sous la forme d'une équation de Langevin, qui s'écrit :

$$\begin{aligned} du_i &= a_i(\boldsymbol{x}, \vec{\boldsymbol{u}}, t)dt + b_{i,j}(\boldsymbol{x}, \vec{\boldsymbol{u}}, t)d\xi_j \\ dx_i &= u_i dt = (\bar{u}_i + U_i)dt \end{aligned} \tag{4.1}$$

où :

- $\boldsymbol{x} = (x, y, z)$ est la position de la PL définie par un repère spécifique : x suit l'axe du vent, y suit l'axe perpendiculaire au vent, et z désigne l'élévation verticale classique.
- $\vec{\boldsymbol{u}}$ est le champ d'écoulement auquel est soumise la PL,
- a_i et $b_{i,j}$ sont des fonctions spécifiques de $(\boldsymbol{x}, \boldsymbol{u}, t)$ respectivement appelées *drift term* et *random forcing*.
- $d\xi_j$ est un incrément aléatoire suivant une distribution gaussienne de moyenne nulle et de variance dt.
- $\bar{u}_i$ représente le vent moyen et U_i sa composante stochastique.

L'expression des fonctions a_i et $b_{i,j}$ varie selon les hypothèses que l'on se fixe sur la nature de la turbulence : une présentation plus détaillée de leur calcul est disponible dans [Wilson and Sawford, 1996]. Une fois que ceux-ci sont définis, l'équation (4.1) est discrétisée et sa résolution permet de calculer un ensemble de trajectoires de PL émanant d'une source dont les paramètres sont connus. Les concentrations volumiques moyennes simulées sont alors obtenues par la somme des particules présentes dans un volume élémentaire $d\boldsymbol{x}$ autour du point d'observation $\boldsymbol{x}$ durant un certain temps de résidence. En d'autres termes, en reprenant la définition énoncée dans [Flesch et al., 1995], on peut écrire la concentration moyenne au point $\boldsymbol{x}$ et à l'instant t comme étant :

$$C(\boldsymbol{x}, t) = \int_{-\infty}^{t} \int_{\Omega} S(\boldsymbol{x}', t')p(\boldsymbol{x}, t|\boldsymbol{x}', t')d\boldsymbol{x}'dt' \tag{4.2}$$

où Ω est le volume défini par le domaine d'étude, $S(\boldsymbol{x}', t')$ est la distribution de la source, et $p(\boldsymbol{x}, t|\boldsymbol{x}', t')$ est la densité de probabilité sur la position $\boldsymbol{x}$ et l'instant t des PL de position initiale $\boldsymbol{x}'$ à l'instant t'.

1. Afin d'éviter toute confusion avec les particules statistiques dont il est fait référence dans le cadre de l'algorithme AMIS, nous utiliserons la dénomination de *particule lagrangienne* abréviée par PL dans la suite du texte. Nous conservons le terme de *particule* pour désigner les échantillons issus de l'AMIS.

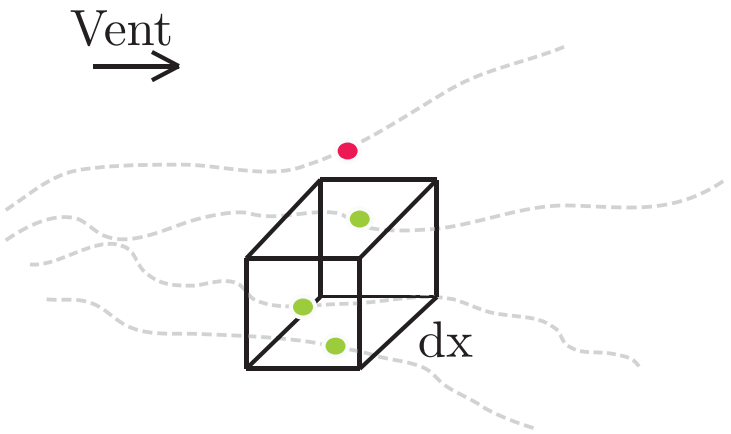


FIGURE 4.1 – Principe du modèle lagrangien : la concentration en $\boldsymbol{x}$ s'obtient par la somme des PL (en vert) traversant le volume élémentaire $d\boldsymbol{x}$ durant un certain temps de résidence.

Afin de calculer le champ d'écoulement $\boldsymbol{u}$ auquel sont soumises les PL, plusieurs méthodes sont disponibles. Il est par exemple possible de résoudre les équations de la mécanique des fluides via un outil de simulation de type CFD, ce qui permet une modélisation fine des phénomènes physiques mis en jeu. Une autre possibilité consiste à avoir recours à une simulation dite *CFD simplifiée*, où le champ de vent est interpolé à partir des mesures d'une ou plusieurs stations météorologiques tout en prenant en compte la topographie du terrain. C'est ce type d'approche qui est présenté dans le paragraphe suivant.

4.1.2 La chaîne de calcul SWIFT-SPRAY

Parallel Micro-SWIFT-SPRAY (PMSS) est une chaîne de calcul constituée de deux éléments distincts : un outil de CFD simplifiée (SWIFT) et un modèle de dispersion lagrangien stochastique (SPRAY). PMSS est conçu pour être appliqué dans des études micro-météorologiques en milieu bâti (site industriel, milieu urbain). La parallélisation rend possible son utilisation sur des domaines plus grands, à l'exemple du cas AIRCITY[Moussafir et al., 2014] où le modèle a été exécuté sur l'ensemble de la ville de Paris.

PSWIFT

Le modèle PSWIFT permet de produire des champs de vent 3D en exploitant différents types de données météorologiques sur un même site (profils de vent et de température, stations de mesures, sorties de modèles météorologiques de prévision). Il permet notamment de prendre en compte la topographie du milieu, la présence d'obstacles tels que des bâtiments, l'occupation des sols ou encore l'influence de la stabilité atmosphérique. Son fonctionnement est illustré par la figure 4.2, et peut être résumé en quatre étapes :

1. Dans un premier temps, les mesures météorologiques reçues en entrée sont interpolées sur les différents points constituant une version discrétisée du domaine.
2. Dans un deuxième temps, l'effet des obstacles présents dans le domaine sur l'écoulement est

modélisé via la création de zones spécifiques dans le voisinage de ces obstacles où le champ de vitesse est calculé par des relations analytiques.

3. La troisième étape consiste à ajuster le champ de vent en appliquant un principe de conservation de la masse.
4. Enfin, la dernière étape consiste à calculer la turbulence intrinsèque à l'écoulement modélisé.

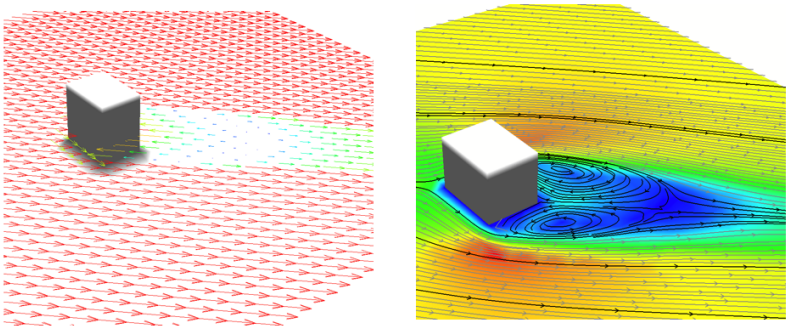


FIGURE 4.2 – Exemple de calcul d'un champ de vent autour d'un obstacle avec SWIFT, avant (à gauche) et après (à droite) l'ajustement du champ

En sortie de cet enchaînement de calculs, on obtient un champ de vent 3D qui peut alors directement être exploité par le modèle de dispersion PSPRAY.

PSPRAY

PSPRAY est un modèle de dispersion lagrangien stochastique dont les principes de base suivent les mécanismes présentés à la section 4.1.1. L'implémentation de PSPRAY repose sur le critère dit de *well-mixed condition* permettant de donner une formulation explicite aux termes a_i et $b_{i,j}$ de l'équation (4.1) , et présenté en détail dans les travaux de [Thomson, 1987].

En pratique, plusieurs fonctionnalités supplémentaires sont implémentées dans PSPRAY, telles que :

— le "rebond" des particules sur les obstacles,

— le calcul de doses pour les sources radioactives,

— la prise en compte des différents types de dépôts (secs ou humides).

Ces fonctionnalités permettent de représenter de façon réaliste le panache d'un rejet en présence d'obstacles, comme présenté sur la figure 4.3.

La combinaison de PSWIFT et PSPRAY permet ainsi de calculer un champ de concentration sur le domaine étudié, connaissant les paramètres du terme source qui sont soumis en entrée du modèle PSPRAY. Dans le contexte de ce chapitre, la chaîne de calcul PMSS permet de :

— générer un jeu d'observations synthétiques en simulant un rejet induit par une source que l'on va chercher à retrouver : pour cela, on calcule les valeurs du champ de concentrations

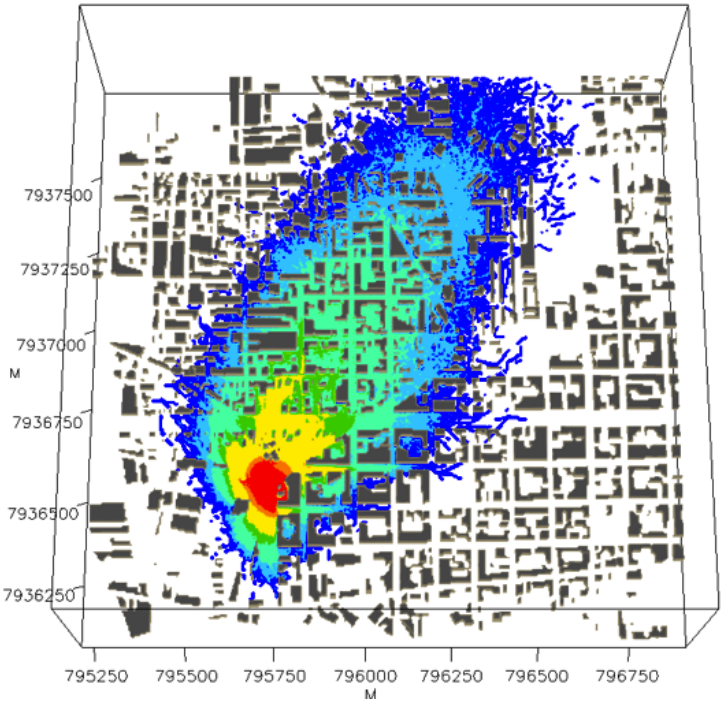


FIGURE 4.3 – Exemple de champ de concentration calculé par SPRAY dans un domaine de type urbain

en un nombre fini de points du domaine que nous définirons comme étant les observations fournies par les capteurs,

— construire les matrices source-récepteur lors de l'exécution de l'algorithme AMIS, nécessaires au processus d'estimation du terme source.

4.1.3 Dualité direct-rétrograde

L'optimisation du calcul des matrices source-récepteur suivant le modèle de l'équation (3.5) est un point important : en effet dans une approche directe telle que présentée dans le chapitre précédent, chaque construction de matrice source-récepteur nécessite une exécution complète du calcul de dispersion, ce qui rend l'opération d'estimation du terme source très coûteuse en temps de calcul.

Nous proposons dans la suite de ce chapitre une amélioration du processus d'estimation privilégiant le calcul des matrices source-récepteur par une approche de type rétrograde. Cette méthodologie a initialement été introduite dans [Keats et al., 2007] pour ensuite être appliquée sur un algorithme d'estimation de type MCMC, nous en reprenons les notations pour en rappeler les bases ci-après.

On se place dans le domaine spatio-temporel $\Omega \times [0, T]$, et on considère une source Q située au point $\boldsymbol{x}_s$ de l'espace, de débit massique constant q_s, et de temps d'activation et d'arrêt respectifs t_{on} et t_{off} :

$$Q(\boldsymbol{x},t) = q_s\delta(\boldsymbol{x}-\boldsymbol{x}_s)\big(H(t-t_{on}) - H(t-t_{off})\big) \tag{4.3}$$

où :

— δ est la distribution de Dirac définie par :

$$\delta(x) = \begin{cases} +\infty, & \text{si } x = 0 \\ 0, & \text{si } x \neq 0 \end{cases} \tag{4.4}$$

— H est la fonction de Heaviside définie par :

$$H(x) = \begin{cases} 0, & \text{si } x < 0 \\ 1, & \text{si } x \geq 0 \end{cases} \tag{4.5}$$

On note C le champ de concentration moyen induit par cette source. Soit R_i la concentration mesurée au i-ème capteur du réseau considéré : c'est une valeur qui dépend à la fois du champ C et de la "fonction de réponse" h du capteur, cette dernière étant définie par :

$$h = h(\boldsymbol{x}-\boldsymbol{x}_{R_i}, t-t_{R_i}) \tag{4.6}$$

pour un capteur renvoyant la concentration au point $\boldsymbol{x}_{R_i}$ à l'instant t_{R_i}. La valeur de R_i s'obtient alors avec la relation suivante :

$$\begin{aligned} R_i &= \int_0^T \int_\Omega C(\boldsymbol{x},t)h(\boldsymbol{x}-\boldsymbol{x}_{R_i}, t-t_{R_i})dtd\boldsymbol{x} \\ &= \langle C, h\rangle \end{aligned} \tag{4.7}$$

[Keats et al., 2007] stipule l'existence d'une relation de dualité entre C et le champ de rétro-concentrations C^*, qui résulte de l'expression suivante de R_i en fonction de C^* :

$$\begin{aligned} R_i &= \int_0^T \int_\Omega Q(\boldsymbol{x},t)C^*(\boldsymbol{x},t)dtd\boldsymbol{x} \\ &= \langle Q, C^*\rangle \end{aligned} \tag{4.8}$$

La relation de dualité direct-rétrograde peut alors être écrite sous la forme suivante :

$$\langle C, h\rangle = \langle Q, C^*\rangle \tag{4.9}$$

En pratique, C^* est simulé par un modèle de dispersion rétrograde. Cette relation de dualité permet de traiter le cas particulier où :

— sous l'hypothèse d'un "capteur idéal", la fonction de réponse h devient une distribution de Dirac centrée sur le point $(\boldsymbol{x},t)$:

$$\begin{aligned} \langle C, h\rangle &= \int_0^T \int_\Omega C(\boldsymbol{x},t)\delta(\boldsymbol{x}-\boldsymbol{x}_{R_i}, t-t_{R_i})dtd\boldsymbol{x} \\ &= C(\boldsymbol{x}_{R_i}, t_{R_i}) \end{aligned} \tag{4.10}$$

— si Q est une source instantanée dont le débit de rejet est unitaire (comme c'est le cas dans l'algorithme AMIS pour chaque particule dont on construit la matrice source-récepteur associée), alors elle est assimilable à une impulsion de type Dirac, avec une masse égale à 1 et centrée sur $(\boldsymbol{x}_s, t_s)$ où t_s est l'instant d'émission. Autrement dit :

$$\begin{aligned}\langle Q, C^* \rangle &= \int_0^T \int_\Omega \delta(\boldsymbol{x} - \boldsymbol{x}_s, t - t_s) C^*(\boldsymbol{x}, t) dt d\boldsymbol{x} \\ &= C^*(\boldsymbol{x}_s, t_s)\end{aligned} \tag{4.11}$$

Sous ces hypothèses, on obtient :

$$C(\boldsymbol{x}_{R_i}, t_{R_i}) = C^*(\boldsymbol{x}_s, t_s) \tag{4.12}$$

4.1.4 Intégration d'un modèle rétrograde au processus d'estimation

Utilisation de RetroSPRAY dans l'AMIS

PMSS contient une implémentation rétrograde du modèle SPRAY, appelée Retro-SPRAY, qui procède à la résolution en mode inverse des équations de Langevin (le détail des calculs est développé dans [Flesch et al., 1995] et [Wilson et al., 2009]).

Si on examine plus en détail la construction de la matrice source-récepteur, on peut réécrire l'équation (3.5) sous une forme plus explicite : notons $C(R_i, t_j|\boldsymbol{\theta}, t'_n)$ la concentration moyenne au capteur R_i à l'instant t_j résultant d'une source située à la position $\boldsymbol{\theta}$ et ayant émis un rejet unitaire à l'instant t'_n. La version directe de la matrice source-récepteur pour $\boldsymbol{\theta}$ s'écrit :

$$\boldsymbol{C}^f(\boldsymbol{\theta}) = \begin{pmatrix} C(R_1, t_1|\boldsymbol{\theta}, t'_1) & C(R_1, t_1|\boldsymbol{\theta}, t'_2) & \cdots & C(R_1, t_1|\boldsymbol{\theta}, t'_{T_s}) \\ C(R_1, t_2|\boldsymbol{\theta}, t'_1) & C(R_1, t_2|\boldsymbol{\theta}, t'_2) & \cdots & C(R_1, t_2|\boldsymbol{\theta}, t'_{T_s}) \\ \vdots & \vdots & & \vdots \\ C(R_1, t_{T_c}|\boldsymbol{\theta}, t'_1) & C(R_1, t_{T_c}|\boldsymbol{\theta}, t'_2) & \cdots & C(R_1, t_{T_c}|\boldsymbol{\theta}, t'_{T_s}) \\ C(R_2, t_1|\boldsymbol{\theta}, t'_1) & C(R_2, t_1|\boldsymbol{\theta}, t'_2) & \cdots & C(R_2, t_1|\boldsymbol{\theta}, t'_{T_s}) \\ \vdots & \vdots & & \vdots \\ \vdots & \vdots & & \vdots \\ C(R_{N_c}, t_{T_c}|\boldsymbol{\theta}, t'_1) & C(R_{N_c}, t_{T_c}|\boldsymbol{\theta}, t'_2) & \cdots & C(R_{N_c}, t_{T_c}|\boldsymbol{\theta}, t'_{T_s}) \end{pmatrix} \tag{4.13}$$

En appliquant la relation de dualité direct-rétrograde, dans l'espace dual où le modèle rétrograde opère, les sources deviennent des "rétro-capteurs", et les capteurs deviennent des "rétro-sources". On définit alors $C^*(\boldsymbol{\theta}, t'_n|R_i, t_j)$ comme la rétro-concentration mesurée au point $\boldsymbol{\theta}$ à l'instant t'_n provenant d'une rétro-source située à la position R_i et ayant émis un rétro-rejet unitaire à l'instant t_j. La version rétrograde de $\boldsymbol{C}^f$ s'écrit alors :

$$\boldsymbol{C}^b(\boldsymbol{\theta}) = \begin{pmatrix} C^*(\boldsymbol{\theta}, t'_1|R_1, t_1) & C^*(\boldsymbol{\theta}, t'_2|R_1, t_1) & \cdots & C^*(\boldsymbol{\theta}, t'_{T_s}|R_1, t_1) \\ C^*(\boldsymbol{\theta}, t'_1|R_1, t_2) & C^*(\boldsymbol{\theta}, t'_2|R_1, t_2) & \cdots & C^*(\boldsymbol{\theta}, t'_{T_s}|R_1, t_1) \\ \vdots & \vdots & & \vdots \\ C^*(\boldsymbol{\theta}, t'_1|R_1, t_{T_c}) & C^*(\boldsymbol{\theta}, t'_2|R_1, t_{T_c}) & \cdots & C^*(\boldsymbol{\theta}, t'_{T_s}|R_1, t_{T_c}) \\ \vdots & \vdots & & \vdots \\ \vdots & \vdots & & \vdots \\ C^*(\boldsymbol{\theta}, t'_1|R_{N_c}, t_{T_c}) & C^*(\boldsymbol{\theta}, t'_2|R_{N_c}, t_{T_c}) & \cdots & C^*(\boldsymbol{\theta}, t'_{T_s}|R_{N_c}, t_{T_c}) \end{pmatrix} \tag{4.14}$$

Dans l'algorithme AMIS, au moment de calculer la vraisemblance de chaque particule échantillonnée, on peut ainsi faire désormais intervenir $\boldsymbol{C}^b$ comme matrice source-récepteur.

Avantages

L'utilisation de l'approche rétrograde permet de n'avoir à faire qu'un seul calcul de rétro-dispersion par capteur, qui donne alors l'ensemble des valeurs de rétro-concentrations sur le domaine. Ces opérations sont faites en amont du schéma itératif de l'AMIS : le domaine est divisé en plusieurs mailles sur lesquelles sont pré-calculées valeurs de C^*. A partir de ces résultats, une série de fichiers binaires est créée puis stockée sur le disque. Ainsi, au moment de l'estimation du terme source, les valeurs à insérer dans les matrices source-récepteur sont déjà disponibles : il suffit d'associer les positions des particules tirées aux centres des mailles les plus proches sur lesquelles les valeurs de C^* ont été obtenues. Pour cela, il suffit alors d'effectuer une opération de lecture sur les fichiers binaires contenant les rétro-concentrations, ce qui permet ainsi de ne plus avoir à lancer le modèle de dispersion à chaque nouvelle particule tirée lors de la phase d'échantillonnage de l'AMIS.

Même si cette opération n'est plus directement intégrée à l'algorithme d'estimation, elle demande néanmoins une certaine quantité de calculs. Cette dernière peut toutefois être réduite si on choisit de paralléliser l'opération de génération des fichiers binaires de rétro-concentration. Comme il s'agit de tâches parfaitement indépendantes les unes par rapport aux autres, la parallélisation est facilement réalisable (*embarrasingly parallel jobs*). Si on doit générer n_b fichiers binaires en allouant n_p coeurs de calcul par opération de génération, cela requiert la disponibilité de $n_b \times n_p$ coeurs. Si une telle quantité de ressources n'est pas immédiatement disponible, il peut être plus judicieux de créer séquentiellement ces fichiers binaires. L'approche séquentielle est également à privilégier si on ne dispose pas d'une architecture de type *cluster* permettant une parallélisation massive.

4.2 Exemple d'application sur un terrain rural

Dans cette section, nous présentons un première application sur une situation simple dans un contexte non-urbain. Il s'agit d'un exemple synthétique dont les données et les paramètres sont issus d'une simulation ayant permis la validation du modèle Retro-SPRAY.

4.2.1 Présentation du cas-test

Nous considérons ici un cas-test en milieu rural reproduisant une émission accidentelle depuis un site industriel dans une zone située près de la commune de Beaune, en Bourgogne, en présence d'une topographie réelle vallonnée (figure 4.4).

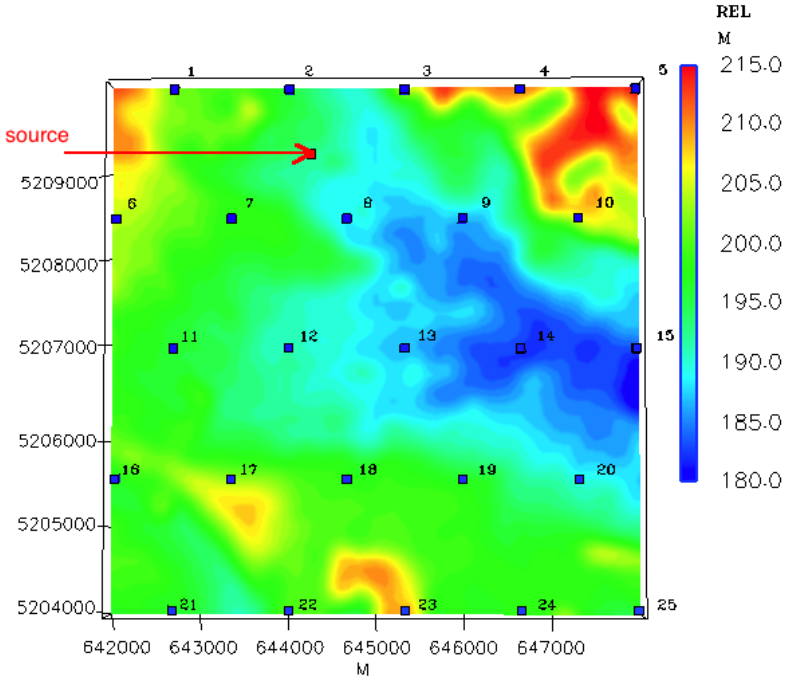


FIGURE 4.4 – Superposition du relief, de l'emplacement des capteurs et de la source du cas-test Beaune

Caractéristiques du domaine

Le domaine considéré couvre une surface de $6\,\mathrm{km}^2$ avec une source unique et un réseau relativement dense de 25 capteurs disposés en quinconce et couvrant toute la superficie du domaine (figure 4.4).

Pour la simulation, le domaine est discrétisé en une grille de 300×300 mailles, avec une résolution du maillage en x et y de $20\,\mathrm{m}$.

Paramètres météorologiques

Sur toute la durée de la simulation, nous considérons un vent constant en vitesse ($1{,}5\,\mathrm{m\,s^{-1}}$) et en direction ($330^\circ$). Le champ de vent produit par SWIFT est ainsi relativement homogène, comme observé en figure 4.5.

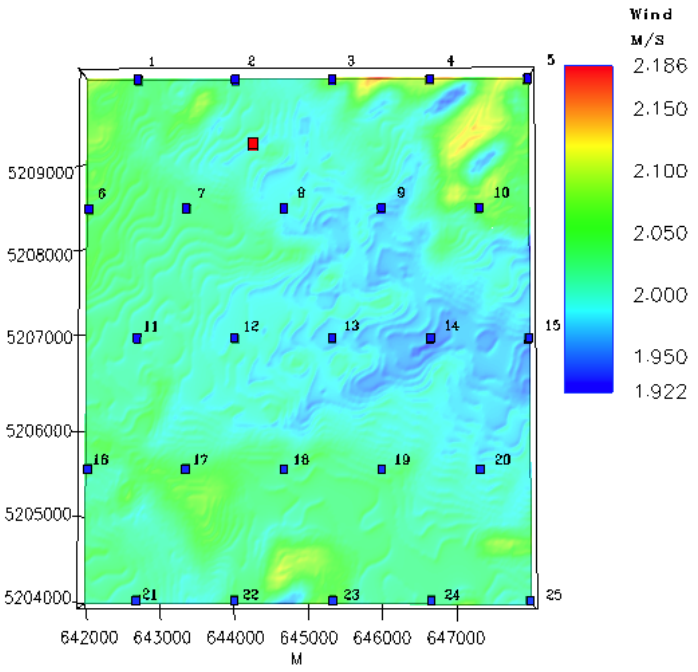


FIGURE 4.5 – Champ de vent calculé par SWIFT pour le cas-test Beaune

Capteurs, source et simulation des observations

On considère un réseau de 25 capteurs disposés de façon à couvrir tout le domaine, et placés à une hauteur de 10 m, égale à celle de la source. Chaque capteur fournit des observations de concentrations moyennes entre 10h05 et 12h00, avec une plage de moyennage de 5 minutes.

Concernant la source, celle-ci est placée dans la partie nord (voir figure 4.4) afin que le panache résultant couvre une partie suffisante du domaine. Elle est située à une altitude de 10 m, identique à celle des capteurs : le processus d'estimation est limité à la reconstruction de la position aux coordonnées (x, y). La source émet un rejet unique entre 10h15 et 11h, avec un débit constant de 1850 unités/s. Lors de la simulation des observations, elle est modélisée par un volume de $(15\,\text{m} \times 15\,\text{m} \times 10\,\text{m})$, les valeurs obtenues sont représentées en figure 4.6.

Il est à noter que pour générer les observations de la figure 4.6, un *run* spécifique du modèle de dispersion a été exécuté, sans réutiliser les valeurs pré-calculées présentes dans les fichiers binaires. Cela entraîne une marge d'incertitude sur les mesures, qu'il est possible d'associer à l'introduction d'une erreur de modèle non-nulle.

Premiers résultats

Un *run* préliminaire de l'AMIS servant de *benchmark* a été lancé avec les paramètres suivants :

— 10 itérations,

— 100 particules générées par itération,

— $\sigma^2_{obs} = 2 \times 10^{-6}$, qui fixe un degré d'incertitude suffisant autour des observations de la figure 4.6,

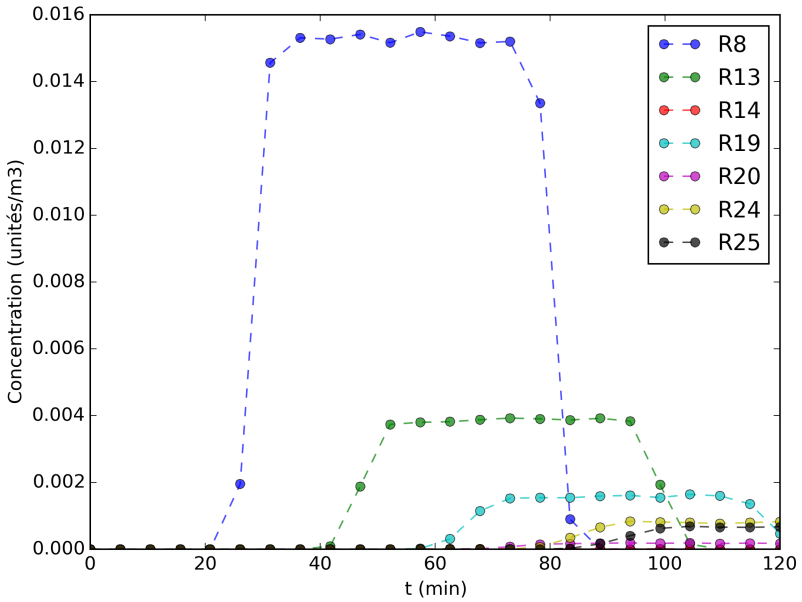


FIGURE 4.6 – Cas-test Beaune : concentrations mesurées aux capteurs

— une discrétisation temporelle de la source par paliers de 5 minutes, avec les paramètres a priori $\boldsymbol{\mu}_q = (0, \cdots, 0)$ et $\sigma_q^2 = 2 \times 10^6$.

Avec ces paramètres, l'algorithme d'estimation parvient à fournir une reconstruction relativement correcte des paramètres, aussi bien concernant la localisation de la source (figures 4.7a et 4.7b) que la reconstruction de son profil d'émission (figure 4.7c).

Pour mesurer l'erreur relative d'une estimation ponctuelle par rapport à la vraie position de la source, on utilise la métrique suivante :

$$r_d = \frac{d(\boldsymbol{x}_s, \widehat{\boldsymbol{x}}_{MMSE})}{L_{\mathcal{D}}} \tag{4.15}$$

où :

— $L_{\mathcal{D}}$ est la diagonale du domaine $\mathcal{D}$ considéré, autrement dit la plus grande distance possible entre deux points : cela permet de borner r_d entre 0 et 1,

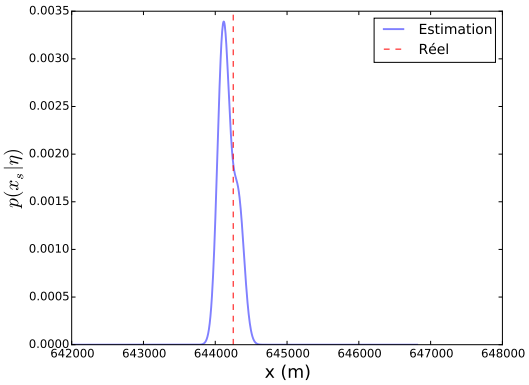
— $\boldsymbol{x}_s$ est le centre de la maille où est située la source,

— $\widehat{\boldsymbol{x}}_{MMSE}$ est l'estimation ponctuelle par MMSE de la position de la source obtenue via les particules et les poids d'importance fournis en sortie de l'AMIS.

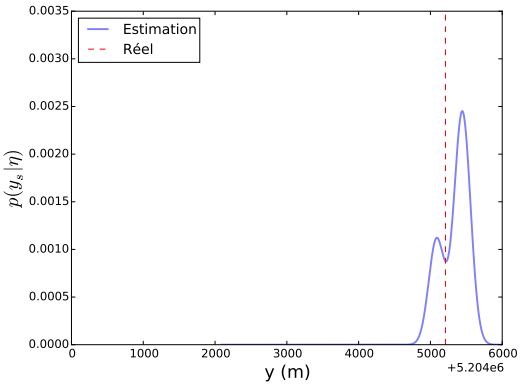
Pour mesurer l'erreur d'estimation du débit, on utilise la même métrique (MSE) que dans le chapitre précédent ((3.29)).

Dans l'exemple de la figure 4.7, on obtient ainsi une erreur relative de $r_d = 0.017$ pour la position, et une erreur d'estimation du profil d'émission valant $\text{MSE}(\widehat{\boldsymbol{q}}, \boldsymbol{q}) = 296.909$.

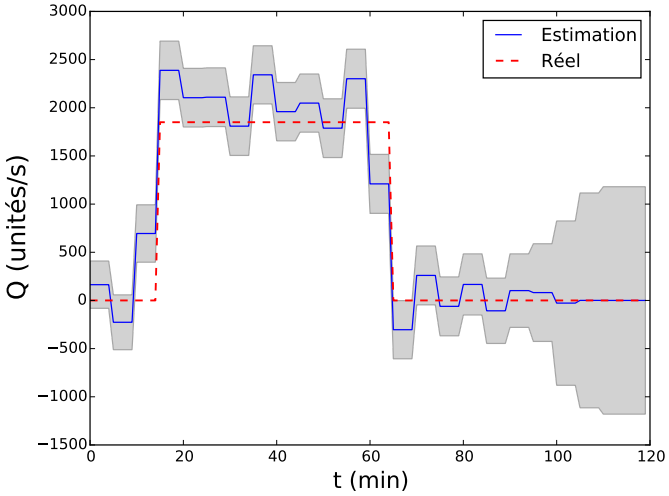
Nous étudions dans les paragraphes suivants l'influence des paramètres de variance d'observation σ_{obs}^2 et de variance a priori σ_q^2 du profil d'émission $\boldsymbol{q}$ sur la qualité de l'estimation.



(a) Position en x



(b) Position en y



(c) Profil d'émission avec intervalle de confiance à $\pm 2\widetilde{\sigma}_q^2$ (gris)

FIGURE 4.7 – Résultats du *benchmark* de l'algorithme d'estimation sur le cas-test Beaune

Une telle étude paramétrique permet en effet de donner du sens à ces paramètres, ceux-ci devant être spécifiés par l'utilisateur parmi les variables d'entrée de l'algorithme d'estimation.

4.2.2 Influence de la variance d'observation

La variance d'observation σ_{obs}^2 est le paramètre qui reflète la "confiance" donnée aux observations $\boldsymbol{\eta}$. Comme expliqué au Chapitre 3, elle caractérise l'ensemble des erreurs à l'origine de l'écart entre les valeurs issues du modèle de données et la réalité physique. Comme il a été expliqué précédemment, l'erreur d'observation est une agrégation de diverses sources individuelles d'erreur (modèle, instrumentation...). Sans chercher à mener une analyse approfondie sur la quantification des incertitudes autour des observations, nous cherchons plutôt ici à avoir une vision d'ensemble de l'influence du paramètre σ_{obs}^2 sur la qualité de l'estimation du terme source. Pour cela, on garde les mêmes paramètres que le *benchmark* de la figure 4.7, on choisit ensuite une plage de valeurs de σ_{obs}^2 à tester, et pour chacune de ces valeurs, on lance 100 *runs* de l'AMIS. Ce dernier faisant intervenir des tirages aléatoires, le fait de considérer les résultats issus d'un nombre suffisant de

runs permet de voir si la qualité des estimations n'est pas perturbée par ces aspects stochastiques.

Concernant la localisation de la source, on observe sur les figures 4.10 et 4.11 que les résultats sont réguliers : l'estimation de la position est bonne pour les valeurs inférieures à 5×10^{-6} puis se dégrade progressivement jusqu'à ce qu'une zone précise de l'espace devienne difficile à définir (4.10f et 4.11f).

Pour le profil d'estimation du débit de rejet sur la figure 4.12, celui-ci est relativement bien estimé pour $\sigma^2_{obs} = 10^{-7}$, les variations étant de plus en plus importantes au fur et à mesure que la valeur de la variance d'observation augmente. Les courbes d'erreur de la figure 4.8 donnent un aperçu de l'influence générale de la variance d'observation.

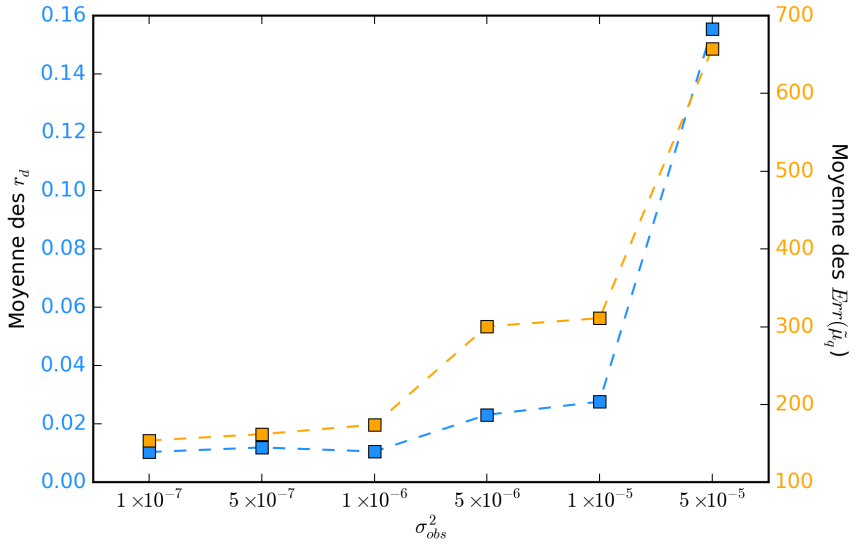


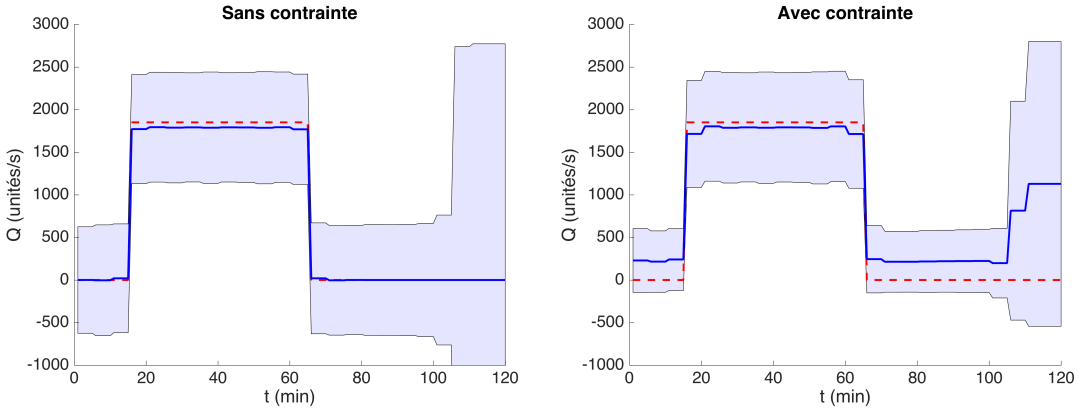
FIGURE 4.8 – Courbes d'erreur pour l'analyse paramétrique de σ^2_{obs}

Certaines des valeurs estimées pour le débit sont négatives, or il a été constaté que l'application de la contrainte de positivité a tendance à introduire un biais sur les valeurs de $\tilde{\boldsymbol{\mu}}_q$, comme le montrent les figures 3.3 et 4.9. Il a donc été décidé de ne pas implémenter cette contrainte, car malgré la présence de quelques valeurs négatives, le sens physique des résultats d'estimation est globalement respecté.

En examinant ces résultats, on pourrait considérer la valeur $\sigma^2_{obs} = 10^{-7}$ comme étant celle qui donne les meilleurs résultats. Il faut cependant rappeler que dans notre cas d'étude, nous utilisons des observations synthétiques non-bruitées, ce qui permet d'accorder une plus grande confiance à ces observations et donc de réduire la valeur de σ^2_{obs}. On ne peut pas forcément en faire autant si des sources d'incertitudes supplémentaires entrent en jeu, par exemple avec des valeurs de concentrations expérimentales issues de mesures réelles, car cela reviendrait à diminuer la marge d'incertitude sur les mesures alors que le caractère aléatoire de ces dernières est plus accentué.

4.2.3 Influence de la variance a priori du profil d'émission

La variance a priori σ^2_q peut être vue comme une hypothèse de départ sur l'amplitude possible des valeurs du profil d'émission. Son influence s'applique à la fois sur la qualité de la reconstruction



(a) (b)

FIGURE 4.9 – Comparaison de l'estimation de $\tilde{\boldsymbol{\mu}}_q$ sur une particule dans la maille de la source sans (à gauche) et avec (à droite) la contrainte de positivité

de $\boldsymbol{q}$, mais également sur celle de la localisation de la source, car σ_q^2 intervient dans l'algorithme AMIS lors du calcul de la vraisemblance des particules.

Pour cette analyse, on utilise le même procédé que pour l'étude de σ_{obs}^2, en reprenant les paramètres du *benchmark* et en exécutant 100 *runs* de l'AMIS sur différentes valeurs de σ_q^2.

Pour l'estimation de la position, on constate que plus la valeur de σ_q^2 considérée est grande, plus la source aura tendance à être estimée en amont par rapport à la direction du vent. En effet, pour la même position potentielle de la source, des valeurs plus élevées pour les quantités de polluant rejetées impliquent des mesures de concentration plus importantes aux capteurs : l'algorithme va donc ajuster cette position en remontant l'axe du vent pour avoir une meilleure vraisemblance par rapport aux observations. Ce phénomène de décalage spatial est bien visible sur la figure 4.13, où les *boxplots* des estimations ponctuelles sont tracées pour chacune des valeurs de σ_q^2 testées.

La meilleure estimation du profil d'émission est obtenue pour $\sigma_q^2 = 2 \times 10^6$ (voir figure 4.17, ce qui coïncide avec la meilleure position estimée dans la figure 4.13. Pour des valeurs plus faibles de σ_q^2 le débit est sous-estimé, à l'inverse pour des valeurs plus élevées le débit est surestimé. On peut résumer ce comportement grâce aux courbes d'erreur de la figure 4.14.

4.2.4 Influence de la densité du réseau de capteurs

Le nombre et la répartition des capteurs constituent un facteur important :

— du point de vue théorique, son étude permet de mieux comprendre le comportement de l'algorithme d'estimation et quantifier son importance par rapport aux autres variables,

— en pratique, l'analyse des résultats d'estimation en fonction de la configuration des capteurs permet un meilleur dimensionnement du réseau de mesure.

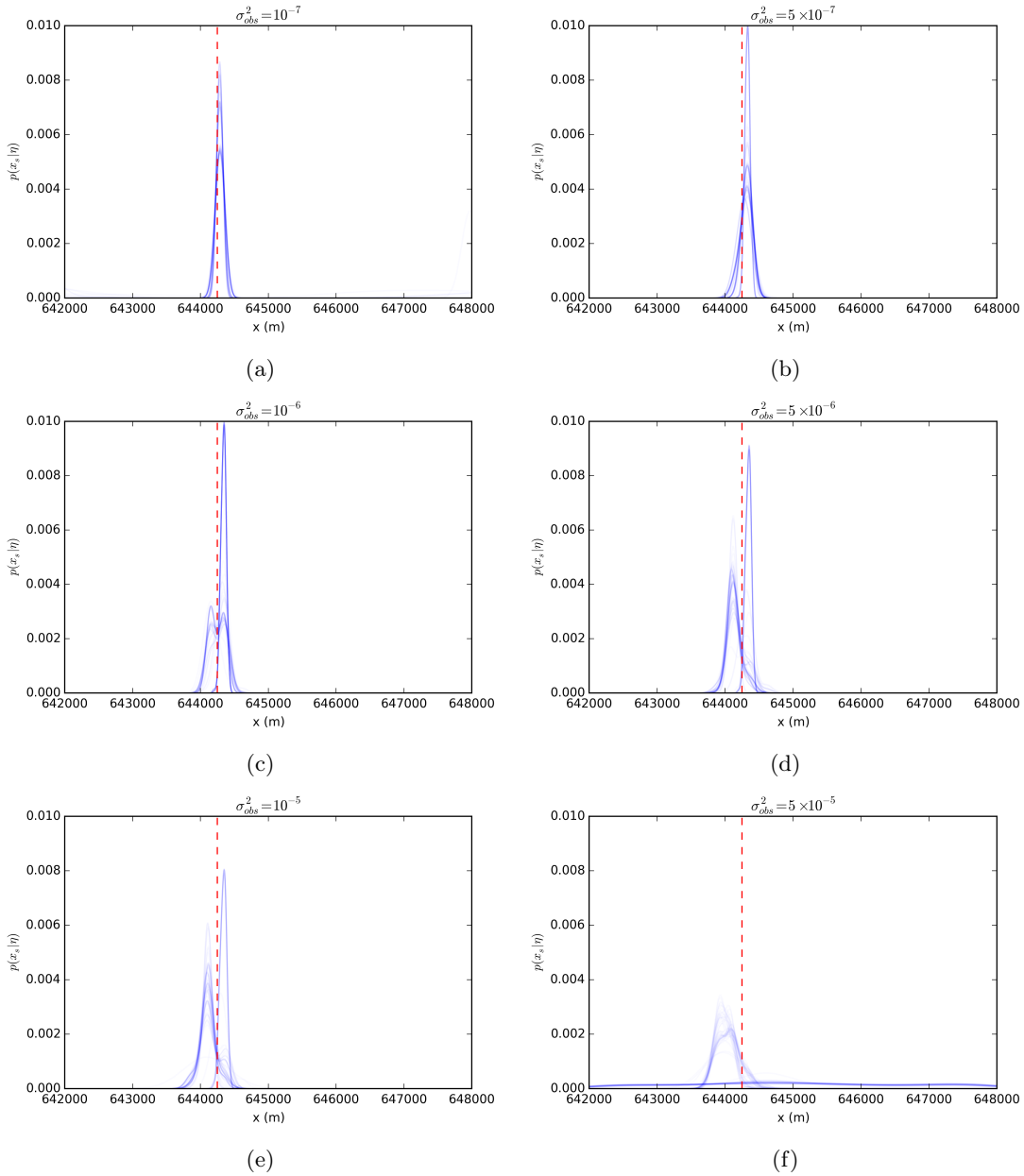


FIGURE 4.10 – Analyse paramétrique sur la variance d'observation σ^2_{obs} pour le cas-test Beaune (25 capteurs) : localisation en x de la source

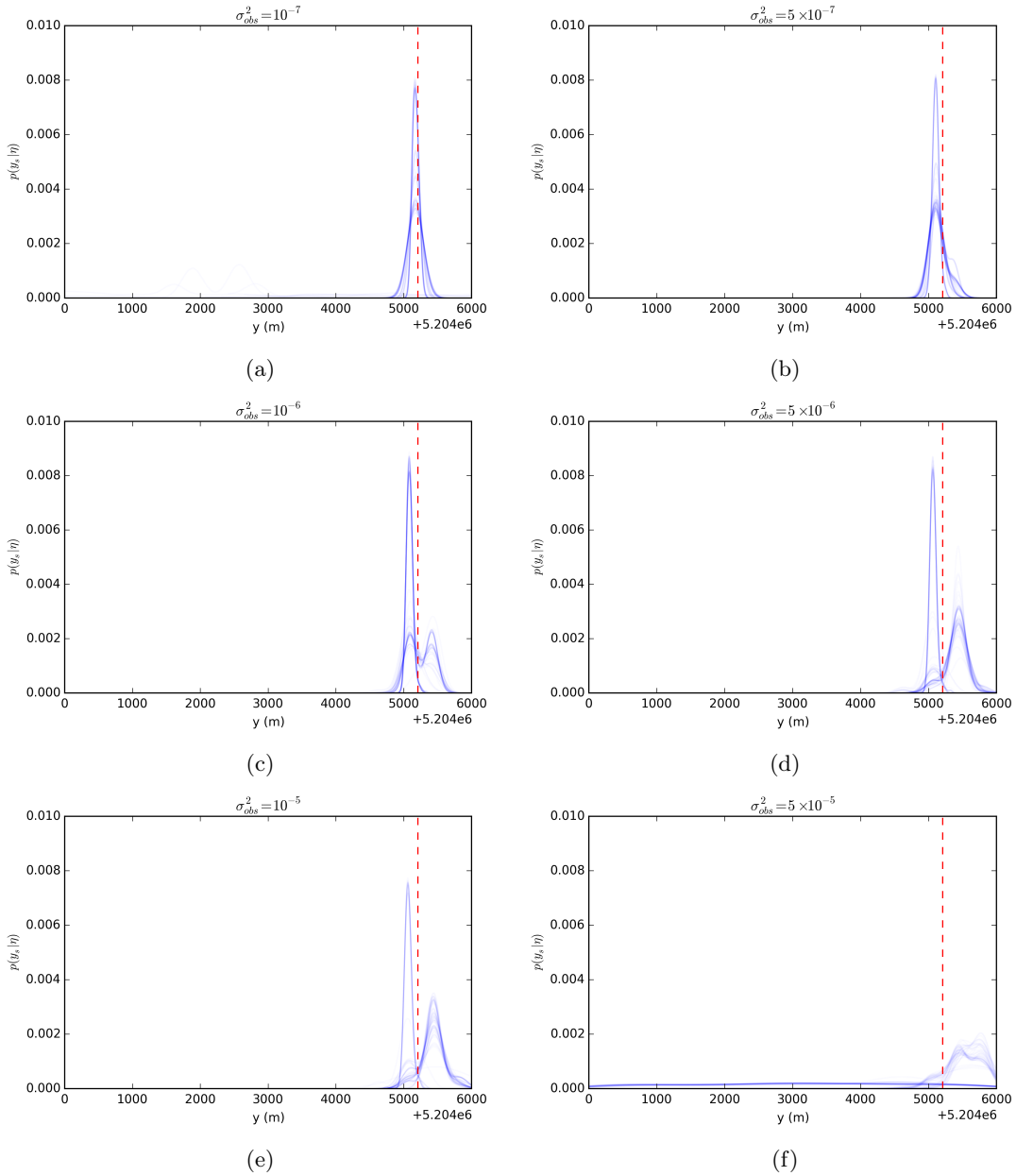


FIGURE 4.11 – Analyse paramétrique sur la variance d'observation σ^2_{obs} pour le cas-test Beaune (25 capteurs) : localisation en y de la source

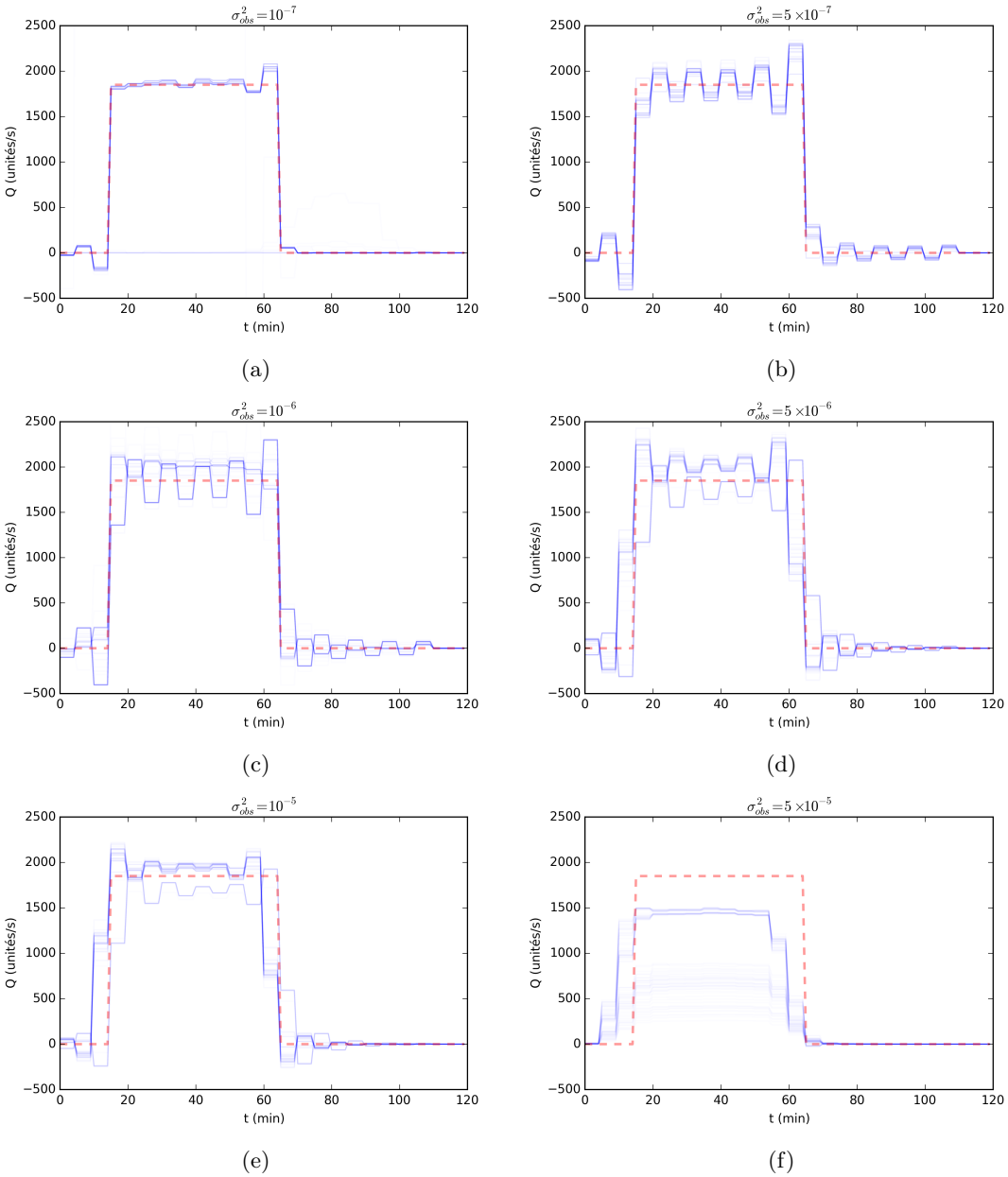


FIGURE 4.12 – Analyse paramétrique sur la variance d'observation σ^2_{obs} pour le cas-test Beaune (25 capteurs) : reconstruction du profil d'émission

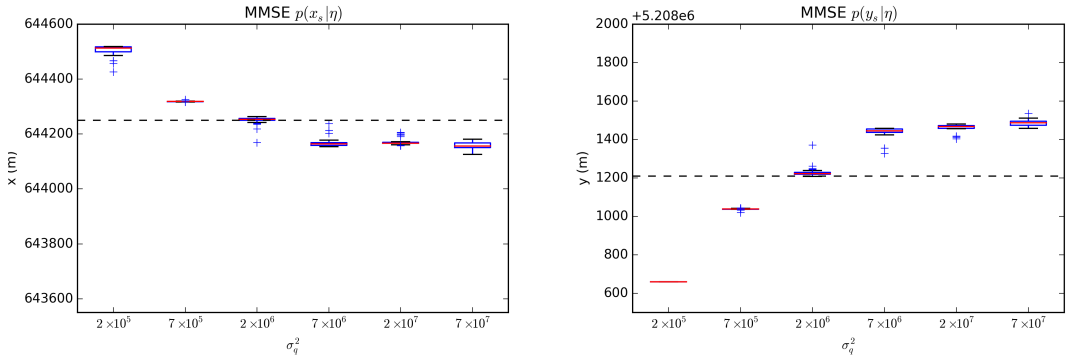


FIGURE 4.13 – *Boxplots* des estimations par MMSE de la position de la source sur 100 runs, comparaison avec les valeurs réelles (en pointillés noirs)

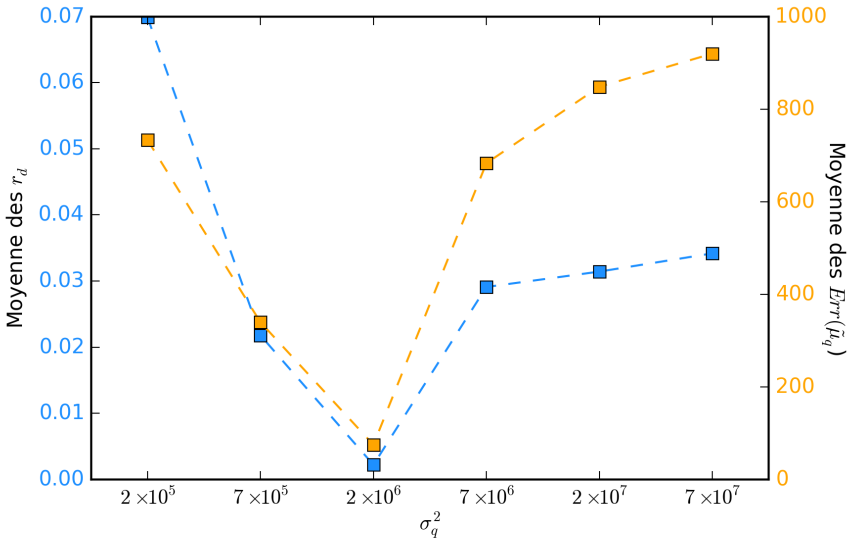


FIGURE 4.14 – Courbes d'erreur pour l'analyse paramétrique de σ_q^2

Impact du capteur R8

Dans un premier temps, nous nous intéressons à l'influence du capteur R8, qui est le récepteur le plus proche de la source à mesurer des concentrations non-nulles. Pour cela, nous le retirons du réseau, et nous effectuons l'opération d'estimation du terme source avec les observations issues des 24 capteurs restants, en suivant les paramètres du *benchmark*. En pratique, cela reviendrait par exemple à simuler la panne du capteur R8 durant la fenêtre temporelle d'observation.

A paramètres identiques, la source est vue plus en amont sur l'axe du vent par rapport à sa position réelle (figures 4.18a et 4.18b). De plus, sans le capteur R8, le débit est relativement sous-estimé par rapport aux valeurs attendues, et le rejet reconstitué commence et s'achève plus tôt que prévu (figure 4.18c). On observe ainsi que les mesures fournies par le capteur R8 apportent une information non-négligeable permettant d'accroître la précision de l'estimation, à la fois sur

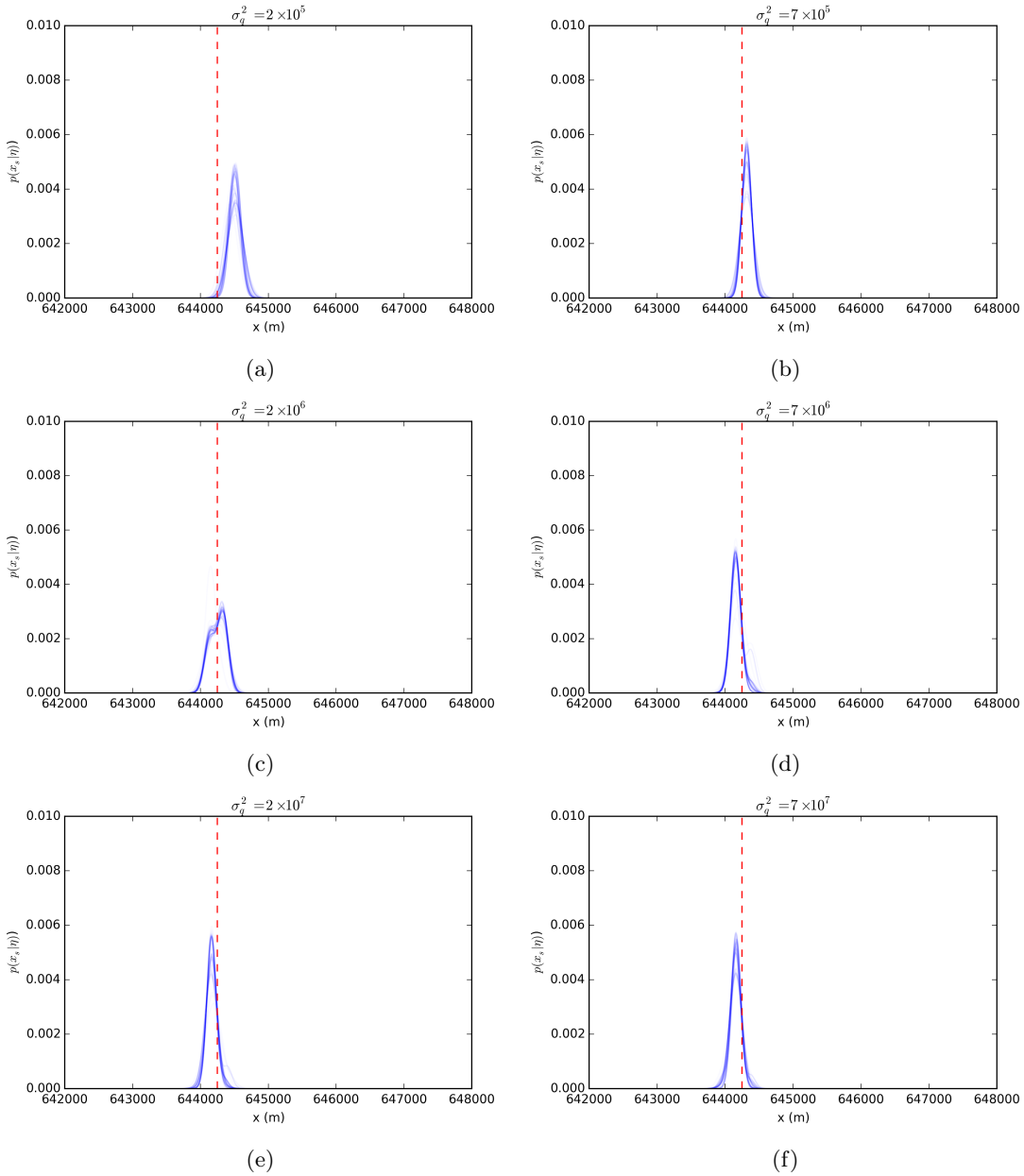


FIGURE 4.15 – Analyse paramétrique sur la variance a priori σ_q^2 pour le cas-test Beaune (25 capteurs) : localisation en x de la source

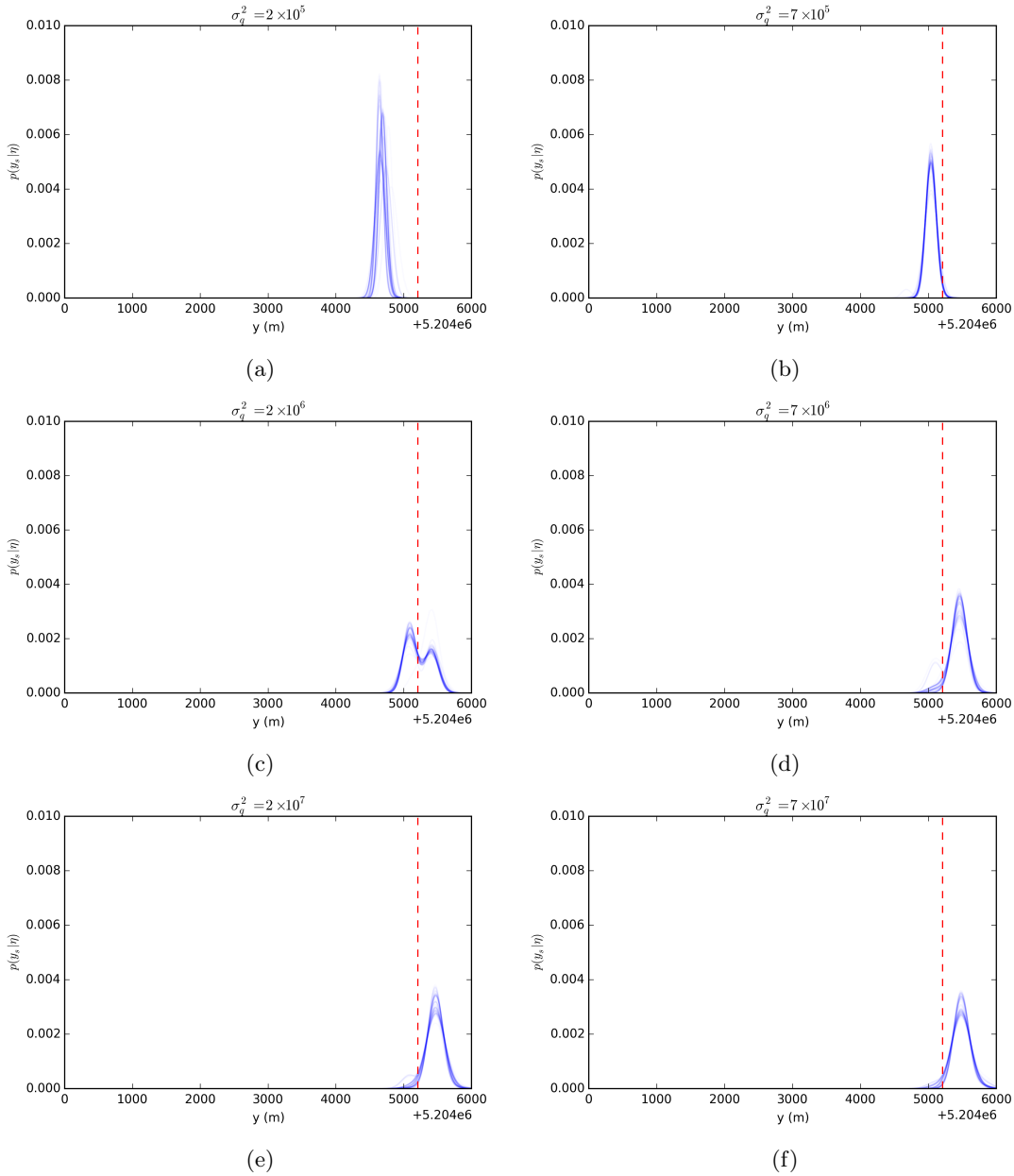


FIGURE 4.16 – Analyse paramétrique sur la variance a priori σ_q^2 pour le cas-test Beaune (25 capteurs) : localisation en y de la source

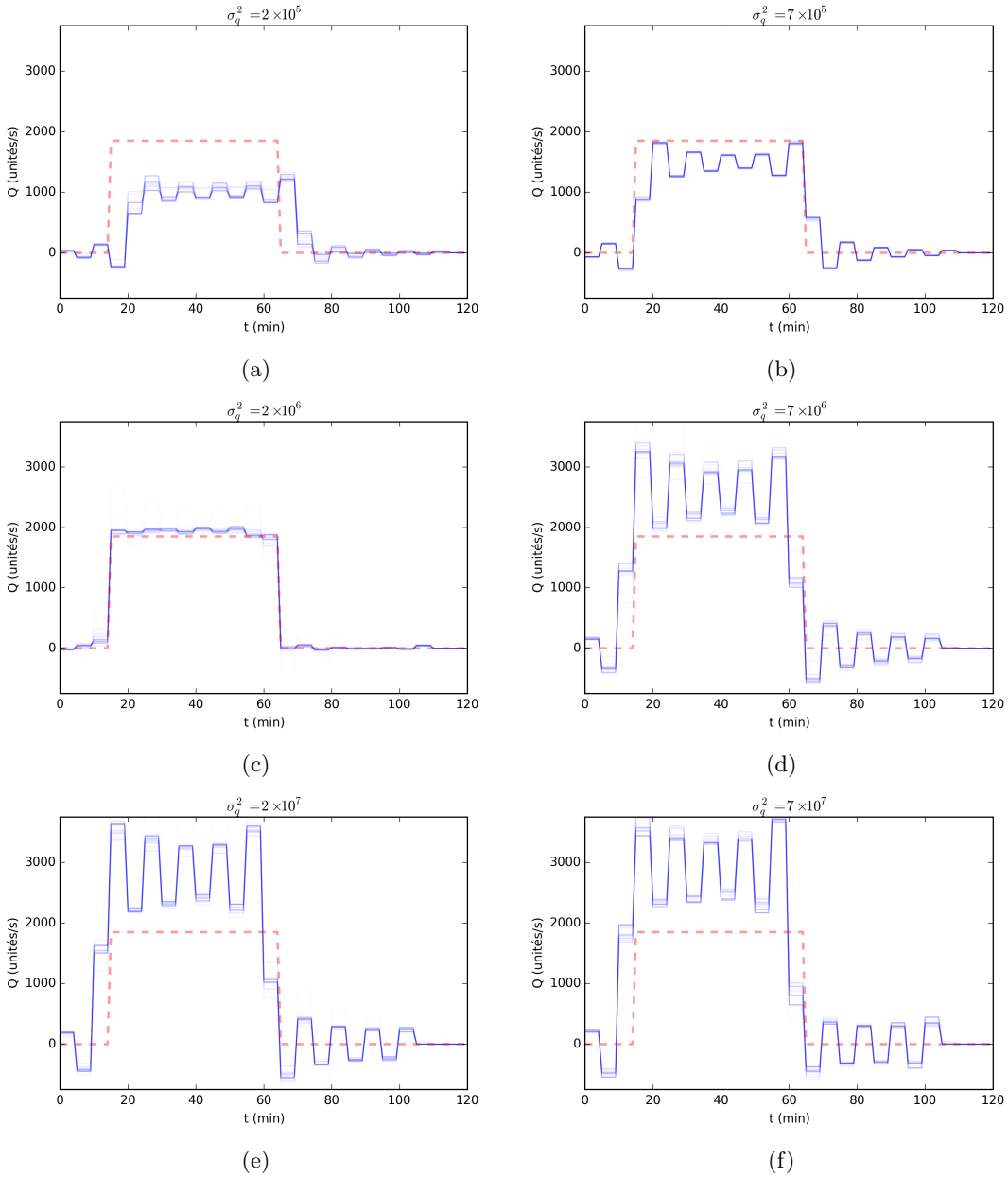
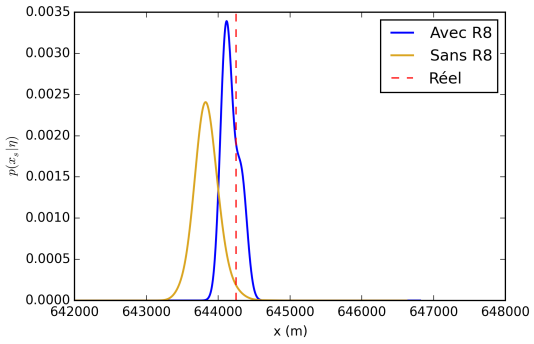
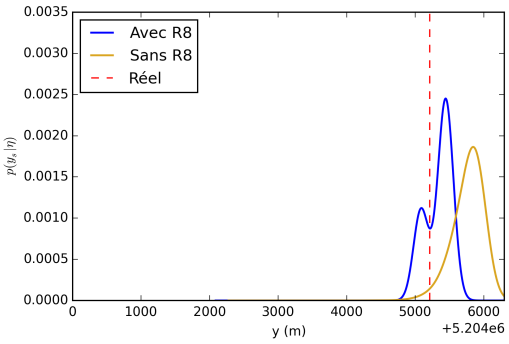


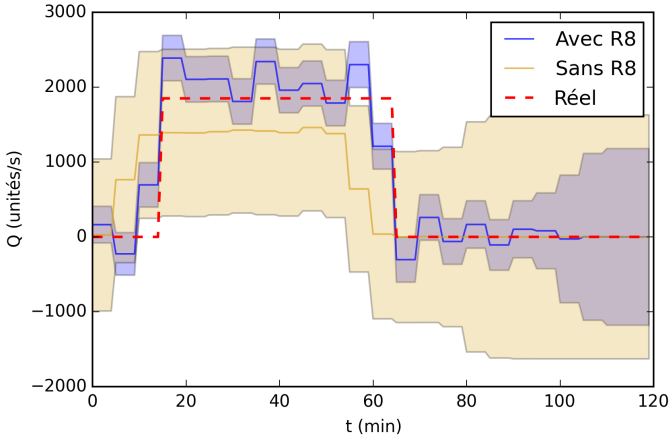
FIGURE 4.17 – Analyse paramétrique sur la variance a priori σ_q^2 pour le cas-test Beaune (25 capteurs) : reconstruction du profil d'émission



(a) Position en x



(b) Position en y



(c) Profil d'émission avec intervalle de confiance à $\pm 2\tilde{\sigma}_q^2$

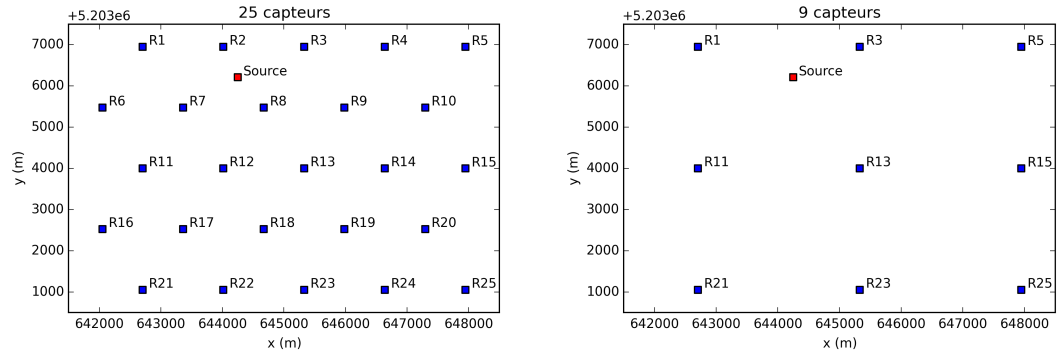
FIGURE 4.18 – Résultats d'un *run* de l'algorithme d'estimation sans (jaune) et avec (bleu) le capteur R8

les aspects spatiaux, temporels, et de quantité émise. De manière générale, on peut donc en déduire que la présence d'un capteur suffisamment proche de la source permet d'assurer une bonne reconstruction de celle-ci.

Impact d'un réseau réduit

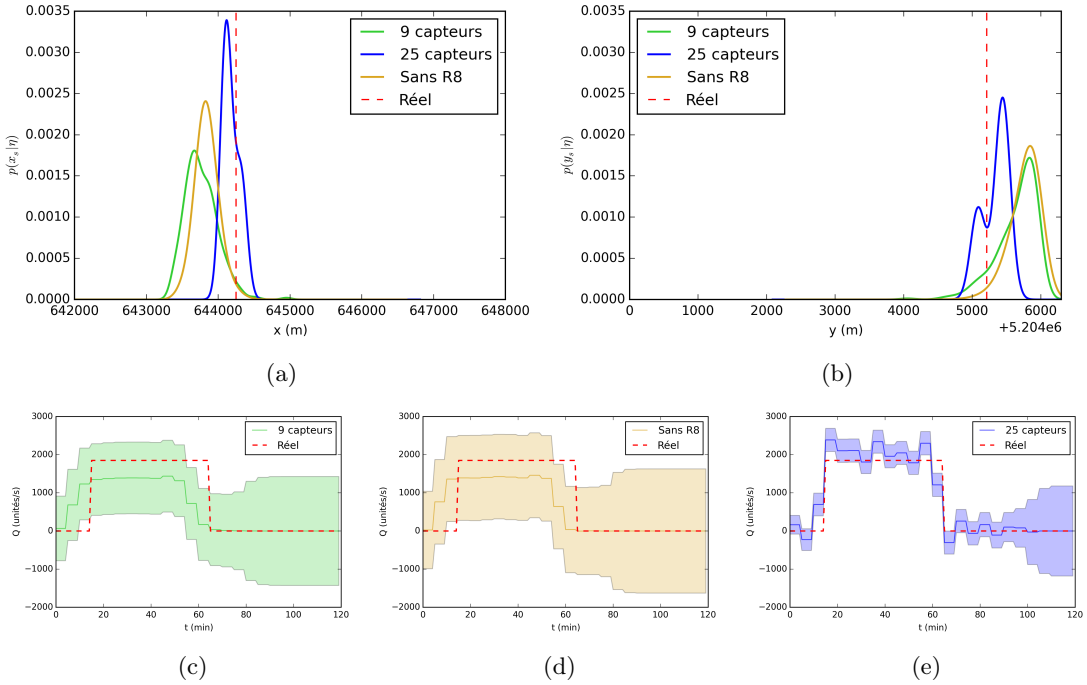
On a vu que le rôle individuel des capteurs peut être important pour pouvoir reconstruire correctement un terme source. Nous cherchons ici à savoir comment varie la qualité de cette reconstruction si l'ensemble du réseau est modifié.

On réduit ainsi la taille de notre réseau à 9 capteurs (figure 4.19), répartis de façon homogène sur le domaine. Cela permet de simuler une situation où le nombre d'instruments de mesure des concentrations est moindre (par exemple, dans le cas où les capteurs sont coûteux à acheter, à déployer et à maintenir en bon état de fonctionnement), et également d'apprécier le comportement de l'algorithme d'estimation lorsque la représentativité spatiale des mesures est limitée. Les paramètres d'entrée de l'algorithme d'estimation sont toujours ceux du *benchmark*.



(a) (b)

FIGURE 4.19 – Réduction de la densité du réseau de capteurs : passage de 25 (gauche) à 9 (droite) capteurs



(a) (b)

(c) (d) (e)

FIGURE 4.20 – Résultats d'un *run* de l'algorithme d'estimation sans (jaune) et avec (bleu) le capteur R8, comparaison avec un réseau réduit (vert)

Les résultats de localisation de la source montrent une dégradation de l'estimation de la position similaire au cas où le capteur R8 n'est pas présent (figures 4.20a et 4.20b). Il en va de même pour pour l'estimation du profil de rejet (figures 4.20c, 4.20d et 4.20e), où la marge d'incertitude est même légèrement moins grande dans la configuration à capteurs.

On peut ainsi en déduire que les capteurs n'observant aucune concentration non-nulle et qui ont été retirés pour établir la configuration à 9 capteurs ont bien un rôle informatif utile au processus d'estimation, et leur apport qualitatif par rapport à la présence du capteur R8 dans le réseau à 25 capteurs semble de même importance.

De façon générale, on peut ainsi voir que l'efficacité de l'estimation du terme source repose sur un certain nombre de paramètres d'entrée pour l'algorithme AMIS. Ceux-ci peuvent dépendre entièrement de la configuration du cas-test (nombre de capteurs et disposition sur le domaine), ou être définis arbitrairement par l'utilisateur. Dans ce dernier cas, on a constaté qu'il est difficile d'établir des critères empiriques pour choisir des valeurs par défaut à affecter aux paramètres de variance σ^2_{obs} et σ^2_q. Pour la variance d'observation, un choix qualitatif a été fait par rapport à l'allure des mesures de la figure 4.6, mais dans d'autres situations pratiques, l'évaluation de σ^2_{obs} peut se révéler plus compliquée :

— si on sait d'avance que certains capteurs du réseau sont plus fiables que d'autres et qu'on choisit d'accorder une plus grande confiance à ces capteurs, alors l'hypothèse d'une loi d'incertitude identiquement distribuée sur les éléments du vecteur d'observation n'est plus valable,

— dans un cas expérimental où on aurait choisi une variance d'observation trop faible,la plage de probabilité non-nulle sur la densité a posteriori estimée peut être réduite au point de ne plus inclure la vraie valeur du paramètre recherché.

Une alternative plausible à cette recherche difficile des paramètres initiaux consisterait à travailler sur un ensemble de scénarios possibles, et ainsi d'étudier plusieurs possibilités quant aux valeurs des variances à fournir. Cela demeure possible grâce au gain d'efficacité de calcul permis par l'approche rétrograde de l'AMIS, et permettrait d'envisager différents niveaux de confiance accordés aux observation (pour σ^2_{obs}), ainsi que plusieurs hypothèses sur le profil du rejet (pour σ^2_q).

4.3 Exemple d'application à un cas urbain

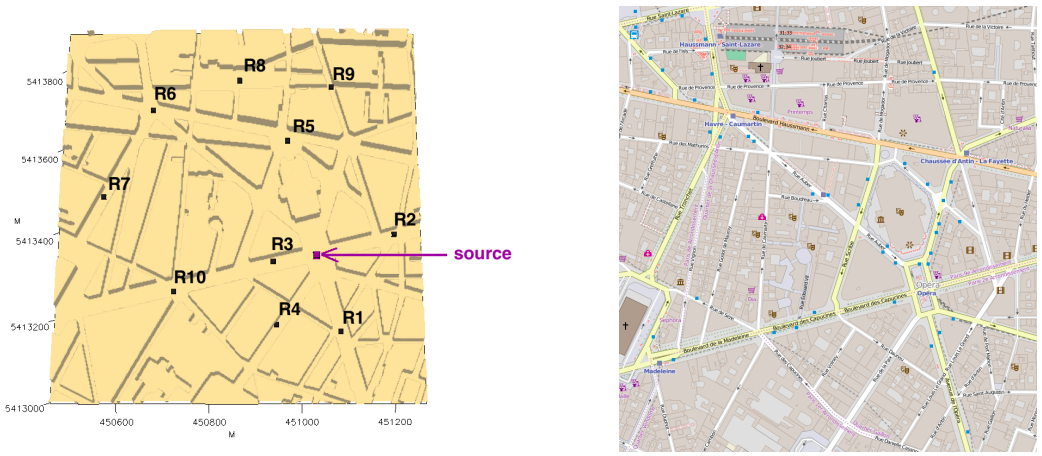
Après avoir utilisé le cas-test Beaune pour tester plusieurs aspects de l'algorithme d'estimation, nous travaillons ici dans un cadre différent, qui est celui du milieu urbain. Plusieurs nouveaux facteurs d'incertitude sont ainsi introduits, et le but de ce paragraphe est d'examiner les résultats de l'AMIS dans un contexte à forte complexité.

4.3.1 Présentation du cas-test

Nous considérons désormais un cas-test en milieu urbain caractérisé par la présence d'obstacles multiples et de géométries variées sur le domaine (bâtiments). Pour cela, on utilise une reconstitution du quartier parisien de l'Opéra (figure 4.21a).

Caractéristiques du domaine

Le domaine couvre une surface de 808 m × 882 m, avec une source unique et un réseau de 10 capteurs, disposés de façon non-régulière. Pour la simulation, le domaine est discrétisé en une grille de 404 × 441 mailles, avec une résolution du maillage en x et y de 2 m : il s'agit d'un espace d'étude plus petit que celui du cas-test Beaune, mais de résolution plus élevée, afin



(a) (b)

FIGURE 4.21 – A gauche : illustration du milieu bâti, du réseau de capteurs (en noir) et de la source (en magenta) utilisés pour le cas-test Opéra. A droite : carte OpenStreetMap du cas-test Opéra.

d'assurer une meilleure précision des calculs de dispersion en présence d'une topographie complexe.

Paramètres météorologiques

Pour ce cas-test, on choisit des paramètres météorologiques instationnaires, à savoir un vent de vitesse constante ($3\,\mathrm{m\,s^{-1}}$), mais dont la direction change toutes les heures :

Heure	11 :00	12 :00	13 :00
Direction du vent	230°	180°	45°

La combinaison de ces variations temporelles ainsi que de la présence d'obstacles sur le domaine fait que les champs de vent 3D diagnostiqués par SWIFT sont relativement complexes, comme l'illustre la figure 4.22.

Capteurs, source et simulation des observations

Le domaine contient un réseau de 10 capteurs, qui sont placés aux centres de diverses intersections de rues, ainsi que sur des places publiques. Ils sont situés à une hauteur de 2m, et délivrent des valeurs de concentrations moyennées sur des plages de 5 minutes entre 11h35 et 13h.

La source est également située à 2m du sol, positionnée au niveau d'une grande intersection, et émet un rejet bref d'une durée de 10 minutes entre 12h10 et 12h20, avec un débit constant de 10^4 unités/s. Ces caractéristiques se rapprochent de celles d'un rejet d'origine malveillante, par exemple suite à l'explosion d'une "bombe sale".

On suit le même raisonnement que pour le cas-test Beaune et on simule le vecteur d'observations à partir d'une matrice source-récepteur *backward* afin d'obtenir les mesures simulées de la figure 4.23. Les dimensions de cette source sont celles d'une maille du domaine, à savoir un volume de 2m × 2m × 2m.

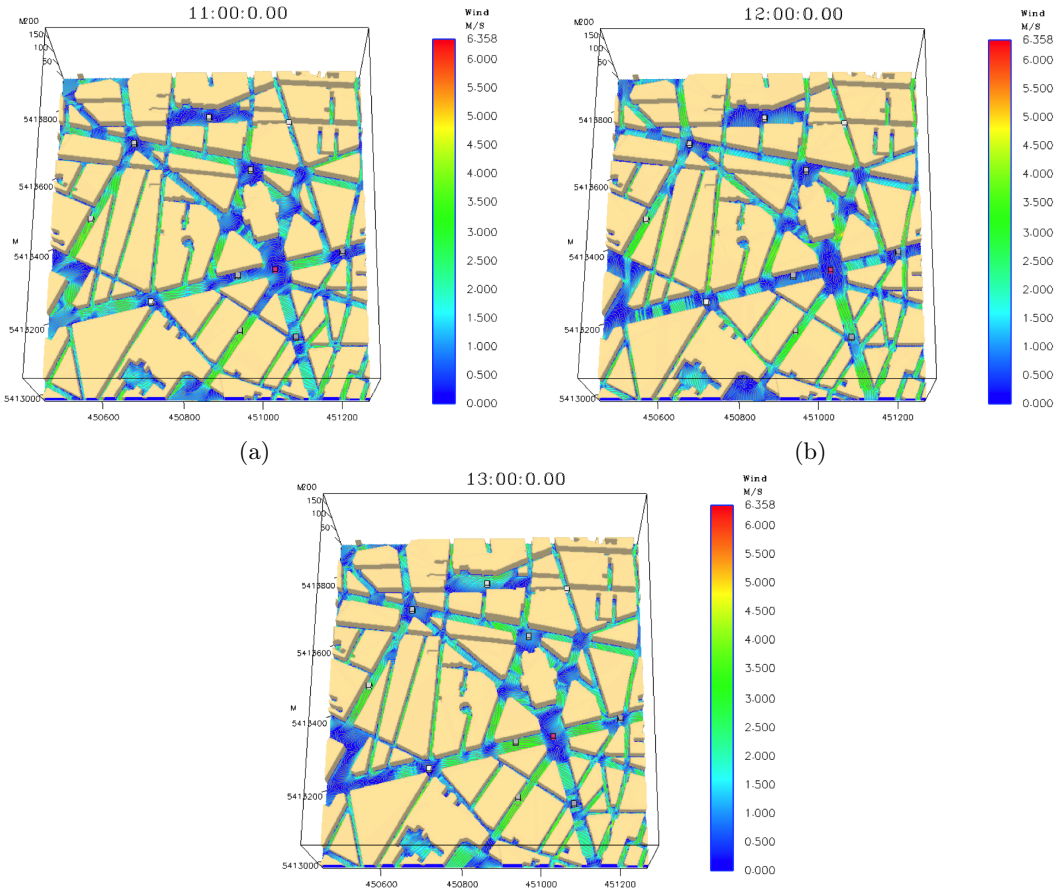


FIGURE 4.22 – Champs de vent à 2 m du sol produits par SWIFT aux trois échéances météorologiques considérées

Etant donné la taille réduite du domaine ainsi que les conditions météorologiques instationnaires auquel le cas-test est confronté, on observe que malgré un nombre total de capteurs inférieur à celui du cas-test Beaune, la proportion de récepteurs mesurant une concentration non-nulle est ici plus élevée (figure 4.23).

4.3.2 Initialisation optimisée de la loi de proposition

Avec le cas-test Opéra, il est vite apparu que sans apport préalable d'information, l'algorithme aurait des difficultés à converger vers une solution correcte étant donné la complexité de la configuration à traiter (milieu obstrué et météorologie variable). Une procédure préliminaire a donc été mise en place visant à prédéterminer une zone à favoriser pour la localisation de la source. Pour cela, le but est d'initialiser les paramètres de la loi de proposition de l'AMIS autrement que par une simple répartition uniforme des particules sur le domaine. Un ajustement de ces paramètres en amont de l'exécution de l'AMIS permettrait alors d'explorer des zones potentiellement intéressantes plus rapidement, augmentant ainsi l'efficacité de l'algorithme d'estimation.

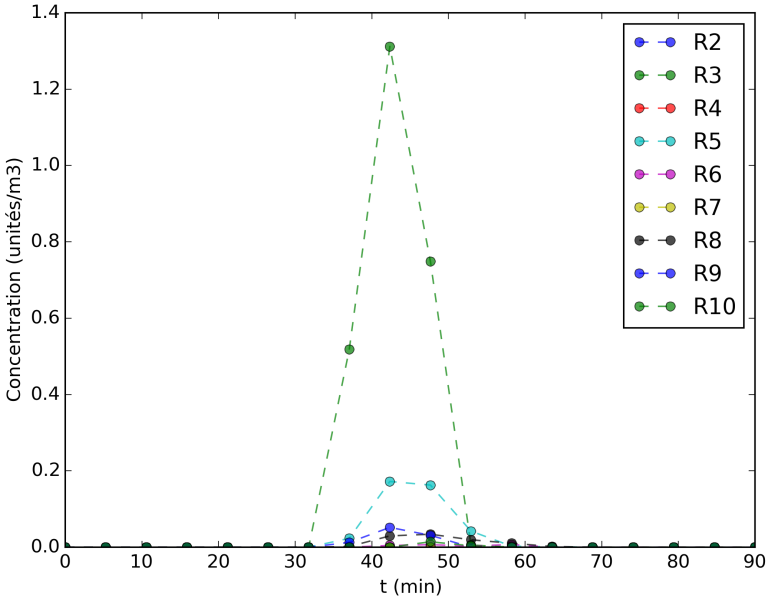


FIGURE 4.23 – Cas-test Opéra : concentrations mesurées aux capteurs

Pour cela, nous proposons d'utiliser le résultat des calculs de rétro-propagation en modélisant par une mixture de K gaussiennes la densité de probabilité sur la localisation obtenue par rétro-propagation à D/K instants choisis uniformément répartis sur un intervalle $[t_0, t_{T_s}]$, où t_0 est l'instant durant lequel est observé le maximum de concentration et t_{T_s} est l'instant final de la rétro-propagation, défini par l'utilisateur. Plus précisément, pour un instant $t_l \in [t_0, t_{T_s}]$, en notant $\boldsymbol{\theta} = (x, y)$ un point du domaine maillé suivant une grille de dimensions (N_x, N_y), on considère la distribution suivante :

$$p(\boldsymbol{\theta}|t_l, \boldsymbol{\eta}) = \sum_{i=1}^{N_x} \sum_{j=1}^{N_y} \widetilde{\omega}^l_{x_i,y_j} \; \mathcal{U}_{[x_{i-1},x_i]\times[y_{j-1},y_j]}(\boldsymbol{\theta}) \tag{4.16}$$

où :

— pour tout point (x_i, y_j) avec $1 \leq x_i \leq N_x$ et $1 \leq y_j \leq N_y$, la pondération $\widetilde{\omega}^l_{x_i,y_j}$ est définie par :

$$\begin{aligned} \omega^l_{x_i,y_j} &= \max\left(0, \sum_{k=1}^{N_c} \mathbb{1}(C^*([x_i, y_j]|R_k, t_l) \geq \varepsilon_{RP}) \times \text{sign}(\mathbb{E}(\boldsymbol{\eta}_k) - \varepsilon_{RP})\right) \\ \widetilde{\omega}^l_{x_i,y_j} &= \frac{\omega^l_{x_i,y_j}}{\sum_{i=1}^{N_x} \sum_{j=1}^{N_y} \omega^l_{x_i,y_j}} \end{aligned} \tag{4.17}$$

$\widetilde{\omega}^l_{x_i,y_j}$ peut être vu comme la probabilité que la source soit dans la zone définie par les segments $[x_{i-1}, x_i]$ et $[y_{j-1}, y_j]$ si l'instant d'émission se situait dans l'intervalle $[t_{l-1}, t_l]$. Cette grandeur est obtenue via un calcul de rétro-propagation des concentrations conjuguées $C^*([x_i, y_j]|R_k, t_l)$ pour chaque capteur R_k mesurant les plus grandes valeurs de concentra-

tions. L'équation (4.17) revient à comptabiliser le nombre de rétro-propagations en un point de l'espace et du temps si la valeur moyenne mesurée par les capteurs est supérieure à un seuil ε_{RP} prédéfini. $\boldsymbol{\eta}_k$ désigne l'ensemble des observations fournies par le capteur R_k.

— $\mathcal{U}_{[x_{i-1},x_i]\times[y_{j-1},y_j]}(\boldsymbol{\theta})$ est la loi de probabilité uniforme en deux dimensions sur la surface définie par les segments $[x_{i-1}, x_i]$ et $[y_{j-1}, y_j]$.

L'objectif de la procédure d'initialisation consiste à adapter une mixture de K gaussiennes sur la distribution (4.16), cette mixture pouvant alors s'écrire :

$$\begin{aligned} \psi_{\alpha,\nu}(\boldsymbol{\theta}) &= \sum_{k=1}^{K} \alpha_k \mathcal{N}(\boldsymbol{\theta}|\nu_k) \\ \nu_k &= (\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \end{aligned} \tag{4.18}$$

Pour cela, on choisit de minimiser la divergence de Kullback-Leibler, la procédure d'optimisation de ce critère permettant de déterminer les paramètres recherchés pour la loi de proposition. En appliquant la définition de l'équation (2.34), on peut écrire la forme explicite de cette divergence :

$$D\left(p(\boldsymbol{\theta}|t_l, \boldsymbol{\eta}) \;||\; \psi_{\alpha,\nu}(\boldsymbol{\theta})\right) = \int \log\left(\frac{p(\boldsymbol{\theta}|t_l, \boldsymbol{\eta})}{\psi_{\alpha,\nu}(\boldsymbol{\theta})}\right) \psi_{\alpha,\nu}(\boldsymbol{\theta}) d\boldsymbol{\theta} \tag{4.19}$$

Minimiser (4.19) revient de façon équivalente à maximiser le terme suivant :

$$\underset{\alpha_k,\nu_k}{\operatorname{argmax}} \int \log\left(\sum_{k=1}^{K} \alpha_k \mathcal{N}(\boldsymbol{\theta}|\nu_k)\right) p(\boldsymbol{\theta}|t_l, \boldsymbol{\eta}) d\boldsymbol{\theta} \tag{4.20}$$

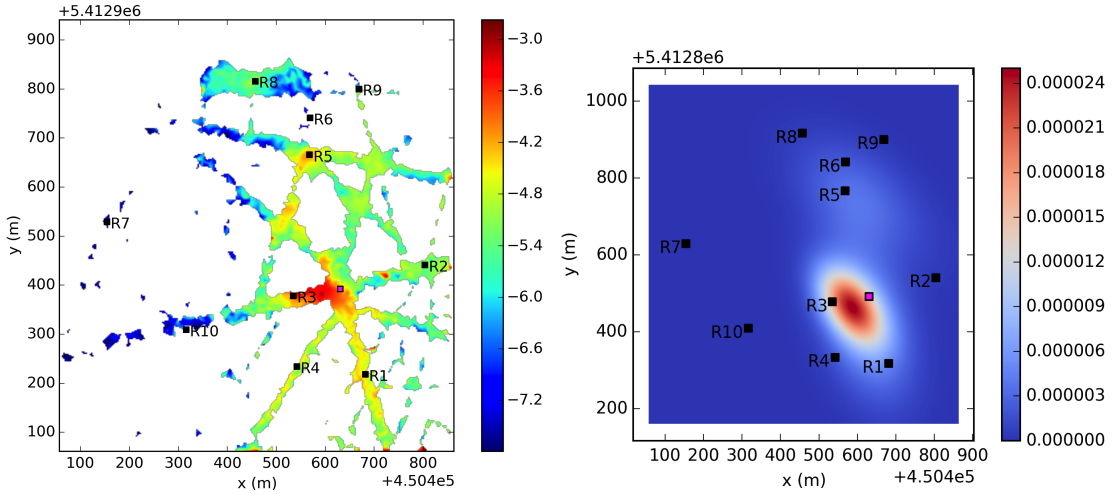
Cette maximisation ne peut pas se faire de façon analytique, mais il reste possible d'appliquer un raisonnement similaire à celui suivi pour établir les équations (2.36) à (2.39) en résolvant le problème de façon itérative à la façon d'un algorithme EM. Ainsi, à l'itération m de l'algorithme de maximisation, en définissant :

$$\rho_k(\boldsymbol{\theta}|\alpha_k^m, \nu_k^m) = \frac{\alpha_k^m \mathcal{N}(\boldsymbol{\theta}|\nu_k^m)}{\sum_{k=1}^{K} \alpha_k^m \mathcal{N}(\boldsymbol{\theta}|\nu_k^m)} \tag{4.21}$$

on peut écrire les règles de mise à jour des paramètres de la k-ième mixture de la façon suivante :

$$\begin{aligned} \alpha_k^{m+1} &= \sum_{i=1}^{N_x}\sum_{i=1}^{N_y} \widetilde{\omega}_{x_i,y_j}^{l} \rho_k(\boldsymbol{\theta}|\alpha_k^m, \nu_k^m) \\ \boldsymbol{\mu}_k^{m+1} &= \frac{\sum_{i=1}^{N_x}\sum_{i=1}^{N_y} \widetilde{\omega}_{x_i,y_j}^{l} \rho_k(\boldsymbol{\theta}|\alpha_k^m, \nu_k^m)\boldsymbol{\theta}^T}{\alpha_k^{m+1}} \\ \boldsymbol{\Sigma}_k^{m+1} &= \frac{\sum_{i=1}^{N_x}\sum_{i=1}^{N_y} \widetilde{\omega}_{x_i,y_j}^{l} \rho_k(\boldsymbol{\theta}|\alpha_k^m, \nu_k^m)(\boldsymbol{\theta} - \boldsymbol{\mu}_k^{m+1})(\boldsymbol{\theta} - \boldsymbol{\mu}_k^{m+1})^T}{\alpha_k^{m+1}} \end{aligned} \tag{4.22}$$

La figure (4.24) présente le résultat de l'initialisation avec l'ensemble des données de rétro-propagation utilisées (figure 4.24a) pour déterminer les paramètres de la loi de proposition, ainsi que la densité de probabilité résultante (figure 4.24b), qui sert donc de point de départ pour l'algorithme AMIS. On observe déjà que la zone à privilégier est bien au voisinage de la source à retrouver.



(a) (b)

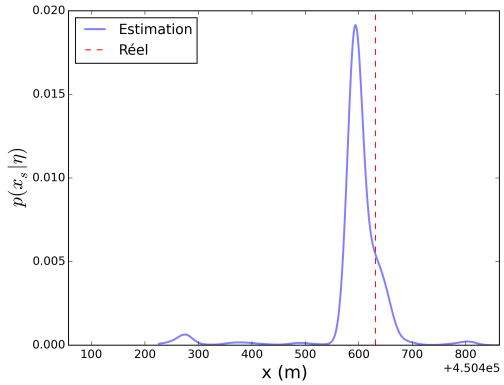
FIGURE 4.24 – Illustration de la procédure d'initialisation optimisée de la loi de proposition, avec la carte des rétro-concentrations (gauche) et la densité de probabilité suivant les paramètres initialisés (à droite)

4.3.3 Résultats

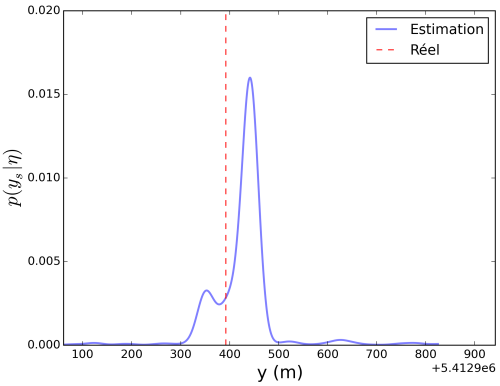
La figure 4.25 présente le résultat d'un *run* de 10 itérations de l'AMIS avec 100 particules tirées par itération, avec une initialisation optimisée et en utilisant les paramètres $\sigma^2_{obs} = 7 \times 10^{-2}$ et $\sigma^2_q = 5 \times 10^7$. Comme prévu, après l'ajustement initial de la loi de proposition, la localisation de la source est satisfaisante, de même, le profil de rejet est bien estimé. Les scores d'erreur sont de $r_d = 0.048$ pour l'estimation ponctuelle et $Err(\widetilde{\boldsymbol{\mu}}_q) = 628.974$ pour le débit.

Bien que l'initialisation optimisée ait permis une reconstruction correcte du terme source, il est à noter que la procédure d'échantillonnage depuis la loi de proposition elle-même n'est pas complètement efficace dans le cas urbain. En effet, durant cette procédure, un certain nombre de particules est tiré à l'intérieur des bâtiments, or une des hypothèses de départ stipule que la source se situe en extérieur. Pour toute particule tirée dans un bâtiment, la rétro-concentration qui lui est attribuée est automatiquement nulle, ce qui par conséquent réduit considérablement le nombre de particules pouvant avoir un poids d'importance représentatif, et réduit ainsi l'efficacité du processus d'échantillonnage d'importance.

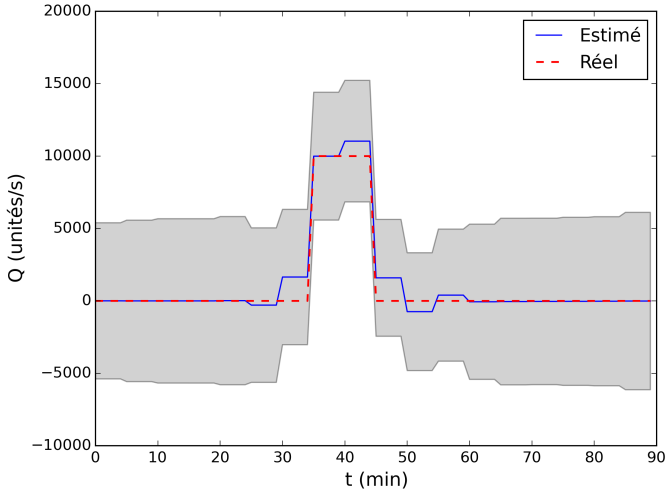
Une solution envisagée a été d'implémenter le ré-échantillonnage systématique de toute particule tirée dans un bâtiment, tant que celle-ci n'est pas en extérieur. Cependant, une telle démarche peut grandement allonger le temps de calcul, car la surface couverte par les bâtiments est supérieure à celle de la partie extérieure du domaine de simulation. De plus, l'échantillon



(a) Position en x



(b) Position en y



(c) Profil d'émission avec intervalle de confiance à $\pm 2\widetilde{\sigma}_q^2$ (gris)

FIGURE 4.25 – Résultats de l'algorithme d'estimation sur le cas-test Opéra avec une initialisation optimisée

obtenu après ces ré-échantillonnages ne représente plus mathématiquement la loi de proposition, et fausse donc le déroulement de l'algorithme d'estimation.

Une autre alternative plus simple consisterait à augmenter le nombre de particules par itération, afin d'accroître la proportion d'échantillons tirés en extérieur. Grâce au pré-calcul des rétro-concentrations, la charge supplémentaire de calcul générée par une quantité plus importante de particules reste raisonnable, bien moins élevée que si l'on avait voulu faire la même chose avec une approche utilisant un modèle direct.

Ce chapitre a permis d'illustrer les apports de l'utilisation en mode rétrograde du modèle de dispersion. La génération des matrices source-récepteur à partir de données de rétro-concentration pré-calculées a permis une utilisation optimale du système PMSS, ne nécessitant pas de multiples calculs de dispersion. Les exemples d'application sur lesquels cette approche a été utilisée ont montré des résultats corrects de l'estimation des paramètres de la source, à la fois sur sa localisation et sur la reconstruction de son débit d'émission. Le cas-test Beaune a permis

d'illustrer l'influence des paramètres σ^2_{obs} et σ^2_q, ainsi que de souligner l'importance prise par un capteur suffisamment proche de la source pour obtenir une bonne estimation. Le cas-test Opéra a confronté l'algorithme à une situation plus complexe, mais qui a pu être bien traitée grâce à l'initialisation améliorée de l'AMIS, en réutilisant les résultats de rétro-propagation afin d'isoler une zone d'intérêt sur laquelle sont ajustés les paramètres de départ de la loi de proposition.

Une question importante est toutefois soulevée par ces résultats, et concerne le réglage de σ^2_{obs} et σ^2_q. Dans ce chapitre, les valeurs par défaut choisies pour les différents cas-tests ont été obtenues par essais et erreurs successifs, mais en pratique il est nécessaire de développer une procédure mieux élaborée. Une solution serait de considérer ces hyperparamètres comme des *variables latentes*, afin de leur affecter une loi de probabilité a priori, et ainsi de considérer l'ensemble du processus d'estimation comme un *modèle hiérarchique bayésien* (voir par exemple les récents travaux présentés dans [Smidl and Hofman, 2015]).

Chapitre 5

Conclusion

Dans le cadre d'un rejet de polluant dans l'atmosphère, que celui-ci soit d'origine accidentelle ou malveillante, des outils de modélisation numérique sont utilisés pour en quantifier l'impact. La qualité des résultats issus de ces modèles dépend des variables d'entrée qui sont fournies, notamment concernant les paramètres de la source à l'origine du rejet. Or, dans la plupart des cas, l'information sur cette source n'est pas immédiatement disponible, et nécessite donc d'être estimée à l'aide de méthodes appropriées.

Plusieurs solutions pour résoudre ce problème sont abordées dans la littérature scientifique, utilisant des méthodologies variées telles que l'optimisation convexe, les algorithmes génétiques, ou l'inférence bayésienne pour retrouver les caractéristique d'une source à partir des mesures fournies par un réseau de capteurs. L'approche bayésienne aborde la question de l'estimation du terme source sous un aspect probabiliste, en cherchant à construire la densité de probabilité a posteriori des paramètres à retrouver. Un tel choix est justifié par l'incertitude qui entoure les observations fournies par les détecteurs, mais est rendu difficile à cause de l'aspect non-linéaire des phénomènes physiques qui rentrent en jeu dans l'atmosphère. Il devient alors nécessaire d'avoir recours à des méthodes de type Monte-Carlo, basées sur la simulation stochastique, et permettant d'approximer numériquement la loi a posteriori recherchée.

Dans les cas où l'optimisation de la fonction-coût s'effectue par une exploration systématique des points d'un maillage sur l'espace et le temps, on se ramène à la résolution de problèmes pouvant potentiellement être de très grande dimension selon la densité du maillage.

L'approche stochastique privilégie un parcours itératif de l'espace des paramètres, soit en suivant les états successifs d'une chaîne de Markov guidée par une procédure d'acceptation-rejet, soit en échantillonnant une population de particules qui sont ensuite pondérées suivant leur cohérence vis-à-vis des observations. Dans les deux cas, il est nécessaire d'exécuter un modèle de dispersion pour évaluer l'état ou la particule candidat(e), et un tel modèle peut se révéler coûteux en temps et en charge de calcul en fonction de son niveau de complexité.

Dans cette thèse, nous avons donc cherché à construire une chaîne de calcul basée sur une méthode bayésienne stochastique permettant d'estimer à la fois la position de la source, ainsi que son profil temporel d'émission, autrement dit l'historique des quantités rejetées au fil du temps.

5.1 Résultats

Le premier chapitre de ce manuscrit permet de définir le contexte dans lequel s'inscrit le travail de thèse, et de faire une synthèse des méthodes existantes, avec leurs avantages et leurs inconvénients respectifs.

Dans le chapitre 2, nous rappelons les principes généraux qui régissent l'inférence bayésienne, avant de souligner l'intérêt des méthodes de Monte-Carlo pour la résolution des problèmes où la formulation analytique de la loi a posteriori est impossible. Parmi ces méthodes, nous évoquons deux approches différentes : les algorithmes de type Markov Chain Monte-Carlo (MCMC) et l'échantillonnage d'importance (IS). Nous étudions ensuite plus en détail l'introduction d'une phase d'adaptation dans les méthodes IS, permettant ainsi à la loi de proposition d'évoluer au fil des itérations en fonction des particules et des poids d'importance précédemment calculés. Cet aspect adaptatif est étendu aux poids d'importance, dont la totalité est tout d'abord recyclée, puis réutilisée dans l'implémentation de l'algorithme Adaptive Multiple Importance Sampling (AMIS), qui constitue une amélioration efficace de l'IS standard.

Le chapitre 3 illustre l'application de l'algorithme AMIS au contexte de l'estimation de source, en utilisant les résultats de la campagne de mesures expérimentales FFT07. Dans la formulation mathématique du problème, nous introduisons un a priori gaussien sur les éléments du profil temporel de rejet, ce qui a permis d'en obtenir une estimation analytique, et de limiter les calculs de simulation stochastique à la partie "localisation" de la source. Dans l'implémentation de cette dernière, nous présentons un modèle de loi de proposition basé sur une mixture de gaussiennes. Le principal avantage de ce type de distribution réside dans la phase de mise à jour des paramètres, dont les équations sont directement données par l'algorithme Expectation-Maximization (EM). Un tel modèle de mélange permet également une certaine flexibilité lors de l'exploration de l'espace des paramètres, en donnant la possibilité de favoriser ou d'exclure facilement une zone prédéfinie. Les résultats d'estimation obtenus sont satisfaisants pour les exemples synthétiques et corrects pour les cas expérimentaux, mais l'exécution de la chaîne de calcul met clairement en évidence la forte dépendance de la durée de calcul avec le nombre d'appels au modèle de dispersion.

Dans le chapitre 4, nous proposons une solution à ce problème, en substituant au modèle direct une version rétrograde pour simuler la dispersion lors du calcul des matrices source-récepteur. Cette modification permet de soustraire au processus d'estimation toute la charge de calcul due au modèle de dispersion, en pré-calculant sur un maillage couvrant tout le domaine les valeurs requises pour la création des matrices source-récepteur avant de lancer l'algorithme AMIS. Cela est rendu possible grâce au principe de dualité direct-rétrograde, qui permet lors de l'estimation de la source, de remplacer les concentrations calculées pour chaque particule par des rétro-concentrations préalablement disponibles sur le domaine maillé. Une amélioration supplémentaire a été apportée à la méthodologie utilisée, à travers l'implementation d'une initialisation améliorée de la loi de proposition. Cette nouvelle procédure consiste à ajuster les paramètres d'une mixture de gaussiennes sur une carte de rétro-concentrations obtenue grâce à un calcul de rétro-dispersion.

La mise en oeuvre de l'ensemble de ces modifications a permis d'obtenir de bons résultats, à la fois dans un milieu rural et dans une configuration urbaine plus complexe, avec l'utilisation d'un

modèle de dispersion lagrangien rétrograde intégré au système Parallel Micro-SWIFT-SPRAY (PMSS). Dans le cas urbain, l'utilisation de l'initialisation améliorée a notamment mené à une bonne orientation de l'algorithme AMIS vers une zone d'intérêt restreinte pertinente dès la première itération, ce qui était impossible avec une simple initalisation uniforme. Les différentes expériences menées sur les cas-tests ont aussi illustré l'influence de la disposition du réseau de capteurs, ainsi que l'impact des valeurs prises par les variances a priori sur le débit d'émission et d'observation. Il faut toutefois noter que ces valeurs nécessitent d'être arbitrairement choisies avant de lancer l'algorithme d'estimation et il n'existe pas, dans l'état actuel des choses, de formule générique permettant de leur affecter des valeurs par défaut. Afin de pallier cette limitation, une amélioration à envisager serait d'intégrer directement ces paramètres à l'ensemble des valeurs à reconstruire par l'algorithme d'estimation.

5.2 Perspectives

En plus des résultats précédemment mentionnés, la synthèse des travaux exposés dans ce manuscrit permet de définir un certain nombre de perspectives potentielles s'inscrivant dans leur continuité.

Une suite logique du chapitre 4 serait l'application de la méthodologie AMIS à des cas réels en présence d'obstacles. Pour cela, il existe plusieurs jeux de données de référence, dont la campagne Mock Urban Setting Test (MUST) [Yee and Biltoft, 2004]. Celle-ci, menée en 2003, a consisté à mesurer, avec des appareils de haute précision, les variables météorologiques et les concentrations lors d'un rejet de gaz traceur dans un environnement où des obstacles (sous la forme de conteneurs) sont disposés de façon régulière. L'utilisation des méthodes d'estimations développées dans ce manuscrit permettrait ainsi de comparer les performances de l'AMIS à celles d'une approche de type MCMC, grâce à l'exploitation des résultats proposés par [Keats et al., 2007] sur le même cas d'application. A partir de là, on peut également envisager des tests dans le cas expérimental urbain, avec par exemple l'utilisation d'une campagne de mesures telle que Joint Urban 2003 [Allwine et al., 2004], où le rejet de gaz traceur a été effectué dans le centre de la ville d'Oklahoma City (Etats-Unis).

Un autre aspect mentionné dans ce manuscrit est celui du dimensionnement d'un réseau de capteurs. Celui-ci peut être associé à la démarche d'estimation du terme source en utilisant cette dernière comme critère d'optimalité pour l'établissement d'un ensemble de points de mesure. Cela pourrait alors constituer un prolongement intéressant des travaux de cette thèse, en analysant de façon plus approfondie l'influence de la densité des capteurs ainsi que de leur résolution temporelle sur la qualité de l'estimation obtenue. Les résultats obtenus permettraient alors :

— de concevoir entièrement un réseau approprié à une situation donnée : par exemple en définissant les zones à instrumenter dans un complexe industriel afin de détecter et d'estimer le plus rapidement possible les paramètres d'une éventuelle fuite, ou en situation opérationnelle, de placer des capteurs mobiles sur un domaine affecté par un rejet ;

— sur un réseau existant, d'améliorer le positionnement des capteurs, voire d'en ajouter dans des zones stratégiques où une meilleure instrumentation permettrait une identification plus

efficace d'un rejet potentiel.

Les méthodes de dimensionnement du réseau de mesures sont déjà exploitées dans le cadre de l'assimilation de données (voir notamment [Abida and Bocquet, 2009]) et constituent ainsi un volet d'étude complémentaire à l'estimation du terme source.

Enfin, le dernier volet prospectif concerne le passage de l'algorithme sur un mode séquentiel. En effet, dans la pratique, il est possible que les informations issues des capteurs ne soient pas immédiatement disponibles, mais deviennent progressivement accessibles au cours du temps. Dans l'état actuel des choses, l'ajout d'éléments supplémentaires au vecteur d'observation nécessite de relancer l'algorithme AMIS depuis le début, ce dernier ne pouvant traiter les observations que de façon "statique". Un moyen de mieux tenir compte d'un caractère "dynamique" de l'acquisition des mesures consisterait à substituer l'AMIS par un nouvel algorithme basé sur le modèle du *SMC sampler* introduit dans [Del Moral et al., 2006]. En suivant cette méthodologie, il est alors possible d'évaluer séquentiellement une suite de distributions de probabilités, et d'ajuster les éléments de cette suite grâce à l'utilisation d'un noyau de transition rétrograde lorsque le vecteur d'observation est mis à jour. Concrètement, cela permet de réutiliser l'information obtenue par les échantillons qui ont été tirés avant le changement du vecteur d'observation, au lieu de devoir reprendre le calcul depuis la première itération en ne tenant plus compte des résultats précédemment obtenus. Une première étude de l'apport de cette approche est présentée dans [Nguyen et al., 2016], où l'objectif est de localiser plusieurs sources dans le cas d'une propagation simple omnidirectionnelle. La transposition de ce cadre méthodologique au contexte de l'estimation du terme source en dispersion atmosphérique constitue ainsi une perspective de prolongement prometteuse des travaux de cette thèse.

Liste des publications

Articles de revues internationales

— Rajaona, H., Septier, F., Armand, P., Delignon, Y., Olry, C., Albergel, A., and Moussafir, J. (2015). An adaptive Bayesian inference algorithm to estimate the parameters of a hazardous atmospheric release. *Atmospheric Environment*, 122:748–762 [Rajaona et al., 2015]

— Nguyen, T., Septier, F., Rajaona, H., Peters, G., Nevat, I., and Delignon, Y. (2016). A Bayesian perspective on multiple source localization in wireless sensor networks. *IEEE Transactions on Signal Processing*, 64(7):1684–1699 [Nguyen et al., 2016]

Articles de conférences internationales

— Rajaona, H., Armand, P., Septier, F., Delignon, Y., Olry, C., and Moussafir, J. (2014). Estimating source term parameters through probabilistic Bayesian inference: An approach based on an adaptive multiple importance sampling algorithm. In *16th International Conference on Harmonisation within Atmospheric Dispersion Modelling for Regulatory Purposes (HARMO 16)* [Rajaona et al., 2014]

— Rajaona, H., Septier, F., Delignon, Y., Armand, P., Makke, L., Olry, C., and Albergel, A. (2016). A Bayesian approach of the source term estimate coupling retro-dispersion computations with a lagrangian particle dispersion model and the adaptive multiple importance sampling. In *17th International Conference on Harmonisation within Atmospheric Dispersion Modelling for Regulatory Purposes (HARMO 17)* [Rajaona et al., 2016]

Bibliographie

[Abida and Bocquet, 2009] Abida, R. and Bocquet, M. (2009). Targeting of observations for accidental atmospheric release monitoring. *Atmospheric Environment*, 43(40) :6312 – 6327.

[Allen et al., 2007] Allen, C., Young, G., and Haupt, S. (2007). Improving pollutant source characterization by better estimating wind direction with a genetic algorithm. *Atmospheric Environment*, 41(11) :2283 – 2289.

[Allwine et al., 2004] Allwine, K. J., Leach, M., Stockham, L., Shinn, J., Hosker, R., Bowers, J., and Pace, J. (2004). Overview of Joint Urban 2003—an atmospheric dispersion study in Oklahoma City. In *AMS Symposium on planning, nowcasting and forecasting in urban zone (on CD)*.

[Arridge, 1999] Arridge, S. (1999). Optical tomography in medical imaging. *Inverse Problems*, 15(2) :R41.

[Backus and Gilbert, 1967] Backus, G. and Gilbert, J. (1967). Numerical applications of a formalism for geophysical inverse problems. *Geophysical Journal International*, 13(1-3) :247–276.

[BARPI-5620, 2008] BARPI-5620 (2008). L'accident de Seveso : rejet à l'atmosphère de dioxines dans une usine chimique - Rapport BARPI-5620 - Ministère de l'Environnement (http ://www.aria.developpement-durable.gouv.fr/wp-content/files_mf/fd_5620_meda_seveso_1976_fr.pdf).

[BARPI-7022, 2014] BARPI-7022 (2014). L'accident de Bhopal : rejet de gaz toxiques dans une usine agrochimique - Rapport BARPI-7022 - Ministère de l'Environnement (http ://www.aria.developpement-durable.gouv.fr/wp-content/files_mf/fd7022bhopal_inde_08122014_fr_pa.pdf).

[Bocquet, 2005] Bocquet, M. (2005). Reconstruction of an atmospheric tracer source using the principle of maximum entropy. I : Theory. *Quarterly Journal of the Royal Meteorological Society*, 131(610) :2191–2208.

[Bocquet, 2008] Bocquet, M. (2008). Inverse modelling of atmospheric tracers : Non-Gaussian methods and second-order sensitivity analysis. *Nonlinear Processes in Geophysics*, 15(1) :127–143.

[Cantelli et al., 2015] Cantelli, A., D'Orta, F., Cattini, A., Sebastianelli, F., and Cedola, L. (2015). Application of a genetic algorithm for the simultaneous identification of atmospheric pollution sources. *Atmospheric Environment*, 115 :36–45.

[Cappé et al., 2004] Cappé, ., Guillin, A., Marin, J., and Robert, C. (2004). Population Monte Carlo. *Journal of Computational and Graphical Statistics*, 13(4).

[Cappé et al., 2008] Cappé, O., Douc, R., Guillin, A., Marin, J., and Robert, C. (2008). Adaptive importance sampling in general mixture classes. *Statistics and Computing*, 18(4) :447–459.

[Cervone and Franzese, 2011] Cervone, G. and Franzese, P. (2011). Non-darwinian evolution for the source detection of atmospheric releases. *Atmospheric Environment*, 41(26) :4497–4506.

[Cheng and Singh, 2008] Cheng, Y. and Singh, T. (2008). Source term estimation using convex optimization. In *11th International Conference on Information Fusion*, pages 1–8. IEEE.

[Chow et al., 2008] Chow, F., Kosovic, B., and Chan, S. (2008). Source inversion for contaminant plume dispersion in urban environments using building-resolving simulations. *Journal of Applied Meteorology and Climatology*, 47(6) :1553–1572.

[Cornuet et al., 2012] Cornuet, J., Marin, J., Mira, A., and Robert, C. (2012). Adaptive Multiple Importance Sampling. *Scandinavian Journal of Statistics*, 39(4) :798–812.

[Courtier et al., 1998] Courtier, P., Andersson, E., Heckley, W., Vasiljevic, D., Hamrud, M., Hollingsworth, A., Rabier, F., Fisher, M., and Pailleux, J. (1998). The ECMWF implementation of three-dimensional variational assimilation (3d-var), I : Formulation. *Quarterly Journal of the Royal Meteorological Society*, 124(550) :1783–1807.

[Del Moral et al., 2006] Del Moral, P., Doucet, A., and Jasra, A. (2006). Sequential Monte-Carlo samplers. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, 68(3) :411–436.

[Dembo and Rosen, 1999] Dembo, R. and Rosen, D. (1999). The practice of portfolio replication : a practical overview of forward and inverse problems. *Annals of Operations Research*, 85 :267–284.

[Douc et al., 2007] Douc, R., Guillin, A., Marin, J.-M., and Robert, C. P. (2007). Convergence of adaptive mixtures of importance sampling schemes. *Ann. Statist.*, 35(1) :420–448.

[Estevan, 2003] Estevan, M. (2003). Consequences of the Algeciras accident, and the Spanish system for the radiological surveillance and control of scrap and the products of its processing. In *Security of radioactive sources, proceedings of an international conference. Vienna Austria : IAEA*, pages 357–62.

[Flesch et al., 1995] Flesch, T. K., Wilson, J. D., and Yee, E. (1995). Backward-time lagrangian stochastic dispersion models and their application to estimate gaseous emissions. *Journal of Applied Meteorology*, 34(6) :1320–1332.

[Gelfand and Smith, 1990] Gelfand, A. and Smith, A. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85(410) :398–409.

[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*, (6) :721–741.

[Hadamard, 1902] Hadamard, J. (1902). Sur les problèmes aux dérivées partielles et leur signification physique. *Princeton University Bulletin*, 13(28) :49–52.

[Halton, 1970] Halton, J. (1970). A retrospective and prospective survey of the Monte Carlo method. *SIAM Review*, 12(1) :1–63.

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

[Haupt, 2005] Haupt, S. (2005). A demonstration of coupled receptor/dispersion modeling with a genetic algorithm. *Atmospheric Environment*, 39(37) :7181 – 7189.

[Haupt et al., 2007] Haupt, S. E., Young, G. S., and Allen, C. T. (2007). A genetic algorithm method to assimilate sensor data for a toxic contaminant release. *Journal of Computers*, 2(6) :85–93.

[Hourdin and Talagrand, 2006] Hourdin, F. and Talagrand, O. (2006). Eulerian backtracking of atmospheric tracers. i : Adjoint derivation and parametrization of subgrid-scale transport. *Quarterly Journal of the Royal Meteorological Society*, 132(615) :567–583.

[Hourdin et al., 2006] Hourdin, F., Talagrand, O., and Idelkadi, A. (2006). Eulerian backtracking of atmospheric tracers. ii : Numerical aspects. *Quarterly Journal of the Royal Meteorological Society*, 132(615) :585–603.

[IRSN, 2012] IRSN (2012). One Year Later : Initial Analyses of the Accident and its consequences. Technical report, IRSN/DG/2012-003. Paris, France : Institut de Radioprotection et de Sûreté Nucléaire.

[Issartel and Baverel, 2003] Issartel, J. and Baverel, J. (2003). Inverse transport for the verification of the Comprehensive Nuclear Test Ban Treaty. *Atmospheric Chemistry and Physics*, 3(3) :475–486.

[Issartel et al., 2007] Issartel, J., Sharan, M., and Modani, M. (2007). An inversion technique to retrieve the source of a tracer with an application to synthetic satellite measurements. In *Proceedings of the Royal Society of London : A Mathematical, Physical and Engineering Sciences*, volume 463, pages 2863–2886.

[Issartel, 2005] Issartel, J.-P. (2005). Emergence of a tracer source from air concentration measurements, a new strategy for linear assimilation. *Atmospheric Chemistry and Physics*, 5(1) :273.

[Jaynes, 1957] Jaynes, E. (1957). Information theory and statistical mechanics. *Physical Review*, 106(4).

[Kathirgamanathan et al., 2002] Kathirgamanathan, P., McKibbin, R., and McLachlan, R. (2002). Source term estimation of pollution from an instantaneous point source. *Research Letters in the Information and Mathematical Sciences*, (3) :59–67.

[Keats et al., 2007] Keats, A., Yee, E., and Lien, F. (2007). Bayesian inference for source determination with applications to a complex urban environment. *Atmospheric Environment*, 41(3) :465–479.

[Kirsch and Kress, 1988] Kirsch, A. and Kress, R. (1988). Two methods for solving the inverse acoustic scattering problem. *Inverse Problems*, 4(3) :749.

[Krysta and Bocquet, 2007] Krysta, M. and Bocquet, M. (2007). Source reconstruction of an accidental radionuclide release at European scale. *Quarterly Journal of the Royal Meteorological Society*, 133(623) :529–544.

[Kumar et al., 2015] Kumar, P., A., F. A., Singh, S. K., Ngae, P., and Turbelin, G. (2015). Reconstruction of an atmospheric tracer source in an urban-like environment. *Journal of Geophysical Research*, 120(24) :12589–12604.

[Long et al., 2010] Long, K., Haupt, S., and Young, G. (2010). Assessing sensitivity of source term estimation. *Atmospheric Environment*, 44(12) :1558–1567.

[Martinez-Camara et al., 2013] Martinez-Camara, M., Dokmanic, I., Ranieri, J., Scheibler, R., Vetterli, M., and Stohl, A. (2013). The Fukushima inverse problem. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4330–4334.

[Matthes et al., 2005] Matthes, J., Groll, L., and Keller, H. (2005). Source localization by spatially distributed electronic noses for advection and diffusion. *IEEE Transactions on Signal Processing*, 53(5) :1711–1719.

[McCormick, 1992] McCormick, N. (1992). Inverse radiative transfer problems : a review. *Nuclear Science and Engineering*, 112(3) :185–198.

[Metropolis et al., 1953] Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6) :1087–1092.

[Metropolis and Ulam, 1949] Metropolis, N. and Ulam, S. (1949). The Monte Carlo method. *Journal of The American Statistical Association*, 44(247) :335–341.

[Moussafir et al., 2014] Moussafir, J., Olry, C., Nibart, M., Albergel, A., Armand, P., Duchenne, C., Mahé, F., Thobois, L., Loaec, S., and Oldrini, O. (2014). Aircity : a very high resolution atmospheric dispersion modeling system for Paris. In *ASME 2014 4th Joint US-European Fluids Engineering Division Summer Meeting collocated with the ASME 2014 12th International Conference on Nanochannels, Microchannels, and Minichannels*, pages 1–10.

[Murphy, 2012] Murphy, K. (2012). *Machine learning : a probabilistic perspective*. MIT Press.

[Nguyen et al., 2016] Nguyen, T., Septier, F., Rajaona, H., Peters, G., Nevat, I., and Delignon, Y. (2016). A Bayesian perspective on multiple source localization in wireless sensor networks. *IEEE Transactions on Signal Processing*, 64(7) :1684–1699.

[Owen and Zhou, 2000] Owen, A. and Zhou, Y. (2000). Safe and effective importance sampling. *Journal of The American Statistical Association*, 95(449) :135–143.

[Park et al., 2003] Park, S., Park, M., and Kang, M. (2003). Super-resolution image reconstruction : a technical overview. *Signal Processing Magazine, IEEE*, 20(3) :21–36.

[Pasquill and Smith, 1983] Pasquill, F. and Smith, F. (1983). *Atmospheric Diffusion*. Wiley.

[Platt and Deriggi, 2010] Platt, N. and Deriggi, D. (2010). Comparative investigation of source term estimation algorithms using FFT07 data. In *8th Conference on Artificial Intelligence Applications to Environmental Sciences, AMS Annual Meeting*.

[Pudykiewicz, 1998] Pudykiewicz, J. A. (1998). Application of adjoint tracer transport equations for evaluating source parameters. *Atmospheric Environment*, 32(17) :3039–3050.

[Rajaona et al., 2014] Rajaona, H., Armand, P., Septier, F., Delignon, Y., Olry, C., and Moussafir, J. (2014). Estimating source term parameters through probabilistic Bayesian inference : An approach based on an adaptive multiple importance sampling algorithm. In *16th International Conference on Harmonisation within Atmospheric Dispersion Modelling for Regulatory Purposes (HARMO 16)*.

[Rajaona et al., 2015] Rajaona, H., Septier, F., Armand, P., Delignon, Y., Olry, C., Albergel, A., and Moussafir, J. (2015). An adaptive Bayesian inference algorithm to estimate the parameters of a hazardous atmospheric release. *Atmospheric Environment*, 122 :748–762.

[Rajaona et al., 2016] Rajaona, H., Septier, F., Delignon, Y., Armand, P., Makke, L., Olry, C., and Albergel, A. (2016). A Bayesian approach of the source term estimate coupling retro-dispersion computations with a lagrangian particle dispersion model and the adaptive multiple importance sampling. In *17th International Conference on Harmonisation within Atmospheric Dispersion Modelling for Regulatory Purposes (HARMO 17)*.

[Rao, 2007] Rao, K. (2007). Source estimation methods for atmospheric dispersion. *Atmospheric Environment*, 41(33) :696–6973.

[Repussard, 2006] Repussard, J. (2006). Les leçons de Tchernobyl. *Environnement, Risques & Santé*, 5(6) :441–442.

[Robert and Casella, 2004] Robert, C. and Casella, G. (2004). *Monte Carlo Statistical Methods, 2nd edition*. Springer.

[Robertson and Langner, 1998] Robertson, L. and Langner, J. (1998). Source function estimate by means of variational data assimilation applied to the etex-i tracer experiment. *Atmospheric Environment*, 32(24) :4219 – 4225.

[Rodriguez et al., 2013] Rodriguez, L., Bieringer, P., and Warner, T. (2013). Urban transport and dispersion model sensitivity to wind direction uncertainty and source location. *Atmospheric Environment*, 64 :25–39.

[Rodriguez et al., 2011] Rodriguez, L., Haupt, S., and Young, G. (2011). Impact of sensor characteristics on source characterization for dispersion modeling. *Measurement*, 44(5) :802–814.

[Rubin et al., 1988] Rubin, D. et al. (1988). Using the SIR algorithm to simulate posterior distributions. *Bayesian Statistics*, 3(1) :395–402.

[Ryall et al., 2001] Ryall, D., Derwent, R., Manning, A., Simmonds, P., and O'Doherty, S. (2001). Estimating source regions of european emissions of trace gases from observations at mace head. *Atmospheric Environment*, 35(14) :2507 – 2523.

[Saunier et al., 2013] Saunier, O., Mathieu, A., Didier, D., Tombette, M., Quélo, D., Winiarek, V., and Bocquet, M. (2013). An inverse modeling method to assess the source term of the Fukushima Nuclear Power Plant accident using gamma dose rate observations. *Atmospheric Chemistry and Physics*, 13(22) :11403–11421.

[Schulze and Turner, 1996] Schulze, R. and Turner, D. (1996). Practical guide to atmospheric dispersion. In *Texas : Trinity Consultants*.

[Senocak et al., 2008] Senocak, I., Hengartner, N., Short, B., and Daniel, W. (2008). Stochastic event reconstruction of atmospheric contaminant dispersion using Bayesian inference. *Atmospheric Environment*, 42(33) :7718–7727.

[Sharan et al., 2009] Sharan, M., Issartel, J., Singh, S., and Kumar, P. (2009). An inversion technique for the retrieval of single-point emissions from atmospheric concentration measurements. In *Proceedings of the Royal Society of London : A Mathematical, Physical and Engineering Sciences*, volume 465, pages 2069–2088.

[Simon and Simon, 2010] Simon, D. and Simon, D. (2010). Constrained Kalman filtering via density function truncation for turbofan engine health estimation. *International Journal of Systems Science*, 41(2) :159–171.

[Singh and Rani, 2014] Singh, S. and Rani, R. (2014). A least-squares inversion technique for identification of a point release : Application to Fusion Field Trials 2007. *Atmospheric Environment*, 92 :104–117.

[Smidl and Hofman, 2015] Smidl, V. and Hofman, R. (2015). Bayesian estimation of prior variance in source term determination. In *EGU General Assembly Conference Abstracts*, volume 17, page 5563.

[Sohn et al., 2002] Sohn, M. D., Reynolds, P., Singh, N., and Gadgil, A. J. (2002). Rapidly locating and characterizing pollutant releases in buildings. *Journal of the Air &amp ; Waste Management Association*, 52(12) :1422–1432.

[Sportisse, 2008] Sportisse, B. (2008). *Pollution atmosphérique : des processus à la modélisation*. Springer.

[Stockie, 2011] Stockie, J. (2011). The mathematics of atmospheric dispersion modeling. *SIAM Review*, 53(2) :349–372.

[Tarantola, 2004] Tarantola, A. (2004). *Inverse Problem Theory and Methods for Model Parameter Estimation*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA.

[Thomson, 1987] Thomson, D. (1987). Criteria for the selection of stochastic models of particle trajectories in turbulent flows. *Journal of Fluid Mechanics*, 180 :529–556.

[Tibshirani, 1996] Tibshirani, R. (1996). Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society, Series B (Methodological)*, pages 267–288.

[Tikhonov, 1963] Tikhonov, A. (1963). Solution of incorrectly formulated problems and the regularization method. In *Soviet Math. Dokl.*, volume 5, pages 1035–1038.

[Veach and Guibas, 1995] Veach, E. and Guibas, L. (1995). Optimally combining sampling techniques for monte carlo rendering. In *Proceedings of the 22nd annual Conference on Computer Graphics and Interactive Techniques*, pages 419–428.

[Wilson et al., 2009] Wilson, J., Yee, E., Ek, N., and D'Amours, R. (2009). Lagrangian simulation of wind transport in the urban environment. *Quarterly Journal of the Royal Meteorological Society*, 135 :1586–1602.

[Wilson and Sawford, 1996] Wilson, J. D. and Sawford, B. L. (1996). Review of lagrangian stochastic models for trajectories in the turbulent atmosphere. *Boundary-layer Meteorology*, 78(1-2) :191–210.

[Winiarek et al., 2012] Winiarek, V., Bocquet, M., Saunier, O., and Mathieu, A. (2012). Estimation of errors in the inverse modeling of accidental release of atmospheric pollutant : Application to the reconstruction of the cesium-137 and iodine-131 source terms from the Fukushima Daiichi power plant. *Journal of Geophysical Research : Atmospheres*, 117(D5).

[Winiarek et al., 2011] Winiarek, V., Vira, J., Bocquet, M., Sofiev, M., and Saunier, O. (2011). Towards the operational estimation of a radiological plume using data assimilation after a radiological accidental atmospheric release. *Atmospheric Environment*, 45(17) :2944 – 2955.

[Yee, 2008] Yee, E. (2008). Theory for reconstruction of an unknown number of contaminant sources using probabilistic inference. *Boundary-Layer Meteorology*, 127(3) :359–394.

[Yee and Biltoft, 2004] Yee, E. and Biltoft, C. (2004). Concentration fluctuation measurements in a plume dispersing through a regular array of obstacles. *Boundary-Layer Meteorology*, 111(3) :363–415.

[Yee et al., 2014] Yee, E., Hoffman, I., and Ungar, K. (2014). Bayesian inference for source reconstruction : a real-world application. *International Scholarly Research Notices*, 2014.