



**HAL**  
open science

# Contributions géostatistiques à la quantification d'incertitudes en géosciences

Thomas Romary

► **To cite this version:**

Thomas Romary. Contributions géostatistiques à la quantification d'incertitudes en géosciences. Statistiques [math.ST]. Sorbonne Université, 2021. tel-03615400

**HAL Id: tel-03615400**

**<https://hal.science/tel-03615400>**

Submitted on 21 Mar 2022

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Sorbonne Université

Mémoire d'Habilitation à diriger des recherches

---

# Contributions géostatistiques à la quantification d'incertitudes en géosciences

---

**Thomas Romary**

Centre de Géosciences  
Mines ParisTech, Université PSL

Soutenue le 28 septembre 2021 devant le jury composé de :

Mme Valérie Monbet	Université Rennes 1	<i>Rapporteur</i>
M. Carlo Gaetan	Université Ca' Foscari de Venise	<i>Rapporteur</i>
M. Bertrand Iooss	EDF R&D	<i>Rapporteur</i>
Mme Mickaele Le Ravalec	IFPEN	<i>Examineur</i>
M. Denis Allard	INRAE	<i>Examineur</i>
M. Gérard Biau	Sorbonne Université	<i>Examineur</i>



## Résumé

Du fait de leur nature, les données que l'on rencontre en géosciences nécessitent des méthodes statistiques adaptées. En effet, contrairement au cadre statistique classique où l'on observe des répétitions indépendantes d'une même expérience, les phénomènes d'intérêt ne sont observés qu'une seule fois. En outre, la connaissance de la physique de ces phénomènes doit être intégrée à l'analyse et la modélisation des données. Dans ce manuscrit, je décris mes contributions au développement et à l'application de méthodes statistiques en sciences de la terre et de l'environnement. Ces travaux visent à caractériser des phénomènes naturels tout en quantifiant leurs incertitudes. Ce document est constitué de cinq chapitres, les quatre premiers couvrant les principales étapes de l'analyse des données issues des géosciences.

Le géostatisticien ne contrôle généralement pas le processus d'échantillonnage. Les plans d'expérience employés sont le plus souvent établis en suivant certaines règles heuristiques et selon les contraintes physiques imposées par le milieu. On peut cependant chercher à optimiser l'emplacement des points de mesure en fonction d'objectifs précis, formalisés par une fonction d'utilité, éventuellement dans un cadre bayésien. Dans ce premier chapitre, je présente différentes applications de techniques d'optimisation d'échantillonnage spatial, d'abord pour optimiser la prédiction et l'estimation des paramètres dans le cadre de campagnes de mesures de la qualité de l'air, puis dans l'évaluation radiologique d'installations nucléaires.

Dans une étude géostatistique, on peut être amené à partitionner le domaine d'étude pour tenir compte de comportements très différents du phénomène d'étude sur le domaine et ainsi simplifier la tâche de modélisation. On peut voir la définition de cette partition comme un problème de classification non supervisée, où le domaine sera séparé en zones homogènes en fonctions des valeurs observées pour les différentes variables dont on dispose. L'application directe d'algorithmes de clustering classiques, développés pour des observations indépendantes, ne permet pas d'assurer la cohérence spatiale des classes obtenues. Dans le second chapitre, je décris deux algorithmes qui reposent sur des adaptations d'approches classiques pour la classification de données géostatistiques multivariées.

Le développement d'outils de mesure facilement accessibles et à moindre coût a engendré des jeux de données spatiales et spatio-temporelles massifs. Ceci engendre des problèmes, essentiellement de nature numérique du fait que les approches classiques ne supportent pas la montée en charge, mais crée également des opportunités dans la mesure où l'abondance de données va permettre de relâcher certaines hypothèses simplificatrices. Le troisième chapitre présente mes travaux autour de ces questions. Il s'achève par une ouverture sur la modélisation des données spatio-temporelles.

Dans de nombreux domaines des géosciences, des modèles phénoménologiques sont développés pour décrire et comprendre des processus physiques sous-jacents. Leur implémentation numérique fournit une évaluation quantitative du phénomène d'étude, qui correspond souvent à une quantité observable, au moins partiellement. Le problème qui se pose est alors de faire correspondre les sorties du modèle aux données observées. Le cadre bayésien offre un formalisme naturel permettant de lier les données aux modèles. Le problème revient à caractériser la loi conditionnelle des paramètres d'entrée du modèle sachant

les observations. Le quatrième chapitre décrit différentes applications de cette approche en ingénierie de réservoir, géophysique et hydrologie.

Le dernier chapitre présente mes perspectives de recherche à court et moyen terme en mettant en particulier le lien sur les apports des méthodes d'apprentissage automatique à la géostatistique.

## Abstract

Geoscientific data require tailored statistical methods due to their nature. The phenomena under study are observed only once instead of repetitively in the classical statistical framework. Moreover, data are to be combined with the physical knowledge that the experts of the application field may have. In this manuscript, I describe my contributions to the development and the application of (geo)statistical methods in earth and environmental sciences. These works aim at characterizing natural phenomena while quantifying their inherent uncertainties. The manuscript consists of five chapters, the four first covering most of the stages of geoscientific data analysis.

The sampling process is generally not controlled by the geostatistician. Sampling designs are usually set following heuristics and under physical constraints over the study area. It is well known however that the sampling design can be optimized so as to achieve different goals. In the first chapter, I present several applications of Bayesian optimal sampling methods to the collection of spatial data, first, to optimize prediction and parameter estimation accuracy in the context of air quality measurement surveys, then to the radiological evaluation of nuclear premises.

It is often necessary to split the domain of study to account for radically different behaviors of the natural phenomenon over the domain and to simplify the subsequent modeling step. The definition of these areas can be seen as a problem of unsupervised classification, or clustering, where the domain is separated into homogeneous areas with respect to the values taken by the observed variables. The application of classical clustering methods, designed for independent observations, does not ensure the spatial coherence of the resulting classes. In the second chapter, I describe two algorithms based on adaptations of classical algorithms to multivariate geostatistical data.

Large spatial datasets are becoming ubiquitous in environmental sciences with the inflating amount of data produced by sensors that monitor and measure the Earth system. Consequently, the geostatistical analysis of these data requires adequate methods. Richer datasets lead to more complex modeling but also prevent from using classical techniques. The challenge of handling such datasets is therefore to extract the maximum of information they contain while ensuring the numerical tractability of the inference and prediction algorithms. The third chapter presents several proposals I made to overcome these difficulties. It ends with an aperture toward the analysis of spatio-temporal data.

In geosciences, phenomenological models are developed to describe and understand underlying physical processes. Their numerical implementation provides a quantitative assessment of the phenomenon under study. Their output is often a quantity that can be measured in the real world, at least partially. These observations are valuable to improve the knowledge on and monitor the natural phenomena. The challenge lies in the incorporation of the data with the numerical models. The Bayesian paradigm is a natural mean to link the observed data to the models, that allows to formulate the question into a conditioning problem, where the unknowns are the input parameters. The fourth chapter is dedicated to applications of this approach in reservoir engineering, geophysics and hydrology.

The last chapter provides some insight into my future works in the short and medium term with an emphasis on the role of machine learning techniques in geostatistics.



# Table des matières

<b>Résumé</b>	<b>iii</b>
<b>Abstract</b>	<b>v</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Fonctions aléatoires . . . . .	1
1.2 Observations . . . . .	2
1.3 Modélisation et conditionnement . . . . .	3
1.4 Plan du manuscrit . . . . .	3
<b>2 Echantillonnage optimal</b>	<b>5</b>
2.1 Problématique . . . . .	5
2.2 Cadre général . . . . .	6
2.3 Echantillonnage optimal de fonctions aléatoires . . . . .	6
2.4 Augmentation d'un réseau existant . . . . .	9
2.5 Discussion . . . . .	11
<b>3 Clustering de données géostatistiques</b>	<b>13</b>
3.1 Problématique . . . . .	13
3.2 Modèle de mélange de fonctions aléatoires . . . . .	14
3.3 Algorithmes GHC et GSC . . . . .	15
3.4 Discussion . . . . .	17
<b>4 Contributions au krigeage</b>	<b>19</b>
4.1 Problématique . . . . .	19
4.2 Krigeage continu . . . . .	20
4.3 Covariances non-stationnaires . . . . .	22
4.4 Krigeage de grands jeux de données . . . . .	24
4.4.1 Approximations creuses . . . . .	25
4.4.2 Approximations de rang faible . . . . .	25
4.4.3 Combinaison . . . . .	27
4.4.4 Approximations de Vecchia . . . . .	30
4.4.5 Equations aux dérivées partielles stochastiques . . . . .	32
4.5 Modélisation spatio-temporelle . . . . .	33



4.6	Discussion . . . . .	34
<b>5</b>	<b>Inversion bayésienne</b>	<b>37</b>
5.1	Problématique . . . . .	37
5.2	Paramétrisation . . . . .	39
5.2.1	Analyse de sensibilité . . . . .	40
5.2.2	Entrées spatiales . . . . .	47
5.3	Méthodes MCMC . . . . .	53
5.3.1	Principe général . . . . .	53
5.3.2	Méthodes avancées . . . . .	54
5.3.3	Chaînes en interaction . . . . .	56
5.4	Assimilation de données . . . . .	59
5.5	Discussion . . . . .	62
<b>6</b>	<b>Perspectives</b>	<b>65</b>
	<b>Bibliographie</b>	<b>69</b>
	<b>Annexes</b>	<b>81</b>
6.1	Curriculum Vitae . . . . .	81
6.2	Liste de publications . . . . .	84
6.2.1	Articles . . . . .	84
6.2.2	Brevet . . . . .	85
6.2.3	Communications . . . . .	85
6.2.4	Thèse . . . . .	88
6.3	Encadrement . . . . .	89
6.4	Enseignement . . . . .	91
6.5	Sélection d'articles . . . . .	92

# Chapitre 1

## Introduction

Après une thèse entre l'IFP (maintenant IFPen) et l'Université Pierre et Marie Curie (désormais Sorbonne Université) soutenue en 2008, j'ai eu la chance d'intégrer l'équipe de géostatistique du centre de Géosciences de l'école des Mines de Paris il y a maintenant 13 ans. Durant cette période j'ai eu l'occasion d'aborder une grande variété de problèmes liés à la caractérisation des incertitudes en sciences de la terre.

Dans ce chapitre introductif, j'expose brièvement un cadre théorique général et quelques outils employés dans la suite qui me permettront de décrire les différents volets de mon activité de recherche. Il vise en particulier à donner une cohérence d'ensemble à mes travaux qui pourraient autrement sembler disparates.

Plus précisément, je décrirai dans un premier temps certains modèles de fonctions aléatoires qui sont à la base de la géostatistique. Dans un second temps, je m'attacherai à formaliser les processus d'observation des variables régionalisées qui sont fondamentaux pour l'inférence et la prédiction dans les modèles de fonctions aléatoires. Enfin, j'introduirai le plan du manuscrit.

### 1.1 Fonctions aléatoires

Bien que son cadre d'application se soit considérablement étendu depuis ses premiers développements, la géostatistique vise fondamentalement à modéliser des phénomènes naturels qui se déploient dans l'espace-temps. La première étape de la modélisation géostatistique consiste à considérer le phénomène naturel en question comme une variable régionalisée  $z(x)$ , c'est-à-dire une fonction de l'espace-temps (ou d'un espace de paramètres donné). Vient en second lieu l'hypothèse probabiliste qui revient à considérer la variable régionalisée comme une réalisation d'une fonction aléatoire  $Z(x, \omega)$  où  $x \in \mathcal{X}$  où  $\mathcal{X} \subset \mathbb{R}^d$  ( $d = 2, 3$  ou  $4$  en général) et  $\omega \in \Omega$  une variable d'un espace probabilisé  $(\Omega, \mathcal{A}, \mathcal{P})$ . Dans la suite, la dépendance en  $\omega$ , ne sera plus mentionnée pour simplifier les écritures. Cette hypothèse n'est en toute rigueur pas contingente au phénomène en question : si le système physique qui le gouverne est parfaitement déterminé, la variable régionalisée peut-être connue en tout point de l'espace temps et sa modélisation par une fonction aléatoire est alors superflue. En pratique cependant, on est amené à modéliser des phénomènes dont la

complexité physique interdit une modélisation déterministe fidèle, de surcroît sur la base de données parcellaires. L'hypothèse probabiliste intervient alors de manière à appréhender la complexité et l'indétermination du phénomène et autorise en outre l'usage d'outils de modélisation permettant d'effectuer, in fine, des prédictions en tout point de l'espace temps à partir de données parcellaires.

Une fois admise l'hypothèse probabiliste, on est amené à choisir un modèle de fonction aléatoire adapté au problème considéré. La plupart des modèles utilisés en géostatistique font l'hypothèse de l'existence d'un moment d'ordre 2 fini et donc d'une fonction de covariance sur l'espace  $\mathcal{X} \times \mathcal{X}$ . Ce sont ainsi les deux premiers moments qui vont conférer sa structure à la fonction aléatoire. Enfin, on peut considérer qu'un vecteur aléatoire est une fonction aléatoire indexée sur un espace discret, ce qui permet d'englober l'ensemble des cas considérés dans ce manuscrit.

## 1.2 Observations

La nature des données que l'on peut être amené à traiter en géostatistique est extrêmement diverse : teneur en métal, température, vitesse du vent, concentration en polluant, etc. Elles sont obtenues à l'aide de capteurs adaptés au phénomène étudié. Bien que les données soient souvent considérées comme ponctuelles (en environnement notamment), les mesures sont généralement effectuées sur un certain élément de volume. En mine, les données de teneur sont issues de moyennes réalisées sur des sections de carottes, les capteurs de pollution atmosphérique réalisent des mesures sur un petit volume d'air, etc. Ce que l'on considère comme un point de données, correspond en réalité à une moyenne locale et il convient donc d'en tenir compte, si possible.

Lors de l'observation de données dites ponctuelles, la fonction aléatoire  $Z(x)$ ,  $x \in \mathcal{X}$  est mesurée en un nombre  $n$  de points de l'espace. On peut ainsi écrire l'opérateur d'observation  $\mu$  comme une mesure discrète sur le domaine  $\mathcal{X}$  :

$$\mu = \sum_{i=1}^n \delta_{x_i}$$

où  $\delta(x)$  est un Dirac au point  $x$  et  $(x_1, \dots, x_n) \in X^n$ . Pour tenir compte du support et du procédé d'observation, on peut plus généralement utiliser une mesure du type :

$$\mu = \sum_{i=1}^n f_{x_i}(x) dx,$$

où  $dx$  représente la mesure de Lebesgue et  $f$  la fonction d'observation, dont le support est généralement compact et centré en  $x_i$ , soit le point d'observation. L'échantillon de données  $Z_n = (Z(x_1), \dots, Z(x_n))$  s'écrit ainsi sous une forme intégrale :

$$Z_n = \int_{\mathcal{X}} Z(x) d\mu(x)$$

En pratique, on ne connaît que rarement la fonction  $f$ , on considère donc des Dirac en les points d'observations et l'erreur commise est incluse dans une erreur de mesure additive, la plupart du temps modélisée par un bruit non structuré. Dans le cas où  $f$  est connue, certains procédés de déconvolution peuvent être employés pour revenir à  $Z$ , voir par exemple Lajaunie et al. [2020].

Il arrive également que le phénomène étudié soit mesuré de façon indirecte au travers d'une fonction complexe de la variable régionalisée. C'est par exemple le cas en ingénierie de réservoir lorsqu'on cherche à caractériser le champ de perméabilité ou de porosité à partir des données de production ou bien en géophysique où l'on cherche à déterminer la valeur du champ de vitesse en tout point du domaine à partir de données issues d'expériences sismiques. Ces sujets seront abordés au chapitre 5.

Les observations dites indirectes sont obtenues par l'application d'une fonction  $F$  à une fonction aléatoire  $Z(x)$ ,  $x \in \mathcal{X}$ . Cette fonction  $F$  est approximée par une résolution numérique du phénomène d'intérêt, imparfaitement connu, appliquée à une discrétisation de  $Z$ . Dans ce cas, on observera une fonction aléatoire  $Y = F(Z) + \varepsilon$ , qui prendra la plupart du temps une forme discrète, où l'erreur  $\varepsilon$  intégrera non seulement l'erreur de mesure mais également l'erreur de modélisation du phénomène.

### 1.3 Modélisation et conditionnement

Dans la plupart des problèmes que j'ai eu à traiter, la résolution s'effectue en deux étapes principales : la modélisation de la fonction aléatoire d'intérêt et son conditionnement aux données disponibles.

La modélisation est cruciale au sens où une erreur majeure à cette étape obliterera les résultats subséquents. Il s'agit d'intégrer l'ensemble des informations disponibles dans la spécification de la fonction aléatoire d'intérêt. Outre les données d'observation, ces informations sont généralement constituées de connaissances sur la physique du phénomène ou encore d'avis d'expert du domaine. Elles vont permettre par exemple de choisir une classe paramétrique de fonctions de covariance, d'identifier un ensemble de variables prédictives ou encore de spécifier une loi a priori sur une variable d'intérêt.

Une fois un modèle arrêté pour la fonction aléatoire d'intérêt  $Z$ , l'étape de conditionnement consiste à caractériser l'impact des données d'observation sur celle-ci. En termes probabilistes, on souhaite idéalement déterminer la loi conditionnelle  $Z|Z_n$  ou  $Z|Y$  selon les cas, de manière à fournir non seulement une prédiction par l'espérance conditionnelle ou bien le maximum a posteriori mais aussi une quantification des incertitudes afférentes, par exemple sous la forme d'un intervalle de confiance.

### 1.4 Plan du manuscrit

Ce manuscrit est constitué de quatre chapitres de longueurs inégales. Le déroulé suit les différentes étapes intervenant dans la pratique des méthodes géostatistiques. Le premier chapitre traite ainsi du recueil de la donnée au travers de la problématique de l'échantillonnage optimal dans un contexte spatial. Le second chapitre a trait à l'analyse de données

où j'aborderai le clustering spatial où l'on cherche à partitionner les données en classes qui devront respecter certaines contraintes spatiales. Le second décrit mes contributions au problème fondamental de la prédiction spatiale. Dans le dernier chapitre, j'exposerai mes différentes activités en lien avec les problèmes inverses.

## Chapitre 2

# Echantillonnage optimal

### 2.1 Problématique

Les données disponibles pour une analyse géostatistique proviennent de campagnes de mesures où le ou les phénomènes que l'on souhaite étudier sont échantillonnés. Dans certains cas de figure, le coût de la mesure peut être élevé, il convient alors d'optimiser le placement de ces points de mesure selon les moyens à disposition.

La problématique de l'échantillonnage optimal consiste à placer les points de mesure de manière à maximiser un certain critère de qualité. Il peut s'agir de minimiser la variance de prédiction sur le domaine d'étude, d'estimer les paramètres du modèle de fonction aléatoire de manière optimale ou encore d'identifier au mieux un ensemble d'excursion.

Deux principaux cas de figure se présentent : il peut s'agir de placer un réseau de capteurs sur une zone non reconnue, c'est-à-dire sur laquelle on ne dispose d'aucune donnée, ou bien d'accroître un réseau préexistant. Dans le second cas, les données disponibles permettent en principe d'inférer un modèle de fonction aléatoire, à partir duquel on réalisera l'optimisation du critère. Dans le premier cas, il sera utile de faire l'hypothèse d'une modélisation a priori, à partir des informations disponibles, par exemple sur des mesures du phénomène d'intérêt obtenues dans un environnement similaire à celui d'étude. On se placera ainsi dans un cadre bayésien.

Ce chapitre synthétise les travaux publiés dans les articles suivants :

- **T. Romary**, L. Malherbe, and C. de Fouquet. *Optimal spatial design for air quality measurement surveys*. *Environmetrics*, 25(1) :16–28, 2014.
- A. Bechler, **T. Romary**, N. Jeannée, and Y. Desnoyers. *Geostatistical sampling optimization of contaminated facilities*. *Stochastic Environmental Research and Risk Assessment*, 27(8) :1967–1974, 2013.
- **T. Romary**, C. de Fouquet, and L. Malherbe. *Sampling design for air quality measurement surveys : An optimization approach*. *Atmospheric Environment*, 45(21) : 3613 – 3620, 2011.

## 2.2 Cadre général

Je décris ici brièvement le principe général de l'échantillonnage (ou planification d'expérience) optimal(e) dans le cadre bayésien ([Chaloner and Verdinelli, 1995]). Le problème se formalise en la maximisation de l'espérance d'une fonction d'utilité définie en fonction d'une modélisation statistique donnée. On optimise cette quantité en fonction du plan  $\eta \in H$ , où  $H$ , continu ou discret, contiendra l'ensemble des plans possibles. L'expérience est définie par un modèle  $p_\eta(Z|\theta)$ , c'est-à-dire la loi des données observables  $Z \in \mathcal{Z}$  conditionnellement aux paramètres  $\theta \in \Theta$ , et dépendant éventuellement du plan  $\eta$ . Le modèle est complété par une loi a priori sur le paramètre  $p(\theta)$ . L'utilité est une fonction, par exemple le coût quadratique négatif  $u = -\{\theta - \mathbb{E}_\eta(\theta|Z)\}^2$ , ou bien une quantité spécifique au problème comme  $u =$  nombre total de patients traités, etc. Puisque le plan d'échantillonnage est choisi avant que l'expérience ne soit menée, il faut maximiser l'espérance de  $u(\cdot)$ . Formellement, le problème se pose de la manière suivante :

$$\eta^* = \arg \max_{\eta \in H} U(\eta), \text{ où } U(\eta) = \int_{\mathcal{Z}} \int_{\Theta} u(\eta, \theta, z) p_\eta(z|\theta) p(\theta) d\theta dz, \quad (2.1)$$

$U(\eta)$  est l'espérance de l'utilité pour le plan  $\eta$ . Autrement dit, l'approche bayésienne suggère l'emploi d'une fonction d'utilité qui reflète l'objet de l'expérience, en liant le choix du plan à un problème de décision.

Il convient de noter qu'à moins que le modèle  $p_\eta(Z|\theta)$  et l'a priori  $p(\theta)$  ne soient choisis de manière à rendre les calculs explicites, la maximisation de  $U$  passe par des méthodes numériques.

## 2.3 Echantillonnage optimal de fonctions aléatoires

L'objectif d'une étude géostatistique est généralement de procéder à la prédiction du phénomène d'intérêt, sous la forme d'une cartographie (support ponctuel) ou sur tout autre support. On pense alors naturellement à construire une fonction d'utilité faisant intervenir une mesure de la qualité de la prédiction. Celle-ci nous est donnée par la variance de krigeage de la quantité prédite lorsqu'on modélise le phénomène par une fonction aléatoire d'ordre 2. On choisit souvent une modélisation en dérive externe :

$$Z(x) = f(x)^t \beta + S(x), \quad (2.2)$$

où  $Z$  représente la variable d'intérêt,  $x \in \mathcal{X}$  est la coordonnée spatiale,  $f$  est un vecteur de variables auxiliaires,  $\beta$  un vecteur de paramètres,  $S(x)$  une fonction aléatoire centrée d'ordre 2 que l'on supposera stationnaire.

Etant donné un  $n$ -échantillon  $z = (z(x_1), \dots, z(x_n))$  et la matrice  $F$  des variables auxiliaires associées, on peut prédire  $Z$  en un point  $x_0$  par krigeage universel :

$$\widehat{Z}(x_0) = (C_0 + F(F'C^{-1}F)^{-1}(F_0 - F'C^{-1}C_0))' C^{-1}z, \quad (2.3)$$

où  $C$  est la matrice de covariance associée de  $z$ ,  $C_0$  est le vecteur des covariances des résidus au points de prédiction et des observations. A cet estimateur est associée une variance de prédiction donnée par la formule suivante :

$$\mathbb{V}(Z(x_0) - \widehat{Z}(x_0)) = C(0) - C'_0 C^{-1} C_0 + (F_0 - F' C^{-1} C_0)' (F' C^{-1} F)^{-1} (F_0 - F' C^{-1} C_0), \quad (2.4)$$

où  $C(0)$  est la variance ponctuelle de  $Z$ . On peut remarquer que (2.4) prend en compte à la fois la variance de l'erreur de prédiction du résidu (deux premiers termes) et la variance d'estimation de la dérive (terme de droite). Elle est en outre indépendante des données, lorsque les paramètres de la fonction de covariance sont fixés. Elle dépend du plan d'échantillonnage  $\eta$  à travers  $C_0$ ,  $C$  et  $F$ . Minimiser cette grandeur intégrée sur le domaine d'étude  $\mathcal{X}$  permet d'obtenir un plan réalisant un bon compromis entre la couverture du domaine spatial et des variables auxiliaires. Il s'agit ainsi de minimiser la variance moyenne sur  $\mathcal{X}$  :

$$O(\eta) = \frac{1}{|\mathcal{X}|} \int_{\mathcal{X}} \mathbb{V}(Z(x) - \widehat{Z}(x)) dx, \quad (2.5)$$

où  $|\mathcal{X}|$  est la surface de  $|\mathcal{X}|$ . On peut aussi considérer une version pondérée de cette quantité :

$$O(\eta) = \frac{1}{|\mathcal{X}|} \int_{\mathcal{X}} \mathbb{V}(Z(x) - \widehat{Z}(x)) w(x) dx, \quad (2.6)$$

où les  $w(x)$  sont des poids. Ces derniers permettent de se prémunir d'indésirables effets de bord (la variance est mécaniquement plus grande aux abords des limites du domaine car les points  $y$  ont moins de voisins) ou d'échantillonner plus précisément certaines zones, par exemple où l'on s'attend à rencontrer des valeurs élevées. C'est le cas notamment en qualité de l'air où l'on cherche à déterminer si la concentration étudiée dépasse ou non certains seuils réglementaires. Il convient ensuite de tenir compte de cette pondération lors de l'estimation des paramètres du modèle afin d'éviter des effets d'échantillonnage préférentiel ([Diggle et al., 2010]).

Il est intéressant de noter que le critère (2.6) est proportionnel à la variance ponctuelle du résidu. Ainsi, le plan optimal selon (2.5) est indépendant d'un facteur multiplicatif de la fonction de covariance. En revanche,  $O(\eta)$  dépend des paramètres  $\theta \in \Theta$  de la fonction de covariance gouvernant le comportement à l'origine et du paramètre d'échelle. Sa minimisation implique ainsi de choisir des valeurs pour ces paramètres avant même d'avoir récolté les données. On peut alors opter pour une approche bayésienne où l'on va choisir une loi a priori pour ces paramètres à partir de la connaissance du phénomène ou des données disponibles sur d'autres domaines. Le critère bayésien prend alors la forme suivante :

$$O(\eta) = \frac{1}{|\mathcal{X}|} \int_{\Theta} \int_{\mathcal{X}} \mathbb{V}(Z(x) - \widehat{Z}(x) | \theta) w(x) p(\theta) dx, \quad (2.7)$$

où  $p(\theta)$  est la loi a priori des paramètres de la fonction de covariance. J'ai développé cette approche dans le cadre de l'optimisation de l'échantillonnage de la concentration de polluants atmosphériques pour des campagnes ponctuelles dans le temps à l'échelle de l'agglomération. Les données de deux précédentes campagnes étaient disponibles afin de calibrer les a priori sur les paramètres, notamment le modèle de fonction de covariance et



la gamme de variation de ses paramètres. Ce travail a été publié dans Romary et al. [2011], où l'optimisation de (2.7), une fois discrétisée sur la grille de prédiction, était réalisée par recuit simulé. J'ai également proposé dans cet article un noyau de transition propositionnel original fait d'un mélange de deux noyaux : le premier déplaçant un ou plusieurs points dans l'espace géographique et le second déplaçant les points dans le nuage des variables auxiliaires. L'avantage de ce noyau est d'autoriser les points à se déplacer dans un espace de dimension supérieure au simple espace géographique, permettant une exploration de  $H$  plus efficace.

Cependant, cette approche ne tient pas compte de l'incertitude liée à l'estimation des paramètres  $\theta$  intervenant dans le calcul de (2.4). En toute rigueur, la variance de prédiction (2.4) devrait être remplacée par sa contrepartie empirique, où l'on intègre le fait que la prédiction est réalisée sur la base d'un estimateur des paramètres de covariance :

$$\mathbb{V}(Z(x) - \widehat{Z}_{\hat{\theta}}(x)|\theta), \tag{2.8}$$

où  $Z_{\hat{\theta}}$  signifie que la prédiction est réalisée avec les paramètres estimés. Cette quantité, qui dépend de la valeur inconnue de  $\theta$ , est malheureusement incalculable. Des approximations ont été proposées dans Zimmerman and Cressie [1992], Zimmerman [2006], Zhu and Stein [2006] mais restent peu praticables. En effet, elles nécessitent de calculer la dérivée des poids de krigeage par rapport aux paramètres de covariance, en chaque point cible, ce qui est particulièrement coûteux en temps de calcul, notamment en vue d'une optimisation. Ces approximations font intervenir la matrice d'information de Fisher  $\mathcal{I}(\theta|\eta)$ , dont l'inverse fournit un estimateur asymptotique de la variance d'estimation des paramètres, sous certaines conditions de régularité de la fonction de covariance. Une manière de prendre en compte la qualité d'un plan d'échantillonnage pour l'estimation des paramètres de covariance, proposée dans Zhu and Stein [2005], est de minimiser la quantité suivante :

$$0_{param}(\eta) = - \int \log \det(\mathcal{I}(\theta|\eta))p(\theta)d\theta. \tag{2.9}$$

Il s'agit de l'espérance du logarithme du déterminant de l'inverse de la matrice de Fisher par rapport à la loi a priori de  $\theta$ . Minimiser cette quantité conduira ainsi à des plans optimaux pour l'estimation des paramètres de covariance. Plutôt que de minimiser (2.8), j'ai ainsi proposé de choisir parmi une population de plan suboptimaux pour (2.6), générés via un algorithme particulière inspiré de l'approche population Monte-Carlo ([Cappé et al., 2004]), le ou les plans qui obtiennent les meilleures performances au regard de (2.9). Cette approche a été développée dans Romary et al. [2014], sur la problématique explorée dans Romary et al. [2011]. Cet ensemble de plans candidats présente également l'avantage de fournir outre un plan optimal, une estimation de la densité d'échantillonnage souhaitable qui peut être utile en pratique pour l'opérateur chargé d'installer les capteurs. Un exemple de plan optimal et de densité d'échantillonnage obtenus pour l'échantillonnage passif de la concentration en benzène sur l'agglomération de Bordeaux sont donnés en figure 2.1.

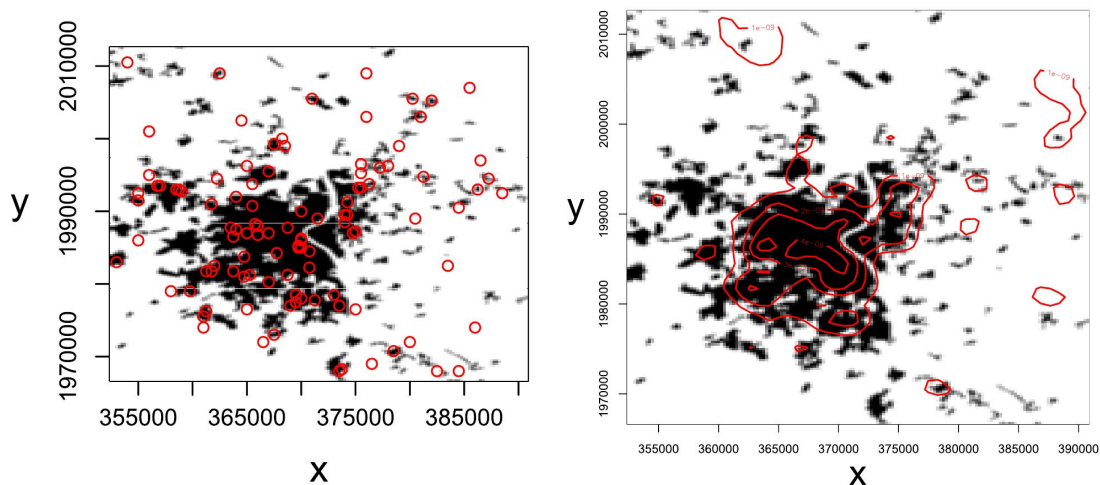


FIGURE 2.1 – Plan optimal et densité d'échantillonnage sur l'agglomération de Bordeaux

## 2.4 Augmentation d'un réseau existant

Dans le cadre de l'échantillonnage dit séquentiel, on dispose d'un échantillonnage initial, dont les mesures permettent d'inférer un premier modèle. On cherche alors à optimiser l'implantation de nouveaux points de mesure selon un critère donné. Cette problématique s'est beaucoup développée dans le cadre de l'analyse de surface de réponse, où une fonction aléatoire gaussienne est employée pour interpoler le comportement d'un modèle physique donné sous la forme d'une fonction d'un certain nombre de paramètres. Dans ce cadre, de nombreux critères ont été proposés afin de trouver les paramètres qui maximisent la sortie du modèle ou encore d'identifier un ensemble d'excursion voir par exemple Gramacy [2020], Santner et al. [2018]. En géosciences en revanche, il est peu courant de pouvoir procéder à des mesures additionnelles en direct. On opère généralement en une voire deux étapes.

Je me suis intéressé à cette problématique dans le cadre de la caractérisation de l'activité radiologique d'installations industrielles en vue de leur décontamination, dans le cadre d'une collaboration avec l'éditeur de logiciel géostatistique Geovariances ([Bechler et al., 2013]). Dans ce cadre, on disposait de deux types de données : des données de surfaces, peu précises, et des mesures issues de prélèvements de bétons. Il s'agissait d'identifier l'emplacement optimal de nouveaux points de prélèvements afin de caractériser au mieux la contamination et en particulier de classer les déchets finaux selon leur activité en vue de leur traitement futur. Les deux types de données étaient traités simultanément par co-krigeage ([Wackernagel, 2003]) après une transformation préalable selon une anamorphose gaussienne, qui consiste à transformer la loi marginale des données. Différents critères ont été évalués notamment des critères de type faux négatif afin de se prémunir d'une classification dans une classe d'activité inférieure, ce qui peut être dommageable du point de vue de la sûreté.

Plus précisément, on souhaite caractériser l'ensemble d'excursion  $\mathcal{C} = \{x, x \in \mathcal{X} \subset \mathbb{R}^2, Z(x) > u\}$ , où  $x$  est la coordonnée,  $Z(x)$  le niveau de radioactivité et  $u$  un seuil donné.

Etant donné un échantillon  $\mathbf{z} = (Z(x_1), \dots, Z(x_n))$  et  $\phi$  la fonction d'anamorphose, telle que  $Y(x) = \phi(Z(x))$  puisse être considérée gaussienne, pour  $x \in \mathcal{X}$ , la probabilité de dépasser le seuil  $u$  est donnée par :

$$\begin{aligned} \mathbb{P}(x \in \mathcal{C}|\mathbf{z}) &= \mathbb{P}(Z(x) > u|\mathbf{z}), \\ &= \mathbb{P}\left(\frac{Y(x) - \widehat{Y}(x)}{\widehat{\sigma}(x)} > \frac{\phi(u) - \widehat{Y}(x)}{\widehat{\sigma}(x)} \middle| \mathbf{z}\right), \\ &= 1 - \Phi\left(\frac{\phi(u) - \widehat{Y}(x)}{\widehat{\sigma}(x)}\right), \end{aligned} \tag{2.10}$$

où  $\widehat{Y}(x)$  est le krigeage de  $Y$  au point  $x$ ,  $\widehat{\sigma}(x)$  l'écart type de krigeage et  $\Phi$  la fonction de répartition d'une gaussienne standard.

Le *critère de faux négatif* proposé dans Bechler et al. [2013] consiste à intégrer sur  $\mathcal{X}$  la distance entre le seuil  $u$  et la borne supérieure de l'intervalle de confiance à 95% :

$$\int_{\mathcal{X}} \left(\phi^{-1}\left(\widehat{Y}(x) + 1.96\widehat{\sigma}(x)\right) - u\right) \mathbb{1}_{\widehat{Y}(x) < \phi(u)} dx. \tag{2.11}$$

On peut remarquer que

$$\begin{aligned} \mathbb{P}\left(Z(x) \in \left(u, \phi^{-1}\left(\widehat{Y}(x) + 1.96\widehat{\sigma}(x)\right)\right) \middle| \mathbf{z}\right) &= \mathbb{P}(\phi(u) < \phi(Z(x)) < \widehat{Y}(x) + 1.96\widehat{\sigma}(x)|\mathbf{z}), \\ &= \mathbb{P}(x \in \mathcal{C}|\mathbf{z}) - 2.5\%, \end{aligned}$$

En pratique, si  $\widehat{Y}(x) \geq \phi(u)$ , le point  $x$  est déclaré contaminé, autrement dit si  $\mathbb{P}(x \in \mathcal{C}|\mathbf{z}) \geq 0.5$ . Si  $\widehat{Y}(x) < \phi(u)$ , alors la valeur  $\phi^{-1}(\widehat{Y}(x) + \beta) - u$  est associée au point  $x$ . En conséquence, ce critère a deux principaux défauts. D'une part, il ne prend pas en compte les points qui ont une probabilité conditionnelle de dépassement de seuil inférieure à 2.5% et d'autre part, il nécessite d'inverser la fonction d'anamorphose pour des valeurs potentiellement élevée, là où elle est peut fiable. En effet, l'anamorphose est modélisée selon une transformation non paramétrique de la loi marginale. J'ai ainsi proposé le critère suivant, directement lié à la probabilité d'appartenir à l'ensemble d'excursion :

$$\int_{\mathcal{X}} \mathbb{P}(x \in \mathcal{C}|\mathbf{z}) \mathbb{1}_{\widehat{Y}(x) < \phi(u)} dx \tag{2.12}$$

Comparativement à (2.11), Ce nouveau critère a le désavantage de ne pas mettre beaucoup plus de poids sur les probabilités de dépassement de seuil proches de 0.5. On peut y remédier en considérant la modification suivante :

$$\int_{\mathcal{X}} f(\mathbb{P}(x \in \mathcal{C}|\mathbf{z})) \mathbb{1}_{\widehat{Y}(x) < \phi(u)} dx,$$

où  $f$  est strictement croissante sur  $[0,0.5]$ , par exemple  $f = \exp$ , ce qui permettra de pénaliser plus fortement les points à la limite tout en évitant l'écueil d'inverser la fonction d'anamorphose dans la queue de distribution. C'est ce dernier critère qui a été retenu pour l'implémentation dans le logiciel Kartotrak, développé par Geovariances.

## 2.5 Discussion

J'ai présenté dans ce chapitre mes contributions à l'échantillonnage optimal de variables environnementales. Les techniques présentées sont valables pour un échantillonnage ponctuel pour des fonctions aléatoires indexées sur un compact de  $\mathbb{R}^2$ . Elles reposent sur la connaissance d'une modélisation a priori. Les plans d'échantillonnage sont obtenus via la maximisation d'une fonction d'utilité que l'on peut adapter suivant le contexte et dont j'ai donné différents exemples. L'optimisation est réalisée grâce à des algorithmes d'optimisation globale à savoir le recuit simulé et un algorithme particulière.

En pratique, on ne dispose pas toujours d'information au préalable. Dans ce cas, on préconise généralement l'utilisation d'un plan uniforme dans l'espace, de manière à obtenir de bons résultats en prédiction auquel on ajoute des points de manière à former des paires proches spatialement ([Diggle and Lophaven, 2006]). Ces paires de points sont utiles pour l'estimation des paramètres, en particulier pour la caractérisation du comportement à l'origine de la fonction de covariance, qui est essentielle pour la prédiction ([Stein, 1999]). Enfin, si des variables auxiliaires sont disponibles, on cherchera également à couvrir le nuage de points de ces variables, notamment en ses extrémités, de manière à minimiser la variance de l'estimateur du coefficient de régression.

Dans le problème de décontamination ci-dessus, on disposait de deux types de mesure : les mesures de surface disposées sur une grille régulière et un premier échantillonnage de mesures destructives. Avec l'essor des sciences participatives et des capteurs à bas coût, il arrive de plus en plus fréquemment de disposer de données alternatives, souvent de moindre qualité, mais dont il convient de tenir compte lors de la mise en place d'un plan d'échantillonnage pour une campagne d'observation menées par les scientifiques. C'est par exemple le cas dans la surveillance de la biodiversité, où un grand nombre d'applications grand public permettent à chacun de renseigner ses observations (de la faune, de la flore, etc.). Une telle approche a été proposée dans Reich et al. [2018] où les données auxiliaires étaient utilisées pour définir une modélisation a priori de l'occupation du territoire par la sittelle à tête brune dans le sud-est des Etats-Unis. En qualité de l'air, les données issues de micro-capteurs, qui évaluent la concentration en polluants dans l'air à faible coût moyennant une erreur de mesure importante, pourraient également aider à l'optimisation du positionnement des stations de mesures automatiques, plus précises ([Gressent et al., 2020]).

Ces derniers exemples s'accompagnent d'un degré de complexité supérieur dans le sens où les phénomènes observés peuvent présenter une dynamique spatio-temporelle, selon l'échelle temporelle considérée. Les travaux portant sur l'échantillonnage optimal sont assez rares dans ce contexte ([Mateu and Müller, 2012]). Le traitement de cette problématique dans le cadre bayésien décliné dans ce chapitre nécessite de disposer de suffisamment d'information au préalable, en particulier sur la dynamique du phénomène d'intérêt, afin de fournir une modélisation a priori adaptée (on reviendra sur la modélisation spatio-temporelle dans le chapitre 4). En particulier, on peut s'interroger sur l'opportunité de disposer de capteurs mobiles afin d'adapter l'échantillonnage au cours du temps selon l'information disponible. Le phénomène spatio-temporel serait ainsi mesuré le long de tra-

jectoires que l'on chercherait à optimiser. Ce cas de figure se présente par exemple en halieutique où des bateaux équipés de sonars cherchent à identifier la présence de bancs de poissons ([Woillez et al., 2012]).

## Chapitre 3

# Clustering de données géostatistiques

### 3.1 Problématique

La classification non supervisée en contexte spatial s'avère nécessaire lorsque le phénomène étudié présente des comportements très différents sur le domaine d'étude, notamment en terme de moments des deux premiers ordres. Dans les applications minières, c'est le cas lorsqu'on a différents lithofaciès qui apparaissent, par exemple zone minéralisée versus non minéralisée. Le problème se pose également en science des sols où plusieurs types de sols présentant des caractéristiques très différentes peuvent apparaître sur le domaine d'étude. La variable régionalisée peut alors être modélisée par un mélange de fonctions aléatoires, où chaque composante modélise le comportement d'un type. Pour identifier ces différents composantes, l'idée que j'ai développée consiste à appliquer un algorithme de classification non supervisée (clustering) aux données disponibles, en s'assurant que les classes obtenues présentent certaines propriétés de contiguïté spatiale. Ces algorithmes ont été développés dans le cadre d'une collaboration avec l'éditeur de logiciel Geovariances en vue de leur application à des données minières.

Ce problème a été largement étudié en analyse d'image ou de données de télédétection où la modélisation par champ de Markov donne de bons résultats, voir par exemple l'ouvrage de Xavier Guyon ([Guyon, 1995]) et Celeux et al. [2003].

La classification de données issues d'un échantillonnage spatial non régulier a été en revanche moins étudiée. Dans le cas de données minières, où l'on dispose de sondages verticaux à l'échelle du gisement, les approches par champs de Markov sont peu adaptées du fait que le graphe sur lequel un tel modèle pourrait être appliqué dépend fortement de l'échantillonnage. Si l'on souhaite appliquer une approche similaire, via l'algorithme EM, le modèle probabiliste le plus adapté à ces données est celui de mélanges de fonctions aléatoires Allard and Guillot [2000], Guillot et al. [2006] qui se prête peu au traitement de grands jeux de données multidimensionnelles : le calcul de la vraisemblance nécessite l'inversion d'une matrice de covariance de très grande taille, dense qui plus est. L'intégration de données catégorielles devient également délicate dans ce cadre.

Dans ce chapitre, je décris dans un premier temps le modèle de mélange de fonctions aléatoires adapté au problème puis deux algorithmes que j'ai développés et qui offrent une

solution pratique. Ce chapitre synthétise les travaux publiés dans les articles suivants :

- **T. Romary**, F. Ors, J. Rivoirard, and J. Deraisme. *Unsupervised classification of multivariate geostatistical data : Two algorithms*. Computers & Geosciences, pages 96–103, 2015.
- **T. Romary**, J. Rivoirard, J. Deraisme, C. Quinones, and X. Freulon. *Domaining by clustering multivariate geostatistical data*. In P. Abrahamsen, R. Hauge, and O. Kolbjørnsen, editors, Geostatistics Oslo 2012, volume 17 of Quantitative Geology and Geostatistics, pages 455–466. Springer Netherlands.

## 3.2 Modèle de mélange de fonctions aléatoires

Lorsque le phénomène étudié prend la forme décrite ci-dessus, on peut modéliser la variable régionalisée associée comme la réalisation d'un mélange de  $p$  fonctions aléatoires :

$$Z(x) = \sum_{i=1}^p t_i(x) Z_i(x), \quad (3.1)$$

où  $t_i(x)$  vaut 1 si le point  $x$  appartient au domaine  $i$ , 0 sinon, et les  $Z_i$  sont les fonctions aléatoires constitutives du mélange. Autrement dit, en tout point de l'espace une et une seule composante est visible et le domaine peut être subdivisé en  $p$  sous domaines disjoints. Le problème de la classification non supervisée consiste alors à identifier le nombre  $p$  de classes et à retrouver les valeurs des  $t_i$  aux points de données.

Pour aller plus loin, on peut vouloir prédire cette valeur ainsi que la valeur des  $Z_i$  en tout point du domaine  $\mathcal{X}$ . Pour ce faire, moyennant une hypothèse sur la loi spatiale des  $Z_i$  et l'indépendance des  $t_i$  d'un site à l'autre, on peut adopter une approche par maximum de vraisemblance. Pour une valeur de la matrice des  $t_i(x_j)$ , qui attribue à chaque donnée sa composante, la vraisemblance d'un échantillon  $(z_1, \dots, z_n)$  issu de (3.1), sous l'hypothèse que les  $Z_i$  sont gaussiennes, prend la forme suivante :

$$L(z_1, \dots, z_n) = \prod_{i=1}^p \frac{1}{(|\Sigma_i| 2\pi)^{n_i/2}} \exp\left(-\frac{1}{2}(\tilde{z}_i - \mu_i)^t \Sigma_i^{-1} (\tilde{z}_i - \mu_i)\right), \quad (3.2)$$

où  $n_i = \sum_{j=1}^n t_i(x_j)$  et  $\tilde{z}_i$  est le sous-échantillon de  $(z_1, \dots, z_n)$  associé à la composante  $i$ . Pratiquement, il convient d'identifier les  $t_i(x_j)$  et d'estimer les paramètres associés à chaque composante. Dans l'approche développée par Allard and Guillot [2000], le problème est résolu via l'algorithme EM, où l'étape d'"expectation" identifie les  $t_i(x_j)$ , considérés comme des variables latentes, et l'étape de "maximization" optimise (3.2). Cependant, un certain nombre de freins empêche la mise en oeuvre de cette approche sur les jeux de données réels. En effet, on est souvent confronté dans les applications à des jeux de données multivariées comprenant des variables catégorielles<sup>1</sup>. Le modèle multigaussien n'est alors pas forcément adapté. Par ailleurs, l'indépendance des  $t_i$  est contestable et le choix du nombre de classe

1. Notons que l'on pourrait éventuellement modéliser des données ordinales à l'aide d'un modèle de fonction aléatoire généralisé, cf. [Diggle and Jr., 2007]

reste arbitraire. Enfin, les jeux de données modernes sont dorénavant de très grande taille ce qui induit un coût computationnel particulièrement élevé pour ce type d’algorithmes. Néanmoins, le modèle (3.1) fournit une grille de lecture théorique pour les algorithmes développés dans la section suivante.

### 3.3 Algorithmes GHC et GSC

Les algorithmes de classification non supervisée que j’ai proposés sont tous deux basés sur la même idée. Elle consiste à structurer les données disponibles selon un graphe (de Delaunay ou basé sur des voisinages) fait d’une unique structure connexe. L’algorithme de classification consiste alors à partitionner ce graphe en autant de classes que désiré. La structure ainsi imposée aux données permet de s’assurer de la cohérence spatiale des classe obtenues : elles sont en effet connexes (au sens du graphe) par construction.

Dans ce cadre, j’ai proposé deux algorithmes : l’un procède de manière ascendante hiérarchique ([Lance and Williams, 1966]) et le second résulte d’une adaptation de l’algorithme de classification spectrale ([Ng et al., 2002, von Luxburg, 2007]). Tous deux dépendent également du choix d’une distance ou dissimilarité adaptée au cas multivarié. Pour un échantillon  $(z_1, \dots, z_n) \in \mathbb{R}^{(n \times (p+d))}$ , qui contient également les  $d$  variables coordonnées, on définit la distance  $D$  :

$$D(z_i, z_j) = \sum_{k=1}^{p+1} w_k d^{(k)}(z_i^{(k)}, z_j^{(k)}), \quad (3.3)$$

où les  $w_k$  sont des pondérateurs définis par l’utilisateur.  $D$  est ainsi une somme pondérée de distances individuelles qui sont choisies en fonction de la variable associée : si celle-ci est quantitative, on utilise généralement la distance euclidienne au carré, s’il s’agit d’une variable catégorielle, une distance adaptée est employée, par exemple une distance qui prend la valeur 0 lorsque les deux observations prennent deux valeurs égales pour cette variable et 1 sinon. En ce qui concerne les coordonnées, une distance particulière peut être utilisée, c’est pourquoi les termes concernés sont groupés dans (3.3). On peut par exemple considérer une distance anisotrope, ou plus généralement une déformation quelconque de l’espace (cf. [Sampson and Guttorp, 1992, Fouedjio et al., 2015]). L’algorithme de classification hiérarchique géostatistique (GHC) se déroule selon le pseudo-code 1, où la notation  $i \leftrightarrow \{kl\}$  signifie que l’élément  $i$  est connecté à la classe  $\{kl\}$ , c’est-à-dire  $i \leftrightarrow k$  ou  $i \leftrightarrow l$ .



---

**Algorithme 1** Geostatistical Hierarchical Clustering algorithm (GHC)

---

- 1: Construire la matrice des distances  $D \in \mathbb{R}^{n \times n}$ , telle que  $D_{ij} = d(x_i, x_j)$ ,  $j < i$ , si  $i \leftrightarrow j$ , 0 sinon
- 2: **répéter**
- 3: Trouver  $k$  et  $l$ ,  $k < l$ , tels que  $D_{lk} = \min_{\{i,j,i \leftrightarrow j\}} D_{ij}$
- 4: Fusionner  $k$  et  $l$  en  $\{kl\}$ , et mettre  $D$  à jour de la manière suivante

$$D_{ki} = \max(D_{ki}, D_{li}) \text{ if } i \leftrightarrow \{kl\} \text{ and } i < k$$

$$D_{ik} = \max(D_{ki}, D_{li}) \text{ if } i \leftrightarrow \{kl\} \text{ and } i > k$$

supprimer la ligne et la colonne  $l$  de  $D$

- 5: **jusqu'à**  $D$  ne contient plus qu'un élément
- 

Dans l'algorithme 1, la valeur  $D_{lk}$  peut être interprétée comme la dissimilarité interne à la classe obtenue lorsque l'on fusionne les classes  $k$  et  $l$ . La notation  $i \leftrightarrow \{kl\}$  signifie que l'élément  $i$  est connecté à la classe  $\{kl\}$ , c'est-à-dire  $i \leftrightarrow k$  ou  $i \leftrightarrow l$ .

Alors que l'on peut utiliser différentes métriques pour calculer la dissimilarité intra classe dans le cas où les observations sont indépendantes, la dissimilarité maximale est préconisée lorsque les données sont structurées par un graphe. Ainsi, les dissimilarités intra classe sont nécessairement croissantes, ce qui autorise la représentation du processus d'agglomération par un dendrogramme, ce dernier permet alors de simplifier l'identification du nombre de clusters.

L'algorithme de classification spectrale géostatistique (GSC) est une simple application de l'algorithme classique de clustering spectral où on utilise le graphe d'échantillonnage et est décrit en pseudo-code ci-dessous 2.

**Algorithme 2** Geostatistical Spectral Clustering algorithm (GSC)

---

1: Construire la matrice de similarité  $W$  :

$$W_{ij} = \begin{cases} \exp\left(-\frac{d(x_i, x_j)}{\sigma^2}\right) & \text{if } i \leftrightarrow j \\ 0 & \text{sinon} \end{cases}$$

2: Construire la matrice diagonale  $H$  :

$$H_{ii} = \sum_{j=1}^n W_{ij}$$

3: Construire la matrice Laplacienne pondérée du graphe

$$L = H^{-1/2}WH^{-1/2}$$

4: Calculer les  $K$  plus grandes valeurs propres de  $L$  et construire la matrice  $V \in \mathbb{R}^{n \times K}$  dont les colonnes sont les  $K$  premiers vecteurs propres de  $L$

5: Appliquer l'algorithme des  $K$ -means aux lignes de  $V$

6: Assigner l'observation  $z_i$  à la classe à laquelle la ligne  $i$  de  $V$  a été assignée

---

La valeur du paramètre  $\sigma$  peut être fixée selon différentes heuristiques proposées dans von Luxburg [2007], par exemple en prenant la valeur moyenne des distances sur l'échantillon ou encore la distance moyenne d'un point à ses  $k$  plus proches voisins avec  $k$  de l'ordre de  $\log n$ .

### 3.4 Discussion

Le développement d'algorithmes de clustering est une tâche relativement frustrante. On peut en effet les tester et valider sur des réalisations de modèles de mélanges de fonctions aléatoires gaussiennes du type (3.1). En revanche, la validation sur données réelles est plus délicate puisqu'on ne dispose pas d'une référence. De fait, l'algorithme GHC a finalement été retenu pour l'implémentation dans le logiciel Isatis.neo de Géovariations en partie car il permet de visualiser le processus d'agglomération au travers du dendrogramme. Par ailleurs, les résultats obtenus sur données minières étaient corroborés par les interprétations des géologues, ce qui n'était pas le cas avec le GSC.

L'ingrédient fondamental de ces deux algorithmes est la distance (3.3). Pour obtenir de bons résultats, il est nécessaire de choisir des poids  $w_k$  qui vont assurer une discrimination efficace entre les classes, c'est-à-dire générant de fortes dissimilarités interclasses et de faibles dissimilarités intraclasses. Il serait notamment intéressant de rechercher les poids qui optimisent cette discrimination dans le modèle (3.1). En particulier, le poids associé à la distance géographique apparaît primordial. En effet, on peut s'apercevoir que si la

proportion de ce poids est de 100%, les deux algorithmes effectueront une classification du plan d'échantillonnage. Si la proportion est nulle, en revanche, les longueurs des connexions n'influeront plus sur la classification, ce qui peut engendrer des résultats étonnants, notamment quand le graphe contient de longues connexions. Les résultats expérimentaux indiquent qu'une proportion de 20% à 40% offre des résultats satisfaisants. Il est en effet préférable de privilégier les faibles distances géographiques : pour les variables structurées, on peut voir que les distances individuelles font alors intervenir le variogramme quand la distance euclidienne au carré est employée. Fouedjio [2016] propose d'employer une distance basée sur un estimateur à noyau des variogrammes simples et croisés. Une comparaison avec mes algorithmes est disponible dans Fouedjio [2020].

L'efficacité numérique de ces deux algorithmes est contrastée. Dans le cas du GHC, seules les distances entre points connectés sont nécessaires au lancement de l'algorithme, mais les distances entre points non connectés sont nécessaires lors de l'étape 3 de l'algorithme et sont donc calculées au fur et à mesure. Par conséquent, sa complexité est relativement élevée. L'implémentation disponible dans `Isatis.neo` propose de réaliser le clustering sur un sous-échantillon puis d'en extrapoler les résultats par SVM ([Hastie et al., 2009]). Dans le cas du GSC, le graphe induit une structure creuse dans toutes les matrices impliquées dans l'algorithme, ce qui permet d'employer les algorithmes d'algèbre linéaire adaptés, notamment pour extraire rapidement les seules  $K$  plus grandes valeurs propres où  $K$  est fixé par l'utilisateur. Ainsi, on peut choisir en première intention une borne supérieure pour  $K$  puis comparer les résultats des clustering obtenus avec au plus  $K$  classes via le calcul de critères dédiés.

L'approche que j'ai développée ici, basée sur un graphe d'échantillonnage, est finalement très similaire à la problématique de détection de communautés dans les réseaux dont on trouve des applications dans des domaines aussi variés que la physique, la biologie moléculaire ou les sciences sociales. L'article de revue Leger et al. [2014] propose un certain nombre d'algorithmes adaptés. Parmi ceux-ci, l'approche par stochastic block model (SBM), un modèle de mélange qui repose sur des variables latentes définissant l'appartenance à une classe et une matrice de probabilités interclasses, semble prometteuse. Cependant, quelques essais réalisés à l'aide du package `blockmodels` de R ([Leger et al., 2020]), qui propose une inférence du SBM par EM variationnel, ne se sont pour l'instant pas avérés concluants.

# Chapitre 4

## Contributions au krigeage

### 4.1 Problématique

Le produit final d'une étude géostatistique est souvent une cartographie du phénomène étudié. Les données disponibles sont interpolées par krigeage sur le domaine d'étude. A cette prédiction est associée une carte de variance ou d'écart type qui fournit une mesure de l'incertitude associée.

On rappelle brièvement le principe du krigeage dans le cas de données ponctuelles. Etant donné des observations  $z = (z(x_1), \dots, z(x_n))^t$  de la variable étudiée, on considère qu'elles sont issues d'une réalisation d'une fonction aléatoire d'ordre 2, que l'on notera  $Z$ , de fonction de covariance  $C$ . Il s'agit dans un premier temps d'estimer les paramètres du modèle, par ajustement d'un variogramme ([Desassis and Renard, 2013]) ou bien par maximum de vraisemblance ([Mardia and Marshall, 1984]), suivant une hypothèse en loi, généralement gaussienne.

Le krigeage au point  $x_0$  est ensuite défini comme le meilleur prédicteur linéaire sans biais de  $Z(x_0)$  au sens de la variance de prédiction. Autrement dit, on cherche un prédicteur  $Z^* = \Lambda^t z$  où

$$\Lambda = \arg \min_{\Lambda \in \mathbb{R}^n} \{\mathbb{V}(Z(x) - Z^*)\} \text{ tel que } \mathbb{E}(Z^*) = \mathbb{E}(Z(x_0)). \quad (4.1)$$

Dans le cas du krigeage universel ([Chilès and Delfiner, 2012]), les contraintes de non biais se traduisent dans la fonction objectif (4.1) par l'ajout d'un terme de la forme  $2 \sum_{l=1}^L \mu_l (\Lambda^t F_l - f_l(x_0))^2$ , où  $F_l = (f_l(x_1), \dots, f_l(x_n))^t$  et  $f_l$  est la  $l^{\text{ème}}$  fonction de dérive. En notant  $C$ , la matrice de covariance associée au vecteur  $z$ ,  $C_0$  le vecteur des covariances entre les points de données et le point cible,  $F$  la matrice des fonctions de dérive dont les colonnes sont constituées des  $F_l$  et  $f_0 = (f_1(x_0), \dots, f_L(x_0))^t$ , le prédicteur de krigeage universel est de la forme :

$$Z^* = C_0^t C^{-1} z + (f_0^t - C_0^t C^{-1} F)(F^t C^{-1} F)^{-1} F^t C^{-1} z, \quad (4.2)$$

et sa variance de prédiction s'écrit

$$\sigma^* = C(0) - C_0^t C^{-1} C_0 + (f_0^t - C_0^t C^{-1} F)(F^t C^{-1} F)^{-1} (f_0^t - C_0^t C^{-1} F)^t. \quad (4.3)$$

On rappelle que sous l'hypothèse gaussienne, (4.2) et (4.3) correspondent respectivement à l'espérance et à la variance conditionnelles sachant  $z$ .

En pratique cependant, il n'est pas toujours possible ni même souhaitable d'utiliser directement ces formules. En effet, en présence de dérive, il peut parfois être intéressant de ne considérer que les points de données appartenant à un voisinage du point cible. Ceci offre plus de flexibilité à l'ajustement intrinsèque de la dérive effectué dans le krigeage en l'autorisant à varier sur le domaine d'étude. On obtient ainsi un ajustement non stationnaire de la dérive. Cette approche présente cependant l'inconvénient de générer des discontinuités. On verra dans la section 4.2 comment corriger ce défaut.

Par ailleurs, avec le déploiement toujours plus important de capteurs géoréférencés et de moyens de télédétection, la taille des jeux de données à traiter a considérablement augmenté. Ces grands jeux de données laissent souvent apparaître des non stationnarités dont il convient de tenir compte dans la structure de covariance de la fonction aléatoire, par exemple des anisotropies variables. Ces aspects sont abordés en section 4.3. Dans le cadre de grands jeux de données, il est impossible de calculer directement (4.2) et (4.3). On verra en section 4.4 différentes approches adaptées au traitement de grands jeux de données.

Enfin, ce déploiement a lieu non seulement dans l'espace mais aussi dans le temps, ce qui nécessite des modèles permettant de rendre compte de dynamiques temporelles particulières. Ces questions seront abordées en section 4.5.

Ce chapitre synthétise les travaux publiés dans les articles suivants :

- **T. Romary**, N. Desassis. *Low rank plus sparse covariance decomposition for a scalable non stationary kriging*, working paper
- R. E. S. S. T. E. *Network Analyzing spatio-temporal data with R : Everything you always wanted to know—but were afraid to ask*. Journal de la Société Française de Statistique, 158(3), 124-158. (2017)
- F. Fouedjio, N. Desassis, **T. Romary**. *Estimation of space deformation model for non-stationary random functions*, Spatial Statistics, 13, 45-61. (2015)
- **T. Romary**, *Incomplete Cholesky decomposition for the kriging of large datasets*, Spatial Statistics, 5, pp 85-99 (2013)
- J. Rivoirard and **T. Romary**, *Continuity for kriging with moving neighborhood*. Mathematical Geosciences, 43 (4), 469-481 (2011)

## 4.2 Krigeage continu

Dans Rivoirard and Romary [2011] nous nous sommes intéressés à la prédiction en voisinage glissant. Le principe est de n'utiliser que les données situées dans un voisinage restreint autour du point cible. Cela permet de limiter la taille du système à résoudre et d'autoriser la dérive à varier dans l'espace. On peut en effet noter que l'on peut réécrire (4.2) sous la forme :

$$Z^* = f_0^t (F^t C^{-1} F)^{-1} F^t C^{-1} z + C_0^t C^{-1} (I - F^t (F^t C^{-1} F)^{-1} F^t C^{-1}) z, \quad (4.4)$$

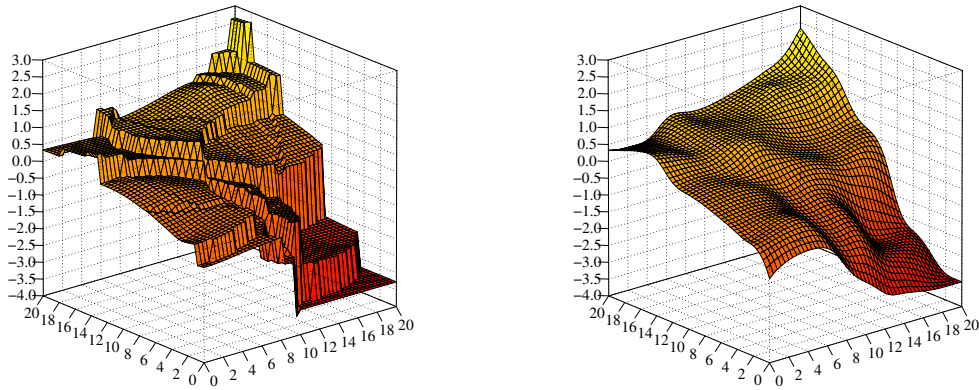
qui peut se lire comme la somme de la dérive en  $f_0$  et du krigeage simple où les coefficients de dérive sont estimés par moindres carrés généralisés. De fait, dans une approche locale, les paramètres de la dérive varient ponctuellement, ce qui leur confère une grande flexibilité.

Cependant, le krigeage en voisinage glissant souffre d'un défaut majeur : lorsqu'on visite les différents points cibles au moment de la prédiction, il arrive qu'un ou plusieurs points de données quittent le voisinage. Il s'ensuit une discontinuité de la surface prédite.

Dans Rivoirard and Romary [2011], nous avons proposé une approche permettant de s'assurer que les poids de krigeage tendent vers 0 lorsque les points de données quittent le voisinage. Elle consiste en la pénalisation des poids de krigeage associés aux points aux bords du voisinage. Typiquement, on ajoute dans (4.1) les contraintes suivantes :

$$\sum_{i=1}^{n_0} l(|x_i - x_0|) \lambda_i^2, \quad (4.5)$$

où  $n_0$  est le nombre de points dans le voisinage de  $x_0$  et  $l$  est une fonction nulle au voisinage de 0 et qui tend continûment vers l'infini en  $R$ , le rayon du voisinage considéré, par exemple  $l(h) = \left(\frac{h-r}{R-h}\right)^2 \mathbb{1}_{]r,R[}$ , où  $0 < r < R < \infty$ . Cette pénalité se traduit dans le système de krigeage en l'ajout d'un terme positif sur la diagonale de  $C$ , autrement dit par un effet pépète tendant continûment vers l'infini lorsque le point se rapproche de la frontière du voisinage. Mécaniquement, les poids de krigeage associés tendent vers 0. La figure 4.1 illustre l'apport de cette méthode.



(a) Krigeage ordinaire en voisinage glissant      (b) Krigeage ordinaire en voisinage glissant continu

FIGURE 4.1 – Comparaison des surfaces krigées

La solution envisagée ci-dessus n'est cependant pas praticable lorsque le nombre de points de données devient très grand. En effet, le krigeage en voisinage glissant nécessite

la résolution d'un système en chaque point cible. D'autre part, la variance de prédiction générée par le krigeage en voisinage glissant est surestimée du fait de l'utilisation d'un nombre réduit de point de données. Il convient ainsi de travailler en voisinage unique ce qui entraîne principalement deux désavantages : la nécessaire résolution d'un système linéaire de grande taille et la perte de la flexibilité liée à un ajustement global de la dérive.

### 4.3 Covariances non-stationnaires

Le caractère non stationnaire d'une fonction aléatoire peut également se traduire dans son moment d'ordre 2. Il n'est pas rare d'observer des anisotropies ou bien une régularité variables dont on peut tenir compte grâce à des modèles de covariance adaptés. De nombreuses approches de modélisation de covariances non stationnaires ont été proposées dans la littérature, parmi lesquelles on peut citer les approches par déformation de l'espace ([Sampson and Guttorp, 1992]) et par convolution ([Fouedjio et al., 2016, Porcu et al., 2009]). Dans Fouedjio et al. [2015], nous avons proposé une approche pour estimer la fonction de déformation de l'espace ne nécessitant pas d'observer des répétitions de la fonction aléatoire d'intérêt, rendant ainsi l'approche de Sampson and Guttorp [1992] praticable à partir d'une unique réalisation. On pourra se référer à la revue de Francky Fouedjio [2017] pour un panorama complet de ces approches.

A titre d'exemple, on donne ici la fonction de covariance de Matérn généralisée proposée dans Fouedjio et al. [2016]. Elle autorise non seulement une anisotropie mais également un paramètre de régularité variables.

$$\text{cov}(Z(x), Z(y)) = \phi_{xy} \frac{2^{1-\nu(x,y)}}{\sqrt{\Gamma(\nu(x))\Gamma(\nu(y))}} M_{\nu(x,y)} \left( \sqrt{Q_{xy}(x-y)} \right), \quad (4.6)$$

où  $\Sigma_x$  est une matrice définie positive qui tient compte de l'anisotropie en  $x$ ,  $\phi_{xy} = |\Sigma_x|^{1/4} |\Sigma_y|^{1/4} \left| \frac{\Sigma_x + \Sigma_y}{2} \right|^{-1/2}$ ,  $Q_{xy}(h) = h^t \left( \frac{\Sigma_x + \Sigma_y}{2} \right)^{-1} h$ ,  $M_{\nu(x,y)}(h) = h^{\nu(x,y)} K_{\nu(x,y)}(h)$ ,  $K_{\nu(x,y)}$  est la fonction de Bessel modifiée de seconde espèce et  $\nu(x, y) = \frac{\nu(x) + \nu(y)}{2}$ .

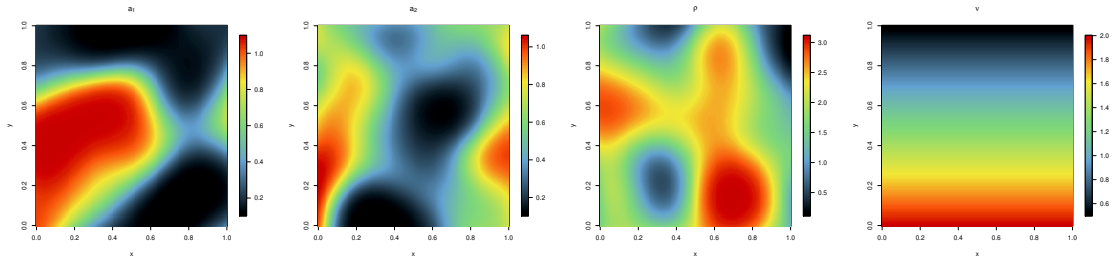
L'estimation des paramètres de ces modèles est souvent délicate. En effet, au lieu d'estimer quelques paramètres : une variance, un paramètre d'échelle et éventuellement un paramètre de régularité avec un paramètre d'échelle supplémentaire et une direction en cas d'anisotropie (en 2D) ; on doit ici estimer des fonctions des coordonnées. Cela passe par des algorithmes coûteux en temps de calcul par exemple basés sur des variogrammes locaux ([Fouedjio et al., 2016]) ou bien des vraisemblances locales pondérées ([Anderes and Stein, 2011]). Ceux-ci ne sont donc pas adaptés au traitement de grands jeux de données dans lesquels on rencontre justement le plus souvent ce type de non-stationnarité et qui seront abordés au paragraphe suivant.

On peut cependant générer assez facilement des réalisations de fonctions aléatoires gaussiennes admettant ce type de covariance. Dans Emery and Arroyo [2017], une adaptation de la méthode spectrale ([Shinozuka and Jan, 1972]) est proposée. La réalisation est approximée par une somme pondérée de fonctions cosinus dont les fréquences aléatoires sont générées selon une densité spectrale stationnaire instrumentale et où les phases

### 4.3. COVARIANCES NON-STATIONNAIRES

aléatoires sont uniformes. Un poids d'importance dépendant des coordonnées permet de respecter la densité spectrale locale de (4.6).

En figure 4.2, on a représenté des champs de paramètres qui vont nous permettre de construire un modèle du type (4.6) et de générer la réalisation de la figure 4.3. Les paramètres d'échelle et la direction d'anisotropie sont obtenus comme des transformées de fonctions aléatoires gaussiennes tandis que le paramètre de régularité varie linéairement dans la direction verticale. Ce choix de modélisation offre une paramétrisation des fonctions gouvernant la non stationnarité. On pourrait d'ailleurs l'utiliser dans un algorithme d'estimation des paramètres d'une covariance non stationnaire, où ces champs apparaîtraient comme des champs latents, moyennant une paramétrisation adaptée. On notera que les champs employés sont très lisses, de manière à éviter une variation trop abrupte des paramètres de (4.6). En effet, ils doivent respecter une contrainte de stationnarité locale, bien que celle-ci n'ait jamais été rigoureusement formalisée ([Fouedjio et al., 2016, Matheron, 1978]). Enfin, on notera qu'on préfère souvent paramétrer l'anisotropie par la décomposition de Cholesky des matrices d'anisotropie  $\Sigma_x$  (produit d'une matrice de rotation d'une matrice diagonale contenant l'inverse des paramètres d'échelle et de la rotation inverse). En effet, les 3 paramètres (en 2D, 6 en 3D) qui interviennent ne sont alors pas contraints et il est donc plus aisé de les interpoler sur le domaine d'étude qu'un angle d'anisotropie par exemple. Cette interpolation est nécessaire lorsque l'estimation des paramètres est réalisée à l'aide d'une méthode locale.



(a) Premier paramètre d'échelle (b) Second paramètre d'échelle (c) Direction d'anisotropie (d) Paramètre de régularité

FIGURE 4.2 – Champs de paramètres de la fonction de covariance

Pour générer la réalisation représentée sur la figure 4.3, on a utilisé une base de 10000 cosinus et la densité spectrale de la covariance exponentielle comme densité instrumentale (on utilise la densité correspondant à la plus faible régularité de manière à mieux représenter les hautes fréquences). On voit que les variations d'anisotropie et de régularité sont bien représentées. On peut noter cependant des artefacts liés à la méthode de simulation dans les zones de moindre régularité. Il est en effet difficile de bien représenter les hautes fréquences avec ce type de méthodes ([Lantuéjoul, 2002]).



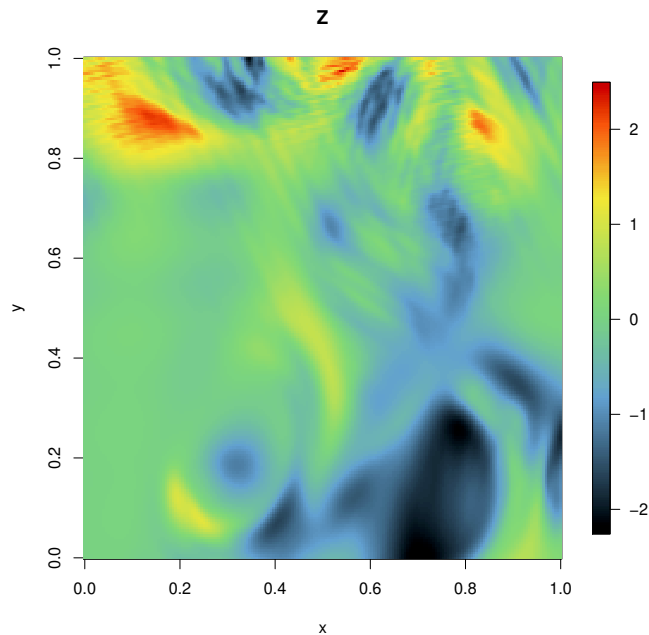


FIGURE 4.3 – Simulation d’une fonction aléatoire gaussienne centrée de covariance non stationnaire de type (4.6)

#### 4.4 Krigeage de grands jeux de données

Comme indiqué en introduction, la taille typique des jeux de données spatiales est passée ces dernières décennies de quelques dizaines de points à plusieurs centaines de milliers voire au-delà. Aussi, les approches géostatistiques développées dans un contexte minier où les données étaient rares ([Matheron, 1965]) doivent être adaptées à ce nouveau paradigme.

Le principal verrou est évidemment la résolution du système linéaire intervenant dans (4.2), où la matrice de covariance est généralement dense. La résolution de ce système intervient également dans le calcul de la log-vraisemblance pour l’estimation de paramètres. La résolution de systèmes linéaires de grande taille passe par l’emploi de formes matricielles adaptées. C’est un problème qui se retrouve dans de nombreux domaines des mathématiques appliquées et un grand nombre d’approches ont été développées pour rendre les calculs praticables. On peut citer l’usage de matrices creuses développé pour la résolution numérique d’équations aux dérivées partielles, les approximations de rang faibles ([Markovsky and Usevich, 2012]) ou encore les matrices hiérarchiques ([Hackbusch, 2015]). L’usage de ces approches numériques a cependant des conséquences en terme de modélisation puisque modifier la matrice de covariance revient à modifier le modèle de covariance et donc la structure de la fonction aléatoire. Le principe de ces méthodes revient donc à chercher une approximation satisfaisante  $\tilde{Z}$  de la fonction aléatoire cible  $Z$  sous une forme

qui permet des calculs efficaces. Dans la suite, sans perte de généralité, on supposera que  $Z$  est centrée.

#### 4.4.1 Approximations creuses

En règle générale, les matrices de covariance qui interviennent dans le krigeage sont denses. Une manière d’y introduire suffisamment de zéros pour rendre ces matrices creuses consiste à employer un modèle de covariance à support compact ([Gneiting, 2002a]), dont la portée est suffisamment faible. Dans Furrer et al. [2006], les auteurs ont montré qu’un modèle de covariance obtenu en multipliant la fonction de covariance initiale par une fonction de covariance à support compact (covariance tapering) est asymptotiquement optimal pour la prédiction. On obtient ce résultat en vérifiant que ce modèle vérifie le critère d’optimalité de Stein ([Stein, 1999]), à savoir que les densités spectrales sont équivalentes dans les hautes fréquences. L’approximation employée est donc une fonction aléatoire de covariance  $C \cdot C_{\text{tap}}$  où  $C_{\text{tap}}$  est une fonction de covariance à support compact. En pratique, cette approche tend à négliger les dépendances à grande échelle. La densité de la matrice de covariance est en effet contrôlée par la portée de la fonction de covariance à support compact employée et celle-ci doit être réduite quand la taille du jeu de données augmente. Cette méthode sera donc efficace pour des phénomènes dont la portée est relativement faible. A noter que la même approximation peut-être employée pour l’estimation des paramètres par maximum de vraisemblance sous hypothèse gaussienne ([Kaufman et al., 2008]). Sur la figure 4.4, on peut voir l’effet d’une portée trop courte sur un krigeage réalisé à partir d’un échantillon de 5000 données uniformément réparties issu de la réalisation représentée en figure 4.3 : quand on cherche à prédire trop loin des données, on prédit simplement l’espérance.

#### 4.4.2 Approximations de rang faible

Pour introduire les approches de rang faible, on peut revenir à certaines propriétés générales des processus du second ordre. En particulier, une fonction aléatoire d’ordre 2, centrée, de fonction de covariance  $C$  et définie sur un compact  $K$ , a ses trajectoires dans un espace de Hilbert  $\mathcal{H}$  dans lequel l’espace vectoriel engendré par les  $C(x, \cdot)$ ,  $x \in K$  est dense (voir par exemple ([Revuz, 1997])). En particulier, si la fonction de covariance est continue sur  $K \times K$ , ce qui est le cas si  $Z$  est continue en moyenne quadratique, le théorème de Karhunen-Loève ([Loève, 1955]) assure que les fonctions propres  $(\phi_i)_{i \in I}$  de  $C$ , où  $I$  est dénombrable, forment une base orthonormale de  $\mathcal{H}$  et que  $Z$  admet la représentation :

$$Z(x) = \sum_{i \in I} \xi_i \phi_i(x), \quad (4.7)$$

où les  $\xi_i$  sont des variables aléatoires orthogonales de variance  $\lambda_i$ , la valeur propre associée à  $\phi_i$ . On peut ainsi obtenir une approximation de rang faible de  $Z$  en tronquant cette somme. Bien sûr, en pratique, on ne sait calculer les  $\phi_i$  que dans des cas particuliers et leur évaluation numérique est extrêmement coûteuse. Toute la difficulté de l’approche réside dans le choix d’une base de fonctions convenables. Un grand nombre de familles

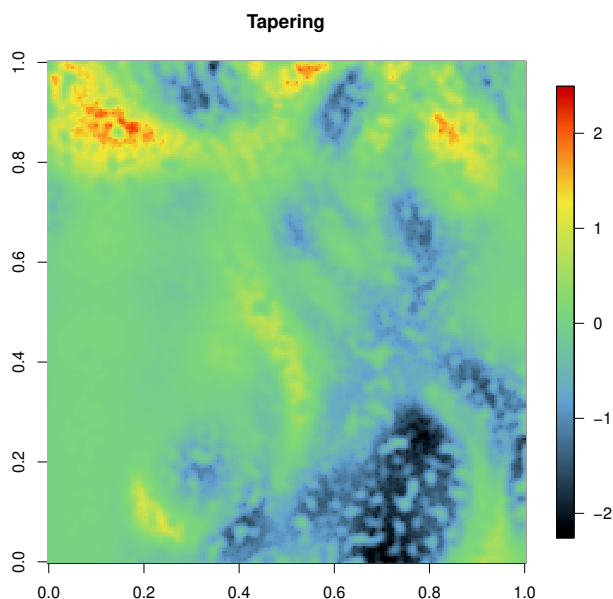


FIGURE 4.4 – Krigeage obtenu par covariance tapering à partir d'un échantillon de 5000 points répartis uniformément

de fonctions de base a été exploré dans la littérature : Cressie and Johannesson [2008] propose l'emploi de fonctions bi-carré, des ondelettes ont été proposées dans Nychka et al. [2002], Stein [2008] utilise des polynômes de Legendre sur la sphère. Dans l'approche des "predictive processes" ([Banerjee et al., 2008]), les fonctions de bases employées sont les  $C(x^*, \cdot)$  pour des  $x^*$  judicieusement choisis. Dans ce cas, les coefficients des fonctions de bases ne sont plus orthogonaux mais on peut montrer que leur covariance est la matrice  $C^{*-1}$ , où  $C^*$  est la matrice de covariance du vecteur des valeurs de  $Z$  en les  $x^*$  ([Sang and Huang, 2012]). Ainsi, l'approximation obtenue est donnée par

$$\tilde{Z}(x) = \sum_{x^*} \eta_i C(x^*, x),$$

où les  $\eta_i$  sont des variables aléatoires centrées de covariance  $C^{*-1}$ .

Dans Romary [2013], j'ai proposé une approche basée sur une décomposition de Cholesky incomplète ([Bach and Jordan, 2005]), qui fournit une approximation de rang faible de la matrice de covariance. Cette approche présente certaines similarités avec les predictive processes. En particulier, on peut s'apercevoir qu'elle revient à choisir les  $x^*$  parmi les points de données, et donc de réaliser un sous-échantillonnage, qui va permettre de conserver une part donnée de la variance globale et donc de contrôler l'erreur d'approximation. On verra au paragraphe 5.2.2 que cet algorithme permet également de fournir des simulations de champs gaussiens de dimension réduite.

Dans la pratique, on ne rencontre que rarement des modèles de covariance continus : on a généralement une discontinuité à l'origine, l'effet pépité, qui permet de modéliser une erreur de mesure par exemple. On décompose alors  $Z$  en une composante continue en moyenne quadratique  $Y$  et un effet de pépité pur  $\varepsilon$  et on considère que les deux composantes sont indépendantes. Si  $Y$  admet une représentation de rang faible,  $Z$  s'écrit alors sous la forme d'un modèle à effets aléatoires ([Gelfand et al., 2011]) :

$$Z(x) = \sum_{i=1}^K \eta_i P_i(x) + \varepsilon(x), \quad (4.8)$$

où les  $P_i$  sont les fonctions de base et les  $\eta$  des coefficients aléatoires. La matrice de covariance d'un vecteur d'observations s'écrit  $C = \sigma^2 I + P^t B P$ , où  $\sigma^2$  est la variance de  $\varepsilon$  supposée constante sur le domaine et  $B$  est la matrice de covariance des  $\eta_i$ . On peut inverser une telle matrice à moindre coût via la formule de Woodbury-Morrison :

$$C^{-1} = \sigma^{-2} I - \sigma^{-2} P \{ \sigma^2 B^{-1} + P^t P \}^{-1} P^t. \quad (4.9)$$

L'estimation des paramètres de ce modèle  $\sigma^2$  et  $B$  peut être réalisée selon une méthode des moments ([Cressie and Johannesson, 2008]) ou par maximum de vraisemblance via un algorithme EM ([Katzfuss and Cressie, 2011, Braham et al., 2017]) qui sera détaillé plus loin.

À l'opposé du covariance tapering, les approximations de rang faibles vont bien représenter les fluctuations à grande échelle mais mal celles à petite échelle, puisque ces dernières seront uniquement représentées par  $\varepsilon$  qui n'est pas structuré. En conséquence, les prédictions vont avoir tendance à être trop lisses. La figure 4.5 illustre ce phénomène sur le jeux de données précédent.

#### 4.4.3 Combinaison

En pratique, selon les cas de figure, l'une ou l'autre de ces approches peut être privilégiée : une approximation de rang faible convient lorsque le phénomène est très régulier, on peut alors le représenter comme une combinaison linéaire d'un nombre réduit fonctions de bases, une approximation creuse est raisonnable lorsque la variable étudiée a une covariance qui décroît rapidement avec la distance. Ce dernier cas de figure s'est présenté dans un travail auquel j'ai pris part portant sur la cartographie de la qualité de la couverture des réseaux téléphoniques ([Alam et al., 2018]). On peut remarquer cependant que les deux approches sont complémentaires et chercher à les combiner. En effet, en admettant que  $\varepsilon$  admet une structure à petite échelle modélisée par une fonction de covariance à support compact, la matrice de covariance des données prend la forme  $C = A + P^t B P$ , où  $A$  est une matrice creuse et la formule de Woodbury-Morrison donne :

$$C^{-1} = A^{-1} - A^{-1} P (B^{-1} + P^t A^{-1} P)^{-1} P^t A^{-1}. \quad (4.10)$$

Dans Sang and Huang [2012], les auteurs proposent d'appliquer le covariance tapering au résidu de l'approximation de rang faible offerte par l'approche des predictive processes.

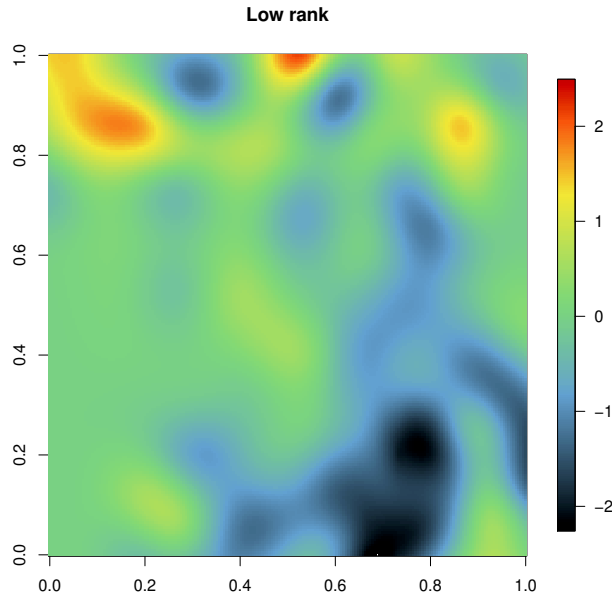


FIGURE 4.5 – Krigeage obtenu par une approximation de rang faible à partir d’un échantillon de 5000 points répartis uniformément

Ils obtiennent ainsi une représentation valable à toutes les échelles. La fonction de covariance prend la forme :

$$\tilde{C}(x, y) = (C(x, y) - C(x, x^*)C^{*-1}C(x^*, y))C_{\text{tap}}(x, y) + C(x, x^*)C^{*-1}C(x^*, y) \quad (4.11)$$

Bien que la forme générale de (4.11) soit non stationnaire, en particulier la composante de rang faible, cette représentation repose sur une fonction de covariance parente  $C$  stationnaire et ne permet pas en l’état de tenir compte d’éventuelles non stationnarités, notamment des anisotropies variables.

Pour pallier ce manque, j’ai proposé dans Romary and Desassis [2018a] de guider le choix des fonctions de base par les données. En effet, le modèle à effets aléatoires (4.8) peut s’interpréter comme un modèle de régression. J’ai ainsi appliqué une méthode de sélection de variables pour résoudre le problème du choix des fonctions de bases. Plus précisément, l’approche consiste à construire un grand nombre de fonctions de bases candidates, à support compact afin de générer des matrices  $P$  creuses, et permettant de capturer d’éventuelles anisotropies locales, puis à les sélectionner via le LASSO ([Hastie et al., 2009]). L’estimation des paramètres du modèle se fait en plusieurs étapes et repose sur une décomposition de la fonction aléatoire selon ses échelles de fluctuations. Dans un premier temps, on calcule le variogramme des données à petite distance et on y ajuste un modèle de covariance à support compact, incluant éventuellement un effet pépète, qui modélisera les fluctuations à petite échelle et dont la portée est fixée de manière à assurer la spar-

sité de la matrice de corrélation associée, notée  $A_\theta$ . Cet ajustement fournit en particulier une valeur d'initialisation de la variance de la composante de petite échelle. Ensuite, on sélectionne les fonctions de base qui vont permettre d'expliquer la variance résiduelle via le LASSO. Une fois les fonctions de bases fixées, l'estimation des paramètres se fait par maximum de vraisemblance via l'algorithme EM. En faisant apparaître les coefficients  $\eta$  comme variables latentes, on peut écrire la loi du couple  $(z, \eta)$  en randomisant par rapport à  $\eta$ . En effet, on a :

$$\begin{aligned}\eta &\sim \mathcal{N}(0, B) \\ Z|\eta &\sim \mathcal{N}(P^t\eta, \sigma^2 A_\theta).\end{aligned}$$

On peut donc estimer la matrice  $B$  et la variance de la composante à petite échelle  $\sigma^2$ , dont on note le couple  $\Theta$  en maximisant la fonction objectif suivante :

$$\begin{aligned}Q(\Theta, \Theta_l) &= -\frac{1}{2\sigma^2}z^t A_\theta^{-1}z + \frac{1}{\sigma^2}z^t A_\theta^{-1}P\mu - \frac{1}{2}\ln \det(\sigma^2 A_\theta) \\ &\quad - \frac{1}{2}\text{Tr}\left(\left(\frac{1}{\sigma^2}P^t A_\theta^{-1}P + B^{-1}\right)(C + \mu\mu^t)\right) - \frac{1}{2}\ln \det(B),\end{aligned}\quad (4.12)$$

où  $\text{Tr}(X)$  est la trace de la matrice  $X$ ,  $C = \mathbb{V}(\eta|Z = z, \Theta_l) = \left(\frac{1}{\sigma_l^2}P^t A_\theta^{-1}P + B_l^{-1}\right)^{-1}$  et  $\mu = \mathbb{E}(\eta|Z = z, \Theta_l) = \frac{1}{\sigma_l^2}C P^t A_\theta^{-1}z$ .

Les formules de mise à jour des paramètres  $\sigma^2$  et  $B$  sont :

$$\begin{aligned}\sigma_{\text{new}}^2 &= \frac{1}{n}\mathbb{E}\left((Z - P\xi)^t A_\theta^{-1}(Z - P\xi)|Z = z, \Theta_l\right) \\ &= \frac{1}{n}\left(z^t A_\theta^{-1}z - 2z^t A_\theta^{-1}P\mu + \text{Tr}\left((P^t A_\theta^{-1}P + B^{-1})(C + \mu\mu^t)\right)\right),\end{aligned}\quad (4.13)$$

$$B_{\text{new}} = C + \mu\mu^t. \quad (4.14)$$

A l'issue de l'algorithme EM, on peut calculer la prédiction en un nouveau point  $x_0$  selon :

$$\mathbb{E}(Z(x_0)|Z = z) = P_0^t\mu + A_0 A^{-1}(z - P^t\mu), \quad (4.15)$$

où  $\mu = \mathbb{E}(\eta|Z = z)$ ,  $P_0$  est le vecteur des fonctions de base au point  $x_0$  et  $A_0$  est la corrélation de la structure à petite échelle entre le point cible et les observations. De même, la variance de krigeage s'obtient en calculant :

$$\mathbb{V}(Z(x_0)|Z = z) = \sigma^2 - A_0 A^{-1}A_0 - 2A_0 A^{-1}P^t C P_0 + P_0^t C P_0, \quad (4.16)$$

où  $C = \mathbb{V}(\eta|Z = z)$ . Ces formules sont peu coûteuses numériquement puisqu'elles ne nécessitent que l'inversion des matrices  $A$ , qui est creuse, et  $C$ , de taille  $p \times p$  avec  $p$  le nombre de fonction de base sélectionnées. Bien que la méthode ait été développée de manière à inverser la matrice de covariance par la formule de Woodbury-Morrison (4.10), on peut remarquer qu'elle n'est finalement pas explicitement employée pour calculer le prédicteur.

Cet algorithme a été appliqué à l'exemple précédent et la carte prédite est affichée en figure 4.6. On peut voir en particulier que l'approche permet de gommer certains défauts du covariance tapering et d'offrir plus de détail que l'approximation de rang faible. Par ailleurs, le fait de sélectionner les fonctions de base permet de réduire leur nombre et ainsi d'accélérer les calculs par rapport à une approche où elles sont fixées arbitrairement et en général plus nombreuses.

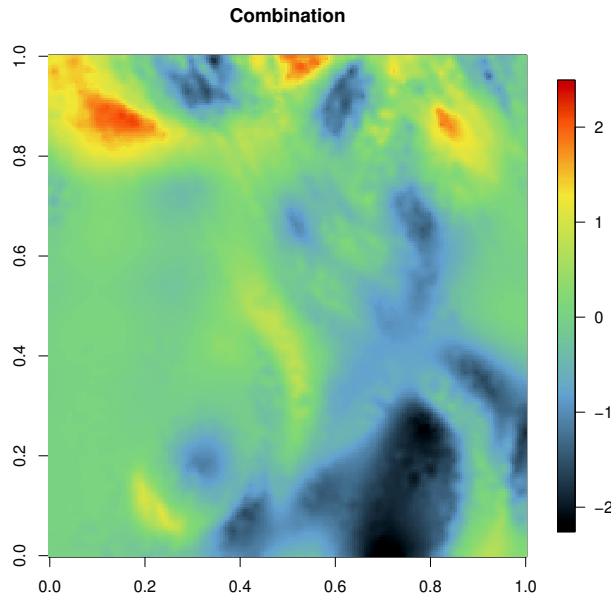


FIGURE 4.6 – Krigeage obtenu par une combinaison de covariance tapering et d'approximation de rang faible à partir d'un échantillon de 5000 points répartis uniformément

#### 4.4.4 Approximations de Vecchia

Plus récemment, des auteurs ([Datta et al., 2016, Katzfuss and Guinness, 2021]) ont remis au goût du jour les approximations de Vecchia ([Vecchia, 1988, Stein et al., 2004]) qui offrent une approximation de la vraisemblance adaptée au traitement de grands jeux de données en la décomposant en produit de vraisemblances conditionnelles :

$$f(\mathbf{z}) = f(z_1) \prod_{i=2}^n f(z_i | \mathbf{z}_{(i-1)}). \quad (4.17)$$

On obtient l'approximation en limitant le nombre de points conditionnant qui augmente en principe avec  $i$  :

$$\tilde{f}(\mathbf{z}) = f(z_1) \prod_{i=2}^n f(z_i | \mathbf{z}_{b(i-1)}), \quad (4.18)$$

TABLE 4.1 – Score de prédiction des différentes approches

model	RMSE
Covariance tapering	0.22
Approximation de rang faible	0.19
Combinaison	0.09
Approximation de Vecchia	0.05
Espérance conditionnelle	0.04

où  $b(i - 1)$  constitue un sous-ensemble de  $\{1, \dots, i - 1\}$ . On peut voir cette approche comme un cas particulier des méthodes de vraisemblance composite ([Varin et al., 2011]). Tandis que ces approximations étaient restreintes à l'estimation de paramètres, Datta et al. [2016] ont montré que (4.18) permet de définir une fonction aléatoire gaussienne et proposent une méthode de prédiction associée. L'idée est d'inclure les valeurs à prédire dans (4.18) et est décrite en détail dans Katzfuss et al. [2020]. L'avantage est que les matrices impliquées auront une structure creuse, assurant des calculs efficaces tout en assurant une bonne approximation. De plus, Katzfuss and Guinness [2021] montrent que l'on peut calculer directement la décomposition de Cholesky de la matrice de précision (inverse de la matrice de covariance), ce qui simplifie encore les calculs. Ils montrent en outre qu'une extension de l'approximation de Vecchia permet de comparer dans un cadre unifié un grand nombre d'approches dédiées aux grands jeux de données. Les principaux paramètres de ces approximations sont l'ordre dans lequel les données sont arrangées et la fonction  $b$  définissant le voisinage employé. Datta et al. [2016] préconise d'utiliser les  $m$  points les plus proches tandis que Guinness [2018] propose un schéma plus élaboré, qu'il valide en comparant la distance de Kullback-Leibler vis-à-vis de la densité cible avec différents concurrents. Ces développements sont accessibles dans le package R GpGp où une optimisation par fisher scoring est également disponible pour l'estimation de paramètres ([Guinness, 2021]). J'ai appliqué cet algorithme à l'exemple précédent en postulant un modèle de Matérn isotrope. La carte prédite est représentée en figure 4.7. Le paramètre de régularité estimé est proche de 1, ce qui induit que la prédiction est un peu trop lisse en haut de la figure mais le résultat est globalement de très bonne facture.

Le tableau 4.1 présente la racine carrée de l'erreur quadratique moyenne obtenue par les différentes approches sur l'exemple proposé, où l'espérance conditionnelle est calculée sans approximation avec le modèle ayant servi à générer les données. On peut voir que bien que la combinaison que j'ai proposée ci-dessus améliore largement les résultats obtenus par covariance tapering et approximation de rang faible, l'approximation de Vecchia obtient un score encore meilleur, très proche de celui de l'espérance conditionnelle. On peut donc se poser la question de l'intérêt de tenir compte de la non-stationnarité dans ce type d'exercice ([Fuglstad et al., 2015]).



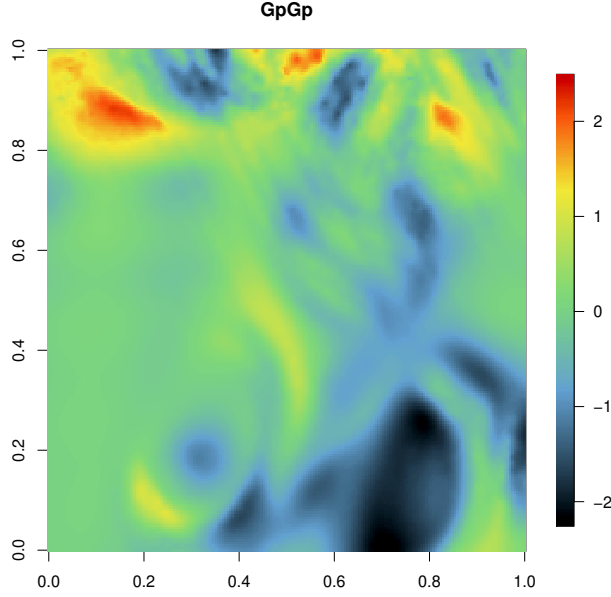


FIGURE 4.7 – Krigeage obtenu par approximation de Vecchia à partir d'un échantillon de 5000 points répartis uniformément

#### 4.4.5 Equations aux dérivées partielles stochastiques

Une dernière approche dédiée au traitement de grands jeux de données est celle des équations aux dérivées partielles stochastiques (EDPS) qui a été popularisée par Lindgren et al. [2011]. Deux thèses ont eu lieu récemment sur ce sujet dans l'équipe de géostatistique ([Vergara, 2018, Pereira, 2019]). Sur ce sujet, j'ai exploré les possibilités offertes par l'emploi d'éléments finis d'ordre supérieur ([Romary and Desassis, 2018b]). Le point de départ de cette approche est un résultat dû à Whittle [1963] qui stipule que si  $Z$  est une fonction aléatoire stationnaire en dimension  $d$  de fonction de covariance de Matérn de paramètres de régularité  $\nu$ , d'échelle  $\kappa$  et de variance  $\sigma^2$ , alors  $Z$  est solution de l'EDPS :

$$(\kappa^2 - \Delta)^{\alpha/2} Z(x) = \mathcal{W}(x), \quad x \in \mathcal{X} \quad (4.19)$$

où  $\mathcal{X} \subset \mathbb{R}^d$  est le domaine de définition de  $Z$ ,  $\Delta = \sum_{i=1}^d \frac{\partial^2}{\partial x_i^2}$  est le Laplacien,  $\mathcal{W}(x)$  est une fonction aléatoire gaussienne standard structurée ou non. Si  $W$  est un bruit blanc gaussien, en notant,  $\alpha = \nu + d/2$  et  $\sigma^2 = \frac{\Gamma(\nu)}{\Gamma(\nu+d/2)(4\pi)^{d/2}\kappa^{2\nu}}$ , la fonction de covariance de Matérn associée s'écrit :

$$C(h) = \frac{\sigma^2}{2^{\nu-1}\Gamma(\nu)} (\kappa h)^\nu \mathcal{K}_\nu(\kappa h), \quad (4.20)$$

où  $\mathcal{K}_\nu$  est la fonction de Bessel modifiée de second espèce.

On peut généraliser (4.19) en remplaçant le terme de gauche par un opérateur différentiel de la forme  $P^{1/2}(-\Delta)$  avec  $P$  un polynôme. En discrétisant cette équation selon un schéma aux éléments finis, on obtient une approximation de la matrice de précision  $Q$  des coefficients associés aux fonctions de base qui sera creuse par construction. Ainsi, le vecteur  $Z$  des coefficients est un champ gaussien markovien ([Guyon, 1995, Rue and Held, 2005]).

Pour la prédiction, en considérant que les points de données et les points cibles sont des sommets de la discrétisation, on forme le vecteur  $Z = (Z_D, Z_T)$  contenant les données  $Z_D$  et les points cibles  $Z_T$ . La matrice de précision de  $Z$  s'écrit :

$$Q = \begin{pmatrix} Q_{DD} & Q_{DT} \\ Q_{TD} & Q_{TT} \end{pmatrix}. \quad (4.21)$$

On obtient une prédiction aux points cibles en calculant l'espérance conditionnelle :

$$Z_T^* = -Q_{TT}^{-1}Q_{TD}Z_D \quad (4.22)$$

tandis que la covariance conditionnelle est donnée par  $Q_{TT}^{-1}$ . Cette approche est très efficace en terme de temps de calcul puisqu'elle ne fait intervenir que des matrices creuses. Elle se prête également très bien à la modélisation d'anisotropies locales que l'on peut introduire aisément dans (4.19) en remplaçant l'opérateur  $\Delta$  par  $\nabla \cdot H(x)\nabla$ , où  $H(x)$  est une matrice d'anisotropie locale [Pereira, 2019]). En contrepartie, il est délicat de travailler avec des exposants  $\alpha$  non entiers qui font apparaître des opérateurs différentiels fractionnaires et induisent des champs non markoviens pour lesquels des approximations sont alors nécessaires ([Rozanov, 1977, Bolin and Kirchner, 2017]).

## 4.5 Modélisation spatio-temporelle

Lorsqu'on travaille avec des données spatio-temporelles, les équations du krigeage (4.2) et de sa variance (4.3) ne diffèrent en rien du cas spatial. La principale difficulté est de disposer d'un modèle adapté. En spatial pur en effet, les modèles de covariances disponibles permettent d'ajuster une anisotropie et une régularité, éventuellement variables localement, ce qui autorise une large gamme de structures spatiales. La dimension temporelle n'est cependant pas une simple dimension additionnelle, les phénomènes que l'on souhaite modéliser en géostatistique présentent une structure spatio-temporelle dépendant de la physique qui les décrit, en particulier des interactions complexes entre les dimensions spatiale et temporelles. Il convient donc d'employer des modèles adaptés pour décrire ces structures. J'ai abordé cette thématique au travers de ma participation au réseau RESSTE (RESeau Statistiques pour données Spatio-TEmporelles). Dans ce cadre, j'ai notamment participé à la rédaction de Network [2017] qui propose une revue des méthodes d'analyse de données spatio-temporelles disponibles sous R.

La plupart des modèles de covariance spatio-temporels employés dans la pratique sont construits en modifiant ou en combinant des modèles de covariances génériques. A ce titre, on peut citer les modèles séparables, les modèles somme-produit ([De Iaco et al.,

2001]), Iaco-Césaire [De Iaco et al., 2002] ou encore de type Porcu ([Porcu et al., 2009]). Les modèles dits de Gneiting ([Gneiting, 2002b]) proposent une famille flexible de modèles non-séparables

$$C_{ST}(h, u) = (\psi_T(u^2) + 1)^{-\delta/2} C_S(\|h\|^2 / \{\psi_T(u^2) + 1\}), \quad (4.23)$$

où  $h$  est la distance spatiale,  $u$  la distance temporelle,  $\psi_T$  est une fonction de la forme  $\psi_T(u) = c + \gamma_T(u)$  avec  $c > 0$  et  $\gamma_T(\cdot)$  un variogramme temporel,  $C_S(\cdot)$  une fonction de covariance qui doit appartenir à la famille des mélanges de covariances gaussiennes (exponentiel, Matérn, Cauchy, ...). On pourra se référer à Porcu et al. [2021] pour une revue exhaustive, incluant notamment les modèles valides sur la sphère, utiles dans les applications à l'échelle du globe.

Le modèle de Gneiting, que l'on peut considérer aujourd'hui comme le couteau suisse de la modélisation spatio-temporelle au même titre que le modèle de Matérn en spatial, de même que les différents modèles évoqués ci-dessus, est une construction théorique qui ne permet pas de tenir compte de la physique des phénomènes que l'on cherche à modéliser, en particulier d'une dynamique temporelle particulière. Une voie prometteuse dans cette optique est l'utilisation de SPDE spatio-temporelles qui permettent d'introduire des contraintes physiques dans les modèles, notamment des phénomènes de diffusion et d'advection ([Bakka et al., 2020, Vergara, 2018, Sigrist et al., 2015]). Une SPDE d'advection-diffusion prend par exemple la forme suivante ([Vergara et al., 2021]) :

$$\frac{\partial Z}{\partial t} + \kappa^2 Z + v^t \nabla Z - \nabla \cdot (Q \nabla Z) = W_S \otimes W_T, \quad (4.24)$$

où  $v$  est une vitesse,  $Q$  est une matrice d'anisotropie,  $W_T$  est un bruit blanc spatial et  $W_S$  est une fonction aléatoire structurée, ce qui permet de contrôler la structure de la trace spatiale de  $Z$ . Sigrist et al. [2015], propose une discrétisation de (4.24) en éléments finis spectraux, ce qui permet de simuler aisément une telle fonction aléatoire mais rend délicat le conditionnement.

## 4.6 Discussion

Ce chapitre a permis de décrire un certain nombre d'aspects de la modélisation spatiale et spatio-temporelle en géostatistique dans lesquels je me suis efforcé d'apporter une contribution. En ce qui concerne la prédiction spatiale en particulier, on a pu voir qu'un grand nombre d'approches coexistent. Les implémentations récentes de l'approximation de Vecchia semblent aujourd'hui se démarquer en proposant des algorithmes très performants aussi bien pour l'estimation de paramètres que pour la prédiction.

Dans ce qui précède, j'ai considéré que le terme d'espérance ou du moins les variables explicatives pertinentes étaient connues. Ce n'est évidemment que rarement le cas en pratique. Il peut être en particulier intéressant d'employer des approches de machine learning pour ajuster la dérive. On peut par exemple effectuer une sélection de variables par LASSO par exemple si celles-ci sont nombreuses ou bien employer des prédicteurs plus complexes comme les random forests ([Breiman, 2001, Biau and Scornet, 2016]).

Certains travaux récents explorent la manière d'effectuer directement une prédiction spatiale au travers des random forests ([Fouedjio and Klump, 2019]) ou bien des réseaux de neurones ([Li et al., 2020, Wang et al., 2019a]). Outre les variables explicatives disponibles, les coordonnées et d'autres prédicteurs, par exemple des fonctions arbitraires des coordonnées, sont construits pour alimenter ces algorithmes. J'ai pu tester de manière expérimentale ce type d'approches et ai été assez déçu des résultats obtenus. On perd d'une part naturellement la propriété d'interpolation exacte inhérente au krigeage et on s'expose par ailleurs à des comportements imprévisibles en extrapolation. La question de combiner idéalement les approches de type machine learning avec la géostatistique reste ouverte.

On a vu dans l'exemple ci-dessus qu'un modèle de covariance isotrope (approché) permettait finalement d'obtenir de meilleurs résultats que lorsqu'on cherchait à capturer les non stationnarités. Ce phénomène est assez typique de ce qu'on observe en pratique quand on cherche à estimer des non-stationnarités du second ordre. En effet, si on applique une méthode d'estimation de covariance non-stationnaire à un jeu de données généré selon un modèle stationnaire, on risque de détecter des non stationnarités. Autrement dit, on risque d'interpréter comme des non stationnarités des comportements liés à la variabilité naturelle de la fonction aléatoire, ce qui s'apparente à un phénomène de surapprentissage ([Fuglstad et al., 2015]). Aussi, je préconise de réserver ces approches aux cas où la non-stationnarité du second ordre est flagrante et/ou lorsque celle-ci peut-être expliquée par des variables auxiliaires ([Schmidt et al., 2011]). Pour guider ces choix de modélisation, si un test, basé sur la représentation spectrale de la fonction aléatoire, a été proposé pour s'assurer de la stationnarité des données ([Fuentes, 2005]), il ne permet pas d'indiquer s'il convient d'en tenir compte dans la modélisation.

Lorsqu'on traite des jeux de données spatio-temporelles, on se retrouve souvent confronté à des problèmes liés à leur taille. Aussi, les approximations évoquées dans la partie 4.4 sont également adaptées à ce contexte. En particulier, les approches de rang faible, basées sur le modèle à effet aléatoire spatial (4.8), se prêtent bien à une extension spatio-temporelle où la dynamique sera portée par les coefficients des fonctions de base ([Katzfuss and Cressie, 2011]). Un tapering spatial appliqué au modèle de Gneiting par exemple peut aussi être adapté, où on restreindra également le voisinage temporel. Enfin, les approximations de Vecchia peuvent être étendues au contexte spatio-temporel en adaptant la fonction de voisinage. Cependant, toutes ces approches ne permettent pas d'imposer une dynamique qui respectera des contraintes physiques. A ce titre, Cressie and Wikle [2011] et Wikle et al. [2019] distinguent les modèles descriptifs ou statistiques des modèles dynamiques où l'évolution de la fonction aléatoire est décrite au travers de l'application d'un noyau, linéaire ou non. L'approche SPDE spatio-temporelle décrite ci-dessus entre dans cette seconde catégorie. L'avantage de l'approche dynamique est de pouvoir tenir compte de la physique du phénomène et d'offrir ainsi l'espoir de meilleures performances prédictives.



# Chapitre 5

## Inversion bayésienne

### 5.1 Problématique

En sciences de la terre, on est fréquemment amené à développer des modèles phénoménologiques pour quantifier, comprendre et prévoir les manifestations de la nature. Parmi les applications auxquelles je me suis intéressé, on peut citer l'ingénierie de réservoir, où si on cherche à prédire la production de pétrole d'un gisement à un horizon de temps, on aura besoin de simuler les écoulements des fluides (eau, gaz, pétrole) au sein du réservoir. En géophysique, on cherchera à modéliser la propagation des ondes dans le sous-sol. En qualité de l'eau, on pourra s'intéresser au métabolisme des cours d'eau afin de suivre l'évolution de leurs caractéristiques (taux d'oxygène dissous par exemple) dans le temps.

Ces modèles, une fois implémentés, nécessitent la plupart du temps un certain nombre de variables d'entrée. Dans les exemples précédents, la répartition spatiale des propriétés pétrophysiques du réservoir (porosité, perméabilité) est requise en entrée du simulateur d'écoulement. De même, la vitesse de propagation des ondes dans le sous-sol doit être renseignée pour simuler leur propagation. Enfin, un certain nombre de paramètres gouvernant le métabolisme des cours d'eau ainsi que les conditions aux limites du système sont nécessaires pour simuler l'évolution des quantités d'intérêt.

Dans l'ensemble de ces problèmes, les variables d'entrée, que nous noterons  $Z$ , à valeurs dans  $\mathcal{Z}$  sont imparfaitement connues, soit de manière parcellaire quand il s'agit de variables régionalisées, soit de manière incertaine lorsqu'il s'agit de variables mesurées en laboratoire. On peut vouloir simplement analyser l'influence de la variation des entrées sur la sortie du modèle, dans le cadre d'une analyse de sensibilité, afin de mieux comprendre son fonctionnement et éventuellement corriger des erreurs. On peut vouloir également caractériser ces variables de manière à ce que le modèle soit capable de reproduire des données observées. On parle alors généralement d'ajustement ou bien de calage aux données.

Les données  $Y \in \mathcal{Y}$  observées dans la pratique sont souvent celles que l'on sait simuler par les modèles phénoménologiques, de manière plus ou moins précise. On peut les considérer comme des sorties bruitées du modèle, où le bruit  $\varepsilon$  tiendra compte à la fois de l'erreur de mesure, due au capteur, et d'une éventuelle erreur de modélisation. Le problème consiste alors à tirer parti des données pour mieux caractériser les entrées et donc l'état

du système étudié, d'où le nom de problème inverse.

On formalise ce problème de la manière suivante :

$$Y = F(Z) + \epsilon,$$

où  $F$  représente le modèle phénoménologique qu'on appellera également le modèle direct. Le modèle  $F$  nous permet de calculer  $Y$  connaissant  $Z$ , à une erreur près. Pour caractériser  $Z$  à partir de  $Y$ , on doit donc inverser la fonction  $F$ . Si  $F$  est bijective, d'inverse continue, alors on peut estimer la valeur de  $Z$  en présence d'une erreur  $\epsilon$ . On dit que le problème est bien posé. Ce cas de figure est malheureusement très rare en pratique et on doit traiter des problèmes mal posés. Dans ce cas,  $F$  est souvent surjective, c'est-à-dire que la dimension de l'espace des variables inconnues est plus grande que celui des données.

Classiquement, ce type de problème est traité par l'optimisation d'une fonction objectif qui prend par exemple la forme suivante :

$$J(Z) = \|Y - F(Z)\|^2 + \lambda O(Z) \tag{5.1}$$

où  $\|\cdot\|$  est la norme euclidienne,  $O(Z)$  est une fonctionnelle régularisante et  $\lambda$  un paramètre. La régularisation est nécessaire pour stabiliser le problème lors d'une trop grande sensibilité aux variations de  $Z$  et en cas de non unicité de la solution.

En passant par une modélisation probabiliste, on s'aperçoit que ce problème revient à une estimation du maximum a posteriori dans un cas particulier. En effet, en supposant que  $Z$  et  $Y$  admettent une densité (strictement positive pour  $Y$ ), la formule de Bayes donne :

$$f(Z|Y) = \frac{f(Y|Z)f(Z)}{f(Y)} \tag{5.2}$$

Faisant l'hypothèse que  $Y|Z \sim \mathcal{N}(F(Z), I)$  et prenant  $\lambda O(Z) = \log(f(Z))$ , le log de cette expression permet bien de retrouver (5.1). Ainsi, le cadre probabiliste permet d'étendre l'approche par optimisation et de viser un résultat sous la forme d'une densité conditionnelle. Cette loi conditionnelle résume l'information apportée par  $Y$  sur la connaissance de  $Z$ . On peut dans certains cas obtenir une expression analytique de cette densité conditionnelle (cas gaussien où  $F$  est linéaire bijectif par exemple), mais en général, on est seulement capable d'évaluer (5.2) en certains points et à la constante de normalisation  $f(Y)$  près. On va donc chercher à approximer  $f(Z|Y)$ . Pour ce faire, on passe par des méthodes de Monte-Carlo. L'échantillon ainsi généré pourra être utilisé dans des tâches de prédiction. Cette approche a été popularisée dans les années 80 par Alberto Tarantola [Tarantola, 2005] et c'est dans ce cadre que se place un certain nombre de mes travaux.

Ce chapitre décrit mes contributions dans la résolution de certains problèmes inverses en géosciences. La démarche que je préconise et que j'ai appliquée dans les différentes applications présentées consiste à simplifier le problème au préalable par des techniques de réduction de la dimension, décrites dans la section 5.2, puis à appliquer des algorithmes d'inférence bayésienne afin d'échantillonner la loi a posteriori. Ceux ci sont présentés dans le paragraphe 5.3. Enfin, je décris en 5.4 un type particulier de problème inverse dans lequel le modèle direct suit une dynamique markovienne, ce qui permet de le traiter séquentiellement. Ce chapitre synthétise les travaux publiés dans les articles suivants :

- J. Langanay, **T. Romary**, X. Freulon, V. Langlais, G. Petit, V. Lagneau *Uncertainty quantification for uranium production in mining exploitation by In Situ Recovery*, Computational Geosciences, pages 1–20, 2021.
- Wang, S., Flipo, N., & **Romary, T.** Oxygen data assimilation for estimating microorganism communities' parameters in river systems. *Water research*, 165, 115021. (2019)
- S. Wang, N. Flipo, **T. Romary**. *Time-dependent global sensitivity analysis of the C-RIVE biogeochemical model in contrasted hydrological and trophic contexts*, *Water research* 144, 341-355. (2018)
- J. Belhadj, **T. Romary**, A. Gesret, M. Noble, B. Figliuzzi. *New parameterizations for Bayesian seismic tomography*, *Inverse Problems*. (2018)
- A. Bottero, A. Gesret, **T. Romary**, M. Noble, C. Maisons. *Stochastic seismic tomography by interacting Markov chains*. *Geophysical Journal International*, 207(1), 374–392. (2016)
- A. Gesret, N. Desassis, M. Noble, **T. Romary**, C. Maisons, *Propagation of the velocity model uncertainties to the seismic event location*, *Geophysical Journal International* 200 (1), 52-66 (2015)
- S. Girard, **T. Romary**, J.-M. Favennec, P. Stabat, H. Wackernagel, *Sensitivity analysis and dimension reduction of a steam generator model for clogging diagnosis*, *Reliability Engineering & System Safety*, 113, 143-153 (2013)
- **T. Romary**, *History matching of approximated lithofacies models under uncertainty*. *Computational Geosciences* 14(2), 343-355 (2010)
- **T. Romary**, *Bayesian inversion by parallel interacting Markov chains*. *Inverse Problems in Science and Engineering* 18(1), 111-130 (2010)
- **T. Romary**, *Integrating production data under uncertainty by parallel interacting Markov chains on a reduced dimensional space*. *Computational Geosciences* 13(1), 103-122 (2009)

## 5.2 Paramétrisation

La simulation selon (5.2) est d'autant plus complexe que la dimension de  $Z$  est grande. Aussi, il convient de chercher à réduire la dimension de  $Z$ , en employant une *paramétrisation* aussi parcimonieuse que possible.

Les modèles phénoménologiques développés en géosciences disposent de jeux d'entrées qui leur sont propres. Il est donc nécessaire de définir la paramétrisation au cas par cas. Certains principes généraux peuvent cependant être tirés de la pratique de ces modèles. On distingue ici deux cas que j'ai fréquemment traités dans mes travaux. Dans le premier, le modèle phénoménologique dépend d'un certain nombre de paramètres scalaires, modélisés par des variables aléatoires indépendantes, dont on peut chercher à hiérarchiser l'influence sur la sortie par des méthodes d'analyse de sensibilité. Dans le second, l'entrée d'intérêt du modèle est une fonction aléatoire, dont on va chercher à réduire la dimension.



### 5.2.1 Analyse de sensibilité

Une première approche peut être de hiérarchiser l'influence des différents paramètres d'entrée sur la sortie du modèle, afin de sélectionner les plus pertinents. Les méthodes d'analyse de sensibilité ([Saltelli et al., 2000]) ont justement été développées dans cette optique. Elles permettent de quantifier l'influence des différentes entrées sur la sortie d'un modèle. De nombreuses méthodes existent ([Iooss and Lemaître, 2015]), et le choix dépend de la complexité du modèle (non linéarité, discontinuités) et du nombre de paramètres. J'ai employé ces méthodes dans deux articles Girard et al. [2013] et Wang et al. [2018]. Dans ce second article, une analyse de sensibilité a été conduite en préalable à un travail d'assimilation de données (que l'on peut voir comme un type particulier de problème inverse et que nous décrirons en fin de chapitre), afin de sélectionner les paramètres les plus influents qui seront ceux que l'on cherchera ensuite à caractériser. Ce travail est résumé ci-après afin d'illustrer les méthodes.

Dans Wang et al. [2018], nous avons conduit une analyse de sensibilité de la librairie bio-géochimique C-RIVE, dont le fonctionnement est illustré sur la figure 5.1. Cette librairie est basée sur le modèle communauté centré RIVE [Billen et al., 1994, Garnier et al., 1995] (<https://www.fire.upmc.fr/rive>). Le modèle RIVE simule le cycle des nutriments, du carbone et de l'oxygène dissous au sein de la colonne d'eau ainsi que dans une couche de sédiments non consolidés. Ce modèle comprend 31 paramètres d'entrée que l'on peut séparer en 3 familles :

- les paramètres physiques : les coefficients de réaération due au vent et à la navigation, le coefficient d'extinction de la lumière, etc.
- les paramètres bactériens : les taux de croissance et de mortalité, le rendement de croissance, etc.
- les paramètres phytoplanctoniques : la température optimale de croissance, le taux de photosynthèse, etc.

Ces différents paramètres gouvernent en particulier l'évolution de la concentration en oxygène dissous qui est un bon indicateur de la qualité de l'eau d'une rivière. Le grand nombre de paramètres d'entrée nous a conduit à employer dans un premier temps une technique de screening, dite méthode de Morris ([Morris, 1991, Campolongo et al., 2007]). Cette méthode est basée sur le calcul d'effets élémentaires qui sont ensuite moyennés pour évaluer l'influence globale de chaque entrée du modèle. Elle permet ainsi de trier rapidement les paramètres selon leur influence et d'éliminer ceux qui contribuent le moins à la variabilité de la réponse.

Dans la méthode de Morris, on fait l'hypothèse que les paramètres sont indépendants et uniformément distribués sur  $[0,1]$ , quitte à les transformer par l'inverse d'une fonction de répartition donnée le cas échéant. Chacun des paramètres est discrétisé en  $n$  niveaux.

L'effet élémentaire  $E_j$  du paramètre  $Z_j$ ,  $j = 1, \dots, p$ , est défini de la manière suivante :

$$E_j = \frac{F(z + \Delta e_j) - F(x)}{\Delta e_j}$$

où  $\Delta$  est une valeur choisie parmi  $\{1/(n-1), \dots, 1-1/(n-1)\}$  et  $e_j$  est le  $j^{\text{ème}}$  vecteur de la base canonique.

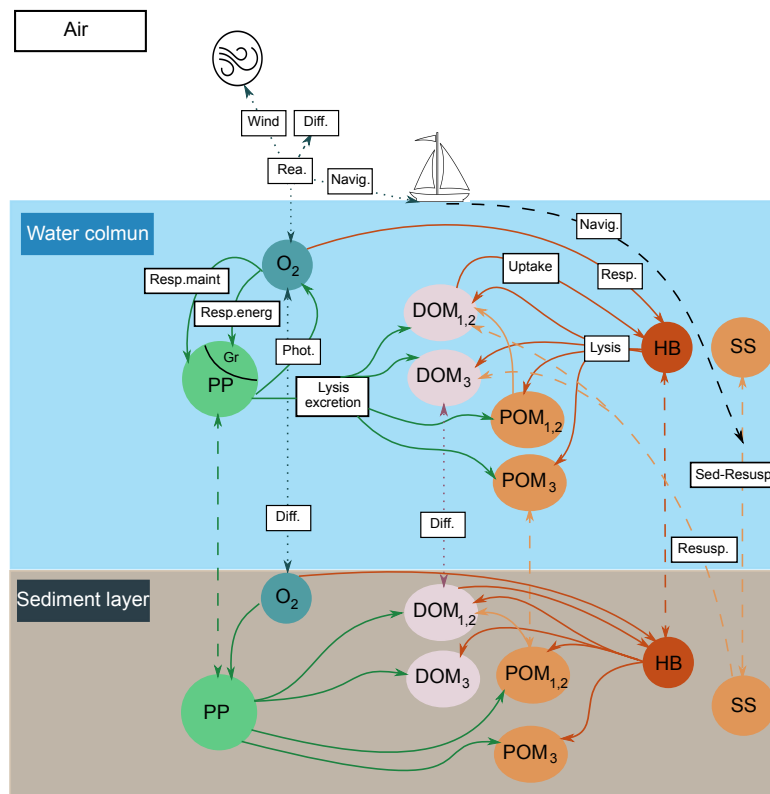


FIGURE 5.1 – Fonctionnement schématique du modèle C-RIVE.

Les effets élémentaires sont calculés pour  $r$  répétitions d'une trajectoire aléatoire dans  $[0, 1]^p$  discrétisé, où chacun des paramètres est modifié une fois à la fois. On parle de plan d'expérience "one at a time" (OAT).

A partir des effets élémentaires, on calcule deux indices qui permettent de mesurer l'influence de chacun des paramètres. La moyenne des valeurs absolues des effets élémentaires :

$$\mu_j^* = \frac{1}{r} \sum_{i=1}^r |E_j^i|$$

et leur écart-type

$$\sigma_j = \sqrt{\frac{1}{r} \sum_{i=1}^r (E_j^i - \frac{1}{r} \sum_{i=1}^r E_j^i)^2}$$

$\mu_j^*$  permet d'estimer l'impact global du paramètre  $Z_j$  sur la sortie du modèle tandis que  $\sigma$  mesure un éventuel effet non linéaire et/ou des effets d'interaction avec d'autres paramètres.

Pour interpréter les résultats, on représente  $\mu_j^*$  en fonction de  $\sigma_j$ . La méthode de Morris permet de classer les paramètres en trois familles selon leur influence sur la sortie du modèle :

- les paramètres ayant un effet négligeable avec de faibles valeurs de  $\mu^*$  et  $\sigma$ ,
- les paramètres ayant des effets non linéaires et/ou des effets d'interaction avec d'autres paramètres,
- Les paramètres ayant des effets linéaires et additifs avec une valeur forte de  $\mu^*$  et un  $\sigma$  faible.

Nous avons appliqué cette approche pour étudier l'évolution de la concentration en oxygène dissous d'un volume d'eau se déplaçant le long de la Seine simulée avec C-RIVE, selon différents états trophiques, à savoir un régime de hautes eaux hivernal, lors d'un bloom (prolifération) algal ainsi qu'avant et après ce bloom.

On peut remarquer que la méthode de Morris, de même que la méthode de Sobol' présentée plus loin, permet de traiter des sorties scalaires. Cependant, la sortie qui nous intéresse ici est un vecteur de concentrations en oxygène dissous à différents pas de temps. Il est donc nécessaire de réduire la dimension. En pratique, nous avons réalisée une ACP sur les sorties et pratiqué l'analyse de sensibilité sur la somme des coefficients de chacune des trajectoires dans les trois premières composantes, qui représentaient 99% de la variation totale. (cf. [Lamboni et al., 2011]). Il est bien évidemment possible de calculer les indices de sensibilité à chaque pas de temps mais leur interprétation est alors difficile. L'approche par réduction de dimension permet au contraire d'évaluer une contribution globale de chacun des paramètres d'entrée à la sortie du modèle.

Les résultats présentés en figure 5.2 indiquent qu'en dehors des périodes de bloom, les paramètres les plus influents sont les paramètres bactériens (en marron) qui présentent des effets non-linéaires et/ou d'interaction avec d'autres paramètres ainsi que les paramètres de réaération due au vent et à la navigation qui présentent un effet linéaire ( $K_{wind}$  et  $K_{navig}$ , en bleu). Durant les périodes de bloom, un certain nombre de paramètres phytoplanctoniques (en vert) deviennent en toute logique plus influents, puisque ce sont alors essentiellement les

## 5.2. PARAMÉTRISATION

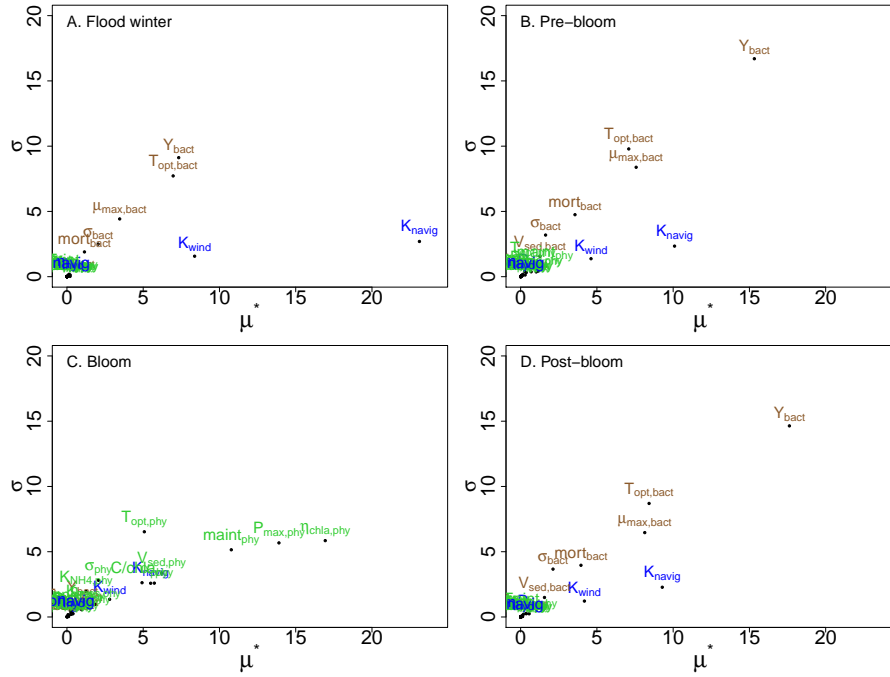


FIGURE 5.2 – Résultats de la méthode de Morris selon différents états trophiques du système

algues qui vont gouverner la concentration en oxygène dissous. Sur les quatre graphiques, on peut voir qu'un grand nombre de paramètres ont un effet négligeable sur la sortie avec des valeurs très faibles de  $\mu^*$  et  $\sigma$ .

Ces résultats nous ont permis d'identifier les paramètres ayant un effet négligeable sur la sortie du modèle. Pour quantifier plus précisément l'influence de chacun des paramètres, on peut nous avons procédé à une analyse de sensibilité selon la méthode de Sobol' [Sobol', 1993].

Cette méthode est basée sur la décomposition de la variance de la sortie du modèle et en particulier sur la représentation suivante :

$$F(Z) = Y = f_0 + \sum_{i=1}^p f_i(Z_i) + \sum_{i=1}^{p-1} \sum_{j=i+1}^p f_{ij}(Z_i, Z_j) + \dots + f_{1,\dots,p}(Z_1, \dots, Z_p), \quad (5.3)$$

où les  $f_{\dots}$  vérifient

$$\int_0^1 f_{i_1,\dots,i_s}(x_{i_1}, \dots, x_{i_s}) dx_{i_k} = 0, \forall k = 1, \dots, s, \forall \{i_1, \dots, i_s\} \subseteq \{1, \dots, p\}. \quad (5.4)$$

Si les paramètres  $(Z_1, \dots, Z_p)$  sont mutuellement indépendants et chaque fonction élémentaire est de carré intégrable et de moyenne nulle, alors  $f_0$  est l'espérance de la sortie du modèle et cette décomposition est unique [Sobol', 1993].

Sous ces hypothèses, on obtient une décomposition de la variance de  $Y$  sous la forme suivante :

$$V(Y) = \sum_{i=1}^p V_i(Z_i) + \sum_{i=1}^{p-1} \sum_{j=i+1}^p V_{ij}(Z_i, Z_j) + \dots + V_{1,\dots,p}(Z_1, \dots, Z_p) \quad (5.5)$$

où  $V_i = V(\mathbb{E}(Y|Z_i))$ ,  $V_{ij} = V(\mathbb{E}(Y|Z_i, Z_j)) - V(\mathbb{E}(Y|Z_i)) - V(\mathbb{E}(Y|Z_j))$ , etc. On appelle ces termes les variances partielles :

$$V_{i_1, \dots, i_s} = \int_0^1 \dots \int_0^1 f_{i_1, \dots, i_s}^2(x_{i_1}, \dots, x_{i_s}) dx_{i_1}, \dots, dx_{i_s}$$

où  $1 \leq i_1 \leq \dots \leq i_s \leq p$  et  $s = 1, \dots, p$ .

On peut alors calculer les indices de Sobol' qui sont donnés par le rapport de la variance partielle sur la variance totale :

$$S_i = \frac{V_i}{V}, S_{ij} = \frac{V_{ij}}{V}, \dots$$

En pratique, on se limite généralement à l'estimation des indices du premier ordre  $S_i$ , pour des raisons d'ordre numérique. En effet, le nombre d'indice croît exponentiellement avec le nombre  $p$  de paramètres. Pour compléter, on estime également les indices totaux :

$$S_{T_i} = S_i + \sum_{j \neq i} S_{ij} + \dots + S_{1,\dots,p}$$

que l'on peut obtenir à moindre coût à partir de plans d'échantillonnage adaptés (cf [Saltelli et al., 2000]).

Les résultats de l'application de la méthode de Sobol' au programme C-RIVE, issus de Wang et al. [2018], sont représentés en figure 5.3. Nous avons omis dans l'analyse les paramètres jugés négligeables dans l'analyse par la méthode de Morris.

L'avantage des indices de Sobol' est qu'ils s'expriment directement comme une part de variance expliquée par chacun des paramètres. Ils permettent ainsi de quantifier précisément leur influence. On s'aperçoit notamment sur la figure 5.3 que le paramètre  $\sigma_{\text{bact}}$ , qui correspond à l'écart-type de la fonction de température des communautés bactériennes, a une influence négligeable. On peut voir également plus finement les changements entre les états pre et post-bloom. En particulier, on s'est rendu compte que l'état pre-bloom fait figure d'intermédiaire entre les hautes eaux hivernales et un régime dominé par l'activité bactérienne. Cette observation nous a conduit à estimer les indices de Sobol' du premier ordre des paramètres les plus influents en fonction de la température et dans deux conditions de débit.

La figure 5.4 représente les indices de Sobol' du premier ordre normalisés pour sommer à un. On voit que lorsque la température est faible, la concentration en oxygène dissous est

## 5.2. PARAMÉTRISATION

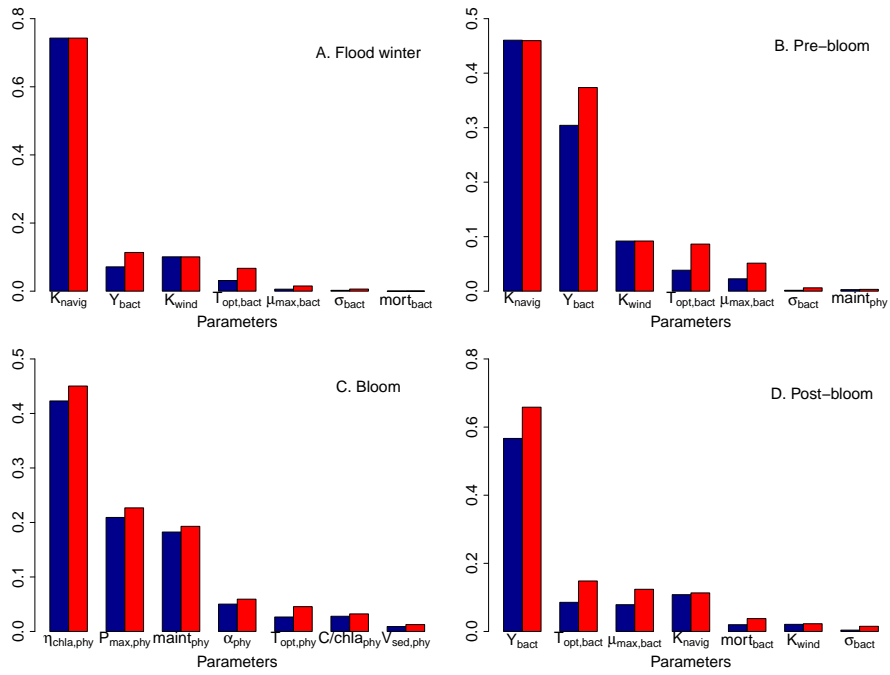


FIGURE 5.3 – Indices de Sobol' selon différents états trophiques du système

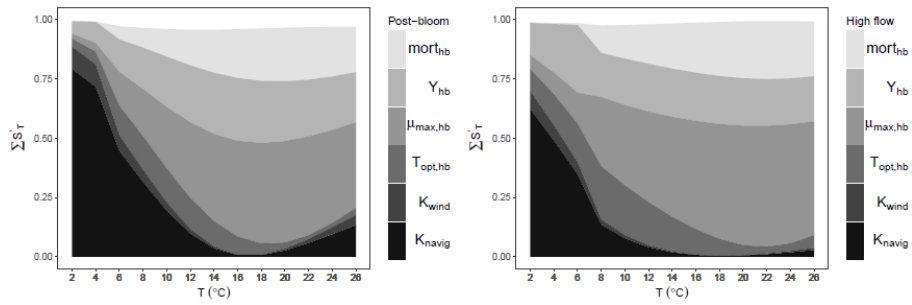


FIGURE 5.4 – Indices de Sobol' en fonction de la température selon deux conditions de débit

principalement dominée par les paramètres de réaération dus au vent et à la navigation, tandis que lorsqu'elle augmente, les paramètres bactériens voient leur influence s'accroître. Les conditions de débit ne semblent en revanche pas avoir une grande influence sur les effets relatifs de ces paramètres.

En résumé, ces analyses de sensibilité nous ont permis d'identifier un sous-ensemble de paramètres influents sur la sortie de C-RIVE et ainsi de réduire la dimension du problème inverse lié à l'assimilation des mesures de concentration en oxygène dissous le long de la Seine qui sera abordée dans le paragraphe 5.4. En outre, ces analyses nous ont permis d'identifier certains comportements aberrants de C-RIVE et de les corriger.

La capacité d'éliminer des paramètres non influents dans les problèmes inverses offerte par les méthodes d'analyse de sensibilité est particulièrement intéressante dans le sens où cela va simplifier le problème de simulation conditionnelle. Les hypothèses induisent cependant un certain nombre de limitations à l'application de ces méthodes.

D'une part, quand la sortie du modèle n'est pas scalaire, il est nécessaire d'en réduire la dimension. Dans l'exemple ci-dessus, une simple ACP permettait de régler le problème. Il arrive souvent dans les applications en géosciences que les sorties soient plus complexes, par exemple sous la forme de champs spatio-temporels pour lesquels une telle compression de l'information est difficile, voire illusoire. On peut alors travailler sur des objectifs de dimension moindre, par exemple l'évolution temporelle en un point du domaine, ou bien l'état du système à un temps donné. En tout état de cause, les résultats deviennent plus difficiles à interpréter. D'autre part, une estimation précise des indices de Sobol' nécessite un grand nombre de simulations du modèle d'intérêt ( $N \times (p + 2)$ , où  $N$  doit être fixé par l'utilisateur) ce qui interdit l'utilisation de ces méthodes pour des modèles complexes, par exemple pour les problématiques de transport réactif à grande échelle qui se rencontrent pour l'exploitation de gisements d'uranium par lixiviation in-situ (Langanay et al. [2021]). Dans ce cas, on doit se limiter à quelques dizaines de simulations qui ne donneront qu'une vue parcellaire des contributions de chacun des paramètres à la variabilité de la sortie.

L'hypothèse d'indépendance des paramètres est également très forte puisque, selon les modèles considérés, certaines combinaisons de valeurs peuvent être physiquement aberrantes. Des méthodes existent pour traiter ces cas de figure aussi bien pour la méthode de Morris ([Ge and Menendez, 2017]) que pour les indices de Sobol' ([Kucherenko et al., 2012, Chastaing et al., 2012]) mais l'interprétation des résultats devient délicate puisque la décomposition de la variance (5.5) n'est plus valable. On peut également chercher à reparamétriser le modèle de manière à ne faire apparaître que des variables indépendantes, au détriment de l'interprétation des résultats. Plus récemment, les indices de Shapley ([Shapley, 1953]), couramment employés pour quantifier l'importance des variables explicatives dans les random forests par exemple, ont été introduits dans un contexte d'analyse de sensibilité ([Owen, 2014, Iooss and Prieur, 2019]). Dans certains cas de figures, enfin, les entrées sont trop complexes pour être appréhendées par des méthodes d'analyse de sensibilité. C'est le cas des codes de calcul prenant en entrée des fonctions aléatoires, fréquent dans les applications que j'ai abordées et qui est décrit dans le paragraphe suivant.

### 5.2.2 Entrées spatiales

Dans de nombreuses applications en géosciences, certaines entrées prennent la forme de fonctions aléatoires en 2 ou 3 dimensions. Dans les applications en géophysique et en ingénierie de réservoir auxquelles je me suis intéressé, il s'agit de caractériser les propriétés du sous-sol à partir de données indirectes. La première étape consiste donc à définir une paramétrisation parcimonieuse du modèle de fonction aléatoire. La dimension de l'espace des paramètres dépendra du modèle choisi, de la grille de discrétisation et/ou de la méthode de simulation employée.  $Z$  désignera désormais une fonction aléatoire définie sur un domaine  $\mathcal{X} \subset \mathbb{R}^d$ , où  $d \in \mathbb{N}^*$ . On peut généralement exprimer  $Z$ , ou son approximation, sous une forme fonctionnelle :

$$Z(x) = \sum_{i=1}^N A_i \phi_i(x), \quad x \in \mathcal{X}, \quad (5.6)$$

où  $N \in \mathbb{N}^*$ , les  $A_i$  sont des variables aléatoires, non nécessairement indépendantes, et les fonctions  $\phi_i$ , dites de base, peuvent dépendre de paramètres aléatoires. Le choix de la paramétrisation d'un modèle spatial porte ainsi sur les variables  $A_i$ , les fonctions de base et la valeur de  $N$ , ainsi que les lois de probabilités associées. On remarquera que ce modèle est semblable au modèle spatial à effets aléatoires utilisé dans le krigeage de grands jeux de données vu au chapitre 4.

**Champs gaussiens** Dans ma thèse [Romary, 2008], dans une problématique d'ingénierie de réservoir, je me suis intéressé à la paramétrisation de (transformations de) champs gaussiens. Ceux-ci permettent de décrire des champs de perméabilité ou bien des lithofaciés aux propriétés pétrophysiques différentes, en considérant l'ensemble d'excursion au delà d'un seuil donné.

Lorsqu'on s'intéresse à la simulation de champs gaussiens sur une grille, un certain nombre de méthodes sont disponibles. On peut citer la méthode de Cholesky qui revient à calculer la décomposition de Cholesky de la matrice de covariance du champ discrétisé puis à la multiplier par un vecteur de gaussiennes standards indépendantes et la méthode de circulant embedding ([Dietrich and Newsam, 1993]) basée sur une transformée de Fourier rapide d'une extension de la matrice de covariance sur un tore. Ces deux approches nécessitent un nombre de réalisations de variables aléatoires indépendantes au moins supérieur au nombre de noeuds de la grille de discrétisation, ce qui les disqualifie pour la résolution de problèmes inverses. Citons enfin la méthode des bandes tournantes proposée par G. Matheron dans les années 70, elle aussi gourmande en réalisations de variables aléatoires ([Matheron, 1973, Lantuéjoul, 2002]).

On est ainsi contraint d'employer une approximation pour obtenir des représentations parcimonieuses de champs aléatoires gaussiens. Dans mes travaux de thèse, j'ai employé la décomposition de Karhunen-Loève ([Loève, 1955]) qui permet de représenter les champs gaussiens dans la base  $\{\phi_i\}_{i \in I}$  des vecteurs propres de leur opérateur de covariance, avec  $I$  au plus dénombrable. Les variables  $A_i$  sont alors gaussiennes, centrées, indépendantes et de variance donnée par la valeur propre correspondante. En tronquant cette représentation,



on peut conserver une part de variance globale égale à la somme des valeurs propres correspondant aux composantes conservées. Cette représentation est en outre optimale au sens de l'erreur quadratique moyenne intégrée sur le domaine  $\mathcal{X}$  ([Ghanem and Spanos, 1991]). Plutôt que de fixer un seuil de conservation d'une certaine part de variance, j'ai montré qu'il est pertinent de fixer le niveau de troncature selon un critère dynamique, de manière à reproduire convenablement les écoulements dans la réalisation approximée ([Romary, 2009a]). Dans le cas des lithofaciès, j'ai suggéré l'étude de la reproduction de la fonction de connexité ([Romary, 2009b]), qui est un bon proxy de la simulation des écoulements à un coût numérique moindre.

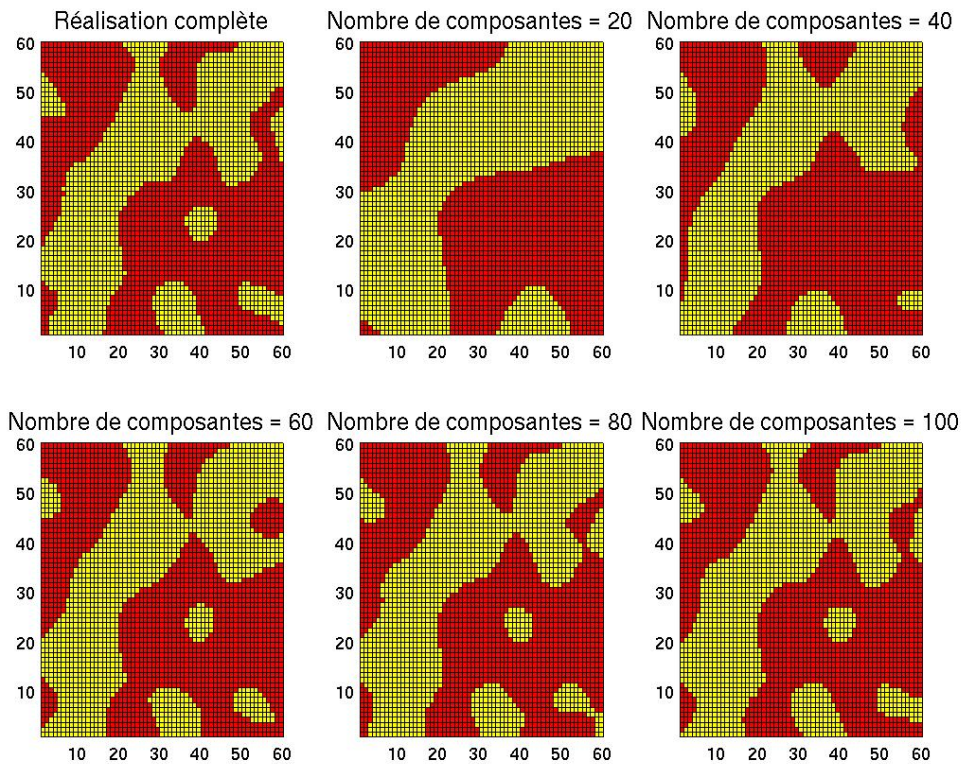


FIGURE 5.5 – Exemple d'approximation d'un champ 2D en facies à différents ordres de troncature

L'usage de la décomposition de Karhunen-Loève est cependant coûteux numériquement. Il nécessite en effet de diagonaliser la matrice de covariance du champ gaussien discrétisé, opération dont la complexité est proportionnelle au cube du nombre de points de discrétisation, et n'est donc pas envisageable pour des applications de grande dimension. Plusieurs autres méthodes de réduction de la dimension de champs gaussiens ont été envisagées, bien que moins efficaces, parmi lesquelles on peut citer la représentation dans

des bases de cosinus ([Jafarpour et al., 2009]) ou d'ondelettes ([Elliott Jr et al., 1997]). Enfin, la décomposition de Cholesky incomplète, que j'ai proposée pour le krigeage de grands jeux de données (cf. [Romary, 2013] et section 4.4.2), offre également une représentation d'un champ gaussien discrétisé de dimension réduite et dont la qualité de l'approximation est quantifiable en terme de variance reproduite, comme illustré sur la figure 5.6, pour un champ gaussien de covariance sphérique. La complexité de l'algorithme de Cholesky incomplet est linéaire en le nombre de points de discrétisation, ce qui la rend attractive pour ce type d'applications.

Il importe de garder à l'esprit que ces approximations de champs gaussiens reviennent généralement à éliminer les variations de haute fréquence, autrement dit les variations à petite échelle. On se retrouve alors avec des réalisations plus lisses que le modèle initial. Il convient donc de contrôler la dégradation dans les sorties des modèles phénoménologiques appliqués à ces approximations, notamment afin de choisir un nombre de composantes suffisant. Dans la pratique, en particulier avec les modèles d'écoulement avec lesquels j'ai pu travailler, je me suis rendu compte qu'un nombre restreint de composantes suffisait à bien reproduire les phénomènes simulés numériquement. On peut supposer que cela soit dû à un phénomène d'homogénéisation, autrement dit que les fluctuations du milieu à petite échelle (celles qui sont gommées dans les approximations), n'ont que peu d'influence dans les processus physiques à l'oeuvre.

**Modèles de vitesse en tomographie sismique** Dans le cadre d'une collaboration avec l'équipe de géophysique du centre de géosciences, je me suis intéressé à la tomographie sismique dans un cadre bayésien. Il s'agit de caractériser le champ de vitesse de propagation des ondes sismiques dans le sous-sol à partir de données de temps de première arrivée, obtenues via un réseau de sources et de récepteurs, que l'on détaillera plus loin.

Le modèle le plus simple utilisé en tomographie sismique est le modèle dit "à couche" qui correspond à une superposition de couches de vitesse constante. L'hypothèse d'un modèle à couches est bien adaptée pour décrire les milieux sédimentaires, comme ceux rencontrés dans les réservoirs pétroliers peu complexes ou de gaz de schiste. Dans la représentation (5.6), ce modèle prend la forme suivante :

$$Z(x) = \sum_{i=1}^N A_i \mathbb{1}_{[z_{i-1}, z_i]}(x), \quad x \in \mathcal{X}, \quad (5.7)$$

où les  $A_i$  représentent la vitesse constante au sein des couches délimitées par les profondeurs des interfaces  $z_i$ , avec  $z_0 = 0$ . On peut éventuellement ajouter un pendage pour chaque interface que l'on peut paramétrer par deux angles  $(\alpha, \beta)$ . Le choix des valeurs des angles doit assurer le non recoupement des couches, ce qui impose une contrainte sur la loi a priori de ces paramètres. Enfin, le nombre de couches  $N$  peut lui même être une des inconnues à caractériser ([Malinverno, 2002]). La figure 5.7 donne une représentation schématique de ce type de modèles.

Les modèles à couches sont cependant limités en terme de complexité. On peut notamment vouloir ajouter de l'information a priori. Dans Bottero et al. [2016], la donnée d'un

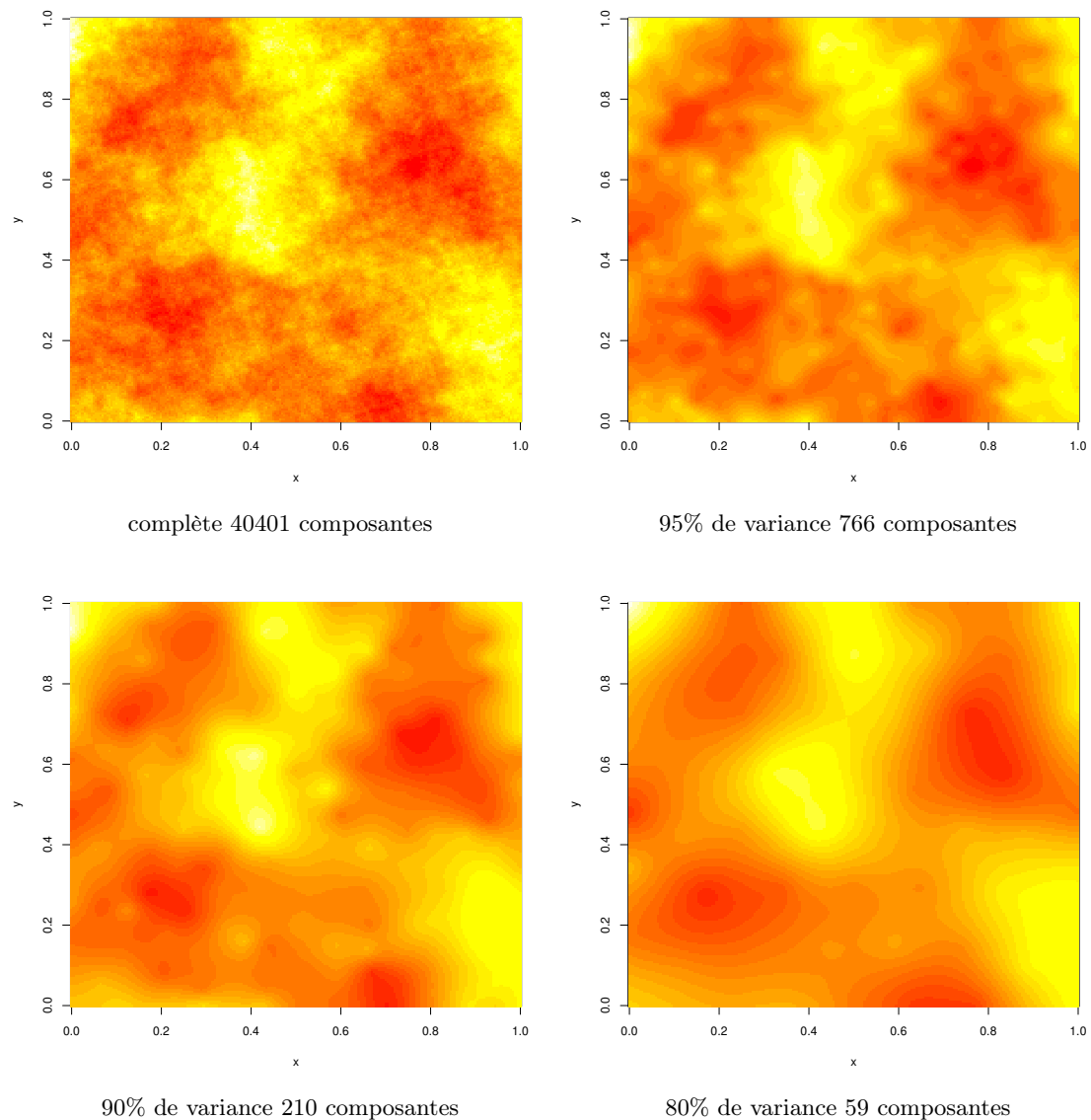


FIGURE 5.6 – Simulation complète (en haut à droite) et approximée à différents ordre

log sismique (vertical) nous a permis de proposer une paramétrisation dans une base d'ondelettes, que nous avons étendue dans l'espace grâce à l'hypothèse d'un milieu gouverné par la stratigraphie.

L'hypothèse stratigraphique n'est malheureusement pas vérifiée dans tous les cas. Dans la thèse de Jihane Belhadj ([Belhadj, 2016]), nous avons exploré un certain nombre d'autres paramétrisations du modèle de vitesse. En particulier, nous nous sommes intéressés aux

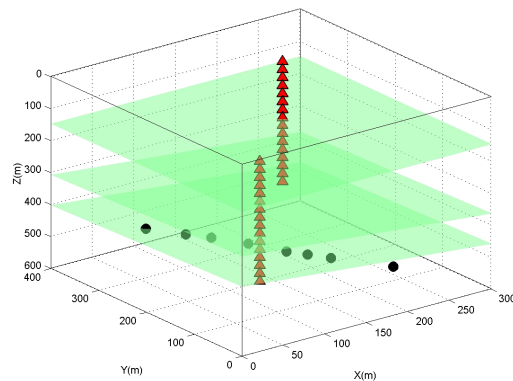


FIGURE 5.7 – Représentation d’un modèle à couche et d’un réseau de capteurs/récepteurs.

modèles de partitionnement, très populaires en tomographie sismique ([Bodin et al., 2009]). Il s’agit d’associer une valeur constante de vitesse à chaque cellule d’une partition. Ainsi, dans la représentation (5.6), les fonctions de base sont les indicatrices des cellules et les paramètres sont les valeurs prises par le champ de vitesse dans chacune des cellules. Pour étendre le partitionnement basé sur des cellules de Voronoï proposé dans Bodin et al. [2009], nous avons proposé une partition basée sur les cellules de Johnson-Mehl (voir par exemple [Møller, 1992]). Les cellules de Johnson-Mehl peuvent être considérées comme une généralisation des cellules de Voronoï : au cours du temps, des germes apparaissent en différents endroits du domaine et grossissent de manière isotrope à vitesse constante. Les nouveaux germes n’apparaissent que dans les endroits qui ne sont pas encore occupés par les cellules en cours de croissance. La croissance des frontières des cellules s’interrompt dès que deux frontières se rencontrent. La mosaïque finale est obtenue lorsque tout l’espace est occupé par les cellules. La figure 5.8 montre les différentes étapes de la construction de la mosaïque de Johnson-Mehl.

Contrairement à la mosaïque de Voronoï, les cellules de Johnson-Mehl ne sont pas forcément convexes et sont bornées par des portions d’hyperboloïdes, offrant ainsi des frontières lisses entre les cellules. L’intérêt essentiel du modèle de Johnson-Mehl, notamment en comparaison du modèle de Voronoï, est qu’il permet, d’une part l’obtention de formes de cellules plus générales, et d’autre part des frontières plus lisses entre cellules.

**Vers des paramétrisations plus réalistes** Les modèles de fonctions aléatoires gaussiennes, et leurs dérivés, ne sont pas adaptés à toutes les situations. Les modèles utilisés pour la tomographie sismique sont eux aussi limités : les modèles reposant sur une hypothèse stratigraphique ne sont pas toujours adaptés tandis que les modèles de partitionnement ne permettent d’obtenir des représentations géologiquement réalistes que lorsqu’on calcule une moyenne d’un échantillon de réalisations, de la loi a posteriori (5.2) par exemple.

Aussi, trouver des paramétrisations parcimonieuses de modèles spatiaux qui assurent un certain réalisme physique reste un problème ouvert. Dans les années 2000, les méthodes

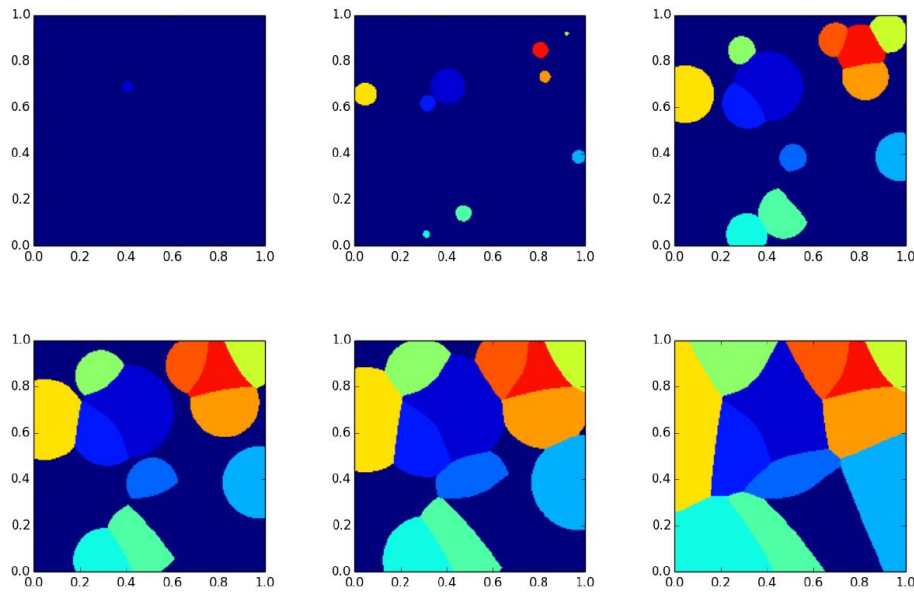


FIGURE 5.8 – Représentation d’une mosaïque de Johnson-Mehl.

ditions de géostatistique multipoint ont été introduites ([Mariethoz and Caers, 2014]). Ces algorithmes permettent de générer des images réalistes capables de reproduire des structures complexes, telles que des chenaux par exemple. Leur principe est de générer des images qui vont reproduire la texture d’une image d’entraînement donnée, qui doit être représentative du phénomène étudié. Si l’on s’intéresse à la répartition spatiale des lithofaciés dans un réservoir chenalisé méandriforme, un modèle génétique, basé sur la simulation des processus de sédimentation, peut être employé pour générer une telle image d’entraînement en 3D (voir par exemple [Lopez et al., 2009]). Cependant, ces algorithmes ne reposent pas sur un modèle de fonction aléatoire et ne sont donc pas adaptés à la résolution de problèmes inverse dans un cadre bayésien. En particulier, il n’y a aucun moyen de définir une marche aléatoire dans l’espace des réalisations.

Plus récemment, d’importants développements ont été réalisés dans la communauté du machine learning pour construire des modèles génératifs à partir de réseaux de neurones profonds. A partir d’une large base de données, ces modèles permettent de construire de nouvelles réalisations aléatoires qui vont reproduire les caractéristiques des données d’entraînement. A titre d’exemple grand public, on peut citer la génération de visages de personnalités. Parmi ces méthodes on peut distinguer les variational autoencoders (VAE), voir par exemple Mattei and Frellsen [2018], et les generative adversarial networks (GANs) ([Goodfellow et al., 2014]. Les VAE sont une classe d’algorithmes basés sur la représentation des données d’entraînement par des variables latentes, généralement prises sous la forme d’un vecteur gaussien. On va notamment chercher à apprendre la vraisemblance des données sachant les variables latentes ainsi que la loi a posteriori des variables latentes par des réseaux de neurones profonds. Les GANs reprennent le principe des VAE à la différence que l’on va entraîner simultanément deux réseaux de neurones profonds : le premier réseau,

dit générateur produit les images en prenant en entrée un certain nombre de variables aléatoires indépendantes, le second réseau est un discriminateur qui classe les images (générées ou réelles). Dans les deux cas, en cas de succès de l'apprentissage, on obtient en sortie un réseau de neurone générateur entraîné qui pourra produire rapidement de nouvelles réalisations en simulant uniquement des variables dans l'espace latent et en les donnant en entrée au réseau générateur.

Quelques expérimentations prometteuses ont été réalisées en géosciences, parmi lesquelles on peut citer Mosser et al. [2017] où des GANs ont été utilisés pour générer de nouvelles réalisations de milieux poreux particuliers, ou encore Laloy et al. [2017] où ils sont employés comme loi a priori dans la résolution de problèmes inverses. Il reste cependant un certain nombre de difficultés à l'application pratiques de ces approches dans l'inversion bayésienne : d'une part, le conditionnement à des données ponctuelles n'est pas immédiat car il s'apparente à un nouveau problème inverse. En effet, il faut alors conditionner les variables latentes aux données observées, ce qui revient à inverser le réseau de neurones. D'autre part, selon les applications, il peut être difficile de disposer d'une base suffisamment large d'images d'entraînement.

## 5.3 Méthodes MCMC

Une fois arrêtée la paramétrisation et choisie une loi a priori adaptée, la résolution du problème inverse dans le cadre bayésien revient à caractériser la loi conditionnelle des paramètres sachant les observations, dont on rappelle ici l'expression dans le cas de lois à densité :

$$\pi(Z) = f(Z|Y) = \frac{f(Y|Z)f(Z)}{f(Y)}, \quad (5.8)$$

où  $Z$  représente le vecteur des paramètres aléatoires de l'état du système. Comme indiqué en introduction, cette densité n'a que très rarement une forme analytique et on cherche donc à construire un échantillon de réalisations de cette loi.

Les approches directes de simulation, par exemple la méthode du rejet ([Robert and Casella, 2004]), échouent dès que la taille de  $Z$  devient grande et/ou que la bonne reproduction des observations est très contraignante. Je me suis beaucoup intéressé à l'application des méthodes MCMC à ce type de problèmes.

### 5.3.1 Principe général

Le principe de ces méthodes consiste à construire une suite de variables aléatoires  $\{Z_n, n \geq 0\}$  de probabilité de transition  $P$  telle que  $\pi(Z)$  est la densité stationnaire de cette chaîne, c'est-à-dire telle que  $\forall A \in \mathcal{B}(\mathcal{Z})$  :

$$\int_{\mathcal{Y}} P(z, A)\pi(z)dz = \pi(A).$$

et telle que la suite  $(Z_n)$  converge en loi vers la loi cible. Pour les principaux résultats théoriques concernant les conditions de convergence, on pourra se référer à Robert and Casella [2004] et Meyn and Tweedie [1993].

De nombreux algorithmes permettent de construire des chaînes de Markov qui vont converger en théorie vers cette loi, en particulier la construction de Metropolis-Hastings ([Hastings, 1970]). Basée sur le choix d'une loi instrumentale  $q$  qui permet de proposer une transition d'un état  $z$  à un état  $z^*$ , la probabilité d'acceptation est construite de manière à assurer la convergence de la chaîne vers la loi stationnaire :

$$\alpha(z, z^*) = \min \left( 1, \frac{\pi(z^*)q(z|z^*)}{\pi(z)q(z^*|z)} \right) \quad (5.9)$$

Les schémas classiques de l'échantillonneur indépendant, où la loi instrumentale est exactement la loi a priori, et la marche aléatoire à incréments symétriques, où l'on perturbe l'état courant par un bruit centré, sont construits selon ce principe.

On a vu au paragraphe 5.2.2 que certaines paramétrisations ont un nombre de paramètres variables, comme le modèle à couches ou les modèles de partitionnement employés en géophysique. Dans ce cas, la résolution du problème inverse nécessite l'emploi d'un schéma MCMC capable de générer des simulations d'une loi a posteriori dans des espaces de dimension variable. C'est le cas du reversible-jump MCMC, proposé par Green [1995] pour des problèmes de sélection de modèles dans un cadre bayésien. Cet algorithme a été largement employé en tomographie sismique bayésienne ([Bodin et al., 2009, Piana Agostinetti et al., 2015, Malinverno, 2002, Belhadj et al., 2018]).

### 5.3.2 Méthodes avancées

Malgré la simplicité d'implémentation de ces algorithmes et leurs bonnes propriétés théoriques, un certain nombre de difficultés pratiques impliquent d'employer des méthodes plus sophistiquées. D'une part, le réglage des paramètres de l'algorithme s'avère souvent délicat, notamment lorsqu'on utilise une marche aléatoire. Il se fait généralement par essai erreur et nécessite de respecter un compromis entre l'exploration de l'espace d'état et le maintien d'un taux d'acceptation raisonnable. On peut alors utiliser des algorithmes adaptatifs (voir par exemple [Roberts and Rosenthal, 2006]), parmi lesquels le schéma de Haario et al. [2001], où l'on optimise les propositions gaussiennes dans une marche aléatoire symétrique en estimant leurs paramètres le long de la chaîne.

Une seconde difficulté qui se rencontre fréquemment dans la résolution de problèmes inverses est la lenteur de la chaîne à entrer dans son régime stationnaire. En effet, à moins d'être particulièrement chanceux, le point de départ de la chaîne, tiré aléatoirement selon la loi a priori, ne concorde pas avec les observations, autrement dit le terme  $f(Y|Z)$  a une valeur très faible. Il faut parfois un très grand nombre d'itérations pour atteindre des réalisations vraisemblables relativement aux observations et que l'on peut ainsi considérer comme issues de la loi stationnaire. Différentes stratégies existent pour accélérer la convergence parmi lesquelles on peut distinguer deux familles. Dans la première, on va tirer parti de l'information offerte par le gradient du problème direct, s'il est disponible. Dans ce cas,

on peut par exemple implémenter l'échantillonneur de Langevin, basé sur une construction de Metropolis Hastings et dont la loi instrumentale est donnée par :

$$q(z, z^*) \sim \mathcal{N}\left(z + \frac{h^2}{2} \nabla \log(\pi(z)), h^2 I_d\right),$$

où  $z$  représente l'état courant et  $z^*$  l'état proposé. Cet algorithme rencontre cependant des difficultés dans le cas d'une loi cible multimodale puisqu'il visera généralement un seul mode, dont les états de la chaîne seront captifs. Pour atténuer ce comportement, on pourra considérer la méthode de simulation hamiltonienne (HMC) qui repose sur une simulation de la dynamique hamiltonienne de l'état courant où un terme d'énergie cinétique aléatoire est introduit. Ceci autorise la chaîne à s'écarter davantage de la trajectoire décrite par le gradient ([Neal, 2010]). On notera que ce type d'algorithme est similaire à l'algorithme d'optimisation ADAM, largement utilisé en deep learning ([Kingma and Ba, 2014]). L'ajustement des paramètres est cependant délicat. Des stratégies très efficaces ont été proposées pour l'automatiser en utilisant notamment une adaptation locale ([Girolami and Calderhead, 2011]). Elles reposent cependant sur la différentielle du second ordre de l'opérateur direct qui est en général difficile à obtenir dans les problèmes complexes. Plus récemment, le No-U-Turn Sampler (NUTS [Hoffman and Gelman, 2014]) propose d'automatiser le choix du pas de discrétisation de la trajectoire Hamiltonienne ainsi que le nombre de pas à effectuer à chaque itération. Les algorithmes HMC, en particulier NUTS, ont montré d'excellentes performances dans des problèmes d'inférence bayésienne en grande dimension. Ils restent toutefois gourmands en nombre d'appels au modèle direct et nécessitent le gradient ce qui relativise la pertinence de leur utilisation dans le cadre de l'inversion bayésienne.

Une seconde stratégie, complémentaire de la précédente, repose sur une relaxation de la contrainte imposée par les observations via l'introduction d'un terme de température. C'est l'idée mise en œuvre dans le recuit-simulé ([Metropolis et al., 1953, Kirkpatrick et al., 1983]). Le principe consiste à cibler une version de la loi cible où le terme de vraisemblance  $f(Y|Z)$  prend la forme :

$$f_T(Y|Z) \propto \exp\left(\frac{-E(Z, Y)}{T}\right), \quad (5.10)$$

où  $E(Z, Y) = -\log(f_T(Y|Z))$  représente l'énergie du système en l'état  $Z$  et  $T$  la température. Dans le cas d'une erreur gaussienne, on voit que  $T$  s'interprète comme un terme multiplicatif de la variance. Autrement dit, plus  $T$  sera grand, moins l'observation  $Y$  sera contraignante. Lorsque la température est suffisamment élevée, il est aisé d'explorer l'espace d'état puisque la plupart des transitions sont acceptées. En la faisant décroître vers 1, on se rapproche mécaniquement des niveaux d'énergie correspondant à la loi a posteriori. Cependant, en faisant simplement décroître la température, on risque tout de même de rester bloquer dans un mode de la loi a posteriori, comme on peut s'en convaincre à la lecture de la figure 5.9.

C'est justement la troisième difficulté posée par l'usage des méthodes MCMC. Dans l'idéal, on souhaite que la chaîne soit capable d'explorer largement la loi cible, sans rester bloquée dans un mode local, tout en produisant des réalisations peu corrélées les unes



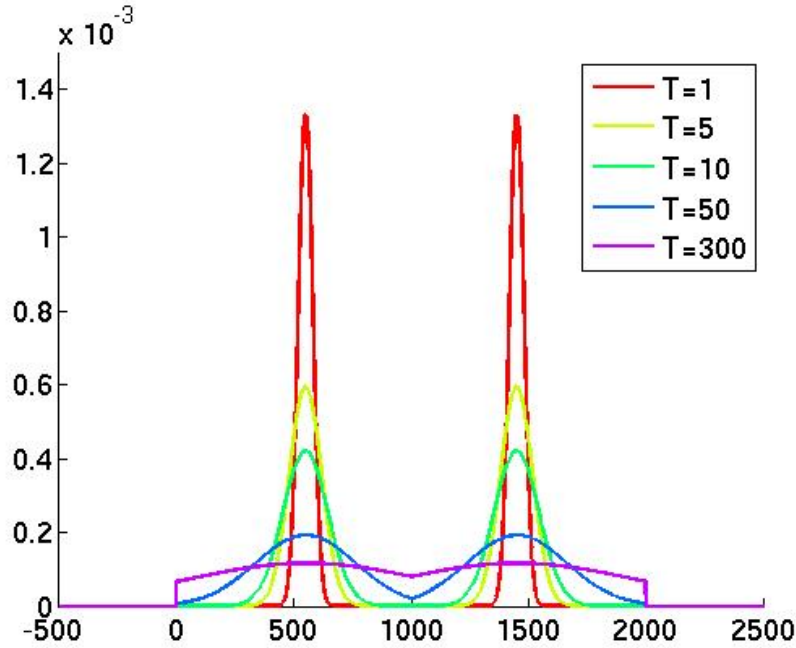


FIGURE 5.9 – Effet de la température sur une densité bimodale

aux autres. On dit d'une telle chaîne qu'elle a de bonnes propriétés de mélange. C'est dans cette optique qu'a été introduite la trempe simulée ([Marinari and Parisi, 1992, Geyer and Thompson, 1995]), où le fait de pouvoir retourner vers des températures plus élevées permet de s'échapper de modes locaux. Cependant la mise en œuvre est délicate puisque la probabilité d'acceptation dépend des constantes de normalisation des  $f_T$ , qui sont généralement inaccessibles. Le schéma proposé dans Celeux et al. [1999] permet de passer outre cet écueil mais requiert un nombre trop important d'appels au modèle direct.

### 5.3.3 Chaînes en interaction

Une autre solution que j'ai employée à plusieurs reprises est l'usage de chaînes de Markov en interaction d'abord proposées dans Geyer [1991] puis formalisées dans Andrieu et al. [2011], dont le principe repose sur l'approximation de noyaux de Markov non linéaires :

$$P_\pi(y, y^*) = \theta P(y, y^*) + (1 - \theta) \int P'(u, y^*) d\pi(u), \quad 0 < \theta < 1 \quad (5.11)$$

où  $P$  et  $P'$  sont deux noyaux de Markov de loi limite  $\pi$ . L'avantage du terme de droite est qu'il va permettre de régénérer la chaîne en proposant un nouvel état indépendant et potentiellement éloigné de l'état courant. Pour simuler un état selon ce terme, avec probabilité  $(1 - \theta)$ , on est censé savoir simuler selon  $\pi$ , qui est inconnue. La mise en œuvre de cette approche repose sur l'approximation de  $\pi$  à l'aide d'états générés par une autre chaîne, dont la loi cible est plus facile à simuler, via l'échantillonnage d'importance. On

utilise en pratique une cascade de telles approximations. L'idée étant que les chaînes aux hautes températures exploreront largement l'espace d'état, tandis que les chaînes aux basses températures exploreront en détail les alentours des réalisations les plus intéressantes. C'est la démarche qui est employée dans l'algorithme du parallel tempering (voir par exemple [Iba, 2001]) et dans d'autres algorithmes de chaînes en interaction ([Kou et al., 2006, Andrieu et al., 2011]) dont j'ai implémenté des versions durant ma thèse ([Romary, 2009a,b, 2010]) puis appliqué plus récemment en tomographie sismique ([Bottero et al., 2016, Belhadj et al., 2015]).

En pratique, on construit d'abord une échelle de  $K + 1$  températures  $T_0 = 1 < T_1 < \dots < T_K = T_{\max}$  et on aura  $K + 1$  chaînes qui cibleront chacune

$$\pi^{(l)}(Z) = \exp\left(-\frac{E(Z, Y)}{T_l}\right) f(Z), \quad l = 0, \dots, K \quad (5.12)$$

Tandis qu'à la température maximale un noyau de Markov classique est employé, la chaîne à la température  $T_l$  ( $0 \leq l < K$ ) évolue selon le noyau suivant :

$$P_{y_{0:n-1}}^{(l)}(y^{(l)}; y^{(l)*}) = \theta P^{(l)}(y^{(l)}, y^{(l)*}) \quad (5.13)$$

$$+ (1 - \theta) \int_{\mathcal{Y}} \nu_{y_{0:n-1}}^{(l)}(y^{(l)}, du) T^{(l)}(u, y^{(l)}; y^{(l)*}) \quad (5.14)$$

qui approxime (5.11) et où  $y_{0:n-1}^{(l+1)}$  désigne les états précédemment simulés à la température  $T_{l+1}$ ,  $y^{(l)}$  l'état courant,  $y^{(l)*}$  l'état cible,  $\nu_{y_{0:n-1}}^{(l)}(y^{(l)}, du)$  représente l'approximation de  $\pi^{(l)}$  obtenue à partir des états passés de la chaîne à la température  $l + 1$  et  $T^{(l)}(u, y^{(l)}; y^{(l)*})$  est un noyau de Markov.

Plusieurs possibilités existent pour le choix de l'approximation  $\nu^{(l)}$ . Dans le parallel tempering, il s'agit d'une mesure de Dirac en l'état courant de la chaîne à la température adjacente supérieure, dans l'equi-energy sampler, on choisira uniformément parmi les états passés de la chaîne à la température adjacente supérieure qui présentent des niveaux d'énergie similaires à l'état courant. Dans l'algorithme dont je préconise l'utilisation,  $\nu^{(l)}$  prend la forme suivante :

$$\nu_{y_{0:n-1}}^{(l)}(y^{(l)}, du) = \frac{\sum_{i=0}^{n-1} r^{(l)}(y_i^{(l+1)}) \delta_{y_i^{(l+1)}}(du)}{\sum_{i=0}^{n-1} r^{(l)}(y_i^{(l+1)})}, \quad (5.15)$$

où  $r^{(l)}(y) = e^{-(E_l(y) - E_{l+1}(y))} \propto \frac{\pi^{(l)}(y)}{\pi^{(l+1)}(y)}$ . Autrement dit, lors d'une interaction, le nouvel état est tiré parmi les états passés de la chaîne à la température adjacente supérieure par échantillonnage d'importance.

Enfin, ce nouvel état est accepté avec la probabilité  $\min\left(1, \frac{r^{(l)}(y^{(l)*})}{r^{(l)}(y^{(l)})}\right)$ , ce qui est représenté dans (5.13) par le noyau  $T^{(l)}(u, y^{(l)}; y^{(l)*})$ . On remarquera que cette construction permet d'évacuer les constantes de normalisations.

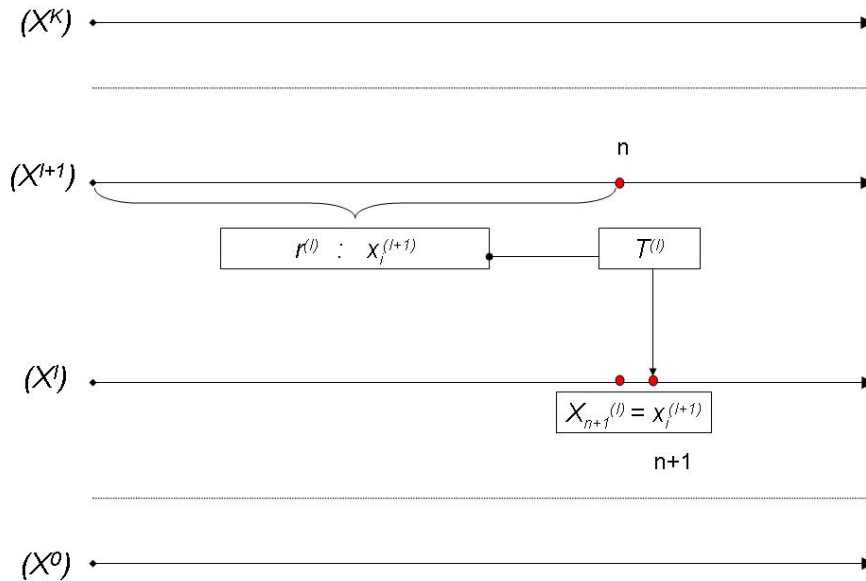


FIGURE 5.10 – Représentation schématique d’une interaction

Au fur et à mesure des itérations, le noyau  $P_{y_{0:n-1}}^{(l)}(y^{(l)}; A)$  va converger vers

$$\theta P^{(l)}(y^{(l)}, A_l) + (1 - \theta)R^{(l)}(y^{(l)}, A_l)$$

où  $R^{(l)}$  est un échantillonneur indépendant dont la loi instrumentale est  $\pi^{(l)}$ , c’est-à-dire exactement la loi cible. Cet algorithme, dont le fonctionnement est résumé en figure 5.10, permet ainsi de visiter des états intéressants qui n’auraient pas pu être atteints par un noyau classique.

En pratique, un certain nombre de paramètres doivent être ajustés pour que cet algorithme fonctionne de manière satisfaisante, à savoir l’échelle de températures, le nombre de chaînes, les noyaux de chacune des chaînes et la probabilité de proposer une interaction. Ces paramètres doivent être choisis de manière à assurer une bonne exploration de l’espace d’états aux hautes températures ainsi qu’une description fine des zones intéressantes à  $T_0$ , tout en permettant des interactions efficaces. J’ai donné un certain nombre de règles heuristiques dans Romary [2010].

Ces algorithmes de chaînes en interaction peuvent être employés avec tout type de noyaux individuels. Ils permettent aux différentes chaînes d’atteindre rapidement le régime stationnaire et rendent possible l’échantillonnage de lois fortement multimodales. De fait, ils sont particulièrement bien adaptés à la résolution de problèmes inverses dans le cadre bayésien. Dans Romary [2009a] et Romary [2009b], je les ai appliqués pour générer des réalisations de la loi a posteriori dans un problème d’history matching en ingénierie de

réservoir. Ils ont également permis d'obtenir des résultats très intéressants en tomographie sismique ([Bottero et al., 2016, Belhadj et al., 2015, Romary, 2017]), en particulier pour l'identification de comportements multimodaux.

## 5.4 Assimilation de données

Dans de nombreux problèmes en géosciences, les données sont mesurées séquentiellement dans le temps et issues de l'observation d'un système dynamique. C'est par exemple le cas en météorologie où les méthodes d'assimilation de données ont été introduites mais aussi en océanographie, glaciologie, sismologie, qualité de l'air, de l'eau, etc. On peut formaliser ce problème sous la forme d'un modèle espace-état :

$$\begin{aligned} Z_t &= F_\theta(Z_{t-1}, \eta_t), \\ Y_t &= H(Z_t) + \varepsilon_t, \end{aligned} \tag{5.16}$$

où  $Z_t$  représente l'état du système au temps  $t$ ,  $\eta_t$  est une variable aléatoire indépendante de  $Z_{t-1}$ ,  $F_\theta$  est une application connue représentant la dynamique du système et dépendant du paramètre  $\theta$ ,  $Y_t$  représente l'observation au temps  $t$  effectuée au travers de l'opérateur d'observation  $H$  et  $\varepsilon_t$  représente l'erreur d'observation. On considère de plus que  $Z_0$  est initialisé selon une loi de densité connue  $f(z_0)$ . On peut représenter alternativement (5.16) sous la forme d'un chaîne de Markov cachée :

$$\begin{aligned} Z_0 &\sim f(z_0) \text{ et } Z_t | (Z_{t-1} = z_{t-1}) \sim f(z_t | z_{t-1}) \\ Y_t | Z_t &\sim g(y_t | z_t) \end{aligned}$$

Le principe est le même qu'en inversion bayésienne, à savoir que l'on va chercher à caractériser l'état du système  $Z_t$  grâce aux observations  $Y_t$ . En effet, pour  $T \in \mathbb{N}^*$ , si on considère une trajectoire  $Z_{1:T} = (Z_1, \dots, Z_T)$  et les observations associées  $Y_{1:T} = (Y_1, \dots, Y_T)$ , on cherche dans ce problème à caractériser  $f(Z|Y)$ . On pourrait donc utiliser les approches décrites plus haut. Cependant, pour simplifier le problème et s'autoriser à traiter des observations sans imposer un temps  $T$  final, on va tirer parti de la dynamique Markovienne du système (5.16) pour décomposer séquentiellement la densité a posteriori. En effet, on a :

$$p(z_{1:t} | y_{1:t}) = f(z_{1:t-1} | y_{1:t-1}) \frac{g(y_t | z_t) f(z_t | z_{t-1})}{p(y_t | y_{t-1})}. \tag{5.17}$$

Ce problème se décompose ainsi en une série de problèmes inverses dont la dimension est divisée par  $t$ . Par marginalisation, on peut résoudre des problèmes de filtrage où on cherche à identifier  $p(z_t | y_{1:t}) \forall t$ , de prédiction avec  $\mathbb{P}(z_{t+n} | y_{1:t})$  pour  $n \in \mathbb{N}^*$ , ou de lissage avec  $\mathbb{P}(z_t | y_{1:T})$ ,  $0 < t \leq T$ .

En pratique, on ne sait résoudre analytiquement (5.17) que dans le cas où  $F_\theta$  est linéaire et les lois des erreurs sont gaussiennes. On retrouve alors le filtre de Kalman et on peut exprimer les lois conditionnelles. Quand ce n'est pas le cas, un grand nombre de méthodes existent parmi lesquelles on peut citer le filtre de Kalman étendu, qui procède

par linéarisation de la dynamique et le filtre de Kalman d'ensemble, qui donne une solution approchée, basée sur une approximation gaussienne et est adapté aux problèmes non linéaires en grande dimension ([Bertino et al., 2003, Evensen, 2009]).

Les méthodes de filtrage particulaire offrent une autre manière d'approximer ces lois conditionnelles ([Doucet et al., 2000, Doucet and Johansen, 2009]). L'avantage de ces méthodes est qu'elles ne reposent ni sur une approximation abrupte de la dynamique, ni sur des hypothèses en loi contraignantes ou rarement vérifiées. En revanche, le prix à payer pour cette flexibilité est d'ordre numérique : le coût computationnel de ces méthodes est élevé, ce qui limite la dimension des problèmes qu'elles peuvent traiter.

Leur principe repose sur l'approximation des lois conditionnelles de l'état du système par un ensemble de réalisations (les particules) pondérées. A  $t = 0$ , on initialise un nombre  $N$  de particules  $z_0^i$  selon  $f(z_0)$ , puis on les propage selon une densité instrumentale  $q(z_t|z_{t-1}, y_t)$ . On construit alors les poids d'importance qui sont mis à jour selon la récursion (5.17) :

$$\omega_t^i = \omega_{t-1}^i \frac{g(y_t|z_t^i)f(z_t^i|z_{t-1}^i)}{q(z_t|z_{t-1}, y_t)p(y_t|y_{t-1})}. \quad (5.18)$$

Puisque la constante  $p(y_t|y_{t-1})$  n'est en général pas calculable, on l'ignore en pratique puis on renormalise les poids. Il arrive souvent que les poids dégénèrent au fur et à mesure des itérations, c'est-à-dire qu'un grand nombre de particules se voit attribuer un poids proche de zéro. Pour pallier ce problème, on va rééchantillonner dans la loi empirique

$$\hat{\pi}_t(z_t) = \sum_{i=1}^N \omega_t^i z_t^i, \quad (5.19)$$

ce qui va avoir pour effet de dupliquer les particules ayant un poids important et de supprimer celles ayant des poids faibles. Les poids sont alors réinitialisés à  $1/N$ . Pour maintenir au maximum la diversité, on procède généralement par échantillonnage stratifié à cette étape. Bien qu'utile, le rééchantillonnage a un effet délétère dans le sens où il va augmenter la variance des estimateurs que l'on va former à partir de la loi empirique (5.19). Aussi, on procédera à un rééchantillonnage de manière parcimonieuse. Généralement, on contrôle la diversité de la population en calculant le critère d'effective sample size :

$$ESS = \left( \sum_{i=1}^N (\omega_t^i)^2 \right)^{-1}. \quad (5.20)$$

Lorsque ce dernier est en-dessous d'un certain seuil (typiquement  $N/2$  ou  $N/3$  selon les sources), on procédera à un rééchantillonnage. Sinon, on mettra simplement à jour les poids selon (5.18) ([Doucet and Johansen, 2009]). Cette stratégie permet de restreindre les rééchantillonnages tout en limitant la dégénérescence des particules. Cependant, ces algorithmes sont sujet au fléau de la dimension. En effet, dès que la dimension du problème, c'est-à-dire de  $z_t$ , devient trop importante, il devient difficile d'échantillonner la loi a posteriori de manière satisfaisante et toutes les particules obtiennent des poids très faibles. Dans ce cas, le rééchantillonnage n'est pas efficace pour régénérer le nuage de particules. Une

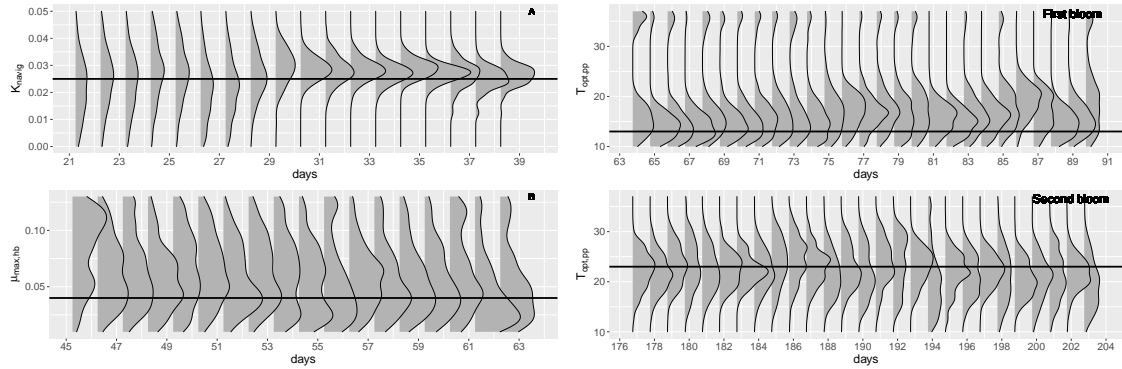


FIGURE 5.11 – Identification de paramètres variables dans Prose-PA

solution consiste à faire évoluer le nuage de particule à chaque itération selon une chaîne de Markov ciblant  $\frac{g(y_t|z_t)f(z_t|z_{t-1})}{p(y_t|y_{t-1})}$  mais ce procédé peut se révéler très lourd en pratique.

J’ai appliqué ce type d’algorithmes dans deux problèmes de natures tout à fait différentes. D’abord pour le suivi des épidémies de grippe saisonnière, où l’évolution de l’épidémie est modélisée au travers d’un modèle de type SIR (suspected-infected-removed [Smith et al., 2004]) spatialisé, où on cherchait en particulier à caractériser un état épidémique gouvernant le taux de contagion du virus. Dans ce modèle, bien qu’il soit possible d’écrire explicitement  $f(z_t|y_t, z_{t-1})$ , nous n’arrivons pas à identifier clairement l’état épidémique dès que la dimension spatiale devenait trop élevée ([Lajaunie and Romary, 2009]). Il est apparu par la suite qu’un moyen très efficace de détecter les épidémies de grippe consistait à analyser les requêtes évoquant les symptômes de la maladie sur les moteurs de recherche ([Ginsberg et al., 2009]).

Plus récemment, dans le cadre de la thèse de Shuaitao Wang, nous avons appliqué un algorithme de filtrage particulaire pour le suivi des paramètres gouvernant le métabolisme des cours d’eau. Après une réduction de dimension effectuée à l’aide d’une analyse de sensibilité déjà évoquée en section 5.2.1, le filtre particulaire nous a permis de caractériser l’évolution des paramètres gouvernant la concentration en oxygène dissous sur un cas synthétique ([Wang et al., 2019b]) puis sur données réelles ([Wang et al., 2021]). La figure 5.11 illustre la bonne identification de différents paramètres en fonction du temps et des différents états trophiques du système, sur le cas synthétique. En particulier, la figure de droite illustre la bonne identification des paramètres phytoplanctoniques durant un épisode de prolifération algale.

L’application aux données réelles de cette approche s’avère plus délicate du fait qu’elle a permis d’identifier des sources d’incertitudes additionnelles qui n’étaient jusqu’alors non prises en compte, notamment dans les apports en matière organique dont se nourrissent les communautés bactériennes. Ces paramètres additionnels correspondent à certaines conditions aux limites du système dont il va maintenant s’agir d’intégrer la caractérisation. Une thèse est en cours sur le sujet.

Lorsque la dimension de l’espace d’état augmente, les performances du filtre particulaire s’en ressentent. En particulier, on observe fréquemment une dégénérescence des poids dont

la plupart des valeurs devient pratiquement nulle. On peut alors considérer une approximation linéaire gaussienne, telle qu'offerte par le filtre de Kalman d'ensemble ([Evensen, 2009]). Ce type de méthodes a déjà été largement appliqué en géosciences, notamment dans les sciences du climat et l'océanographie ([Carrassi et al., 2018]). Cette approximation peut convenir pour certains paramètres évoluant lentement et selon des dynamiques relativement simples mais offre moins de souplesse sur la forme de la loi a posteriori que les méthodes particulières. On peut alors imaginer combiner les deux approches comme proposé dans Katzfuss et al. [2019].

## 5.5 Discussion

Ce tour d'horizon de mes contributions scientifiques sur la problématique de l'inversion bayésienne et de la caractérisation d'incertitudes en géosciences a permis de décrire un certain nombre de techniques couramment utilisées dans ce cadre. Il n'est cependant pas exhaustif.

En particulier, je n'ai jamais employé de techniques basées sur la modélisation de surfaces de réponse. Celles-ci interviennent généralement quand le modèle direct est trop coûteux à évaluer numériquement pour réaliser une analyse de sensibilité ou bien une inversion. En effet, puisque ces approches sont gourmandes en appels au modèle direct, on peut vouloir lui substituer une approximation offerte par un modèle statistique. Les modèles les plus fréquemment employés dans ce cadre sont basés sur des processus gaussiens indexés sur l'espace des paramètres. Ils font ainsi appel au krigeage pour la prédiction de la sortie du modèle direct en de nouveaux points. On peut alors se servir de ce métamodèle pour réaliser une analyse de sensibilité, optimiser un comportement, réaliser une inversion ou encore caractériser des événements particuliers. Il existe une vaste littérature qui présente ces approches, voir par exemple Santner et al. [2018], Gramacy [2020]. Elles sont cependant limitées au cas où la dimension de l'espace des paramètres reste modérée.

Il se peut également que le nombre disponible d'appels au modèle direct soit encore plus limité en raison de son temps de calcul, par exemple de l'ordre d'une petite dizaine. C'est un cas de figure que j'ai rencontré récemment dans le cadre de l'exploitation de l'uranium par lixiviation in-situ qui fait intervenir un code de transport réactif pour simuler la production d'uranium. Ce code prend en entrée une réalisation d'un modèle géostatistique complexe, discrétisé sur une grille contenant quelques centaines de milliers de mailles. Dans ces conditions, il est illusoire de réaliser une inversion bayésienne ou même une analyse de sensibilité. On peut tout de même chercher à propager l'incertitude de l'entrée sur la sortie du modèle afin d'évaluer l'incertitude sur la production d'uranium. Pour ce faire, on peut sélectionner un nombre restreint de réalisations des entrées du modèle qui permet d'estimer quelques quantiles de la production. On appelle cette approche la réduction de scénario ([Scheidt and Caers, 2009]). Elle repose sur une discrimination des entrées du modèle similaire à celle effectuée par le modèle direct via un proxy de ce dernier. Dans le cas de la production d'uranium par lixiviation in-situ, différents proxys ont été explorés dans Langanay et al. [2021]. L'impact de l'incertitude sur la gestion de l'exploitation du gisement a également été analysé dans la thèse de Jean Langanay.

J'ai évoqué plus haut l'usage possible des réseaux de neurones profonds pour fournir des modèles génératifs, autrement dit des loi a priori, plus réalistes. Ce n'est cependant pas la seule application de ces approches dans la résolution de problèmes inverses. Récemment, les Physics-Informed Neural Networks ont été proposés dans Raissi et al. [2019]. Cette construction consiste à contraindre le réseau à vérifier les équations aux dérivées partielles décrivant la physique du phénomène considéré en les intégrant à la fonction coût. Elle permet d'obtenir non seulement des approximations du modèle direct qui vont vérifier la physique du phénomène mais aussi de procéder à l'identification de ses paramètres d'entrées à partir d'observations. Une approche pour la quantification d'incertitudes dans ce cadre à également été proposée dans Yang and Perdikaris [2019]. Ces approches n'en sont qu'à leurs balbutiements et leur potentiel reste à explorer. On peut se poser en particulier la question de leur application à des codes de calculs complexes qui font intervenir un grand nombre d'équations couplées. Elles font cependant déjà l'objet d'applications prometteuses, notamment dans le cadre de l'assimilation de données en océanographie ([Fablet et al., 2021]). Ces approches représentent enfin un changement de paradigme dans le sens où la physique du phénomène étudié n'est plus simulé par une "boîte noire" mais fait désormais partie intégrante de la solution.





## Chapitre 6

# Perspectives

Au fil des discussions de ce manuscrit, j'ai disséminé différents éléments de perspectives que je propose de synthétiser ici tout en présentant les projets de recherche dans lesquels je suis impliqué à plus court terme.

Sur la thématique des problèmes inverses et de l'assimilation de données, j'encadre actuellement avec Nicolas Flipo (Centre de Géosciences) la thèse de Masihullah Hasanyar portant sur le suivi de la qualité de l'eau en Seine, dans la continuité des travaux de Shuaitao Wang évoqués au chapitre 5. Il s'agit de mieux comprendre le rôle de la matière organique dans le métabolisme du fleuve et d'intégrer l'incertitude portant sur ses mesures, en amont du fleuve et aux points de rejet, dans l'algorithme d'assimilation de données existant. Ces travaux permettront d'aller vers un pilotage intégré de la qualité de l'eau en Seine et notamment de fournir une aide à la décision quant aux rejets émis par les stations d'épuration.

Dans le chapitre 4, j'ai mentionné le besoin de disposer de modèles spatio-temporels garantissant de bonnes propriétés de prédiction en intégrant des contraintes physiques sur la dynamique du modèle. Une première piste évoquée est celle des équations différentielles stochastiques spatio-temporelles dont la discrétisation produit des fonctions aléatoires permettant de représenter des phénomènes tels que l'advection ou la diffusion. Dans ce cadre, la dynamique est imposée par l'EDPS et ses paramètres doivent être estimés à partir des données disponibles. L'inférence de ces modèles est rendue difficile par leur complexité. En particulier, on a vu que l'on pouvait facilement intégrer des anisotropies variables dans le terme de diffusion. Il en est de même de l'advection que l'on peut représenter par un champ de vecteurs. Par ailleurs, ces deux champs de paramètres peuvent non seulement varier dans l'espace mais aussi dans le temps. On peut alors imaginer les modéliser selon une approche hiérarchique par des champs latents spatio-temporels, par exemple des champs gaussiens suffisamment réguliers pour qu'une représentation de rang faible soit efficace et afin d'assurer une certaine parcimonie, ou alors via des réseaux de neurones en étendant au cas spatio-temporel l'approche proposée dans Zammit-Mangion et al. [2021]. Ici encore, outre le problème de proposer un algorithme efficace pour l'estimation des paramètres, on risquera de se heurter à des problèmes similaires à du surapprentissage, comme évoqué au chapitre 4. On préférera si possible guider ces non stationnarités par des variables auxi-

liaires. Par exemple, si on s'intéresse à la concentration d'un polluant dans l'atmosphère, le terme d'advection sera naturellement gouverné par le vent. Je coencadre actuellement la thèse de Lucia Clarotto avec Denis Allard (BioSP, INRAe) et Nicolas Desassis (Centre de Géosciences) sur ce sujet.

Dans une SPDE spatio-temporelle, si la différentielle en temps est du premier ordre, on peut s'apercevoir qu'un schéma de discrétisation du même ordre fera apparaître une dynamique temporelle markovienne. Une approche plus générale consiste à supposer l'évolution markovienne d'un champ spatial et à en estimer le noyau de transition ([Cressie and Wikle, 2011]). Des travaux récents proposent de modéliser ce noyau par des réseaux de neurones ([Zammit-Mangion and Wikle, 2020]). Il est possible que cette approche autorise plus de flexibilité que les SPDE, notamment si la physique du phénomène est imparfaitement connue. L'inférence de la dynamique temporelle est d'ailleurs une thématique que l'on retrouve en assimilation de données où les approches par réseaux de neurones sont également privilégiées pour représenter des dynamiques complexes ([Bocquet et al., 2020]). Parmi ces différentes méthodes, il convient d'identifier leur applicabilité selon les cas de figure, notamment la connaissance de la physique du phénomène et le nombre de données disponibles. Je souhaite également explorer cette piste dans le futur.

Par ailleurs, j'ai récemment démarré une collaboration avec Laurent Gerbaud du Centre de Géosciences sur la thématique du forage pour la géothermie. L'idée est de chercher à prédire les dysfonctionnements lors du forage à partir des données disponibles en surface. La physique du phénomène est bien connue, cependant la résolution numérique des équations le régissant est plus lente que le forage en lui-même. Il n'est donc pas possible de mettre en place directement un schéma d'assimilation de données. La solution passera donc par l'intégration de la physique du phénomène par d'autres biais : le développement de proxys, par exemple par des méthodes de modélisation en base réduite ([Hesthaven et al., 2016]), ou bien via les Physics Informed Neural Networks ([Raissi et al., 2019]). Ces travaux seront abordés dans le cadre de la thèse d'Hana Yahia qui démarrera en octobre prochain et à l'encadrement de laquelle je contribuerai en compagnie de Laurent Gerbaud, Florent Di Meglio (Centre Automatique et Systèmes, Mines ParisTech) et Bruno Figliuzzi (Centre de Morphologie Mathématique, Mines ParisTech).

Le logiciel Flumy est développé depuis de nombreuses années au sein de l'équipe géostatistique en collaboration avec l'équipe de géologie du centre de Géosciences. Il s'agit d'un simulateur génétique de réservoir chenalisés méandriformes ou turbiditiques où les processus physiques de sédimentation sont simulés numériquement. Les cubes 3D générés sont très réalistes du point de vue de la géologie et surpassent en cela les modèles à base de fonctions aléatoires gaussiennes couramment employés. Cependant, leur usage est limité car il est difficile voire impossible de les conditionner à des données réelles. Dans une collaboration avec Nicolas Desassis, nous explorons le potentiel des Generative Adversarial Networks (GANs, [Goodfellow et al., 2014]) pour générer des représentations similaires et simplifier leur conditionnement. Ce projet va se poursuivre avec le démarrage de la thèse de Ferdinand Bhavsar en octobre prochain. Dans le cadre de cette thèse, nous chercherons également à appliquer ces méthodes à des données de télédétection.

La plupart de ces projets s'inscrivent dans un cadre plus général abordé en filigrane

---

tout au long du manuscrit qui est celui du positionnement de la géostatistique vis-à-vis des approches machine learning, notamment les réseaux de neurones et leurs différentes applications. On a vu qu'à la différence de l'analyse d'image par exemple où les réseaux de neurones ont pratiquement supplanté les méthodes préexistantes, l'application directe de ces approches à la prédiction spatiale n'était pas efficace. Cependant, comme évoqué, on peut tout à fait imaginer combiner méthodes géostatistique et réseaux de neurones dans des approches hybrides. Ces nouvelles approches offrent un certain nombre d'opportunités dans l'ensemble de mes thématiques de recherche. Elles présentent néanmoins certaines limites qu'il conviendra d'identifier. Je pense en particulier à la manière de caractériser les incertitudes dans l'utilisation des réseaux de neurones. Les prédictions effectuées ne s'accompagnent pas d'une quantification des incertitudes associées. Des travaux récents proposent une approche générale permettant de construire des intervalles de confiance fiables dans ce cadre ([Romano et al., 2019]).



# Bibliographie

- A. M. Alam, S. Benjemaa, and T. Romary. Performance evaluation of covariance tapering for coverage mapping. In *2018 IEEE 87th Vehicular Technology Conference (VTC Spring)*, pages 1–5. IEEE, 2018.
- D. Allard and G. Guillot. Clustering geostatistical data. In *Proceedings of the sixth geostatistical conference*, 2000.
- E. B. Anderes and M. L. Stein. Local likelihood estimation for nonstationary random fields. *Journal of Multivariate Analysis*, 102(3) :506–520, 2011.
- C. Andrieu, A. Jasra, A. Doucet, and P. Del Moral. On nonlinear markov chain monte carlo. *Bernoulli*, 17(3) :987–1014, 2011.
- F. R. Bach and M. I. Jordan. Predictive low-rank decomposition for kernel methods. In *Proceedings of the 22nd international conference on Machine learning, ICML '05*, pages 33–40, New York, NY, USA, 2005. ACM. ISBN 1-59593-180-5. doi : 10.1145/1102351.1102356. URL <http://doi.acm.org/10.1145/1102351.1102356>.
- H. Bakka, E. Krainski, D. Bolin, H. Rue, and F. Lindgren. The diffusion-based extension of the mat\`ern field to space-time. *arXiv preprint arXiv :2006.04917*, 2020.
- S. Banerjee, A. E. Gelfand, A. O. Finley, and H. Sang. Gaussian predictive process models for large spatial data sets. *Journal of the Royal Statistical Society. Series B*, 70(4) :825–848, 2008.
- A. Bechler, T. Romary, N. Jeannée, and Y. Desnoyers. Geostatistical sampling optimization of contaminated facilities. *Stochastic Environmental Research and Risk Assessment*, 27(8) :1967–1974, 2013.
- J. Belhadj. *Parametric models for bayesian seismic tomography*. Theses, PSL Research University, Dec. 2016. URL <https://pastel.archives-ouvertes.fr/tel-01764943>.
- J. Belhadj, T. Romary, A. Gesret, and M. Noble. First arrival travel time tomography-bayesian approach. In *Petroleum Geostatistics 2015*, 2015.
- J. Belhadj, T. Romary, A. Gesret, M. Noble, and B. Figliuzzi. New parameterizations for bayesian seismic tomography. *Inverse Problems*, 34(6) :065007, 2018.
- L. Bertino, G. Evensen, and H. Wackernagel. Sequential data assimilation techniques in oceanography. *International Statistical Review*, 71(2) :223–241, 2003. doi : 10.1111/j.1751-5823.2003.tb00194.x. URL <https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1751-5823.2003.tb00194.x>.
- G. Biau and E. Scornet. A random forest guided tour. *Test*, 25(2) :197–227, 2016.
- G. Billen, J. Garnier, and P. Hanset. Modelling phytoplankton development in whole drainage networks : the RIVERSTRAHLER Model applied to the Seine river system. *Hydrobiologia*, 289 :119–137, 1994.

- 
- M. Bocquet, J. Brajard, A. Carrassi, and L. Bertino. Bayesian inference of chaotic dynamics by merging data assimilation, machine learning and expectation-maximization. *arXiv preprint arXiv :2001.06270*, 2020.
- T. Bodin, M. Sambridge, and K. Gallagher. A self-parametrizing partition model approach to tomographic inverse problems. *Inverse Problems*, 25(5) :055009, 2009.
- D. Bolin and K. Kirchner. The spde approach for gaussian random fields with general smoothness. *arXiv preprint arXiv :1711.04333*, 2017.
- A. Bottero, A. Gesret, T. Romary, M. Noble, and C. Maisons. Stochastic seismic tomography by interacting markov chains. *Geophysical Journal International*, 207(1) :374–392, 2016.
- H. Braham, S. B. Jemaa, G. Fort, E. Moulines, and B. Sayrac. Fixed rank kriging for cellular coverage analysis. *IEEE Transactions on Vehicular Technology*, 66(5) :4212–4222, 2017.
- L. Breiman. Random forests. *Machine learning*, 45(1) :5–32, 2001.
- F. Campolongo, J. Cariboni, and A. Saltelli. An effective screening design for sensitivity analysis of large models. *Environmental Modelling & Software*, 22(10) :1509–1518, 2007. doi : 10.1016/j.envsoft.2006.10.004.
- O. Cappé, A. Guillin, J.-M. Marin, and C. P. Robert. Population Monte-Carlo. *Journal of Computational and Graphical Statistics*, 13(4) :907–929, 2004.
- A. Carrassi, M. Bocquet, L. Bertino, and G. Evensen. Data assimilation in the geosciences : An overview of methods, issues, and perspectives. *Wiley Interdisciplinary Reviews : Climate Change*, 9(5) :e535, 2018.
- G. Celeux, M. Hurn, and C. P. Robert. Computational and inferential difficulties with mixture posterior distributions. Technical Report RR-3627, INRIA, 1999.
- G. Celeux, F. Forbes, and N. Peyrard. Em procedures using mean field-like approximations for markov model-based image segmentation. *Pattern recognition*, 36(1) :131–144, 2003.
- K. Chaloner and I. Verdinelli. Bayesian experimental design : a review. *Statistical Science*, 10(3) :273–304, 1995.
- G. Chastaing, F. Gamboa, and C. Prieur. Generalized hoeffding-sobol decomposition for dependent variables - application to sensitivity analysis. *Electron. J. Statist.*, 6 :2420–2448, 2012. doi : 10.1214/12-EJS749. URL <https://doi.org/10.1214/12-EJS749>.
- J. P. Chilès and P. Delfiner. *Geostatistics, Modeling Spatial Uncertainty*. John Wiley & Sons, New-York, 2nd edition, 2012.
- N. Cressie and G. Johannesson. Fixed rank kriging for very large spatial data sets. *Journal of the Royal Statistical Society. Series B*, 70 :209–226, 2008.
- N. Cressie and C. Wikle. *Statistics for Spatio-Temporal Data*. Wiley Series in Probability and Statistics. Wiley, 2011. ISBN 9780471692744. URL <http://books.google.fr/books?id=-k0C6D0DiNYC>.
- A. Datta, S. Banerjee, A. O. Finley, and A. E. Gelfand. Hierarchical nearest-neighbor gaussian process models for large geostatistical datasets. *Journal of the American Statistical Association*, 111(514) : 800–812, 2016.
- S. De Iaco, D. E. Myers, and D. Posa. Space-time analysis using a general product-sum model. *Statistics & Probability Letters*, 52(1) :21–28, 2001.

## BIBLIOGRAPHIE

---

- S. De Iaco, D. E. Myers, and D. Posa. Nonseparable space-time covariance models : some parametric families. *Mathematical Geology*, 34 :23–42, 2002.
- N. Desassis and D. Renard. Automatic variogram modeling by iterative least squares : Univariate and multivariate cases. *Mathematical Geosciences*, 45(4) :453–470, 2013.
- C. Dietrich and G. Newsam. A fast and exact method for multidimensional gaussian stochastic simulations. *Water Resources Research*, 29(8) :2861–2869, 1993.
- P. J. Diggle and P. J. R. Jr. *Model-based geostatistics*. Springer, 2007.
- P. J. Diggle and S. Lophaven. Bayesian geostatistical design. *Scandinavian Journal of Statistics*, 33(1) : 53–64, 2006.
- P. J. Diggle, R. Menezes, and T.-l. Su. Geostatistical inference under preferential sampling. *Journal of the Royal Statistical Society : Series C (Applied Statistics)*, 59(2) :191–232, 2010. ISSN 1467-9876. doi : 10.1111/j.1467-9876.2009.00701.x. URL <http://dx.doi.org/10.1111/j.1467-9876.2009.00701.x>.
- A. Doucet and A. M. Johansen. A tutorial on particle filtering and smoothing : Fifteen years later. *Handbook of nonlinear filtering*, 12(656-704) :3, 2009.
- A. Doucet, S. Godsill, and C. Andrieu. On sequential monte carlo sampling methods for bayesian filtering. *Statistics and computing*, 10(3) :197–208, 2000.
- F. W. Elliott Jr, D. J. Horntrop, and A. J. Majda. A fourier–wavelet monte carlo method for fractal random fields. *Journal of Computational Physics*, 132(2) :384–408, 1997.
- X. Emery and D. Arroyo. On a continuous spectral algorithm for simulating non-stationary gaussian random fields. *Stochastic Environmental Research and Risk Assessment*, 2017.
- G. Evensen. *Data assimilation : the ensemble Kalman filter*. Springer Science & Business Media, 2009.
- R. Fablet, M. Amar, Q. Febvre, M. Beauchamp, and B. Chapron. End-to-end physics-informed representation learning from and for satellite ocean remote sensing data. In *Proc. XXIV ISPRS Congress*, 2021.
- F. Fouedjio. A hierarchical clustering method for multivariate geostatistical data. *Spatial Statistics*, 18 : 333–351, 2016.
- F. Fouedjio. Second-order non-stationary modeling approaches for univariate geostatistical data. *Stochastic Environmental Research and Risk Assessment*, 31(8) :1887–1906, Oct. 2017. ISSN 1436-3259. doi : 10.1007/s00477-016-1274-y. URL <https://doi.org/10.1007/s00477-016-1274-y>.
- F. Fouedjio. Clustering of multivariate geostatistical data. *WIREs Computational Statistics*, 12(5) :e1510, 2020. doi : <https://doi.org/10.1002/wics.1510>. URL <https://onlinelibrary.wiley.com/doi/abs/10.1002/wics.1510>.
- F. Fouedjio and J. Klump. Exploring prediction uncertainty of spatial data in geostatistical and machine learning approaches. *Environmental Earth Sciences*, 78(1) :38, 2019.
- F. Fouedjio, N. Desassis, and T. Romary. Estimation of space deformation model for non-stationary random functions. *Spatial Statistics*, 13 :45 – 61, 2015. ISSN 2211-6753. doi : <http://dx.doi.org/10.1016/j.spasta.2015.05.001>. URL <http://www.sciencedirect.com/science/article/pii/S221167531500041X>.
- F. Fouedjio, N. Desassis, and J. Rivoirard. A generalized convolution model and estimation for non-stationary random functions. *Spatial Statistics*, 16 :35–52, 2016.



- M. Fuentes. A formal test for nonstationarity of spatial stochastic processes. *Journal of Multivariate Analysis*, 96(1) :30–54, 2005. ISSN 0047-259X. doi : <https://doi.org/10.1016/j.jmva.2004.09.003>. URL <https://www.sciencedirect.com/science/article/pii/S0047259X04001757>.
- G.-A. Fuglstad, D. Simpson, F. Lindgren, and H. Rue. Does non-stationary spatial data always require non-stationary random fields? *Spatial Statistics*, 14 :505–531, 2015. ISSN 2211-6753. doi : <https://doi.org/10.1016/j.spasta.2015.10.001>. URL <https://www.sciencedirect.com/science/article/pii/S2211675315000780>.
- R. Furrer, M. G. Genton, and D. W. Nychka. Covariance tapering for interpolation of large spatial datasets. *Journal of Computational and Graphical Statistics*, 15(3) :502–523, 2006.
- J. Garnier, G. Billen, and M. Coste. Seasonal succession of diatoms and chlorophyceae in the drainage network of the river Seine : Observations and modelling. *Limnology and Oceanography*, 40(4) :750–765, 1995.
- Q. Ge and M. Menendez. Extending morris method for qualitative global sensitivity analysis of models with dependent inputs. *Reliability Engineering & System Safety*, 162 :28 – 39, 2017. ISSN 0951-8320. doi : <https://doi.org/10.1016/j.res.2017.01.010>. URL <http://www.sciencedirect.com/science/article/pii/S0951832017300625>.
- A. E. Gelfand, P. J. Diggle, M. Fuentes, and P. Guttorp, editors. *Handbook of Spatial Statistics*. Chapman & Hall/CRC, 2011.
- C. J. Geyer. Markov Chain Monte Carlo Maximum Likelihood. In *Computing Science and Statistics : Proceedings of 23rd Symposium on the Interface Interface Foundation*, page 156. American Statistical Association, Fairfax Station, New-York, 1991.
- C. J. Geyer and E. A. Thompson. Annealing Markov Chain Monte Carlo with Applications to Ancestral Inference. *Journal of the American Statistical Association*, 90(431) :909–920, 1995.
- R. G. Ghanem and P. D. Spanos. *Stochastic Finite Elements, a Spectral Approach*. Springer-Verlag, 1991.
- J. Ginsberg, M. H. Mohebbi, R. S. Patel, L. Brammer, M. S. Smolinski, and L. Brilliant. Detecting influenza epidemics using search engine query data. *Nature*, 457(7232) :1012–1014, 2009.
- S. Girard, T. Romary, J.-M. Favennec, P. Stabat, and H. Wackernagel. Sensitivity analysis and dimension reduction of a steam generator model for clogging diagnosis. *Reliability Engineering & System Safety*, 113 :143–153, 2013.
- M. Girolami and B. Calderhead. Riemann manifold langevin and hamiltonian monte carlo methods. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, 73(2) :123–214, 2011.
- T. Gneiting. Compactly supported correlation functions. *Journal of Multivariate Analysis*, 83(2) :493–508, 2002a.
- T. Gneiting. Nonseparable, stationary covariance functions for space–time data. *Journal of the American Statistical Association*, 97(458) :590–600, 2002b.
- I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.
- R. B. Gramacy. *Surrogates : Gaussian process modeling, design, and optimization for the applied sciences*. CRC Press, 2020.

## BIBLIOGRAPHIE

---

- P. J. Green. Reversible jump markov chain monte carlo computation and bayesian model determination. *Biometrika*, pages 711–732, 1995.
- A. Gressent, L. Malherbe, A. Colette, H. Rollin, and R. Scimia. Data fusion for air quality mapping using low-cost sensor observations : Feasibility and added-value. *Environment International*, 143 : 105965, 2020. ISSN 0160-4120. doi : <https://doi.org/10.1016/j.envint.2020.105965>. URL <https://www.sciencedirect.com/science/article/pii/S0160412020319206>.
- G. Guillot, D. Kan-King-Yu, J. Michelin, and P. Huet. Inference of a hidden spatial tessellation from multivariate data : application to the delineation of homogeneous regions in an agricultural field. *Journal of the Royal Statistical Society : Series C (Applied Statistics)*, 55(3) :407–430, 2006.
- J. Guinness. Permutation and grouping methods for sharpening gaussian process approximations. *Technometrics*, 60(4) :415–429, 2018.
- J. Guinness. Gaussian process learning via fisher scoring of vecchia’s approximation. *Statistics and Computing*, 31(3) :1–8, 2021.
- X. Guyon. *Random fields on a network*. Springer, 1995.
- H. Haario, E. Saksman, and J. Tamminen. An Adaptive Metropolis Algorithm. *Bernoulli*, 7 :223–242, 2001.
- W. Hackbusch. *Hierarchical matrices : algorithms and analysis*, volume 49. Springer, 2015.
- T. Hastie, R. Tibshirani, and J. Friedman. *The elements of statistical learning*. Springer, 2nd edition, 2009.
- W. K. Hastings. "monte carlo sampling methods using markov chains and their applications". *Biometrika*, 57(1) :97–109, 1970.
- J. S. Hesthaven, G. Rozza, B. Stamm, et al. *Certified reduced basis methods for parametrized partial differential equations*, volume 590. Springer, 2016.
- M. D. Hoffman and A. Gelman. The no-u-turn sampler : adaptively setting path lengths in hamiltonian monte carlo. *J. Mach. Learn. Res.*, 15(1) :1593–1623, 2014.
- Y. Iba. Extended Ensemble Monte Carlo. *International Journal of Modern Physics C*, 12(5) :623 – 656, 2001.
- B. Iooss and P. Lemaître. A review on global sensitivity analysis methods. In *Uncertainty management in simulation-optimization of complex systems*, pages 101–122. Springer, 2015.
- B. Iooss and C. Prieur. Shapley effects for sensitivity analysis with correlated inputs : comparisons with sobol’indices, numerical estimation and applications. *International Journal for Uncertainty Quantification*, 9(5), 2019.
- B. Jafarpour, D. B. McLaughlin, et al. Reservoir characterization with the discrete cosine transform. *SPE Journal*, 14(01) :182–201, 2009.
- M. Katzfuss and N. Cressie. Spatio-temporal smoothing and em estimation for massive remote-sensing data sets. *Journal of Time Series Analysis*, 32(4) :430–446, 2011.
- M. Katzfuss and J. Guinness. A general framework for vecchia approximations of gaussian processes. *Statistical Science*, 36(1) :124–141, 2021.

- 
- M. Katzfuss, J. R. Stroud, and C. K. Wikle. Ensemble kalman methods for high-dimensional hierarchical dynamic space-time models. *Journal of the American Statistical Association*, 0(ja) :1–68, 2019. doi : 10.1080/01621459.2019.1592753. URL <https://doi.org/10.1080/01621459.2019.1592753>.
- M. Katzfuss, J. Guinness, W. Gong, and D. Zilber. Vecchia approximations of gaussian-process predictions. *Journal of Agricultural, Biological and Environmental Statistics*, 25(3) :383–414, 2020.
- C. G. Kaufman, M. J. Schervish, and D. W. Nychka. Covariance tapering for likelihood-based estimation in large spatial data sets. *Journal of the American Statistical Association*, 103(484) :1545–1555, 2008. doi :10.1198/016214508000000959.
- D. P. Kingma and J. Ba. Adam : A method for stochastic optimization. *arXiv preprint arXiv :1412.6980*, 2014.
- S. Kirkpatrick, C. D. Gelatt Jr., and M. P. Vecchi. Optimization by Simulated Annealing. *Science*, 220 (4598) :671–680, 1983.
- S. C. Kou, Q. Zhou, and W. H. Wong. Equi-Energy Sampler with Applications in Statistical Inference and Statistical Mechanics. *The Annals of Statistics*, 34(4) :1581–1619, 2006.
- S. Kucherenko, S. Tarantola, and P. Annoni. Estimation of global sensitivity indices for models with dependent variables. *Computer physics communications*, 183(4) :937–946, 2012.
- C. Lajaunie and T. Romary. 4 particle filters for the SIR model. Technical Report R090717CLAJ, Mines ParisTech, Centre de Géosciences, Equipe Géostatistique, 2009.
- C. Lajaunie, D. Renard, A. Quentin, V. Le Guen, and Y. Caffari. A non-homogeneous model for kriging dosimetric data. *Mathematical Geosciences*, 52(7) :847–863, 2020.
- E. Laloy, R. Héroult, D. Jacques, and N. Linde. Efficient training-image based geostatistical simulation and inversion using a spatial generative adversarial neural network. *arXiv preprint arXiv :1708.04975*, 2017.
- M. Lamboni, H. Monod, and D. Makowski. Multivariate sensitivity analysis to measure global contribution of input factors in dynamic models. *Reliability Engineering & System Safety*, 96(4) :450–459, 2011.
- G. Lance and W. Williams. A generalized sorting strategy for computer classifications. *Nature*, 1966.
- J. Langanay, T. Romary, X. Freulon, V. Langlais, G. Petit, and V. Lagneau. Uncertainty quantification for uranium production in mining exploitation by in situ recovery. *Computational Geosciences*, pages 1–20, 2021.
- C. Lantuéjoul. *Geostatistical simulation*. Springer, Berlin, 2002.
- J.-B. Leger, C. Vacher, and J.-J. Daudin. Detection of structurally homogeneous subsets in graphs. *Statistics and computing*, 24(5) :675–692, 2014.
- J.-B. Leger, P. Barbillon, and J. Chiquet. *blockmodels : Latent and Stochastic Block Model Estimation by a 'V-EM' Algorithm*, 2020. URL <https://CRAN.R-project.org/package=blockmodels>. R package version 1.1.4.
- Y. Li, Y. Sun, and B. J. Reich. Deepkriging : Spatially dependent deep neural networks for spatial prediction. *arXiv preprint arXiv :2007.11972*, 2020.
- F. Lindgren, H. Rue, and J. Lindström. An explicit link between gaussian fields and gaussian markov random fields : the stochastic partial differential equation approach. *Journal of the Royal Statistical Society : Series B*, 73(4) :423–498, 2011. ISSN 1467-9868.

## BIBLIOGRAPHIE

---

- M. Loève. *Probability Theory*. Van Nostrand, New-York, 1955.
- S. Lopez, I. Cojan, J. Rivoirard, and A. Galli. Process-based stochastic modelling : Meandering channelized reservoirs. *Analogue and Numerical Modelling of Sedimentary Systems : From Understanding to Prediction (Special Publication 40 of the IAS)*, 40 :139, 2009.
- A. Malinverno. Parsimonious Bayesian Markov chain Monte Carlo inversion in a nonlinear geophysical problem. *Geophysical Journal International*, 151(3) :675–688, 12 2002. ISSN 0956-540X. doi : 10.1046/j.1365-246X.2002.01847.x. URL <https://doi.org/10.1046/j.1365-246X.2002.01847.x>.
- K. V. Mardia and R. J. Marshall. Maximum likelihood estimation of models for residual covariance in spatial regression. *biometrika*, 71(1) :135–46, 1984.
- G. Mariethoz and J. Caers. *Multiple-point geostatistics : stochastic modeling with training images*. John Wiley & Sons, 2014.
- E. Marinari and G. Parisi. Simulated tempering : A new monte-carlo scheme. *Europhysics Letters*, 19 : 451–458, 1992.
- I. Markovsky and K. Usevich. *Low rank approximation*, volume 139. Springer, 2012.
- J. Mateu and W. G. Müller. *Collecting spatio-temporal data*, 2012.
- G. Matheron. *Les variables régionalisées et leur estimation. Une application de la théorie des fonctions aléatoires aux Sciences de la nature*. Masson, 1965.
- G. Matheron. The intrinsic random functions and their applications. *Advances in applied probability*, 5 (3) :439–468, 1973.
- G. Matheron. *Estimer et choisir : essai sur la pratique des probabilités*. Ecole nationale supérieure des mines de Paris, 1978.
- P.-A. Mattei and J. Frellsen. Leveraging the exact likelihood of deep latent variable models. In *Advances in Neural Information Processing Systems*, pages 3855–3866, 2018.
- N. Metropolis, A. Rosenbluth, M. Rosenbluth, and A. T. M. Teller. Equations of state calculations by fast computing machines. *Journal of Chemical Physics*, 21 :1087–1091, 1953.
- S. P. Meyn and R. L. Tweedie. *Markov Chains and Stochastic Stability*. Springer, Berlin, 1993.
- J. Møller. Random johnson-mehl tessellations. *Advances in applied probability*, 24(4) :814–844, 1992.
- M. Morris. Factorial sampling plans for preliminary computational experiments. *Technometrics*, 33 : 161–174, 1991.
- L. Mosser, O. Dubrulle, and M. J. Blunt. Reconstruction of three-dimensional porous media using generative adversarial neural networks. *Physical Review E*, 96(4) :043309, 2017.
- R. N. Neal. *Handbook of Markov Chain Monte Carlo*, chapter MCMC using Hamiltonian dynamics. Chapman & Hall / CRC Press, 2010. available at <http://www.cs.utoronto.ca/~radford/ftp/ham-mcmc.pdf>.
- R. Network. Analyzing spatio-temporal data with R : Everything you always wanted to know – but were afraid to ask. *Journal de la Societe Française de Statistique*, 158(3) :124–158, 2017. URL <https://hal.archives-ouvertes.fr/hal-01606944>.
- A. Ng, M. Jordan, and Y. Weiss. On spectral clustering : analysis and an algorithm. In T. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems*, volume 14, page 849 – 856. MIT Press, 2002.

- D. Nychka, C. Wikle, and J. A. Royle. Multiresolution models for nonstationary spatial covariance functions. *Statistical Modelling*, 2 :315–331, 2002.
- A. B. Owen. Sobol’indices and shapley value. *SIAM/ASA Journal on Uncertainty Quantification*, 2(1) : 245–251, 2014.
- M. Pereira. *Champs aléatoires généralisés définis sur des variétés riemanniennes : théorie et pratique*. PhD thesis, Mines ParisTech, Nov. 2019. URL <http://www.theses.fr/s172855>.
- N. Piana Agostinetti, G. Giacomuzzi, and A. Malinverno. Local three-dimensional earthquake tomography by trans-dimensional monte carlo sampling. *Geophysical Journal International*, 201(3) :1598–1617, 2015.
- E. Porcu, J. Mateu, and G. Christakos. Quasi-arithmetic means of covariance functions with potential applications to space–time data. *Journal of Multivariate Analysis*, 100(8) :1830–1844, 2009.
- E. Porcu, R. Furrer, and D. Nychka. 30 years of space–time covariance functions. *Wiley Interdisciplinary Reviews : Computational Statistics*, 13(2) :e1512, 2021.
- M. Raissi, P. Perdikaris, and G. E. Karniadakis. Physics-informed neural networks : A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational Physics*, 378 :686–707, 2019.
- B. J. Reich, K. Pacifici, and J. W. Stallings. Integrating auxiliary data in optimal spatial design for species distribution modelling. *Methods in Ecology and Evolution*, 9(6) :1626–1637, 2018. doi : <https://doi.org/10.1111/2041-210X.13002>. URL <https://besjournals.onlinelibrary.wiley.com/doi/abs/10.1111/2041-210X.13002>.
- D. Revuz. *Probabilités*. Hermann, 1997.
- J. Rivoirard and T. Romary. Continuity for kriging with moving neighborhood. *Mathematical geosciences*, 43(4) :469–481, 2011.
- C. P. Robert and G. Casella. *Monte-Carlo Statistical Methods*. Springer, Berlin, 2nd edition, edition, 2004.
- G. O. Roberts and J. S. Rosenthal. Examples of Adaptive MCMC. *Preprint*, 2006.
- Y. Romano, E. Patterson, and E. Candes. Conformalized quantile regression. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d’ Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL <https://proceedings.neurips.cc/paper/2019/file/5103c3584b063c431bd1268e9b5e76fb-Paper.pdf>.
- T. Romary. *Inversion des modèles stochastiques de milieux hétérogènes*. PhD thesis, Paris 6, 2008.
- T. Romary. Integrating production data under uncertainty by parallel interacting Markov chains on a reduced dimensional space . *Computational Geosciences*, 13(1) :103–122, 2009a.
- T. Romary. History matching of approximated lithofacies models by interacting Markov chains . *Computational Geosciences*, 14(2) :343–355, 2009b.
- T. Romary. Bayesian inversion by parallel interacting markov chains. *Inverse Problems in Science and Engineering*, 18(1) :111–130, 2010.
- T. Romary. Incomplete cholesky decomposition for the kriging of large datasets. *Spatial Statistics*, 5 : 85–99, 2013.
- T. Romary. Bayesian tomography by interacting markov chains. *AGUFM*, 2017 :S31E–01, 2017.

## BIBLIOGRAPHIE

---

- T. Romary and N. Desassis. Low rank plus sparse covariance decomposition for a scalable non stationary kriging. *arXiv preprint arXiv :1806.01558*, 2018a.
- T. Romary and N. Desassis. Second order finite elements for the spde approach. *RESSTE SPDE/INLA workshop*, 2018b.
- T. Romary, C. de Fouquet, and L. Malherbe. Sampling design for air quality measurement surveys : An optimization approach. *Atmospheric Environment*, 45(21) :3613 – 3620, 2011.
- T. Romary, L. Malherbe, and C. de Fouquet. Optimal spatial design for air quality measurement surveys. *Environmetrics*, 25(1) :16–28, 2014.
- J. A. Rozanov. Markov random fields and stochastic partial differential equations. *Mathematics of the USSR-Sbornik*, 32(4) :515, 1977.
- H. Rue and L. Held. *Gaussian Markov random fields : theory and applications*. CRC press, 2005.
- A. Saltelli, K. Chan, M. Scott, et al. *Sensitivity analysis. Probability and statistics series*. John and Wiley & Sons, New York, 2000.
- P. D. Sampson and P. Guttorp. Nonparametric estimation of nonstationary spatial covariance structure. *Journal of the American Statistical Association*, 87(417) :108–119, 1992.
- H. Sang and J. Huang. A full scale approximation of covariance functions for large spatial data sets. *Journal of the Royal Statistical Society. Series B*, 74(1) :111–132, 2012.
- T. J. Santner, B. J. Williams, W. I. Notz, and B. J. Williams. *The design and analysis of computer experiments*, volume 1. Springer, 2018.
- C. Scheidt and J. Caers. Representing spatial uncertainty using distances and kernels. *Mathematical Geosciences*, 41(4) :397–419, 2009.
- A. M. Schmidt, P. Guttorp, and A. O’Hagan. Considering covariates in the covariance structure of spatial processes. *Environmetrics*, 22(4) :487–500, 2011. doi : <https://doi.org/10.1002/env.1101>. URL <https://onlinelibrary.wiley.com/doi/abs/10.1002/env.1101>.
- L. S. Shapley. *Contributions to the theory of games II*, chapter A value for n-person games. Princeton University Press, 1953.
- M. Shinozuka and C.-M. Jan. Digital simulation of random processes and its applications. *Journal of Sound and Vibration*, 25(1) :111–128, 1972. ISSN 0022-460X. doi : [https://doi.org/10.1016/0022-460X\(72\)90600-1](https://doi.org/10.1016/0022-460X(72)90600-1). URL <https://www.sciencedirect.com/science/article/pii/0022460X72906001>.
- F. Sigrist, H. R. Künsch, and W. A. Stahel. Stochastic partial differential equation based modelling of large space–time data sets. *Journal of the Royal Statistical Society : Series B : Statistical Methodology*, pages 3–33, 2015.
- D. Smith, L. Moore, et al. The sir model for spread of disease : the differential equation model. *Loci.(originally Convergence.)* <https://www.maa.org/press/periodicals/loci/joma/the-sir-model-for-spread-of-disease-the-differential-equation-model>, 2004.
- I. Sobol’. Sensitivity estimates for non linear mathematical models. *Mathematical Modelling and Computational Experiments*, 1, 1993.
- M. L. Stein. *Interpolation of spatial data : some theory for kriging*. Springer, 1999.
- M. L. Stein. A modeling approach for large spatial datasets. *Journal of the Korean Statistical Society*, 37 : 3–10, 2008.

- M. L. Stein, Z. Chi, and L. J. Welty. Approximating likelihoods for large spatial data sets. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, 66(2) :275–296, 2004.
- A. Tarantola. *Inverse Problem Theory and Model Parameter Estimation*. SIAM, Philadelphia, 2005.
- C. Varin, N. M. Reid, and D. Firth. An overview of composite likelihood methods. *Statistica Sinica*, 21(1) :5–42, 2011.
- A. V. Vecchia. Estimation and model identification for continuous spatial processes. *Journal of the Royal Statistical Society : Series B (Methodological)*, 50(2) :297–312, 1988.
- R. Vergara. *Développement de modèles géostatistiques à l'aide d'équations aux dérivées partielles stochastiques*. PhD thesis, Paris Sciences et Lettres, 2018.
- R. C. Vergara, D. Allard, and N. Desassis. A general framework for spde-based stationary random fields. *Bernoulli*, 2021.
- U. von Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17(4), 2007.
- H. Wackernagel. *Multivariate geostatistics*. Springer Science & Business Media, 2003.
- H. Wang, Y. Guan, and B. Reich. Nearest-neighbor neural networks for geostatistics. In *2019 International Conference on Data Mining Workshops (ICDMW)*, pages 196–205. IEEE, 2019a.
- S. Wang, N. Flipo, and T. Romary. Time-dependent global sensitivity analysis of the c-rive biogeochemical model in contrasted hydrological and trophic contexts. *Water research*, 144 :341–355, 2018.
- S. Wang, N. Flipo, and T. Romary. Oxygen data assimilation for estimating micro-organism communities' parameters in river systems. *Water Research*, 165 :115021, 2019b.
- S. Wang, N. Flipo, and T. Romary. Particle filter for high frequency oxygen data assimilation in river systems. *submitted to Environmental Modelling & Software*, 2021.
- P. Whittle. Stochastic-processes in several dimensions. *Bulletin of the International Statistical Institute*, 40(2) :974–994, 1963.
- C. K. Wikle, A. Zammit-Mangion, and N. Cressie. *Spatio-temporal Statistics with R*. CRC Press, 2019.
- M. Woillez, P. Ressler, C. Wilson, and J. Horne. Multifrequency species classification of acoustic-trawl survey data using semi-supervised learning with class discovery. *The Journal of the Acoustical Society of America*, 131(2) :EL184–EL190, 2012.
- Y. Yang and P. Perdikaris. Adversarial uncertainty quantification in physics-informed neural networks. *Journal of Computational Physics*, 394 :136–152, 2019. ISSN 0021-9991. doi : <https://doi.org/10.1016/j.jcp.2019.05.027>. URL <https://www.sciencedirect.com/science/article/pii/S0021999119303584>.
- A. Zammit-Mangion and C. K. Wikle. Deep integro-difference equation models for spatio-temporal forecasting. *Spatial Statistics*, 37 :100408, 2020. ISSN 2211-6753. doi : <https://doi.org/10.1016/j.spasta.2020.100408>. URL <https://www.sciencedirect.com/science/article/pii/S2211675320300026>. Frontiers in Spatial and Spatio-temporal Research.
- A. Zammit-Mangion, T. L. J. Ng, Q. Vu, and M. Filippone. Deep compositional spatial models. *Journal of the American Statistical Association*, 0(0) :1–22, 2021. doi : 10.1080/01621459.2021.1887741. URL <https://doi.org/10.1080/01621459.2021.1887741>.
- Z. Zhu and M. L. Stein. Spatial sampling design for parameter estimation of the covariance function. *Journal of Statistical Planning and Inference*, 134(2) :583 – 603, 2005. ISSN 0378-3758. doi : DOI: 10.1016/j.jspi.2004.04.017.

## BIBLIOGRAPHIE

---

- Z. Zhu and M. L. Stein. Spatial sampling design for prediction with estimated parameters. *Journal of Agricultural Biological and Environmental Statistics*, 11 :24–44, 2006.
- D. L. Zimmerman. Optimal network design for spatial prediction, covariance parameter estimation, and empirical prediction. *Environmetrics*, 17(6) :635–652, 2006. ISSN 1099-095X. doi : 10.1002/env.769. URL <http://dx.doi.org/10.1002/env.769>.
- D. L. Zimmerman and N. Cressie. Mean squared prediction error in the spatial linear model with estimated covariance parameters. *Annals of the Institute of Statistical Mathematics*, 44 :27–43, 1992. doi : 10.1007/BF00048668. URL <http://dx.doi.org/10.1007/BF00048668>.





# Annexes

## 6.1 Curriculum Vitae

## Thomas Romary

---

30/09/1981

thomas.romary@mines-paristech.fr

35 rue Saint-Honoré

77305 Fontainebleau, FRANCE

+33 (0)1 64 69 47 04

### THEMES DE RECHERCHE

Géostatistique, Statistique spatiale, Planification d'expérience pour des données spatiales, Krigeage, Classification de données spatiales, Inversion bayésienne, Algorithmes MCMC et SMC

### FORMATION

**Université Pierre et Marie Curie**, Paris, FRANCE

Doctorat en mathématiques, soutenue le 19 Décembre 2008

*Inversion des modèles stochastiques de milieux hétérogènes*

Directeurs de thèse : Jean Jacod (UPMC) et Lin Ying Hu (IFP)

**Université Rennes 1**, Rennes, FRANCE

Master 2 de mathématiques, option statistiques, 2005

**ENSAI**, Bruz, FRANCE

Diplôme d'ingénieur de l'ENSAI (Ecole nationale de la statistique et de l'analyse de l'information), 2005

### EXPERIENCE ACADEMIQUE

**Mines ParisTech**, Centre de Géosciences, Equipe Géostatistique, Fontainebleau, France

*Maître assistant*

*Chargé de recherche*

*Assistant de recherche*

**Juillet 2015 -**

**Décembre 2011 - Juin 2015**

**Février 2009 - Novembre 2011**

**Stanford University**, School of Earth, Energy and environmental science, Stanford, CA, USA

*Visiteur scientifique*

**juillet 2013 - octobre 2013**

**Université Pierre et Marie Curie / IFP**, Paris / Rueil-Malmaison, France

*Doctorant*

**Novembre 2005 - Décembre 2008**

### PROJETS

Responsable de la tâche quantification d'incertitudes de la chaire industrielle **ISR-U** Exploitation de l'uranium par récupération in-situ (2017-2020)

Expert scientifique dans le projet **Exploitation of Geo-Located measurements for radio resource management in cellular networks** mené par Orange (2017-18)

Participation au consortium **G2DC Geostatistical Domaining Consortium** mené par Géovariances (2011 - 13)

Participation au projet de collaboration avec l'INERIS sur l'**optimisation de réseaux de mesures dans les campagnes de surveillance de la qualité de l'air** (2010 - 2013)

Participation au projet Carnot **Tomographie Stochastique** (2010-2011)

**SERVICE  
PROFESSIONNEL**

Organisateur de **Geoenv2014**, the 10th conference on geostatistics for environmental applications (2014)

Reviewing pour : WIREs, Mathematical Geosciences, Spatial Statistics, Statistics and Computing, Computational Statistics and Data Analysis, International Statistical Review, Oil & Gas Science and Technology, Computer and Geosciences, Applied Mathematical modelling, Advances in Water Resources

Membre du comité scientifique des conférences: Geoenv, Spatial Statistics Conference, IAMG

Membre de la SFdS, de la SMAI, du réseau RESSTE (Réseau statistiques pour données spatio-temporelles)

**DISTINCTIONS**

2013 bourse du France-Stanford Center for Interdisciplinary Studies pour le projet "geostatistical methods for remote sensing data" avec le Pr. Jef Caers  
Cette bourse a financé un séjour scientifique de 4 mois à l'université de Stanford

**LANGUES**

Français (langue maternelle)  
Anglais  
Italien

## 6.2 Liste de publications

Voir également mes profils Google scholar et Research gate. Les étudiants encadrés sont mentionnés en [bleu](#).

### 6.2.1 Articles

1. [J. Langanay](#), **T. Romary**, X. Freulon, V. Langlais, G. Petit, V. Lagneau *Uncertainty quantification for uranium production in mining exploitation by In Situ Recovery*, Computational Geosciences, pages 1–20, 2021.
2. **T. Romary**, N. Desassis. *Combining covariance tapering and lasso driven low rank decomposition for the kriging of large spatial datasets*, working paper
3. A.M. Alam, S. Benjemaa, **T. Romary**. *Performance Evaluation of Covariance Tapering for Coverage Mapping*, submitted to IEEE Transactions on Communications
4. Bubnova, A., Ors, F., Rivoirard, J., Cojan, I., & **Romary, T.**. Automatic Determination of Sedimentary Units from Well Data. *Mathematical Geosciences*, 52(2), 213-231. (2020)
5. [Wang, S.](#), Flipo, N., & **Romary, T.** Oxygen data assimilation for estimating micro-organism communities' parameters in river systems. *Water research*, 165, 115021. (2019)
6. [S. Wang](#), N. Flipo, **T. Romary**. *Time-dependent global sensitivity analysis of the C-RIVE biogeochemical model in contrasted hydrological and trophic contexts*, *Water research* 144, 341-355. (2018)
7. [J. Belhadj](#), **T. Romary**, A. Gesret, M. Noble, B. Figliuzzi. *New parameterizations for Bayesian seismic tomography*, *Inverse Problems*. (2018)
8. Network, R. E. S. S. T. E. *Analyzing spatio-temporal data with R : Everything you always wanted to know—but were afraid to ask*. *Journal de la Société Française de Statistique*, 158(3), 124-158. (2017)
9. [A. Bottero](#), A. Gesret, **T. Romary**, M. Noble, C. Maisons. *Stochastic seismic tomography by interacting Markov chains*. *Geophysical Journal International*, 207(1), 374–392. (2016)
10. F. Fouedjio, N. Desassis, **T. Romary**. *Estimation of space deformation model for non-stationary random functions*, *Spatial Statistics*, 13, 45-61. (2015)
11. L. Li, **T. Romary**, J. Caers. *Universal kriging with training images*, *Spatial Statistics*, 14, 240-268, (2015)
12. **T. Romary**, F. Ors, J. Rivoirard, J. Deraisme. *Unsupervised classification of multivariate geostatistical data : Two algorithms*, *Computers and Geosciences*, Elsevier, *Statistical learning in geoscience modelling : Novel algorithms and challenging case studies*, 85, 96-103 (2015)
13. A. Gesret, N. Desassis, M. Noble, **T. Romary**, C. Maisons, *Propagation of the velocity model uncertainties to the seismic event location*, *Geophysical Journal International* 200 (1), 52-66 (2015)

14. **T. Romary**, L. Malherbe, C. de Fouquet, *Optimal spatial design for air quality measurement surveys*, *Environmetrics* 25 (1), 16-28 (2014)
15. **T. Romary**, *Incomplete Cholesky decomposition for the kriging of large datasets*, *Spatial Statistics*, 5, pp 85-99 (2013)
16. **A. Bechler**, **T. Romary**, N. Jeannée, Y. Desnoyers, *Geostatistical Sampling Optimization of Contaminated Facilities*, *Stochastic Environmental Research and Risk Assessment*, 27(8), pp 1967-1974 (2013)
17. **S. Girard**, **T. Romary**, J.-M. Favennec, P. Stabat, H. Wackernagel, *Sensitivity analysis and dimension reduction of a steam generator model for clogging diagnosis*, *Reliability Engineering & System Safety*, 113, 143-153 (2013)
18. **T. Romary**, C. de Fouquet, L. Malherbe, *Sampling design for air quality measurement surveys : an optimisation approach*. *Atmospheric Environment*, 45(21), 3613-3620 (2011)
19. J. Rivoirard and **T. Romary**, *Continuity for kriging with moving neighborhood*. *Mathematical Geosciences*, 43 (4), 469-481 (2011)
20. **T. Romary**, *History matching of approximated lithofacies models under uncertainty*. *Computational Geosciences* 14(2), 343-355 (2010)
21. **T. Romary**, *Bayesian inversion by parallel interacting Markov chains*. *Inverse Problems in Science and Engineering* 18(1), 111-130 (2010)
22. **T. Romary**, *Integrating production data under uncertainty by parallel interacting Markov chains on a reduced dimensional space*. *Computational Geosciences* 13(1), 103-122 (2009)

### 6.2.2 Brevet

D. Busby, M. Feraille, **T. Romary**, S. Touzani, *Method for evaluating an underground reservoir production scheme taking account of uncertainties*. US Patent 8,392,164

### 6.2.3 Communications

J. Chautru, **T. Romary**, H. Binet, M. Bourges. *Geostatistical Electrofacies Calculation in Non-Stationary Cases*. 81st EAGE Conference and Exhibition 2019, Jun 2019, London, UK. 10.3997/2214-4609.201900708.

S. Wang, N. Flipo, **T. Romary**. *Assimilating high-frequency data in a hydro-biogeochemical model of river systems, the ProSe-PA approach*. EGU, Apr 2019, Vienne, Austria.

Langanay, J., **Romary, T.**, Lagneau, V., Petit, G. *Scenario Reduction and dimension reduction in uranium ore deposit mining simulations by In Situ Recovery*. GeoEnv 2018, Jul 2018, Belfast UK.

Alam, A.M., Benjemaa, S., **Romary, T.**. *Performance evaluation of covariance tapering for coverage mapping*. 2018 IEEE 87th Vehicular Technology Conference (VTC Spring), 1-5 (2018)

Alam, A.M., Benjemaa, S., **Romary, T.** Clustering for High Accuracy Coverage Mapping. 2018 IEEE International Conference on Communications (ICC), 1-6 (2018)

Wang, S., Flipo, N., **Romary, T.** (2018, April). Sensitivity Analysis of a biogeochemical model on dissolved oxygen during low water periods. In EGU General Assembly Conference Abstracts (Vol. 20, p. 13425).

**T Romary.** Covariance decomposition for the kriging of large datasets. Data science and environment, Jul 2017, Brest, France.

**T. Romary.** Covariance models for large spatial datasets and their inference. TIES-GRASPA 2017, Jul 2017, Bergamo, Italy.

**T. Romary.** Bayesian tomography by interacting Markov chains. AGU Fall Meeting, Dec 2017, New Orleans, United States.

A. Gesret, J. Belhadj, **T. Romary**, M. Noble, N. Desassis. Uncertainty estimation by probabilistic first arrival time tomography using Markov Chain Monte Carlo sampling. 79th EAGE Conference and Exhibition 2017, 2017, Paris, France. 2017, Expanded abstracts.

A. Gesret, J. Belhadj, **T. Romary**, M. Noble, E. Gaucher, et al.. Microseismic monitoring : Hierarchical transdimensional tomography and propagation of velocity model uncertainties to seismic event locations. AGU Fall Meeting, 2016, San Fransico, United States. 2016.

J. Belhadj, **T. Romary**, A. Gesret, M. Noble. Bayesian seismic tomography as a velocity anomaly detection tool. Geostats 2016, Sep 2016, Valencia, Spain.

L. Wagner, D. Renard, **T. Romary**, G. Pilger, C. Carey, et al.. Combining geostatistical hierarchical clustering (GHC) and potential field method for fast and flexible domaining with updating capabilities : a case study with grade shell modelling in iron ore. Geostats 2016, Sep 2016, Valencia, Spain.

**T. Romary**, N. Desassis, F. Fouedjio. Covariance sale decomposition for the kriging of large datasets. Geostats 2016, Sep 2016, Valencia, Spain.

**T. Romary**, F. Ors, J. Rivoirard. Unsupervised learning of multivariate geostatistical data : two algorithms. Geoenv 2016, Jul 2016, Lisbon, Portugal.

J. Belhadj, **T. Romary**, A. Gesret, M. Noble. Bayesian seismic tomography asa velocity anomaly detection tool. Geostats 2016, Sep 2016, Valencia, Spain.

A. Gesret, J. Belhadj, **T. Romary**, M. Noble, E. Gaucher, et al.. Microseismic monitoring : Hierarchical transdimensional tomography and propagation of velocity model uncertainties to seismic event locations. AGU Fall Meeting, 2016, San Fransico, United States. 2016.

J. Belhadj, **T. Romary**, A. Gesret, M. Noble. First Arrival Travel Time Tomography - Bayesian Approach. Petroleum Geostatistics 2015, Sep 2015, Biarritz, France.

**T. Romary**, N. Desassis, F. Fouedjio. A full scale, non stationary approach for the kriging of large datasets. Spatial Statistics Conference 2015, Jun 2015, Avignon, France.

J. Caers, L. Li, **T. Romary.** Spatial Estimation Using Universal Kriging with Training Images. JSM 2014, Aug 2014, Boston, United States.

L. Li, **T. Romary**, J. Caers. Universal Kriging With Training Images. geoENV 2014, Jul 2014, Paris, France.

A. Gesret, A. Bottero, **T. Romary**, M. Noble, N. Desassis. Bayesian seismic tomogra-

phy by parallel interacting Markov chains. EGU General Assembly, 2014, Austria. pp.Vol. 16, EGU2014-6997-1, 2014.

F. Fouedjio, N. Desassis, **T. Romary**, J. Rivoirard. Estimation Of Non-Stationary Random Fields Through Space Deformation. 10th Conference on Geostatistics for Environmental Applications, Jul 2014, Paris, France.

F. Fouedjio, N. Desassis, **T. Romary**. Estimation Of Non-Stationary Random Fields Through Space Deformation. 11th German Probability and Statistics Days, Mar 2014, Ulm, Germany.

F. Fouedjio, Desassis N., **T. Romary**. Estimation de Fonctions Aléatoires Non-stationnaires de Second Ordre par Déformation Spatiale. 46èmes journées de la Société Française de Statistique (SFdS), Jun 2014, Rennes, France. pp.242, 2014.

F. Fouedjio, N. Desassis, **T. Romary**. Estimation of Space Deformation Models for Non-stationary Random Functions. 21st International Conference on Computational Statistics, Aug 2014, Genève, Switzerland.

**T. Romary**, A. Bottero, M. Noble, A. Gesret, N. Desassis. Bayesian first arrival travel time tomography by interacting MCMC. geoENV 2014, Jul 2014, Paris, France.

**T. Romary**. Interacting Markov chains algorithms for Bayesian inversion. GRF - Sim workshop : Simulation of Gaussian and related Random Fields, Nov 2014, Bern, Switzerland.

**T. Romary**, A. Bottero, A. Gesret, N. Desassis, M. Noble. Bayesian first arrival travel time tomography by interacting MCMC. 15th Annual Conference of the International Association for Mathematical Geosciences, Sep 2013, Madrid, Spain.

A. Gesret, M. Noble, N. Desassis, **T. Romary**. A new Bayesian approach of tomography and seismic event location dedicated to the estimation of the true uncertainties. EGU General Assembly 2013, 2013, Austria.

A. Gesret, M. Noble, N. Desassis, **T. Romary**. Microseismic Monitoring - Consequences of Velocity Model Uncertainties on Event Location Uncertainties. 75th EAGE Conference & Exhibition incorporating SPE EUROPEC 2013, 2013, United Kingdom. pp.1, 2013.

**T. Romary**. Deux algorithmes pour la classification non supervisée de données géostatistiques. 45e Journées de Statistique, May 2013, France.

A. Gesret, M. Noble, N. Desassis, **T. Romary**. Estimation of the velocity model uncertainties and of the associated uncertainties on microseismic event locations by a Bayesian approach. AGU 2012 Fall Meeting, Dec 2012, United States.

**T. Romary**, C. De Fouquet, L. Malherbe. Optimization strategies for spatial sampling of air quality measurement surveys. 9 International geostatistics congress, Jun 2012, Oslo, Norway. 8p, 2012.

**T. Romary**, J. Rivoirard, J. Deraisme, C. Quinones, X. Freulon. Domaining by clustering multivariate geostatistical data. Ninth International Geostatistics Congress, 2012, Norway, France. pp.455-466, 2012.

A. Bechler, **T. Romary**, N. Jeannée, Y. Desnoyers. Geostatistical Sampling Optimization of Contaminated Facilities. geoENV2012, 2012, Spain. pp.13, 2012.

**T. Romary**. Incomplete Cholesky decomposition for the kriging of large datasets.



GeoENV 2012, 2012, Spain. pp.259, 2012.

N. Desassis, **T. Romary**, A. Gesret, M. Noble. Inversion sismique dans un milieu aléatoire. 42e Journées de Statistique, 2011, Tunisia.

S. Girard, **T. Romary**, J.-M. Favennec, P. Stabat, H. Wackernagel. Towards a better understanding of clogged steam generators : a sensitivity analysis of dynamic thermohydraulic model output. ICON19 - 19th International Conference on Nuclear Engineering, Oct 2011, Osaka, Japan. 11 p. ; ICON19-43254, 2011.

A. Gesret, M. Noble, N. Desassis, **T. Romary**. Microseismic Monitoring : Consequences of Velocity Model Uncertainties on Location Uncertainties. Third Passive Seismic Workshop, Eur. Ass. of Geoscientists and Engineers, Athens - Greece (2011), Mar 2011, Athens, Greece. pp.PAS31, 2011.

A. Gesret, M. Noble, **T. Romary**, N. Desassis. Microseismic monitoring : new accurate estimate of velocity model uncertainties and consequences on location uncertainties. EGU General Assembly 2011, Apr 2011, Austria.

**T. Romary**, C. De Fouquet, L. Malherbe. Optimal spatial design for air quality measurement surveys : what criteria ?. Spatial Data Methods for Environmental and Ecological Processes - 2nd Edition, Sep 2011, Foggia, Italy.

M. Noble, A. Gesret, N. Desassis, **T. Romary**. Passive Seismic Monitoring - understanding and reducing event location uncertainties. ESC 2010 European Seismological Commission 32nd General Assembly, Sep 2010, Montpellier, France.

**T. Romary**, J. Rivoirard. Continuity for kriging with moving neighborhood : a penalization approach. IAMG 2010, Aug 2010, Hungary.

**T. Romary**, C. De Fouquet, L. Malherbe. Schéma d'échantillonnage pour les campagnes de mesures de la qualité de l'air : une approche par optimisation. 42èmes Journées de Statistique, 2010, Marseille, France, France. 2010.

**T. Romary**. Krigeage de grands jeux de données : une revue. Journées de Géostatistique 2009, Sep 2009, Fontainebleau, France.

**T. Romary**. History matching by interacting MCMC. 4th international EnKF workshop, Jun 2009, Bergen, Norway.

#### 6.2.4 Thèse

**T. Romary**. Inversion des modèles stochastiques de milieux hétérogènes. Mathématiques. Université Pierre et Marie Curie - Paris VI, 2008.

## 6.3 Encadrement

### Doctorants

- Lucia Clarotto (2020-) Prédiction spatio-temporelle dans l'approche par équations différentielles partielles stochastiques  
co-encadrement avec Denis Allard (BioSP, INRAE) et Nicolas Desassis (Mines ParisTech, centre de Géosciences)
- Masihullah Hasanyar (2020-) High Frequency Data Assimilation in the ProSe-PA Water Quality Model  
co-encadrement avec Nicolas Flipo (Mines ParisTech, centre de Géosciences)
- Jean Langanay (2017 - 2021) Quantification d'incertitudes dans l'exploitation de gisements d'uranium par récupération in situ  
co-encadrement avec Vincent Lagneau (Mines ParisTech, centre de Géosciences)
- Shuaitao Wang (2016 -2019) Simulation du métabolisme de la Seine par assimilation de données en continu  
co-encadrement avec Nicolas Flipo (Mines ParisTech, centre de Géosciences)  
2 articles dans Water Research
- Jihane Belhadj (2013-16) Méthodes et modèles stochastiques pour la géophysique  
co-encadrement avec Alexandrine Gesret et Mark Noble (Mines ParisTech, centre de Géosciences)  
un article dans Inverse Problems
- Sylvain Girard (2010-12) Diagnostic du colmatage des générateurs de vapeur à l'aide de modèles physiques et statistiques  
co-encadrement avec Hans Wackernagel (Mines ParisTech, centre de Géosciences)  
un article dans Reliability Engineering & System Safety

### Post doctorants

- Ahmad M. Alam (2017-18) Exploitation of Geo-Located measurements for radio resource management in cellular network, en collaboration avec Orange
- Amar Mokrani (2012) Algorithmes pour l'interpolation de grands jeux de données spatiales, en collaboration avec le BRGM

### Stagiaires

- Jean Langanay (2017), ENSG, Impact de la variabilité spatiale sur les calculs couplés chimie transport d'exploitation d'uranium par Récupération In Situ (ISR)  
co-encadrement avec Vincent Lagneau (Mines ParisTech, centre de Géosciences) et Gwenaële Petit (Orano)
- Julien Barrau (2016), Mines ParisTech, Prédiction du revenu généré par des joueurs  
collaboration entertainment
- Antoine Thomas (2014), Mines ParisTech, Quantification de l'impact de la variabilité des simulations stochastiques sur la volumétrie  
collaboration Total

- Paul-Arthur Jeantaud (2014), Mines ParisTech, Application de techniques de data-mining pour l'amélioration de la production des réservoirs  
collaboration Total
- Alexis Bottero (2013), Master GPX IPGP, Seismic tomography under uncertainties by parallel interacting Markov chains  
en collaboration avec Alexandrine Gesret (Mines ParisTech, centre de Géosciences)  
article dans *Geophysical Journal International*
- Jed Oukmal (2013), Mines ParisTech, Simulation stochastique d'un corps géologique sous contrainte de connectivité  
collaboration Total
- Daria Khmel (2013), Master MSE Mines ParisTech, Classification of spatial data for the evaluation of mineral reserves  
collaboration Geovariances, dans le cadre du consortium G2DC
- Edia Briceno (2013), CFSG, Non Linear Geostatistics applied to a Uranium Deposit  
collaboration Areva
- Charline Pennequin (2013), INSA Rouen, Cartographie de la qualité de l'air à partir de données de mesure et de résultats de modélisation déterministe : sélection des stations de mesure et choix des paramètres de krigeage  
collaboration Ineris
- Douglas Almeida (2012), CFSG, Comparaison des méthodes d'estimation d'un gisement de phosphate dans les carbonatites  
collaboration Vale
- Gunes Ertunc (2011), CFSG, Comparison Of Three Methods For Global Reserve Estimation : Ordinary Kriging, Covariance Matching Constrained Kriging, And Conditional Simulations
- Aurélien Bechler (2011), ENSAI, Geostatistical sampling optimization of contaminated facilities  
collaboration Geovariances  
article dans *SERRA*
- Susana Zapata Pelaez (2010), CFSG, Proposition pour une estimation de la densité dans des latérites nickélifères
- Lee Huey Chyi (2009), A comparison between bootstrap filter and auxiliary particle filter for infectious disease modeling

## 6.4 Enseignement

<b>Mines ParisTech</b> , Paris et Fontainebleau, FRANCE <i>Calcul différentiel intégral et stochastique (maths 1 et 2)</i> 1ère année du cycle ingénieur civil (43h équivalent TD)	2019 - Co-responsable
<i>Calcul intégral (maths 2)</i> 1ère année du cycle ingénieur civil (16h de TD)	2013 - 2019
<i>Enseignement spécialisé de Géostatistique (semaine Athens)</i> 2e et 3e année du cycle ingénieur civil (22h équivalent TD)	2011 - Responsable depuis 2018
<i>Simulations géostatistiques (MS CFSG)</i> Mastère spécialisé (40h équivalent TD)	2010 - 2020
<i>Probabilités</i> 1ère année du cycle ingénieur civil (12h de TD)	2009 - 2012
<i>Les méthodes de la géostatistique</i> Cours d'option de 3e année du cycle ingénieur civil (26h équivalent TD) depuis 2017	2009 - Co-responsable
<b>ENSG</b> (Ecole nationale supérieure de Géologie), Nancy, FRANCE <i>Modèle booléen</i> Cours de 3e année du cycle ingénieur dans l'option géologie numérique (8h équivalent TD)	2012 -
<b>ENSAI</b> , Bruz, FRANCE <i>Introduction à la géostatistique</i> Cours de 3e année du cycle ingénieur dans l'option génie statistique (18h équivalent TD)	2006 -

## 6.5 Sélection d'articles

- **T. Romary**, L. Malherbe, C. de Fouquet, *Optimal spatial design for air quality measurement surveys*, Environmetrics 25 (1), 16-28 (2014)
- **T. Romary**, F. Ors, J. Rivoirard, J. Deraisme. *Unsupervised classification of multivariate geostatistical data : Two algorithms*, Computers and Geosciences, Elsevier, Statistical learning in geoscience modelling : Novel algorithms and challenging case studies, 85, 96-103 (2015)
- **T. Romary**, *Incomplete Cholesky decomposition for the kriging of large datasets*, Spatial Statistics, 5, pp 85-99 (2013)
- A. Bottero, A. Gesret, **T. Romary**, M. Noble, C. Maisons. *Stochastic seismic tomography by interacting Markov chains*. Geophysical Journal International, 207(1), 374–392. (2016)
- Wang, S., Flipo, N., & **Romary, T.** Oxygen data assimilation for estimating micro-organism communities' parameters in river systems. Water research, 165, 115021. (2019)

# Optimal spatial design for air quality measurement surveys

Thomas Romary<sup>a\*</sup>, Laure Malherbe<sup>b</sup> and Chantal de Fouquet<sup>a</sup>

Measurement surveys using passive diffusion tubes are regularly carried out to elaborate atmospheric concentration maps over various areas. Sampling schemes must be designed to characterize both contaminant concentrations (of benzene or nitrogen dioxide for example) and their relations to environmental variables so as to obtain pollution maps as precise as possible. Here, a spatial statistical methodology to design benzene air concentration measurement surveys on the urban scale is exposed. In a first step, an *a priori* modeling is conducted that is based on the analysis of data coming from previous campaigns on two different agglomerations. More precisely, we retain a modeling with an external drift which consists of a drift plus a spatially correlated residual. The statistical analysis performed on available data leads to choose the most relevant auxiliary variables and to determine an *a priori* variogram model for the residual. An *a priori* distribution is also defined for the variogram parameters, whose values appear to vary from a campaign to another. In a second step, we optimize the positioning of the measuring devices on a third agglomeration according to a Bayesian criterion. Practically, we aim at finding the design that minimizes the mean over the urban domain of the universal kriging variance, whose parameters are based on the *a priori* modeling while accounting for the prior distribution over the variogram parameters. Two global optimization algorithms are compared: simulated annealing and a particle filter-based algorithm. Copyright © 2014 John Wiley & Sons, Ltd.

**Keywords:** optimal design; geostatistics; external drift kriging; air quality

## 1. INTRODUCTION

Mapping air pollution as precisely as possible is a major issue for French Local Air Quality Monitoring Agencies (AASQAs) both for regulatory and information purposes and for public health concerns. Seasonal or annual average concentration maps of gaseous pollutants such as nitrogen dioxide and benzene can be obtained from passive sampling data collected at a large number of sites across the area of interest. In addition to continuous monitoring performed by automatic stations, sampling surveys are useful to evaluate the spatial distribution of atmospheric concentrations more precisely and provide information about air quality in any part of the territory, as required by the European legislation. Furthermore, spatial information provided by passive sampling surveys may be very helpful in addressing the question of the spatial representativeness of monitoring stations. This issue has been receiving increasing attention during those last years because of exposure concerns, the need to streamline the assessment means in view of cost constraints, and air quality modeling considerations. Those reflections highlight the necessity of developing efficient sampling design methodologies that are able to integrate various data about the environment characteristics and so that the resulting concentration maps fulfill precision criteria.

With its internal quantification of spatial variability through the covariance function (or variogram), kriging methodology can produce maps of optimal predictions and associated prediction error variance from incomplete and possibly noisy spatial data, see, for example, Chilès and Delfiner (2012) and de Fouquet *et al.* (2007) for an application in atmospheric sciences. Through external drift (or universal) kriging, relevant environmental covariates can be taken into account to improve the prediction accuracy.

Various methods have been proposed in the literature for the problem of collecting spatial data, see Müller (2007) or De Grujter *et al.* (2006) for a review. Specifically, several competing goals can be considered when sampling spatial data. The main interest can be in the quality of the resulting kriging map, in the sense of the prediction variance under known values of the covariance parameters, as in Van Groenigen *et al.* (1999) or Brus and Heuvelink (2007). Otherwise, we may want to estimate the covariance parameters accurately, as in Zhu and Stein (2005), or both covariance and drift parameters as proposed in Müller and Stehlík (2010). Information matrix-based criteria have therefore been proposed. Providing sampling designs that satisfy both criteria is a particularly challenging task, however. The key issue is how to incorporate the parameter uncertainty into design criteria to correctly represent the uncertainty in prediction. Several approaches have been proposed. Zhu and Stein (2006) consider to split the sample into two parts, each of them satisfying one of the criteria. Zimmerman (2006) proposes to minimize the empirical kriging variance, a criterion that endorses both aspects.

\* Correspondence to: Thomas Romary, Geosciences/Geostatistics, Mines ParisTech, France. E-mail: thomas.romary@mines-paristech.fr

<sup>a</sup> Centre de Géosciences, Equipe Géostatistique, Mines ParisTech, France

<sup>b</sup> Chronic Risks Division, INERIS, France

The purpose of this work is to design a sampling scheme for a fixed number of locations, taking into account the available information coming from previous surveys. The quality of the resulting kriging maps being our main goal, we also would like to control the accuracy of the parameter estimation as a small perturbation of a covariance parameter value may have a great impact on the prediction variance. The methodology developed relies on the definition of a quality criterion of the sampling design, based on the external drift kriging variance (see, e.g., Chilès and Delfiner (2012)). A general family of criteria is proposed. In particular, a Bayesian criterion is introduced that allows to account for an *a priori* knowledge on the uncertainty of model parameters such as covariance parameters. The *a priori* distribution is defined from an analysis of the available data and our own knowledge about the phenomenon under study. The uncertainty about the parameters is accounted for afterwards. First, a population of candidates designs is generated, then they are classified according to their performances in estimating the parameters of the covariance accurately.

Optimizing one of these criteria over a grid is a challenging combinatorial problem. Heuristic global optimization is needed, and the adaptation of the simulated annealing (SA) algorithm (Kirkpatrick *et al.* (1983)) to the problem of optimal spatial design proposed by Van Groenigen *et al.* (1999) is generally used. In this paper, we describe a proposal kernel that helps to better explore the space of possible designs and hence, to optimize the criterion faster. Besides the SA algorithm using the aforementioned kernel, we propose an interacting particles algorithm inspired by Cappé *et al.* (2004). The latter allows to generate a whole population of near optimal designs, whereas the SA proposes only one solution. This population can then be used to build an intensity map of the sampling process that can be used as a tool by the operator to effectively install the passive samplers. Indeed, at the agglomeration scale, downtown in particular, physical technical constraints may preclude from using exactly the location proposed by one optimal sampling design.

In Section 2, the modeling of the pollutant concentration is exposed. Within this setting, different criteria are exposed and discussed. Section 3 presents and discusses the two global optimization algorithms that are proposed. The methodology is then applied in Section 4 to the problem of designing a sampling scheme for a survey of the benzene concentration over the agglomeration of Bordeaux. Data from previous surveys were available to conduct the modeling step. Finally, we show how the proposed methodology can be used to dimension the network of passive samplers. This work completes and extends Romary *et al.* (2011).

## 2. MODELING AND CRITERION DEFINITION

### 2.1. Modeling, prediction, and inference

A classical spatial model is adopted for the pollutant concentration. The concentration is decomposed in a drift plus a spatially correlated residual:

$$Z(x) = \beta_0 + Y'(x)\beta + S(x) \quad (1)$$

where  $Z$  is the pollutant concentration variable,  $x \in \mathcal{X} \subset \mathbb{R}^2$  is the spatial coordinate,  $Y$  is the matrix of covariates exhaustively known on  $\mathcal{X}$ ,  $\beta$  is a vector of unknown parameters, and  $S(x)$  is a centered, spatially correlated residual with covariance  $C_\theta$ , with parameters  $\theta$ . No assumption is made over the distribution of  $S$  for the moment.

Once the covariance model has been fitted and given  $\mathbf{Z} = (Z(x_1), \dots, Z(x_n))$ ,  $\mathbf{Y} = (Y(x_1), \dots, Y(x_n))$  and the covariance matrix  $C_\theta$  of  $S$  at locations  $(x_1, \dots, x_n)$ ,  $Z(x_0)$  can be predicted by kriging with external drift (also called universal kriging):

$$\widehat{Z}(x_0) = \left( c_0 + \mathbf{Y} \left( \mathbf{Y}' C_\theta^{-1} \mathbf{Y} \right)^{-1} \left( \mathbf{Y}_0 - \mathbf{Y}' C_\theta^{-1} c_0 \right) \right)' C_\theta^{-1} \mathbf{Z} \quad (2)$$

where  $c_0 = (C_\theta(x_0, x_1), \dots, C_\theta(x_0, x_n))'$  and  $\mathbf{Y}_0 = Y(x_0)$  is the vector of covariates at  $x_0$ . The associated prediction error variance, given by

$$\mathbb{V} \left( Z(x_0) - \widehat{Z}(x_0) \right) = C(0) - c_0' C_\theta^{-1} c_0 + \left( \mathbf{Y}_0 - \mathbf{Y}' C_\theta^{-1} c_0 \right)' \left( \mathbf{Y}' C_\theta^{-1} \mathbf{Y} \right)^{-1} \left( \mathbf{Y}_0 - \mathbf{Y}' C_\theta^{-1} c_0 \right) \quad (3)$$

accounts for the prediction error variance of the residual (second term of (3)) and for the prediction error of the drift as well (third term of (3)). Additionally, the prediction error variance depends on the sampling scheme  $\eta$  through  $c_0$ ,  $C_\theta$ , and  $\mathbf{Y}$  and not on the values of  $Z$ , when  $C_\theta$  is known beforehand. It can be shown that the universal kriging predictor and the associated prediction variance correspond to their Bayesian counterpart when  $Z$  is Gaussian,  $C_\theta$  is known and a non-informative prior is used for  $\beta$ , see Handcock and Stein (1993).

Throughout the previous exposition, it has been assumed that the covariance function  $C_\theta(x, y)$  and its parameters are known. In practice, however, this is almost never the case. There exists a wide range of different techniques for estimating  $C$  from least squares-based methods to maximum likelihood under Gaussian hypothesis through Bayesian methods, see, for example, Gelfand *et al.* (2010). We first describe the classical least squares approach in the case of a stationary covariance. This approach consists of different steps. First, the regression term is estimated by ordinary least squares (OLS). Then, the empirical semivariogram of the residuals is computed by the following formula:

$$\widehat{\gamma}(h) = \frac{1}{2N(h)} \sum_{|x_i - x_j| \approx h} (\widehat{s}(x_i) - \widehat{s}(x_j))^2 \quad (4)$$

where  $N(h)$  is the number of pair of points separated approximately by the distance  $h$ , and the  $\widehat{s}(x_i)$  are the residual from the fitted provisional mean function at location  $x_i$ . A parametric model  $\gamma(h)$  is then fitted by weighted least squares to the empirical variogram. Then, the covariance matrix of the residuals is built upon the following formula:  $C(x, y) = \widehat{C}(0) - \gamma(|x - y|)$ , where  $\widehat{C}(0)$  is given by the

sill ( $\max \gamma(h)$ ) of the fitted semivariogram. The regression parameters are consequently re-estimated by generalized least squares (GLS) or EGLS (where E stands for estimated). If desired, the GLS residuals may be computed and the semivariogram re-estimated from them. This procedure is necessary because the residuals obtained from the OLS estimation do not belong to the correct subspace, as the spatial correlation is not taken into account in OLS.

There exists two main problems with OLS estimation of covariance parameters, however: the need of tuning the window size for the computation of the empirical variogram and the lack of uncertainty quantification of the estimates. The empirical variogram remains a necessary tool, however, to choose the parametric family of covariances that are to be used.

Under the Gaussian assumption and with particular assumption Mardia and Marshall (1984), the maximum likelihood approach allows to build an asymptotic estimate of the variance of the estimated covariance parameters. Indeed, under this model, the log-likelihood function of  $\mathbf{Z}$  takes the form

$$l(\beta, \theta) = -\frac{n}{2} \log(2\pi) - \frac{1}{2} \log(\det(C_\theta)) - \frac{1}{2} (\mathbf{Z} - Y'\beta)' C_\theta^{-1} (\mathbf{Z} - Y'\beta) \tag{5}$$

By maximizing (5), we get the maximum likelihood estimators  $\hat{\beta}_n$  and  $\hat{\theta}_n$ . Then, the Fisher information matrix can be derived to provide an estimate of the inverse covariance matrix of the estimated covariance parameters:

$$\mathcal{I}(\theta) = \mathbb{E}_\theta \left( \frac{\partial}{\partial \theta} l(\beta, \theta) \left( \frac{\partial}{\partial \theta} l(\beta, \theta) \right)' \right) \tag{6}$$

Moreover, the  $(j, k)$ th element of  $\mathcal{I}(\theta)$  takes the following form:

$$\mathcal{I}_{j,k}(\theta) = \frac{1}{2} \left( C_\theta^{-1} C_\theta^j C_\theta^{-1} C_\theta^k \right) \tag{7}$$

where  $C_\theta^j = \frac{\partial C_\theta}{\partial \theta^j}$ . Even if this result is theoretically correct under particular conditions and not in others, especially when the covariance parameters cannot be estimated consistently (see, e.g., Zhu and Stein (2006)), it provides a reasonable indicator of the quality of the estimation of  $\theta$ . We can see that (7) depends on the locations of the points through  $C_\theta$  and its derivatives and not on the sampled values. The problem, however, is that (7) depends on the unknown value of  $\theta$ . Therefore, it can only be used locally, say around an initial guess  $\theta_0$ , or within a Bayesian framework when some prior information over the covariance parameters is known.

The information matrix of the regression parameters can also be derived and is obtained by a formula similar to (6). It is worth noticing that it is orthogonal to (6), as the terms  $\frac{\partial l(\beta, \theta)}{\partial \theta \partial \beta}$  are null. Therefore, introducing the parameter estimation uncertainty in a design quality criterion can be carried out separately for  $\theta$  and  $\beta$ .

Finally, when  $\theta$  is estimated from the same dataset with which the prediction is performed, an additional uncertainty enters the prediction variance (3). It is generally approximated by

$$\text{tr} \left( \mathcal{I}(\theta)^{-1} \left( \frac{\partial \lambda_\theta}{\partial \theta} \right)' C_\theta \frac{\partial \lambda_\theta}{\partial \theta} \right) \tag{8}$$

where  $\lambda_\theta = C_\theta^{-1} c_0$ , that is the kriging weights, see Zimmerman and Cressie (1992). Equation (8) accounts for the mean squared error of the difference between the predictions performed with the estimated  $\theta$  and with the true value of  $\theta$ .

**2.2. Quality criteria of sampling designs**

Equation (3) leads to a straightforward quality criterion of the sampling scheme when the aim of the study is to predict the pollutant concentration as accurately as possible:

$$O(\eta) = \frac{1}{|\mathcal{X}|} \int_{\mathcal{X}} \mathbb{V}(Z(x) - \hat{Z}(x)) dx \tag{9}$$

where  $|\mathcal{X}|$  is the area of  $\mathcal{X}$ . Here, the prediction error variance is averaged over the area under study. Therefore, sampling schemes obtained by minimizing this criterion will lead to accurate mapping of the concentration, performing a balance between the quality of the drift prediction and the spatial prediction. However, as shown in Romary *et al.* (2011), it may be of interest to target preferentially some areas, for instance, to map more precisely the pollutant concentration where it is likely to be high. This weight function can be designed to obtain a more accurate mapping in some areas, for instance, as a function of the covariates, especially where we expect to encounter extreme values. Note that it will also be helpful for the subsequent variographic analysis once the survey has been performed. Indeed, it will make it possible to correct the undesirable effects of preferential sampling for the estimation of covariance parameters by assigning a weight to each datapoint.

Finally, when some parameters of the model cannot be fitted accurately or when their inference presents variations over the different available datasets, we can associate them an *a priori* distribution. Then, a Bayesian version of (9) can be considered:

$$O_{Bayes}(\eta) = \frac{1}{|\mathcal{X}|} \int_{\Theta} \int_{\mathcal{X}} \mathbb{V}(Z(x) - \hat{Z}(x)|\theta) w(x) p(\theta) dx d\theta \tag{10}$$



where  $\theta \in \Theta$  is the set of uncertain parameters, and  $p(\theta)$  is its *a priori* distribution. It is worth noticing that (3) is proportional to the point variance of  $Z$ , and hence, the optimal design (according to (9) or (10)) is independent of a multiplicative factor of the covariance function.

The estimation uncertainty of  $\theta$  is not taken into account in (9) and (10). The error term (8) can be added but will increase the computational burden. A criterion based on the full Bayesian predictive distribution over  $\mathcal{X}$  and  $(\beta, \theta)$  could also be considered but remains infeasible, as each evaluation of such a criterion would involve a long Markov chain Monte Carlo run. As an alternative, we propose to consider the following criterion:

$$O_{param}(\eta) = - \int_{\Theta} \log \det(\mathcal{I}(\theta|\eta)) p(\theta) d\theta \quad (11)$$

where  $\mathcal{I}(\theta|\eta)$  is the Fisher information matrix associated to the design  $\eta$ . It can be seen as an average of the log variance of the likelihood estimates of the covariance parameters over the prior distribution. Minimizing this quantity will lead to optimal designs for the estimation of the covariance parameters. It seems, therefore, reasonable to think that a design that shows good performances in (11) will come along with a small value for (8) over  $\mathcal{X}$ , while reducing the computational burden as (11) does not involve an integral over  $\mathcal{X}$ .

The remaining problem is how to minimize (10) and (11) simultaneously. One could think about building a compound criteria as a weighted average or to split the design into two populations and perform the optimization sequentially on each. These solutions imply however to make arbitrary choices of the weights or the proportion of the population, otherwise to test a wide range of tuning parameters values and choose the one that produce the best results, assuming that the computational burden is not too heavy. Splitting the sampling design into two with varying proportion has been proposed in Zhu and Stein (2006). In their conclusions, they highlighted the fact that the best results were obtained when the proportion of additional points (added with respect to an estimation criterion) was below 10% of the overall population. Therefore, we propose to generate a population of candidate designs with good performances with respect to (10) and then to assign them a weight given by (11). This will makes more sense after the description of the interacting particles algorithm.

### 3. OPTIMIZATION ALGORITHMS

Once the covariance model, the relevant auxiliary variables, and the prior distribution for  $\theta$  have been identified, the criterion (10) has to be optimized with respect to  $\eta$ . Its computation involves an integral over  $\mathcal{X}$  that is evaluated on a grid discretizing the domain. When considering a fine grid, it amounts to choosing the best combination of  $n$  locations among all points of the grid. It is therefore a hard combinatorial problem. Considering all possible combinations is practically out of reach. Moreover, there is no closed-form expression of the gradient of (10), and a steepest descent approach is not feasible. Therefore, we have to rely on heuristic optimization algorithms to perform the global optimization: the standard SA (Kirkpatrick *et al.* (1983)) scheme and an Interacting Particles Algorithm (IPA, inspired by Cappé *et al.* (2004)) are proposed. The variant of the SA algorithm proposed here is depicted in algorithm 1.

---

#### Algorithm 1 Simulated annealing

---

```

Generate  $\eta_0$ 
for  $i = 1 \dots N$  do
  Simulate  $\zeta$  from  $q_{\eta_{i-1}}$ 
  Simulate  $u \sim \mathcal{U}_{[0,1]}$ 
  if  $u < \alpha = \min(\exp(\frac{O(\eta_{i-1}) - O(\zeta)}{T}), 1)$  then
     $\eta_i = \zeta$ 
  else
     $\eta_i = \eta_{i-1}$ 
  end if
  Decrease  $T$ 
end for

```

---

Since the seminal work of Kirkpatrick *et al.* (1983), inspired itself by Metropolis *et al.* (1953), SA has been widely used and is well documented, see, for example, Müller *et al.* (2004) for an application to optimal design. The fundamental idea of this algorithm is that a change of scale, named temperature ( $T$  in algorithm 1), allows larger moves on the surface to optimize and therefore avoid theoretically to remain trapped in a local minimum.

Practically, SA is an iterative optimization algorithm in which a sequence of states ( $\eta_i$ ) is generated by deriving a new solution from randomly modifying the previous state by the inhomogeneous kernel  $q_{\eta_{i-1}}$ . Each time a new sampling scheme is generated, the criterion is evaluated and compared with the value of the previous state. If the criterion value has been improved by the change, the new state is systematically accepted. Otherwise, the SA algorithm may accept changes that worsen the criterion according to a given probability, given in algorithm 1. This helps to avoid remaining trapped in a local minimum.

The main shortcomings of SA are well known, particularly the need of a tedious tuning of the parameters and the uncertainty of finding the global optimum in practical applications. Indeed, theoretical results point out that in most cases, a logarithmic decrease of the temperature is necessary to ensure the asymptotic convergence, a condition that is never met in practice. One generally uses a geometric cooling

schedule with a factor between 0.99 and 0.9. Furthermore, the propositional kernel needs also to be tuned as it may vary with respect to the temperature: at high temperatures, SA authorizes the sampling scheme to vary widely while at low temperatures, only slight modifications of the current state will be accepted at a satisfactory rate.

Finally, because of its stochastic nature, SA will not produce the same results from one run to another, especially if the objective function to optimize presents several local minima. It may be the case in sampling scheme optimization where the objective functions are generally flat with lots of local minima, see, for example, Müller *et al.* (2004) and Amzal *et al.* (2006). Therefore, there should not be *one* but a *population* of optimal designs, corresponding to different minima of the criterion. Going further, this population of optimal designs can be seen as the modes of a distribution over the space of the sampling designs. Simulating from this distribution would make it possible to generate a population of optimum designs or even a density representing the probability of a location to be chosen to position a sensor. On the basis of these tools, the practitioner can then choose the design that best fits with practical constraints

This reflection leads us to propose the algorithm 2 whose goal is to simulate from a distribution of sampling designs rather than to find one optimum design while simplifying the parametrization with respect to SA. Indeed, IPA will produce a whole family of weighted solutions without using a temperature parameter. Here, instead of modifying sequentially one sampling design, a whole population of  $n$  particles ( $n$  sampling designs) is considered. At each iteration, each of the  $n$  particles is modified. Then, the  $n/2$  best particles are selected and resampled according to their weights. The weights used for the resampling are function of the objective function value and of the global distance  $\Delta(\eta^{(j)})$  between the considered particle  $\eta^{(j)}$  and all the others. This distance is introduced so as to hinder the degeneracy of the population and is defined as follows:

$$\Delta(\eta^{(j)}) = \exp\left(\frac{1}{\Omega} \sum_{i \neq j} \sum_{l=1}^n |\eta^{(j)}(l) - \eta^{(i)}(l)|\right)$$

where  $\Omega$  is a normalizing constant, and  $|\eta^{(j)}(l) - \eta^{(i)}(l)|$  is the Euclidean distance between the points  $l$  of the designs  $j$  and  $i$ . This algorithm has the advantage to produce a whole population of candidate designs without any tedious parametrization. It is also straightforwardly parallelizable.

---

**Algorithm 2** Interacting particles algorithm

---

```

Generate  $\eta_0^{(1,\dots,n)}$ 
for  $i = 1 \dots N$  do
  for  $j = 1 \dots n$  do
    Simulate  $\tilde{\eta}_i^{(j)}$  from  $q_{\eta_{i-1}^{(j)}}$ 
  end for
  Select the  $n/2$  best particles  $\tilde{\eta}_i^{(1,\dots,n/2)}$ 
  Generate  $\eta_i^{(1,\dots,n)}$  by resampling  $\tilde{\eta}_i^{(1,\dots,n/2)}$  with weights  $\exp\left(-O(\tilde{\eta}_i^{(j)})\right) \times \Delta(\tilde{\eta}_i^{(j)})$ 
end for

```

---

Once the population of candidate designs has been generated, the criterion for the estimation parameters (11) can be computed for a carefully chosen subsample. By subsampling or reweighting the design population with respect to the value of (11), we can produce the intensity map of the sampling process by a kernel density estimation.

In both algorithms, the following propositional kernel is used:

$$q = pq_1 + (1 - p)q_2, \quad p \in [0, 1] \tag{12}$$

where  $q_1$  replaces one or several uniformly selected points of the current design by one of its neighbors on the grid that is distant from less than  $h_i$ , and  $q_2$  replaces one or several uniformly selected points by one of its  $n_i$  nearest neighbors in the covariates cloud. This kernel helps to explore faster the covariates cloud and therefore accelerates the convergence, with respect to Van Groenigen *et al.* (1999) and Brus and Heuvelink (2007) where only the kernel  $q_1$  is used. In SA and IPA, the search radius  $h_i$ , the number of neighbors  $n_i$ , and the number of points that are affected decrease along the algorithm.

#### 4. APPLICATION

In this section, we apply the methodology described previously to generate an optimal sampling design for a monitoring survey of the atmospheric benzene concentration by passive diffusion tubes over the agglomeration of Bordeaux. In our application, we would like to use all the information about the spatial behavior of the air concentration of pollutants gathered in previous surveys, not necessarily on the same agglomeration the sampling design has to be generated. The estimation of  $\beta$  and  $\theta$  is performed on each available dataset. For the covariance estimation step, it can be useful to consider a single parametric family. Indeed, in the approach described in the succeeding texts, it will be easier to account for the uncertainty of a parameter value than that of a covariance model. Then, the variation range of the estimated covariance parameters is considered to build a (informative) uniform prior. To sum up, the study of the available datasets makes it possible to select the auxiliary variables and to define an *a priori* distribution on the covariance parameters that will be used in the following. In practice,

the estimation value of the regression parameters vary largely from one agglomeration to another, depending on the urban characteristics of each one. Therefore, it makes sense to consider a non-informative prior on  $\beta$ , which is taken into account by the universal kriging-based criterion. The estimation step is described briefly prior to the results. For more information on the latter step, we refer the interested reader to Romary *et al.* (2011).

#### 4.1. Model estimation

A geostatistical analysis of the data collected during previous surveys is performed in order to set up the model (covariates, covariance function and priors) to use when applying the optimization method to another agglomeration. Data from benzene concentration monitoring surveys conducted in two French cities (Lille and Reims) are used to fit the geostatistical model (1). For each survey of each agglomeration, a variable selection is performed among the translated logarithm of land cover variables, population density, and their cross products. Among a preselection obtained by LASSO due to the large number of possible predictors (see Tibshirani (1996) and Friedman *et al.* (2010) for the associated R package), most relevant variables are identified by stepwise backward elimination (see, e.g., Saporta (2006)).

For Lille surveys, a heteroscedastic variance is fitted, accounting for the increase of the benzene concentration variability as a function of the population density: the intraclass variance appeared to be two times larger for population density values higher than the median (Figure 1). Owing to this, weighted least squares were used to estimate the associated linear regression models, where the weights were chosen to be the inverse of the intraclass variance. Urban fabric proportion and population density were finally retained as prior covariates as they appeared relevant in most surveys. Then, for each survey, the variogram of the residuals has been fitted with the automatic variogram fitting function of the package Renard *et al.* (2010) for R. Note that the heteroscedasticity had been previously filtered out from the residuals by the weighted least squares estimation step. Therefore, a model consisting of an heteroscedastic nugget effect plus a linear spatial structure appeared to be the most adequate for all the surveys (Figure 2). Finally, the following covariance model has been retained:

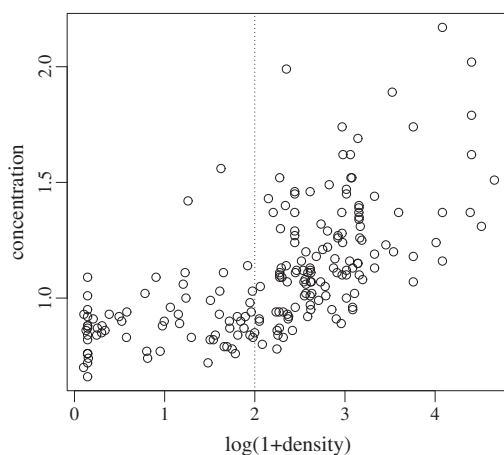


Figure 1. Mean benzene concentration as a function of the population density (Lille)

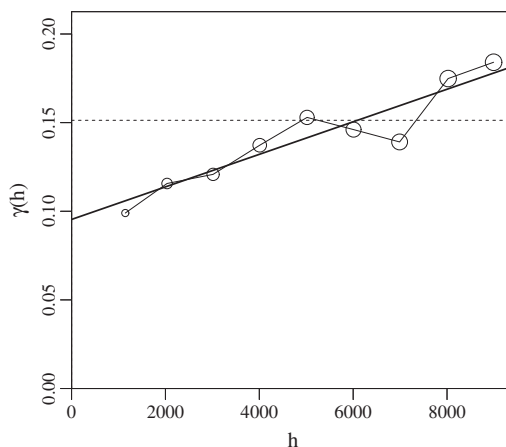


Figure 2. Experimental and fitted variogram of the residuals (Lille, average of two surveys)

$$C(x, y) = \theta_1 (1 + \mathbb{1}_{\{Y_1(x) > m_1\}}) \mathbb{1}_{\{|x-y|=0\}} + \theta_2 \left(1 - \frac{|x-y|}{R}\right), \quad x, y \in \mathcal{X} \tag{13}$$

where  $R = \max_{(x,y) \in \mathcal{X}^2} (|x-y|)$ ,  $Y_1$  stands for the population density,  $m_1$  is its median over the domain, and  $\theta_1$  and  $\theta_2$  are two varying parameters. The spatially varying nugget effect accounts for the heteroscedasticity related previously. It has been observed on Lille but not on Reims. It is retained in the prior modeling as the typology of Bordeaux is closer to that of Lille than to that of Reims, in terms of size, population density, and urban fabric proportion. The second term of (13) corresponds to a linear spatial structure and was relevant for all surveys. Hence,  $S(x)$  can be decomposed in two terms : the first one accounting for a heteroscedastic punctual variance and the second one accounting for a linear spatial structure.

Finally, note that (13) is a generalized covariance function: there is no covariance model associated with the linear variogram, as it corresponds to a non-stationary model. However, when considered on a bounded domain, a generalized covariance function can be built to simplify the computations, see Chilès and Delfiner (2012) for more details on generalized covariances.

The estimates of  $\theta_1$  and  $\theta_2$  were found to vary among the different studied datasets. Therefore, we chose to consider them as uncertain and to use the Bayesian criterion (10). As noted in Section 2.2, this criterion is proportional to a multiplicative factor of the covariance function. This allows us to set  $C(0) = 1$ , where  $Y_1(x) < m_1$ , and  $\theta_1 = 1 - \theta_2$ , so that only one parameter varies. This enables to simplify expression (13) to get

$$C(x, y) = 1 - \frac{|x-y|}{R} + \theta_1 \left( \mathbb{1}_{\{Y_1(x) > m_1\}} \mathbb{1}_{\{|x-y|=0\}} - \frac{|x-y|}{R} \right), \quad x, y \in \mathcal{X} \tag{14}$$

The *a priori* distribution of this parameter has been set to the uniform distribution over  $[0.2, 0.8]$  so as to include all situations encountered in the model estimation step.

Concerning the regression parameters, they were found to vary widely from one survey to another, even on the same agglomeration. Therefore, the choice of a non-informative prior, through the use of the universal kriging variance, is justified.

4.2. Results

The French agglomeration of Bordeaux is taken as an application case. The domain size is  $38 \times 45.5$  km; it is discretized on a  $184 \times 154$  grid with a 250 m step. The criterion (10) is optimized with the following weight function:

$$w(x) = 1 + 9 \mathbb{1}_{\{Y_1(x) > m_1 \cap \{Y_2(x) > m_2\}}(x)$$

where  $m_2$  stands for the median of the urban fabric proportion. Such a criterion assigns a relative weight of 10 to points belonging to areas where both population density and urban fabric proportion exceed the respective median observed on the domain. Furthermore, to speed up the calculation, a coarser grid, with a 500 m step, is used on areas where the weight function value is low. The justification for this is that because a low number of points will lie in those areas, it is unnecessary to offer a large choice of locations. This allows to reduce the number of grid points from 28,336 to 11,940, hence, to reduce the computational burden by a factor 2.4. The weight function is of course modified so as to maintain the desired weights.

Moreover, the evaluation of the integral with respect to  $\theta = \theta_1$  in (10) can be performed with a simple numerical integration scheme as the function that maps  $\theta_1$  to the mean universal kriging variance is linear over the interval  $[0.2, 0.8]$ , as shown in Figure 3, although it is not obvious from (14). This allows to evaluate the integral by using solely the midpoint of the interval to evaluate it. It terms to compute (10) with  $p(\theta_1) = \mathbb{1}_{\{\theta_1=0.5\}}$ .

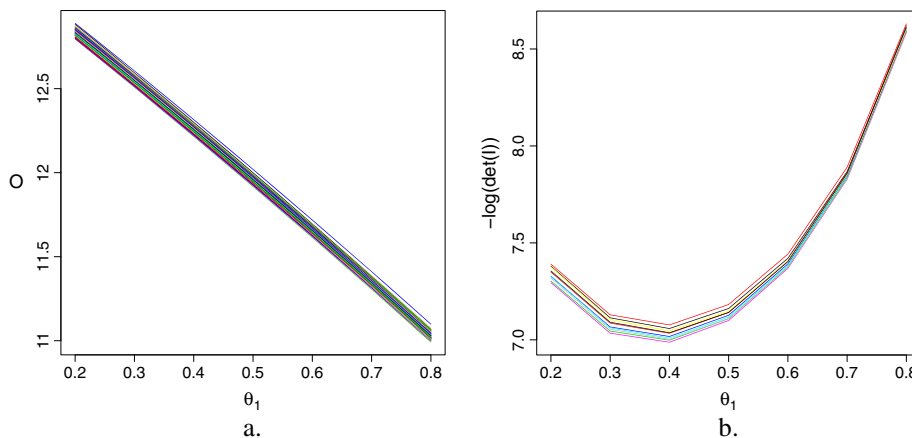


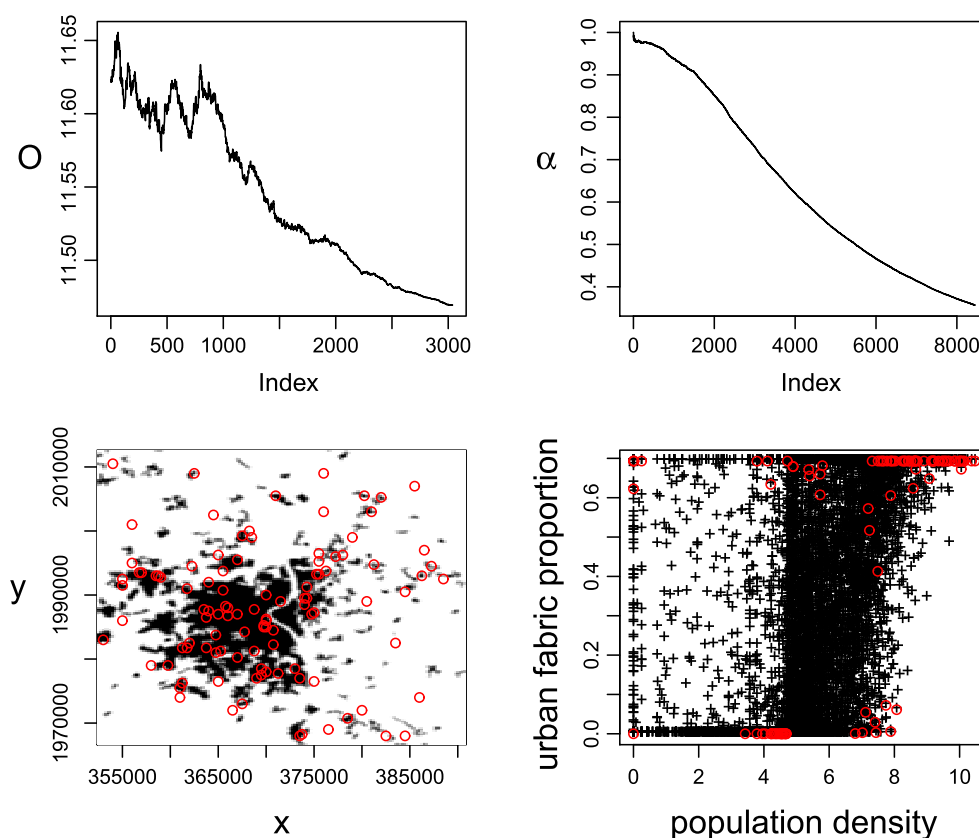
Figure 3. Mean universal kriging variance (a) and -log determinant of the Fisher information matrix as a function of  $\theta_1$  (b)

The determinant of the Fisher information matrix is not linear in  $\theta_1$ , however. This implies that the numerical integration scheme to evaluate (11) requires more than one single point. We therefore decide to compute (11) by the trapezoid formula over  $[0.2, 0.8]$  discretized with a 0.1 step.

The performances of both algorithms are compared, in terms of optimization quality and computing time on the problem of placing 105 sampling points. Finally, the criterion (11) is computed for the optimal designs generated. This number of samplers was chosen as the number of samplers used for previous surveys on the agglomeration of Bordeaux, taking into account the budget constraints.

The SA algorithm was run until no moves were accepted for 200 iterations, that is, until the acceptance rate became less than 0.005 over the last iterations. The initial temperature was chosen such that the initial acceptance rate of worsening proposition was greater than 0.9, following the rule of thumb given in Kirkpatrick (1983). The temperature was lowered exponentially with a geometrical rate of 0.95. The search neighborhood radii, in the physical space and in the covariates cloud, as well as the number of points moved, decreased along the algorithm together with the temperature. The initial design was generated randomly. The algorithm was implemented with R R Development Core Team (2010) and took about 4 h to run on a desktop computer.

Figure 4 presents the results provided by the simulated annealing (SA) algorithm. It is made of four graphs. In the upper left corner is depicted the curve of the criterion value of each accepted design along the algorithm. The profile of this curve is typical of an SA algorithm: when the temperature is high, worsening states are more likely to be accepted than when it is low. Indeed, this curve becomes more regular along the iterations. In the upper right corner is plotted the empirical acceptance rate over all iterations: it starts from values close to 1 and decreases gently toward lower values. Note that over the last 200 iterations, its value is  $1/200$ . The two lower figures represent the optimal sampling design obtained by SA. We can claim that the criterion (10) has been properly optimized. Indeed, the lower left plot, where the sampling design is plotted together with the urban fabric proportion, shows that most of the points of the resulting sampling scheme are placed in urban locations while the others cover uniformly the surroundings of the agglomeration. This indicates that the weighting has been respected while the infill effect of minimizing the average variance over the whole domain is present. The lower right picture represents the sampling design in the covariates cloud. Most of the points are located close to the bounds of each covariate and close on the left to the limit defining the non-stationary nugget effect which is the median. This behavior was expected as it is well known that the best design to estimate a linear regression by OLS is to have one point at each extremity of the regression line. As the drift is implicitly estimated by *weighted* least squares, due to the non-stationary covariance model used here, additional points near the limits defining the non stationary variance have been generated with a number of points on both sides that depends on the weighting function, see Appendix A for more details. Therefore,



**Figure 4.** Optimal sampling design for Bordeaux (SA). The upper left corner panel depicts the decrease of the criterion value along the algorithm; the upper right depicts the decrease of the empirical acceptance rate of a transition. Lower left: sampling design in the physical space and urban fabric proportion. Lower right: sampling design in the covariates cloud

the occurrence of this pattern indicates that the criterion has been properly optimized. Finally, SA has produced an optimal sampling scheme that makes the balance between a good spatial prediction over the domain and the estimation of the drift while respecting the weighting imposed for the densely populated and constructed areas.

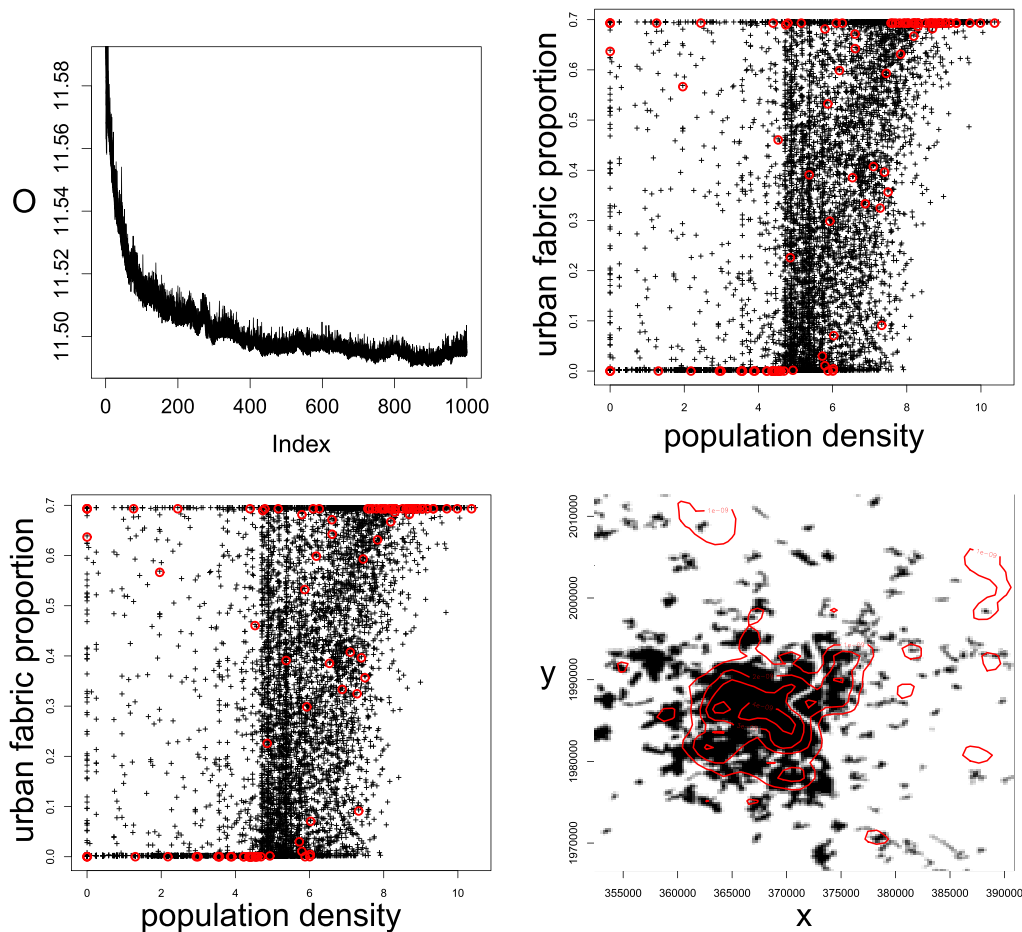
Finally, the computed value of (11) was found equal to 4.49, to be compared with the results obtained by the designs generated by IPA.

Figure 5 represents the results obtained with the IPA. It has been run for 1000 iterations with 50 particles and took about 8 h to run on a dual processor computer. No acceptance rate is plotted because the transitions are always accepted in this algorithm. The graph on the upper left shows the criterion value curves (one for each of the 50 particles) as a function of the number of iterations while that on the upper right represents the *best* (in the sense of the Fisher information matrix-based criterion (11) sampling scheme generated in the physical space, together with the urban fabric proportion over the agglomeration. The decreasing profile of the criterion curves is absolutely different from that observed with the SA (Figure 4). It is very fast during the first 100 iterations, then less fast between 100 and 400 iterations and finally, the curves seem to attain a floor beyond 400 where they exhibit a stationary behavior. Moreover, the minimum value reached for the criterion is significantly greater than the one found by SA.

The *best* design that is represented in both physical and covariates spaces has a value of (10) of about 11.5. The corresponding value of (11) is slightly lower however (4.48). The representation in the space of covariates is given in the lower left corner. It is very similar to that of Figure 4, except that the points are less concentrated near the bounds of the auxiliary variables values. If we look at the design that performs the best in terms of average kriging variance among the whole population, however, it is slightly higher than the one found by SA (4.53). This value is actually one of the worse among the generated population. Overall, the value of (11) is comprised between 4.48 and 4.54.

The graph in the lower right corner of Figure 5 depicts the kernel density estimate of the point locations computed with the *best* 1000 designs generated by the IPA. They have been subsampled from the whole population by selecting the best 1000 designs according to (11), with values between 4.48 and 4.50.

Although this density is estimated from only *near* optimal designs, it gives an insight on the likely locations of design points. Indeed, the optimal design generated by SA shares some patterns with this density: the high concentration of points in the center of the city and the almost uniform location of the points outside the urbanized areas. Some artifacts exist, however: the mode situated on the north edge of



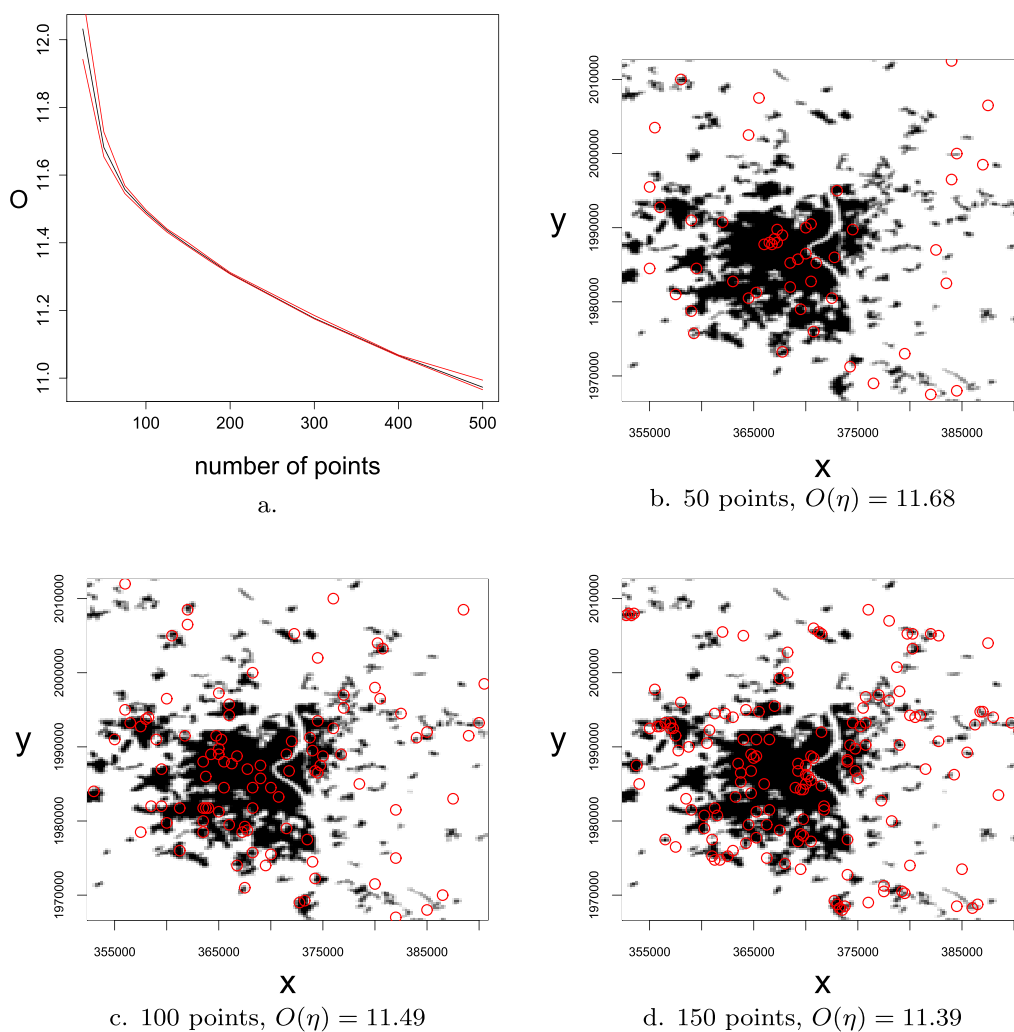
**Figure 5.** Optimal sampling design for Bordeaux (IPA). The upper left corner panel depicts the trajectories of the criterion value along the algorithm for each particle; the upper right depicts the best sampling design in the physical space. Lower left: sampling design in the covariates cloud. Lower right: Contours of the kernel density estimate of the sampling points over the subsampled population

the agglomeration seems odd. This may be due to the small size of the population used. A larger population may have given better results, as it would have performed a better exploration of the set of the designs, but the numerical burden would have been heavier. Nevertheless, this density estimate illustrates that there is not a single optimal sampling design but a whole family of near-optimal designs (as can be seen when running several times SA with different random seed), which shares the patterns of this estimated density.

The fast decrease of the objective function during the first iterations of IPA can be explained by the ability of the algorithm to explore quickly the space of possible designs and to select the best design. Its inefficiency to optimize the criterion properly, however, is certainly due to the fact that no constraint is imposed on the decrease of the criterion value, contrarily to SA. Even if this would drive away from the objective of simplifying the parametrization, the good properties of this algorithm could be exploited in a combination with SA. For instance, several near-optimal designs provided by IPA could be used as starting points for SA.

### 4.3. Dimension of the sampling design

By running several times the SA (with different seeds) with an increasing number of points, a decision support for the size of the sampling design can be determined. In Figure 6(a), the criterion (10) is plotted as a function of the number of points, in black, while the red curves are made of the maximum and the minimum of the criterion among 10 different runs. The criterion decreases quickly from 25 to 100 points then shows an almost linear decrease. Moreover, the range of the variation interval of the criterion value decreases dramatically as the number of points increases except when considering 500 points. This illustrates a good stability of the results for this implementation of the SA. The variability observed for the last case is certainly due to the incomplete convergence of the algorithm: it reached the maximum number of iterations, set to 20,000, prior to fulfill the stopping criterion. While adding points at the beginning decreases drastically the mean prediction variance, there seems to be a threshold (around 100) beyond which adding new points will only have an impact on the prediction variance at a scale inferior to the discretization step. Therefore, beyond this threshold, adding points does not affect so much the perceived prediction quality, the criterion decrease being explained by the mechanistic effect of subtracting one term to the sum that approximates (10). Finally,



**Figure 6.** Criterion as a function of the size of the network (a) and several optimal sampling designs of different sizes (b–d.) generated by simulated annealing

we can conclude from Figure 6(a) that a necessary number of sampling points is to consider, which is about 100 in our case, to get high precision maps at a given resolution. Note that adding more points is still useful if the budget allows it. In that case, it may be interesting to reduce the discretization step. Three resulting designs are depicted in Figure 6(b–d), respectively, for 50, 100, and 150 points. The evolution of the design when the number of points increases is mainly characterized by two patterns: an improvement in the spatial coverage of the low-weighted areas and an increased density of points in high-weighted areas.

Drawing such a figure can be useful for future users. The size of the sampling scheme can be decided as a function of the desired level of precision the forthcoming map should meet. Particularly, the responsible of an AASQAs can see the level of precision that he can afford and decide to amend the budget devoted to the survey accordingly, especially in order to reach the necessary number of points defined previously. The possible losses of sensors because of dysfunction or vandalism (around 10% in average) can also be taken into account *a priori*.

## 5. CONCLUSIONS

The current work exposes an optimal design methodology for the problem of generating sampling designs for measurement surveys of air quality, where the objective is to map the concentration of pollutant as precisely as possible at the considered scale. This methodology allows to generate optimal sampling schemes based on available information, including known relationships between the pollutant and environmental variables as well as uncertainty about the covariance parameters. A discussion about the different possible criteria lead to choose a pragmatic solution: first, optimize the Bayesian weighted average kriging variance over the field, then order several near optimal designs according to their performance in terms of covariance parameters estimation quality. A new proposal kernel for the global optimization algorithms and an adapted interacting particles algorithm have also been introduced. The methodology has been applied here to benzene sampling over urban areas, but it is general and can be straightforwardly extended to the study of other spatial phenomena and to other spatial scales.

We considered the problem of the existence of possible multiple optima of the objective function through the use of an interacting particle algorithm. Estimating the density of the sampling points generated gave an insight of the patterns that a good sampling design should exhibit. It may also be more useful to the end-user to have an indicative map of the locations rather than a single “optimal” design, as practical installation of passive samplers requires flexibility because of physical constraints.

In particular, this map can be obtained by subsampling the population of designs according to different criteria. Here, we chose to select the design with respect to the quality of the forthcoming covariance parameters estimation. It is worth noticing, however, that the *best* design among the population produced by the IPA, chosen according to the Fisher information matrix-based criterion, and the *best* design produced by the SA algorithm are very similar in terms of criteria values and spatial patterns. It seems that the prominent patterns that a good design (according to our criteria) should exhibit are a domain covered uniformly overall with a high concentration of points within the urban centers with some close pairs, points well distributed in the covariates cloud, so as to perform a good estimation of the regression part of the model. These results recall those already shown in Diggle and Lophaven (2006). In that paper, the authors computed a Bayesian criterion based on the predictive posterior distribution for a selection of designs that were built upon geometrical considerations. The results showed that the best performances were obtained for designs that fill the space uniformly with additional close pairs. Optimizing such a criterion remains intractable, as a Markov chain Monte Carlo run is required for each evaluation. A promising solution to speed up the calculations, however, would be to use the recently developed SPDE-INLA methodology (Lindgren *et al.* (2011) and Rue *et al.* (2009)) that decreases drastically the computational burden. The theoretical assumptions may be too restrictive for pollution, however.

The main problem with the Bayesian approach indeed is that it relies on strong assumptions about the distribution of the data that may not be fulfilled. The data from the available surveys were indeed far from being normally distributed with possible extreme values. It remains to define a more suitable probabilistic model to the pollutant concentration distribution and to adapt the criteria accordingly. In this work, we found it reasonable to consider that the larger weight given to urban areas would force the points to better cover the areas where extreme values can occur but this approach can seem insufficient.

We finally studied the effect of the dimension of the sampling design on the quality criterion value of the optimal design generated by SA. The plot of the curve, that represents the optimal value (found by SA) of the criterion as a function of the number of sampling points, leads us to define a *necessary* number of sampling points that corresponds to the number of points beyond which this curve becomes almost linear, indicating that the effect of adding new points becomes less important.

The proposed methodology has been developed with the aim of supplying scientific and technical support to the AASQAs. For the moment, it has been applied to benzene sampling over urban areas, but it can be extended to other pollutants such as NO<sub>2</sub>, to larger spatial domains like regions or even other applications like soil sampling.

## Acknowledgements

This work has been carried out for the French central laboratory for air quality surveillance. It has been supported by the French ministry responsible for the ecology and sustainable development. We also thank the associations AIRAQ, ATMO Nord-Pas-de-Calais, and ATMO Champagne-Ardenne for having supplied the data.

We are finally grateful to two anonymous reviewers who greatly helped to improve the manuscript.



## REFERENCES

- Amzal B, Bois FY, Parent E, Robert CP. 2006. Bayesian-optimal design via interacting particle systems. *Journal of the American Statistical Association* **101**(474): 773–785.
- Brus DJ, Heuvelink GBM. 2007. Optimization of sample patterns for universal kriging of environmental variables. *Geoderma* **138**: 86–95.
- Cappé O, Guillin A, Marin J-M, Robert CP. 2004. Population Monte-Carlo. *Journal of Computational and Graphical Statistics* **13**(4): 907–929.
- Chilès JP, Delfiner P. 2012. *Geostatistics, Modeling Spatial Uncertainty*, 2nd Edition. New-York.
- de Fouquet C, Gallois D, Perron G. 2007. Geostatistical characterization of the nitrogen dioxide concentration in an urban area. Part i: spatial variability and cartography of the annual concentration. *Atmospheric Environment* **41**: 6701–6714.
- De Gruijter J, Brus D, Bierkens M. 2006. *Sampling for Natural Resource Monitoring*. Springer Verlag: Berlin.
- Diggle PJ, Lophaven S. 2006. Bayesian geostatistical design. *Scandinavian Journal of Statistics* **33**(1): 53–64.
- Friedman JH, Hastie T, Tibshirani R. 2010. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software* **33**(1): 1–22. Available from: <http://www.jstatsoft.org/v33/i01>.
- Gelfand AE, Diggle PJ, Fuentes M, Guttorp P. 2010. *Handbook of Spatial Statistics*. Chapman & Hall/CRC Press: Hoboken, NJ. 622 p.
- Handcock M, Stein M. 1993. A Bayesian analysis of kriging. *Technometrics* **89**(426): 368–390.
- Kirkpatrick S. 1983. Optimization by simulated annealing: quantitative studies. *Journal of Statistical physics* **34**: 975–986.
- Kirkpatrick S, Gelatt Jr CD, Vecchi MP. 1983. Simulated annealing. *Science* **220**(4598): 671–680.
- Lindgren F, Rue H, Lindström M. 2011. An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach. *Journal of the Royal Statistical Society: Series B* **73**(4): 423–498.
- Mardia KV, Marshall RJ. 1984. Maximum likelihood estimation of models for residual covariance in spatial regression. *Biometrika* **71**(1): 135–146.
- Metropolis N, Rosenbluth A, Rosenbluth M, Teller ATM. 1953. Equations of state calculations by fast computing machines. *Journal of Chemical Physics* **21**: 1087–1091.
- Müller P, Sansó B, De Iorio M. 2004. Optimal bayesian design by inhomogeneous Markov chain simulation. *Journal of the American Statistical Association* **99**(467): 788–798.
- Müller WG. 2007. Collecting Spatial Data: Optimum Design of Experiments for Random Fields.
- Müller WG, Stehlik M. 2010. Compound optimal spatial designs. *Environmetrics* **21**: 354–364.
- R Development Core Team. 2010. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing: Vienna, Austria. ISBN 3-900051-07-0. Available from: <http://www.R-project.org>.
- Renard D, Bez N, Desassis N, Laporte F. 2010. RGeoS: The Geostatistical package. Mines ParisTech, version 9.0.0. Available from: <http://cg.ensmp.fr/rgeos/>.
- Romary T, de Fouquet C, Malherbe L. 2011. Sampling design for air quality measurement surveys: an optimization approach. *Atmospheric Environment* **45**(21): 3613–3620.
- Rue H, Martino S, Chopin N. 2009. Approximate bayesian inference for latent Gaussian models by using integrated nested laplace approximations. *Journal of the Royal Statistical Society: Series B* **71**(2): 319–392.
- Saporta G. 2006. *Probabilités, Analyses des Données et Statistiques*, 2nd Edition. Editions Technip: France.
- Tibshirani R. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B* **58**(1): 267–288.
- Van Groenigen JW, Siderius W, Stein A. 1999. Constrained optimization of soil sampling for minimisation of the kriging variance. *Geoderma* **87**: 139–259.
- Zhu Z, Stein ML. 2005. Spatial sampling design for parameter estimation of the covariance function. *Journal of Statistical Planning and Inference* **134**(2): 583–603.
- Zhu Z, Stein ML. 2006. Spatial sampling design for prediction with estimated parameters. *Journal of Agricultural Biological and Environmental Statistics* **11**: 24–44.
- Zimmerman DL. 2006. Optimal network design for spatial prediction, covariance parameter estimation, and empirical prediction. *Environmetrics* **17**(6): 635–652. Available from: <http://dx.doi.org/10.1002/env.769>.
- Zimmerman DL, Cressie N. 1992. Mean squared prediction error in the spatial linear model with estimated covariance parameters. *Annals of the Institute of Statistical Mathematics* **44**: 27–43. Available from: <http://dx.doi.org/10.1007/BF00048668>.

## APPENDIX A. OPTIMAL SAMPLING FOR THE LINEAR REGRESSION PARAMETERS ESTIMATION IN PRESENCE OF A HETEROSCEDASTIC VARIANCE

We examine here the simple case of building an optimal design for weighted linear regression with one predictor assuming independent errors (residuals), with two levels for the residual variance (below and above some threshold for the predictor).

We consider the following model:

$$Y_x = \alpha x + \varepsilon_x, \quad x \in [0, 1]$$

where  $\varepsilon_x$  are independent, centered, random variables with  $\mathbb{V}(\varepsilon_x) = 1 + \mathbb{1}_{x>a}$ ,  $0 < a < 1$ . We assume that we have  $n$  measures  $Y = (y_1, \dots, y_n)$  and we form the vector  $X = (x_1, \dots, x_n)$ . We also form the matrix  $\Sigma$  such that

$$\Sigma_{ij} = (1 + \mathbb{1}_{x_i>a}) \delta_{ij}$$

where  $\delta_{ij} = 1$  if  $i = j$ , 0 otherwise.

In that framework, it is well known that the BLUE (best linear unbiased estimator) of  $\alpha$  knowing  $X$  is given by

$$\widehat{\alpha}_X = (X' \Sigma^{-1} X)^{-1} X' Y$$

and its variance by

$$\mathbb{V}(\widehat{\alpha}_X) = (X' \Sigma^{-1} X)^{-1}$$

We can write down the prediction variance as follows:

$$\begin{aligned} \mathbb{V}(Y_x - \widehat{Y}_x) &= \mathbb{V}((\alpha - \widehat{\alpha}_X)x + \varepsilon_x) \\ &= \mathbb{V}(\widehat{\alpha}_X)x^2 + 1 + \mathbb{1}_{x>a} \end{aligned}$$

Minimizing the averaged prediction variance over  $[0, 1]$  terms to minimize:

$$\begin{aligned} O(X) &= \int_0^1 \mathbb{V}(Y_x - \widehat{Y}_x) dx \\ &= \frac{1}{3}\mathbb{V}(\widehat{\alpha}_X) + 2 - a \end{aligned}$$

Therefore, in that case, minimizing the averaged prediction variance over  $[0, 1]$  terms to minimize  $\mathbb{V}(\widehat{\alpha}_X)$ .

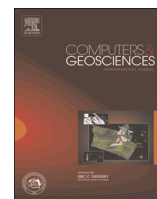
We now give a closer look to this term. We have

$$\begin{aligned} (\Sigma^{-1}X)_i &= \frac{x_i}{1 + \mathbb{1}_{x_i>a}} = x_i - \frac{x_i}{2} \mathbb{1}_{x_i>a} \\ \Rightarrow X'\Sigma^{-1}X &= \sum_{i=1}^n \left( x_i^2 - \frac{x_i^2}{2} \mathbb{1}_{x_i>a} \right) \\ \Rightarrow \mathbb{V}(\widehat{\alpha}_X) &= \left( X'\Sigma^{-1}X \right)^{-1} = \frac{1}{\sum_{i=1}^n \left( x_i^2 - \frac{x_i^2}{2} \mathbb{1}_{x_i>a} \right)} \end{aligned}$$

This quantity is not differentiable with respect to  $x_i$  because of the indicators. Using derivation in the sense of the distributions makes appear Dirac terms, and the quantity obtained is difficult to interpret. Nevertheless, it can be seen that this quantity is minimized when the denominator takes large values, that is, for values close to 1 (the maximum of  $x$ ) or values close to  $a$  (but inferior).

Contents lists available at [ScienceDirect](#)

Computers &amp; Geosciences

journal homepage: [www.elsevier.com/locate/cageo](http://www.elsevier.com/locate/cageo)

## Unsupervised classification of multivariate geostatistical data: Two algorithms

Thomas Romary<sup>a,\*</sup>, Fabien Ors<sup>a</sup>, Jacques Rivoirard<sup>a</sup>, Jacques Deraisme<sup>b</sup><sup>a</sup> MINES ParisTech, PSL Research University, Centre de Géosciences/Géostatistique, 35 rue Saint-Honoré, 77305 Fontainebleau, France<sup>b</sup> Geovariances, 49 Avenue Franklin Roosevelt, 77212 Avon, France

### ARTICLE INFO

#### Article history:

Received 1 September 2014

Received in revised form

12 January 2015

Accepted 28 May 2015

#### Keywords:

Spatial data

Clustering

Multivariate data

Geostatistics

### ABSTRACT

With the increasing development of remote sensing platforms and the evolution of sampling facilities in mining and oil industry, spatial datasets are becoming increasingly large, inform a growing number of variables and cover wider and wider areas. Therefore, it is often necessary to split the domain of study to account for radically different behaviors of the natural phenomenon over the domain and to simplify the subsequent modeling step. The definition of these areas can be seen as a problem of unsupervised classification, or clustering, where we try to divide the domain into homogeneous domains with respect to the values taken by the variables in hand. The application of classical clustering methods, designed for independent observations, does not ensure the spatial coherence of the resulting classes. Image segmentation methods, based on e.g. Markov random fields, are not adapted to irregularly sampled data. Other existing approaches, based on mixtures of Gaussian random functions estimated via the expectation-maximization algorithm, are limited to reasonable sample sizes and a small number of variables. In this work, we propose two algorithms based on adaptations of classical algorithms to multivariate geostatistical data. Both algorithms are model free and can handle large volumes of multivariate, irregularly spaced data. The first one proceeds by agglomerative hierarchical clustering. The spatial coherence is ensured by a proximity condition imposed for two clusters to merge. This proximity condition relies on a graph organizing the data in the coordinates space. The hierarchical algorithm can then be seen as a graph-partitioning algorithm. Following this interpretation, a spatial version of the spectral clustering algorithm is also proposed. The performances of both algorithms are assessed on toy examples and a mining dataset.

© 2015 Elsevier Ltd. All rights reserved.

### 1. Introduction

In mining assessment, a partitioning of the data is often to be conducted prior to evaluate the reserves. This is necessary to design the mineralization envelopes corresponding to several lithofacies where the grades of the ore to be mined may have different spatial behavior, in terms of mean, variability and spatial structure. In remote sensing of environmental variables, a similar problem may be encountered when the characteristics of the variable of interest are governed by a hidden variable, e.g. the component of a mixture model, accounting for a particular local behaviour. A typical example in soil sciences consists in the retrieval of soil classes over a region from the observation of continuous variables.

A natural solution to this problem is to cluster the data.

Clustering a dataset consists in partitioning the observations into subsets (called clusters) so that observations in the same cluster are similar in some sense. Clustering is used in many fields, including machine learning, data mining, pattern recognition, image analysis, information retrieval and bioinformatics (Hastie et al., 2009). It is an unsupervised classification problem where the goal is to determine a structure among the data, with no response variable to lead the process.

While a wide range of methods exist for independent (Hastie et al., 2009) or gridded spatial observations (in the image processing literature), not much attention has been paid to the case of irregularly spaced data. Indeed, in a geostatistical context, one expects to obtain a classification of the data that presents some spatial continuity. This is especially the case with mining data, where the geologist wishes to delineate homogeneous areas in a deposit to facilitate its evaluation and exploitation.

Clustering in a spatial framework has been mainly studied in the image analysis context where the data is organized on a grid.

\* Corresponding author.

E-mail address: [thomas.romary@mines-paristech.fr](mailto:thomas.romary@mines-paristech.fr) (T. Romary).

The model is usually a hidden Markov random field. In this model, label properties and pixel values need only to be conditioned on nearest neighbors instead of all pixels of the map, see e.g. Guyon (1995) for a review and Celeux et al. (2003) for more recent developments. In Ambroise et al. (1995), the authors proposed to use this approach directly to irregularly sampled data using a neighborhood defined by the Delaunay graph of the data. As the lengths of the edges of the graph are not accounted for in the approach, this neighborhood structure does not reflect a structure in the data, rather a structure in the sampling scheme. This disqualifies this approach especially for mining data, where the samples are located along drillholes: two neighbors on a same drillhole are distant a few centimeters while two neighbors from two different drillholes may be distant several decimeters.

Oliver and Webster (1989) were the first to propose a method for the clustering of multivariate non-lattice data. They proposed to modify the dissimilarity matrix of the data, used e.g. in a hierarchical algorithm, by multiplying it by a variogram matrix. This terms to smooth the dissimilarity matrix for close pairs of points. However, this will not enforce the connexity of the resulting clusters, it will rather blur the borders between geologically different areas, making them difficult to differentiate, as our practice showed.

In Allard and Guillot (2000), the authors proposed a clustering method based on a mixture of random functions models where an approximation of the expectation-maximization (EM, see Dempster et al., 1977) algorithm is used to estimate the parameters and the labels. It has been later extended to multivariate data in Guillot et al. (2006). However this method relies on strong assumptions that are not likely to be encountered in practice: the data are assumed to be Gaussian and data belonging to different clusters are assumed independent. Moreover, the estimation algorithm requires the computation of the maximum likelihood estimator of the random function model at each iteration of the EM algorithm, which involves the inversion of the covariance matrix and is not computationally tractable for large, multivariate datasets. Indeed, a single iteration requires several inversions of a  $(n \times p) \times (n \times p)$  matrix, where  $n$  is the number of data and  $p$  is the number of variables. Using composite likelihood techniques (Varin et al., 2011) could be useful to alleviate the computational burden but it will add a degree of approximation while still not allowing to deal with categorical data.

The approaches developed in this paper are model free and do not involve complex computations. Therefore, they are able to process large, multivariate datasets. The first one, already outlined in Romary et al. (2012), is based on an agglomerative hierarchical algorithm with complete linkage (see e.g. Hastie et al., 2009), where the connexity of the resulting clusters is enforced through the use of a graph structuring the data. It only involves the computation of distances along the edges of the graph which has a sparse structure. Its sequential nature makes it practical for reasonable volumes of data. An alternative for large datasets consists however in running first the algorithm on a subsample, then training a supervised classifier and finally applying it to the rest of the data. The second proposed algorithm provides a non-hierarchical alternative to partition the same graph. It is an adaptation of the spectral clustering algorithm (Ng et al., 2002; von Luxburg, 2007) to geostatistical data. The computations involve only sparse matrices, therefore this second algorithm is adapted to large volumes of data.

The paper is organized as follows: in Section 2, we describe both algorithms as well as a method to classify newly available data based on the results of a preceding clustering. In Section 3, we show the performance of each algorithm on a synthetic dataset as well as on a mining dataset.

## 2. Algorithms

Both algorithms proposed rely on the same basic idea. The latter consists in structuring the available data in a graph in the geographical space made of a unique connex component. This graph is then partitioned into clusters either hierarchically or directly by decomposition. The structure thus imposed ensures the spatial coherency of the resulting clusters.

We consider a sample of georeferenced data  $(x_1, \dots, x_n) \in \mathbb{R}^{n \times p}$ , where  $p$  is the number of variables, coordinates included. We also consider that the data have been standardized preliminary to the application of the clustering algorithms. It may also be useful to Gaussianize the variables, e.g. by anamorphosis (Chilès and Delfiner, 2012), for skewed data. This preliminary processing allows to make the variables comparable. We describe in this section the different ingredients required to implement both algorithms as well as their core.

### 2.1. Structuring the data

Being either regular or not, the spatial sampling of a geostatistical dataset defines a geometric set, namely a set of points in the geographical space. From this set, a neighborhood system can be built. This can be represented by an undirected graph where each vertex represents an observation and each edge shows the relation of neighborhood shared by close points (Geman and Geman, 1984). We call this graph the sampling graph. Several methods can be applied to build it such as Delaunay triangulation, Gabriel graph or a graph based on the neighborhood structure defined by moving neighborhood algorithms used in kriging, for instance based on octants (see e.g. Chilès and Delfiner, 2012). Particular shapes can also be obtained by using non-isotropic distances or coordinates transformation. The graph should be parsimonious whilst containing enough edges to support a variety of configurations for the clusters. In our experience, the Delaunay graph and a graph based on a neighborhood selection algorithm give good results.

Once the graph  $G$  has been built, two observations,  $x_i$  and  $x_j$ ,  $i \neq j$ , are said to be connected if their exists an edge in  $G$  linking  $x_i$  and  $x_j$ . This is denoted by  $x_i \leftrightarrow x_j$ .  $G$  can also be represented by its adjacency matrix with general term  $(G_{ij})_{i,j \in \{1, \dots, n\}}$ :

$$G_{ij} = \begin{cases} 1 & \text{if } x_i \leftrightarrow x_j, \\ 0 & \text{otherwise.} \end{cases}$$

Note that an individual is not considered to be connected with itself.

### 2.2. Choice of a distance

The second basic ingredient of both algorithms is a distance or metric measuring the dissimilarity between two observations. The aim of clustering algorithms is to group similar observations, hence the need to define *similar*. We define the distance  $d$  between two observations  $x_i$  and  $x_j$  by

$$d(x_i, x_j) = \sum_{k=1}^p \sum_{l=1}^p \omega_{k,l} d^{(k,l)}(x_i^{(k)}, x_j^{(l)}), \quad (1)$$

where  $(\omega_{k,l})_{(k,l) \in \{1, \dots, p\}^2}$  are the entries of a positive definite matrix  $\Omega$  and  $(d^{(k,l)})_{k=1, \dots, p}$  is a set of coupled distances, each one chosen according to the corresponding couple of variables.  $d$  is therefore a weighted sum of distances. The weights are to be chosen by the user, depending on the relative importance the variables should have and their possible correlation. As noted above, the variables have been preliminary standardized so as to avoid any scale effect

between the variables. In practice,  $\Omega$  is generally chosen to be diagonal and only individual distances are thus involved. The use of the squared Mahalanobis distance, where  $\Omega$  is the inverse of the empirical covariance matrix, could be considered so as to account for possible correlations between variables, but has not proven useful in our experiments. The individual distances are chosen according to the associated variable: if the latter is quantitative, the squared euclidean distance is advocated from its strong relation with the variogram as a measure of the local continuity; if it is a categorical variable, an ad hoc distance is used. Such a distance may take the value 0 when both observations have an equal value for this variable and 1 otherwise. Other options are also available, see e.g. Hastie et al. (2009) for a comprehensive view.

It is worth noting that the coordinates are also included in (1). Indeed, although the spatial location of the data is already accounted for by the graph structure, this allows to account for the length of the edges. By doing this, we promote short connections.

Concerning the setting up of the weights, we generally recommend to put 5–30% of the total on the coordinates and to set the other variables to 1 at a first guess, then to progressively tune the weights of the variables according to the outcome of the algorithm.

### 2.3. Geostatistical hierarchical clustering

The distance defined above is only valid between pairs of observations. Agglomerative hierarchical clustering algorithms require a linkage criterion which specifies the dissimilarity of sets as a function of the pairwise distances of observations in the sets. Lance and Williams formula (Lance and Williams, 1966) enables the use of a unique recurrence formula to update the distances when merging two clusters for a large family of criteria, including the maximum, minimum or average distance, respectively named complete, single or average linkage criteria or Ward's criterion which computes the intra-cluster variance, see e.g. Milligan (1979).

In our context, the spatial continuity needs to be taken into account during the linkage process. In the proposed algorithm, two clusters can merge if and only if they are connected in the graph structure  $G$ . When two clusters merge, the resulting cluster inherits all connections of its components. This point is the only departure from the original hierarchical clustering algorithm.

The geostatistical hierarchical clustering algorithm (GHC) is described in pseudo code in Algorithm 1 under the complete linkage criterion.

**Algorithm 1.** Geostatistical hierarchical clustering algorithm (GHC).

- 1: Compute the distance matrix  $D \in \mathbb{R}^{n \times n}$ , such that  

$$D_{ij} = d(x_i, x_j), j < i, \text{ if } i \leftrightarrow j, 0 \text{ otherwise}$$
- 2: **repeat**
- 3: Find  $k$  and  $l, k < l$ , such that  $D_{lk} = \min_{(i,j,i \leftrightarrow j)} D_{ij}$
- 4: Merge  $k$  and  $l$  in  $\{kl\}$ , and update  $D$  such that

$$D_{ki} = \max(D_{ki}, D_{li}) \text{ if } i \leftrightarrow \{kl\} \text{ and } i < k$$

$$D_{ik} = \max(D_{ik}, D_{li}) \text{ if } i \leftrightarrow \{kl\} \text{ and } k < i < l$$

$$D_{il} = \max(D_{ik}, D_{li}) \text{ if } i \leftrightarrow \{kl\} \text{ and } i > l$$

discard line and column  $l$  from  $D$

- 5: **until**  $D$  is a scalar

In Algorithm 1, the value  $D_{lk}$  can be interpreted as the inner

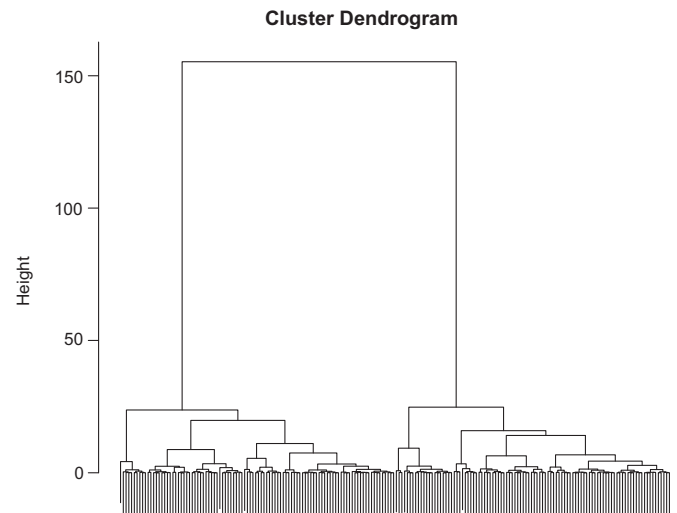


Fig. 1. Example of a dendrogram.

distance or dissimilarity of the cluster obtained when merging clusters  $k$  and  $l$ . The notation  $i \leftrightarrow \{kl\}$  means  $i$  is connected with the cluster  $\{kl\}$ , that is  $i \leftrightarrow k$  or  $i \leftrightarrow l$ .

Since two clusters are merged when they realize the minimum distance among the connected pairs of clusters, they may not realize the minimum distance in absolute, depending on the chosen linkage criterion. In particular, more dissimilar points may merge into clusters before having merged points which are actually more similar but not directly connected. That is why we advocate the use of the complete linkage criterion which is, to our knowledge, the only way to preserve the ultrametric property in our algorithm. The ultrametric property means a monotonic increase of the dissimilarity value of the clusters, see Milligan (1979) for further details. In particular, the ultrametric property allows to build a dendrogram.

The dendrogram is a very practical tool to select the final number of clusters, see an example in Fig. 1. It represents the evolution of the intra-cluster dissimilarity along the agglomeration process. A long branch means that the merge between two clusters leads to a much less homogeneous one. Therefore the tree should be pruned at the level where the branches are long. The number of pruned branches gives the number of clusters to consider, 2 in the example of Fig. 1.

The computational efficiency of this algorithm relies on the graph structure employed and especially on the number of connections. Indeed, only the distances between connected points are required at the beginning of the algorithm, which makes the matrix  $D$  sparse and allows fast computations. Then, the computation of the distances between connected points required at step 4 can be performed on the fly.

### 2.4. Geostatistical spectral clustering

The geostatistical spectral clustering (GSC) is an adaptation of the spectral clustering algorithm where the graph used is the sampling graph defined above instead of a graph based on the similarity. Contrarily to GHC, it requires a preselection of the number  $K$  of desired clusters and does not rely on an iterative procedure. This is not a major drawback however. Once computed the quantities required for a given maximum number of classes, it is straightforward to compute the outcome for a smaller number of classes. The different steps of the algorithm are described in Algorithm 2.

**Algorithm 2.** Geostatistical spectral clustering algorithm (GSC).

1: Compute the similarity or weighted adjacency matrix  $W$ :

$$W_{ij} = \begin{cases} \exp\left(-\frac{d(x_i, x_j)}{\sigma^2}\right) & \text{if } i \leftrightarrow j \\ 0 & \text{otherwise} \end{cases}$$

2: Compute the degree matrix  $D$ :

$$D_{ii} = \sum_{j=1}^n W_{ij}$$

3: Compute the graph Laplacian matrix

$$L = D^{-1/2}WD^{-1/2}$$

4: Compute the  $K$  largest eigenvalues of  $L$  and form the matrix  $V \in \mathbb{R}^{n \times K}$  whose columns are the associated  $K$  first eigenvectors of  $L$

5: Apply the  $K$ -means algorithm to the lines of  $V$

6: Assign observation  $x_i$  to the same class the line  $i$  of  $V$  has been assigned

This algorithm consists in representing the data into an infinite dimensional space (the reproducing kernel Hilbert space associated to the kernel used in (2), here the Gaussian (or radial basis function kernel) where they are easily clustered through  $K$ -means.

The parameter  $\sigma^2$  is chosen as the empirical variance of the variable, following von Luxburg (2007). Note that a local adaptive approach could be considered for the setting of  $\sigma^2$ , as proposed in Zelnik-Manor and Perona (2004). However, this refinement has not proven useful in our practice. Also, the lines of  $V$  can be optionally normalized prior to step 5, as proposed in Ng et al. (2002). The differences when using the normalization or not did not appear sensible in our experimentations.

The number of clusters to consider can be chosen by studying the eigenvalues of  $L$ . A small eigenvalue signifies that the associated eigenvector is not relevant to discriminate the data. In practice, we advocate to compute a given maximum number of eigenvalues (10–20), which corresponds to the maximum number of clusters we want, and then to plot them. A large difference between two eigenvalues means that the smaller one is not so relevant.

As the graph structure is sparse, all the computations required in Algorithm 2 can be carried out using sparse linear matrix algebra, which makes GSC computationally efficient and adapted to large multivariate datasets.

### 2.5. Classifying new data

Sometimes, the sampling of the variables of interest on a domain can be performed in several steps. For instance, new drill-holes can be added to an initial sampling campaign. In the case where a clustering has already been performed, we may want to classify the new data into the classes resulting from that previous run. An other occurrence when we want to classify data upon the results of a previous clustering is when dealing with very large datasets with the GHC. In that case, we propose to run first the algorithm on a subsample, then train a supervised classifier and finally apply the latter to the remaining data.

It is particularly difficult to incorporate new data into the clustering results with simple rules. Indeed, when new data are added, the sampling graph gets modified and the outcome of GHC and GSC may change dramatically. Therefore, the idea developed

here is to learn a classification rule based on the initial clustering results. This can be achieved for instance through support vector machines (SVM, see Hastie et al., 2009). In the case of two classes, the basic principle is to find  $f(x) = \alpha_0 + \sum_{i=1}^N \alpha_i \Phi(x, x_i)$ , where the  $(\alpha_i)_{i=0, \dots, N}$  are scalars and  $\Phi$  a given kernel function, that minimizes

$$(2) \sum_{i=1}^N (1 - y_i f(x_i))_+ + \lambda \alpha^t \Phi \alpha, \quad (3)$$

as a function of  $(\alpha_i)_{i=0, \dots, N}$  and where the underscript  $+$  means the maximum between 0 and the quantity between parenthesis, and  $\lambda$  is a penalty parameter. For multi-class classification, several options are available among which we retain the standard “one versus all” implemented in LIBSVM (Chang and Lin, 2011). The penalty parameter  $\lambda$  is set through cross-validation. Applying the rule to a new observation allows to assign it to an existing class.

## 3. Results

### 3.1. Toy dataset

Here, we describe a 2D example on which we have evaluated the performances of several methods including GHC and GSC. We consider a random function on the unit square which is made of a Gaussian random function with mean 2 and a cubic covariance with range 0.3 and sill 1 on the disk of radius 0.3 and center (0.5, 0.5) and a Gaussian random function with mean 0 and an exponential covariance with range 0.1 and sill 1 elsewhere. This model is made to mimic a mineralization area in a mining deposit, where high grades are more likely to be found within the disk. A realization is shown in Fig. 2a while Fig. 2b corresponds to the sampling performed by picking 650 points out of the 2601 points of the complete realization.

We can clearly see a smooth surface with high values in the central disk in Fig. 2a and this is the area we would like to retrieve from the 650 observations plotted in Fig. 2b. We test the performances of five different methods for this task:  $K$ -means, complete linkage hierarchical clustering (HC), Oliver and Webster's method (O&W), GHC and GSC.

For every method, the three variables are scaled such that the coordinates are given a weight of 10% in the computation of the distance. This preliminary treatment makes the different methods comparable. In HC, O&W, GHC and GSC, we use the squared euclidean distance.  $K$ -means does not need any parameterization. For O&W, several variogram models and sets of parameters have been considered, without much success. The results presented here are obtained with an exponential variogram with range 0.5. The Delaunay graph has been used for both GHC and GSC. Concerning GSC, normalizing the rows of  $V$  (see Algorithm 2) gave similar results as without normalization. Consequently, only the results without normalization are presented.

Figs. 3 and 4 show the results obtained by each five methods on the realization depicted in Fig. 2. Each subpicture represents the dataset on scatterplots with respect to the coordinates ( $x$  and  $y$ ) and the sampled value ( $Z$ ).  $K$ -means (a) identifies well the central area. The result lacks in connexity however. In particular, large values outside of the disk are classified as belonging to the disk and low values within the disk are missclassified as well. It can be seen that the method only discriminates between low and high values of  $Z$ : the limiting value between the two clusters can be read as more or less 0.5. HC (b) also discriminates between low and high value but the limiting value is higher, around 2. To sum up, those two classical methods in an independent observations context fail to produce spatially connected clusters. O&W's approach has been tested with various variograms and variogram

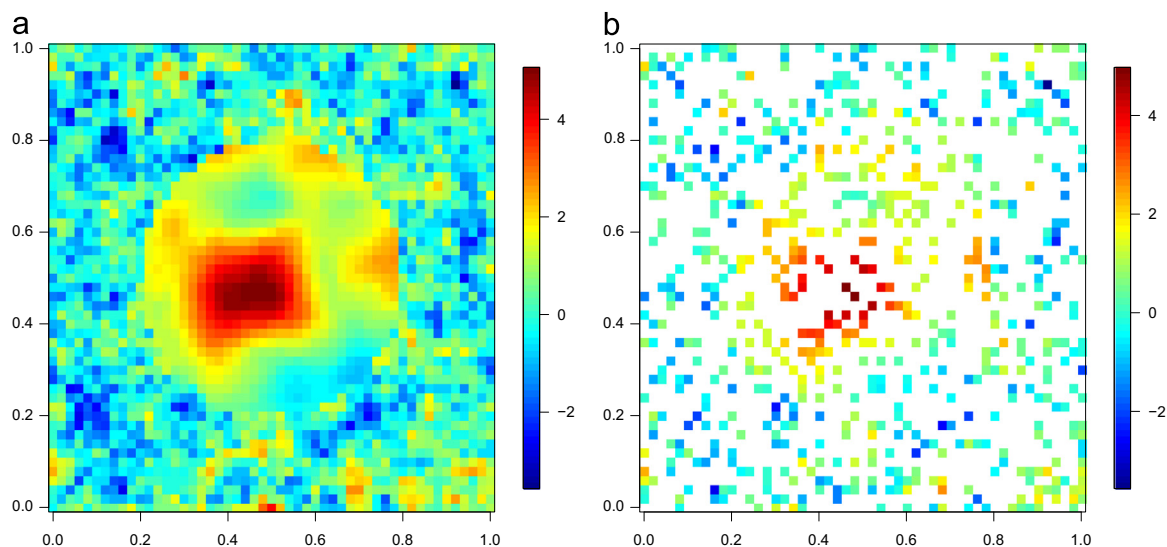


Fig. 2. One realization of the random function (a) and sampling performed (b).

parameter values but it never showed any structured result (c). Our interpretation is that multiplying the dissimilarity matrix by a variogram may erase some dissimilarities, inducing a loss in the structure of the data. The GHC algorithm succeeds in providing a clustering with spatial connexity (d) though non-perfect. A part of the area surrounding the disk is misclassified however. If we turn back to the complete realization in Fig. 2a, we can see that the misclassified area corresponds to high values of the realization around the border of the disk that are very close to the values taken inside the disk and are thus difficult to classify correctly. Finally, the GSC algorithm performed a geometrical classification by making a cut along the axis of the first coordinate, see Fig. 4. However, when looking at the result obtained when asking for five classes, it provided a class delineating the disk fairly well. It seems that this algorithm tends to generate more compact subsets of the sampling graph.

Each of the five algorithms is applied to 100 realizations of the same random function model, each with a different uniform random sampling. Then we compute the mean, median and 90% percentile of the rate of correctly classified points. Results are summarized in Table 1.

GHC exhibits the best performances overall with 85% correctly classified points in average while *K*-means providing similar results in average, GSC performing the worst with HC and O&W in between. If we look at the median however, GHC has the greatest one with a larger margin. The 90% percentile indicates that in the 10% most favorable cases, GHC misclassified only 0.02% of the points, while all the other algorithms perform worse. It can also be seen that the 90% percentile are similar for the *K*-means and the HC. This means that the HC, and GHC (its worse result in this task was a misclassification of almost 50%, seemingly due to a high sensitivity to large values), can sometimes perform really bad, whereas the *K*-means algorithm gives more stable results, being less sensitive to extreme values. Indeed, in the presence of very large or very low value, it occurs that the algorithm comes out with a class made of a single point while the other contains all the other observations. In the favorable cases however, HC algorithm works as well as the *K*-means, while GHC outperforms clearly all other algorithms. Concerning GSC, the results obtained are extremely poor but do not account for the interesting results obtained when considering more classes.

It is worth noting that the drawbacks exhibited by GHC and GSC are far from being prohibitive in practice. Indeed, when applying clustering algorithms to real data the user generally

observes the outcome for several numbers of classes. This can be performed easily with both algorithms with a negligible computational cost.

### 3.2. Mining data example

In this section, we present an application of both geostatistical clustering algorithms to an ore deposit. We describe the different steps and exhibit some results.

The first step is to select the data that will be used for the clustering. The following variables are chosen:

- coordinates,  $X$ ,  $Y$  and  $Z$ ,
- ore grades,
- a geological factor describing the basement vs. a sandy part on top of it,
- the hematization degree.

This choice is made upon an exploratory analysis of the data and discussions with geologists. Some transformations of the data are preliminary performed:

- coordinates are standardized,
- ore grades are log-transformed and standardized,
- the degree of hematization is transformed into a continuous variable, then standardized.

The next step consists of building the sampling graph connections between geographically close samples. The graph is here built from the neighbouring structure induced by the moving neighbourhood kriging algorithm of Isatis<sup>®</sup>2013 (Geostatistics, 2013). At each point, the space is split into 16 hexadecants: 8 above and 8 below the horizon. One neighbor per hexadecant is authorized at most for each point with no more than 2 from the same drillhole. The search ellipse is of infinite size so as to connect even possibly distant points. The angles of the search ellipse are chosen so that to take into account the horizontal shape of the mineralization of the deposit.

Then the dissimilarity matrix is built. All variables listed above are used. A particular distance for the geological factor is considered: it is chosen to be 1 when the samples have different factor values and 0 otherwise. This distance is scaled to maintain the coherency with the other individual distances. Weights are set step by step, as advocated in Section 2.2: we begin by giving an

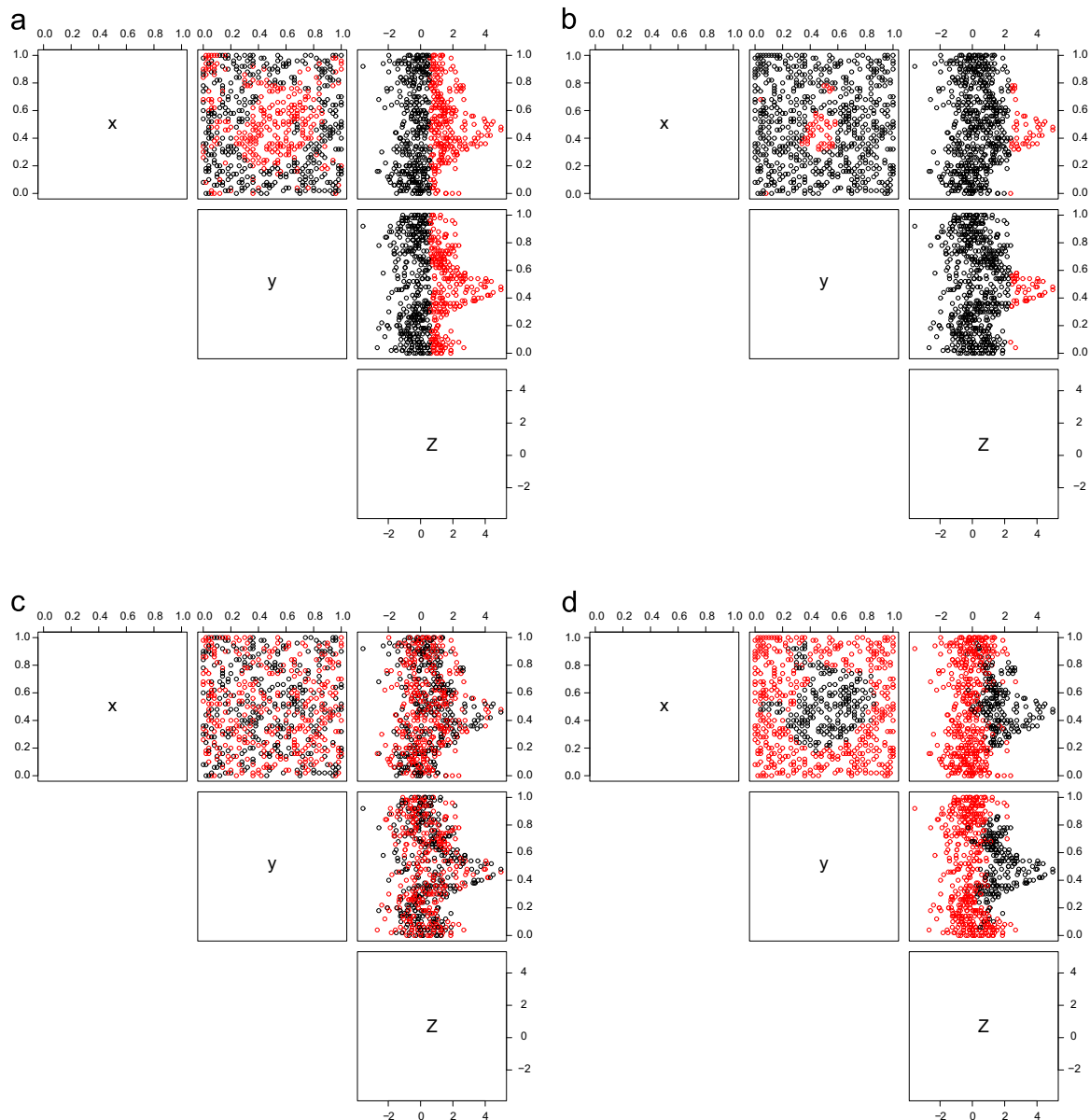


Fig. 3. Results of *K*-means (a), hierarchical clustering (b), Oliver and Webster's method (c) and geostatistical hierarchical clustering (d).

equal weight to all variables with a 30% contribution to the coordinates. Finally, the contribution of the coordinates is lowered to 10% while the other variables are assigned equal weights. The same set of weights is used for both algorithms. Practice shows indeed that setting low weights to the coordinates leads to better results, as the spatial aspect is already somehow taken into account by the sampling graph. However, the coordinates need to be included in the distance so as to account for different length of the edges in the graph. This is especially important for drillholes data where two neighbors along a drillhole are generally much closer than two neighbors belonging to two different drillholes.

Finally, we can run both GHC and GSC algorithms described in Section 2. We choose to represent 6 clusters as the intra cluster dissimilarity at that step of the GHC shows a great increase. The results are depicted in Fig. 5 for GHC and Fig. 6 for GSC.

GHC separates the basement into two classes, the black one being richer than the red one. Note that the black cluster is mainly present in the middle of the deposit. The sandy part on top of the basement is splitted into 3 separate classes plus one single observation (in cyan), see Fig. 5. The discrimination between the

three sandy classes seems to rely on geographical considerations.

As for GSC, it splits the basement into 5 classes and puts every observation on top of it into one single class. Some similarities can be observed between the clustering results obtained with the two algorithms however. In particular, both make a clear distinction between the basement and the sand on top of it, emphasizing the variable 'geology'. They also both exhibit the desired connectivity properties. Both also reveal a high grades area in the center of the deposit (the black cluster in both figures), whose retrieval was the goal of the experimentation. As already noticed in the previous paragraph, GSC tends to produce more compact clusters than GHC who can follow awkward routes along the graph.

#### 4. Conclusion

In this paper, we presented two clustering procedures adapted to irregularly sampled spatial data. Both algorithms allow to process large multivariate datasets. They rely on a partition of a graph structuring the data in the geographical space, thus ensuring the



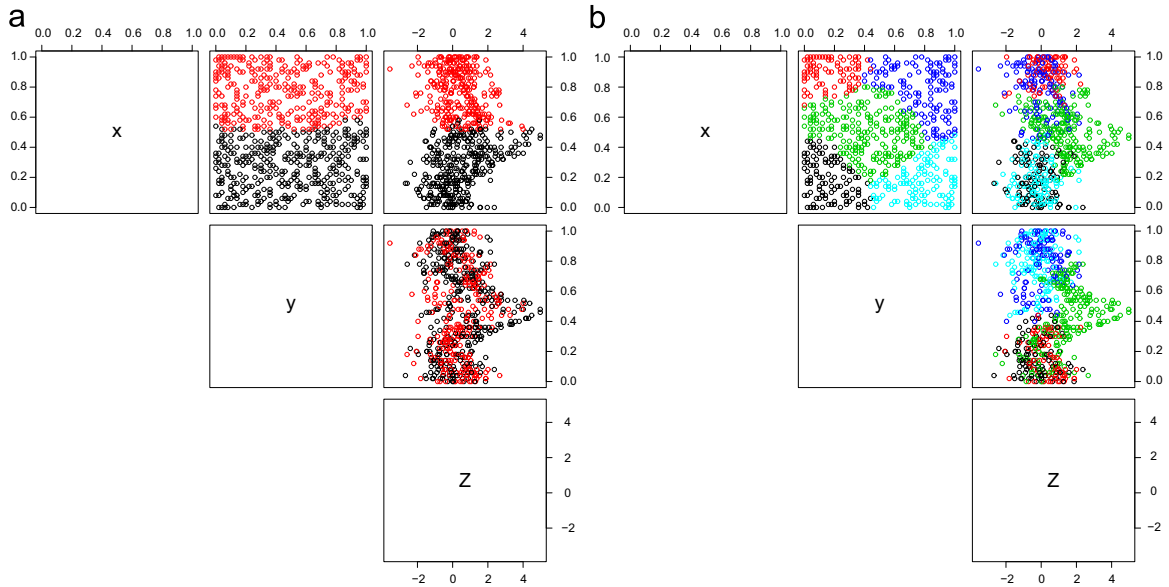


Fig. 4. Results of GSC for two (a) and five classes (b).

Table 1

Rates of correctly classified points for the 5 algorithms.

Score	K-means	HC	O&W	GHC	GSC
Mean	0.86	0.70	0.65	0.85	0.52
Median	0.86	0.64	0.67	0.90	0.52
90% percentile	0.90	0.91	0.72	0.98	0.54

spatial coherency of the resulting clusters. Two applications have been provided, the first one on a toy example and the second on mining data.

The results shown on the toy example validate both algorithms as they are able to produce compact, connected clusters. The results obtained for the mining application are also satisfactory as they highlight a homogeneous area with high grades. Thanks to

the sequential nature of GHC, it generates a whole ensemble of coherent clusterings that can be useful to the user: he can visualize the results at different hierarchical levels which helps the interpretation and the choice of the final number of clusters for the end user. Note that GSC does not enjoy this property as the results may change dramatically from one desired number of clusters to another. The main drawback of GHC is its limitation to datasets of reasonable size. It becomes slow when the number of observations goes beyond 10,000. In the case of large datasets, a two step approach based on subsampling and supervised classification is proposed.

Finally, setting the distance used to compute the graph and the weights associated to each variable allows the practitioner to get different clusterings, according to its knowledge of the geology and the variables he wants to be emphasized in the results. The

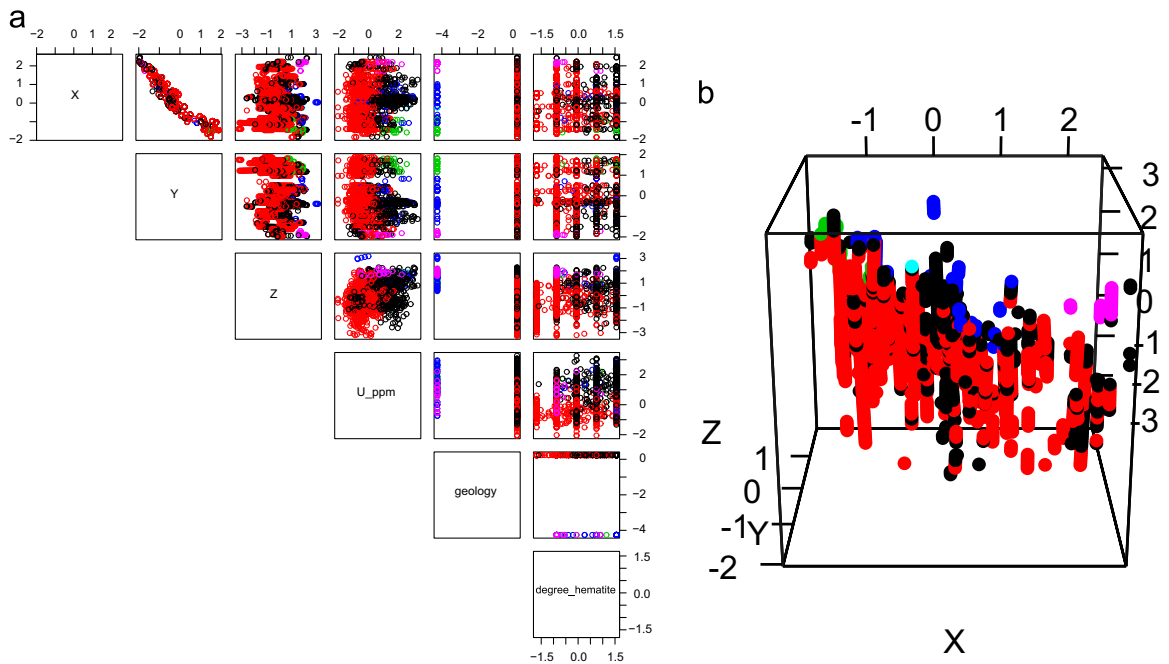


Fig. 5. Resulting clusters for the GHC algorithm from the variables point of view (a) and in 3D (b). (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this paper.)

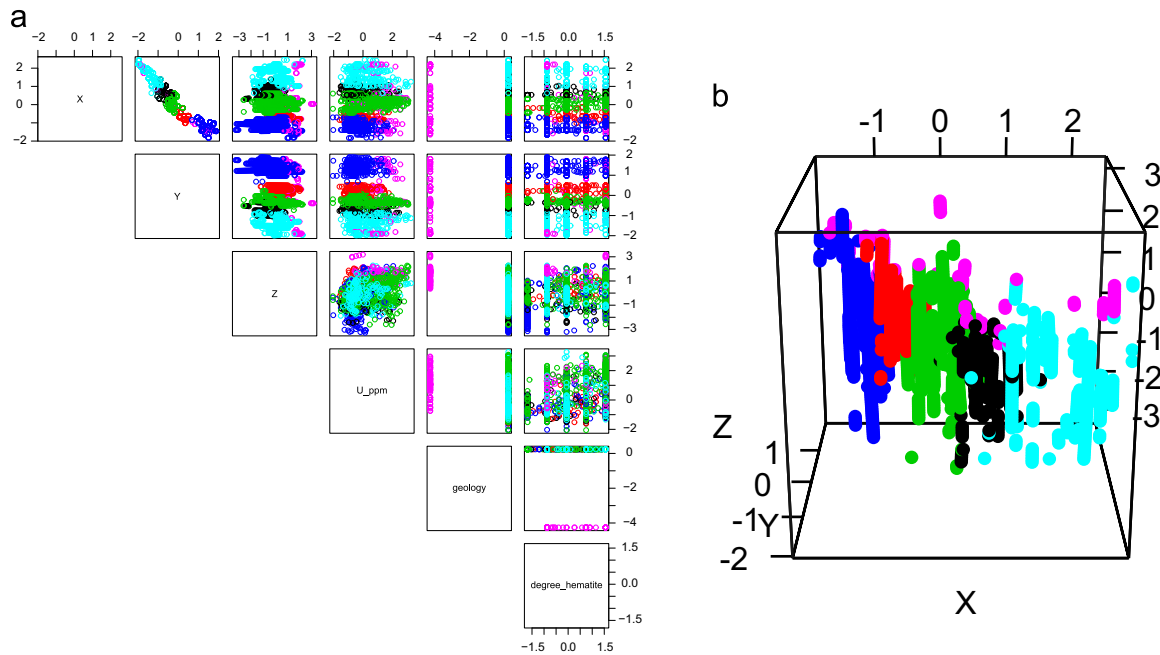


Fig. 6. Resulting clusters for the GSC algorithm from the variables point of view (a) and in 3D (b).

main difficulty in handling these algorithms is their sensitivity to the different parameters used. Moreover the results are difficult to validate except from the computation of indices of compactness of the clusters or of heterogeneity between them. They are mostly to be validated by the eye of the practitioner whose knowledge of the data should guide in the step by step parameterization of the approach.

### Acknowledgements

This work has been partly funded by the Geological and Geostatistical Domainning ( $G^2DC$ ) consortium, conducted by Geovariances. The authors are grateful to all members of the consortium for helpful discussions, namely Anglogold Ashanti, Areva, BHP Billiton, Eramet and Vale.

### References

- Allard, D., Guillot, G., 2000. Clustering geostatistical data. In: Proceedings of the Sixth Geostatistical Conference.
- Ambrose, C., Dang, M., Govaert, G., 1995. Clustering of spatial data by the EM algorithm. In: Soares, Amílcar, Pereira, Maria João, Dimitrakopoulos, Rousos (Eds.), *geoENV I – Geostatistics for Environmental Applications*. Kluwer Academic Publishers, Springer, Netherlands, pp. 493–504.
- Celeux, G., Forbes, F., Peyrard, N., 2003. Em procedures using mean field-like approximations for Markov model-based image segmentation. *Pattern Recognit.* 36 (1), 131–144.
- Chang, C.-C., Lin, C.-J., 2011. LIBSVM: a library for support vector machines. *ACM*

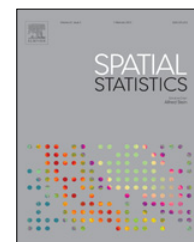
- Trans. Intell. Syst. Technol.* 2, 27:1–27:27. Available at: (<http://www.csie.ntu.edu.tw/~cjlin/libsvm>).
- Chilès, J.P., Delfiner, P., 2012. *Geostatistics, Modeling Spatial Uncertainty*, 2nd edition. John Wiley & Sons, New York.
- Dempster, A.P., Laird, N.M., Rubin, D.B., 1977. Maximum likelihood from incomplete data via EM algorithm (with discussion). *J. R. Stat. Soc., Ser. B* 39, 1–38.
- Geman, S., Geman, D., 1984. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Trans. Pattern Anal. Mach. Intell.* 6, 721–741.
- Geovariances, 2013. *Isatis Technical References*. Version 13.
- Guillot, G., Kan-King-Yu, D., Michelin, J., Huet, P., 2006. Inference of a hidden spatial tessellation from multivariate data: application to the delineation of homogeneous regions in an agricultural field. *J. R. Stat. Soc.: Ser. C (Appl. Stat.)* 55 (3), 407–430.
- Guyon, X., 1995. *Random Fields on a Network*. Springer-Verlag, New York.
- Hastie, T., Tibshirani, R., Friedman, J., 2009. *The Elements of Statistical Learning*, 2nd edition. Springer-Verlag, New York.
- Lance, G., Williams, W., 1966. A generalized sorting strategy for computer classifications. *Nature*, <http://dx.doi.org/10.1038/212218a0>.
- Milligan, G.W., 1979. Ultrametric hierarchical clustering algorithms. *Psychometrika* 44 (3), 343–346.
- Ng, A., Jordan, M., Weiss, Y., 2002. Advances in neural information processing systems. In: Dietterich, T., Becker, S., Ghahramani, Z. (Eds.), *On Spectral Clustering: Analysis and an Algorithm*, 14. MIT Press, pp. 849–856.
- Oliver, M., Webster, R., 1989. A geostatistical basis for spatial weighting in multivariate classification. *Math. Geol.* 21, 15–35. <http://dx.doi.org/10.1007/BF00897238>.
- Romary, T., Rivoirard, J., Deraisme, J., Quinones, C., Freulon, X., 2012. Domainning by clustering multivariate geostatistical data, *Geostatistics Oslo 2012*. Springer, Netherlands, pp. 455–466.
- Varin, C., Reid, N.M., Firth, D., 2011. An overview of composite likelihood methods. *Stat. Sin.* 21 (1), 5–42.
- von Luxburg, U., 2007. A tutorial on spectral clustering. *Stat. Comput.* 17(4).
- Zelnik-Manor, L., Perona, P., 2004. Self-tuning spectral clustering. *Advances in Neural Information Processing Systems*, vol. 17. MIT Press, pp. 1601–1608.



ELSEVIER

Contents lists available at ScienceDirect

# Spatial Statistics

journal homepage: [www.elsevier.com/locate/spasta](http://www.elsevier.com/locate/spasta)

## Incomplete Cholesky decomposition for the kriging of large datasets



CrossMark

Thomas Romary\*

Mines ParisTech, Equipe Géostatistique, Centre de Géosciences, 35 rue Saint Honoré, 77305 Fontainebleau, France

### ARTICLE INFO

#### Article history:

Available online 13 May 2013

#### Keywords:

Covariance function  
Spatial prediction  
Low rank approximation  
Gaussian process  
Geostatistics

### ABSTRACT

Kriging of very large spatial datasets is a challenging problem. The size  $n$  of the dataset causes problems in computing the kriging estimate: solving the kriging equations directly involves inverting an  $n \times n$  covariance matrix. This operation requires  $O(n^3)$  computations and a storage of  $O(n^2)$ . Under these circumstances, straightforward kriging of massive datasets is not possible. Several approaches have been proposed in the literature among which two main families exist: sparse approximation of the covariance function and low rank approaches. We propose here an approach that is built upon a low rank approximation of the covariance matrix obtained by incomplete Cholesky decomposition. This algorithm requires  $O(nk)$  storage and takes  $O(nk^2)$  arithmetic operations, where  $k$  is the rank of the approximation, whose accuracy is controlled by a parameter. We detail the main properties of this method and explore its links with existing methods. Its benefits are illustrated on simple examples and compared to those of existing approaches. Finally, we show that this low rank representation is also suited for inverse conditioning of Gaussian random fields.

© 2013 Elsevier B.V. All rights reserved.

### 1. Introduction

While a spatial datum was expensive to obtain in the traditional application fields of geostatistics (e.g. drilling wells for oil reserve estimation), with the development of remote sensing platforms on satellites or planes, spatial database paradigms have moved from small to massive. Therefore,

\* Tel.: +33 164694704.

E-mail addresses: [thomas.romary@mines-paristech.fr](mailto:thomas.romary@mines-paristech.fr), [thomas.romary@gmail.com](mailto:thomas.romary@gmail.com).

new methods for the geostatistical analysis of large datasets have been developed. Indeed, richer datasets allow for more complex modeling but may also prevent the straightforward use of classical techniques. The challenge of handling such datasets is to extract the maximum of information that they contain while ensuring the numerical tractability of the interpolation algorithms.

Conventional geostatistical interpolation methods such as kriging are made numerically intractable by the size of the dataset. Indeed, solving the kriging equations directly involves the inversion of an  $n \times n$  variance–covariance matrix  $C$ , which requires  $O(n^3)$  computations. Under these circumstances, straightforward kriging of massive datasets is not possible. In practice, kriging is achieved approximately by local approaches that are based on considering only a relatively small number of points that lie close to the query point. Determining the proper neighborhood size is usually solved by ad hoc methods such as selecting a fixed number of nearest neighbors or all the points lying within a fixed radius. Such fixed neighborhood sizes may not work well for all query points, depending on the local density of the point distribution. Local methods also suffer from the problem that the resulting interpolant is not continuous, even though approaches have been proposed to tackle the latter; see [Gribov and Krivoruchko \(2004\)](#) and in a more general context [Rivoirard and Romary \(2011\)](#). Furthermore, the larger the dataset, the lower the prediction variance: the most accurate prediction cannot be obtained with a subsample. Finally, if a non-stationary covariance model is used (see e.g. [Gelfand et al., 2011](#)) a small neighborhood will not be able to capture the complex correlation structure. For these different reasons, it is better to consider and solve the global system for each interpolant. Solving such large dense systems for each query point is impractical however. To tackle this problem, several approaches have been proposed in the literature for both prediction and estimation, and are described below; see also [Sun et al. \(2012\)](#) for a review.

In this paper, we consider a zero-mean spatial random field  $\{Z(x), x \in \mathcal{X} \subset \mathbb{R}^d\}$ ,  $d \in \mathbb{N}^*$ . We denote by  $C(x, y)$  the covariance of  $Z$ ,  $(x, y) \in \mathcal{X}^2$ , the covariance function. We assume throughout the paper that the parameters of  $C$  are known. The kriging problem is to predict  $Z(x_0)$ ,  $x_0 \in \mathcal{X}$ , given the observation of  $Z$  at  $n$  locations  $x_1, \dots, x_n$ . The BLUP at an unobserved location  $x_0$  is

$$Z^*(x_0) = C_0' C^{-1} Z, \quad (1)$$

where  $Z = (Z(x_1), \dots, Z(x_n))'$ ,  $C_{ij} = C(x_i, x_j)$  and  $C_{0i} = C(x_i, x_0)$ . The associated mean squared prediction error is

$$\sigma(x_0) = C(x_0, x_0) - C_0' C^{-1} C_0. \quad (2)$$

Note that  $Z^*(x_0)$  and  $\sigma(x_0)$  are the conditional expectation and variance of  $Z(x_0)$ , given the observations, under the Gaussian hypothesis.

This situation corresponds to the simple kriging with a known zero mean. Nonetheless, the different approaches described in this paper can be used with any kriging technique (ordinary, universal, etc.) by considering a blockwise matrix inversion.

The paper is organized as follows: In [Section 2](#), we review the main existing approaches for the kriging of large datasets. Then in [Section 3](#), we detail the main properties of the incomplete Cholesky decomposition when used for kriging and explore its links with existing methods. We then test the sensitivity of the different approaches to various settings on toy examples in [Section 4](#). Finally, in [Section 5](#), we discuss the use of incomplete Cholesky decomposition for dimension reduction in inverse problems as well as its link with the predictive process approach.

## 2. Existing approaches

Most existing approaches have considered approximating the covariance function  $C$  in an attempt to make the computation of  $C^{-1}$  computationally tractable. These have been classified into two main categories: sparse approximation and low rank approximation. The combination of these is also considered.

## 2.1. Sparse approximation

Introduced in [Furrer et al. \(2006\)](#), covariance tapering gives an approximation to the standard linear spatial predictor. The idea is to taper the spatial covariance function to zero beyond a certain range using a positive definite but compactly supported function. This involves deliberately introducing zeros into the matrix  $C$  to make it sparse. The kriging predictor (1) can then be computed efficiently using sparse matrix inversion algorithms.

How the zeros are introduced is crucial. In particular, the positive definiteness of any sparse modification of the covariance matrix must be maintained. Let  $C_{\text{tap}}$  be a covariance function that is identically zero beyond a particular range described by  $\theta$ . The tapered covariance function is the product of  $C_{\text{tap}}$  with  $C$ :

$$C(x, y) \approx C(x, y)C_{\text{tap}}(x, y). \quad (3)$$

The intuition behind this choice is both that the product  $CC_{\text{tap}}$  preserves *some* of the shape of  $C$  (in particular, its behavior near the origin) and that it is identically zero beyond a fixed range.  $CC_{\text{tap}}$  is a valid covariance function, as the product of two positive definite functions. Finally,  $CC_{\text{tap}}$  allows one to build a sparse covariance matrix that can be inverted using sparse matrix techniques (e.g. sparse Cholesky decomposition), to compute (1) and (2). The numerical complexity and storage of this method both scale with  $n$ .

The theoretical justification of tapering relies on a result of [Stein \(1993\)](#), for infill asymptotics for kriging with misspecified covariance. Note that this result is only valid for the Matérn ( $K$ -Bessel) covariance family. This theoretical framework also implicitly supposes a uniformly dense sampling scheme.

## 2.2. Low rank approximation

Low rank representations of random functions are well-known in spatial statistics; see e.g. [Gelfand et al. \(2011\)](#). In these representations, the covariance matrix takes the following form:

$$C(x, y) \approx \sigma^2 \delta_{xy} + P(x)' B P(y), \quad (4)$$

where  $\sigma^2$  is the variance of the nugget effect,  $P$  is a vector of  $p$  basis functions, and  $B \in \mathbb{R}^{p \times p}$  with  $p \ll n$  is a positive definite matrix. It is necessary for the covariance to contain a nugget effect, so as to make the resulting covariance matrix invertible. Otherwise, a pseudoinverse may be used. Under the form (4), the covariance matrix can be inverted using the Sherman–Morrison–Woodbury formula. The numerical complexity is here  $O(np^2)$  and the storage  $O(np)$ .  $p$  being small with respect to  $n$ , this is similar to what is observed with the tapering.

There exist a large choice of basis functions that can be chosen for  $P$ . If one wants to approximate a full rank covariance matrix, a truncation of the Karhunen–Loève basis ([Loève, 1955](#)) is optimal (in the sense of the mean squared error) for a given  $k$ . It is however necessary to extract the  $k$  largest eigenvalues and associated vector, which is prohibitive for large  $n$ . Note that this corresponds to solving the homogeneous Fredholm equation of the second kind ([Ghanem and Spanos, 1991](#)) for  $C$ , discretized on the grid made up of the locations of the data. It has been shown in [Sang and Huang \(2012\)](#) that the predictive process approach proposed by [Banerjee et al. \(2008\)](#), where the basis functions are kernel basis functions based on the actual covariance function taken at judiciously chosen knots, is equivalent to solving that same integral equation discretized at the knots. In that particular case, denoting by  $\mathbf{x}^*$  the set of knots, the low rank component in (4) takes the form  $C(x, \mathbf{x}^*)' C^{*-1} C(y, \mathbf{x}^*)$ , where  $C^* = C(\mathbf{x}^*, \mathbf{x}^*)$ .

In [Cressie and Johannesson \(2008\)](#) the authors mention a list of available basis functions: smoothing spline functions,  $W$ -wavelet functions (see [Nychka et al., 2002](#)), radial basis functions, local bi-square functions, .... In [Stein \(2008\)](#), normalized Legendre polynomials are used on the sphere.

The rationale behind the low rank approximation is the opposite of that for the sparse approximation. As the basis functions that capture most of the variance are kept, the large scale features of the random function are conserved, while the high frequency features (and hence also the small scale behavior) are discarded.

### 2.3. Combination

The two previous statements directly lead to the idea of combining the sparse and the low rank approximations. In that case, the low rank approximation captures the large scale features while the sparse approximation of the difference between  $C$  and its low rank approximation corrects for the small scale behavior. Moreover, the inversion of the matrix remains tractable, by combining the sparse matrix algorithm and the Sherman–Morrison–Woodbury formula. This is the main idea that is developed in [Sang and Huang \(2012\)](#). Their model can be written in the following way:

$$C(x, y) \approx C_l(x, y) + C_s(x, y), \quad (5)$$

where  $C_l(x, y) = C(x, \mathbf{x}^*)'C^{*-1}C(y, \mathbf{x}^*)$  for a given set of knots  $\mathbf{x}^*$  and  $C_s(x, y) = (C(x, y) - C_l(x, y))C_{\text{tap}}(x, y)$ . These two terms stand respectively for the large and small scale features of  $C$ .

The main problem with all the aforementioned approximations is that their accuracy is not quantified. In the case of the sparse approximation, the choices are guided by the sparsity of the resulting covariance matrix. This may modify dramatically the initial covariance function. In the case of the low rank approximation, the number of components that are kept is also governed by numerical considerations. Moreover, using (4) instead of the full rank covariance matrix will systematically underestimate the variance of the random function at any location. Finally, while the combination approach seems appealing, in bridging between the two other methods, its benefits need to be ascertained practically.

### 3. Incomplete Cholesky decomposition for kriging

The idea behind the use of the incomplete Cholesky decomposition is that instead of trying to approximate the covariance *function*, we could approximate the covariance *matrix* or solve the kriging *system* approximately. The approach by incomplete Cholesky decomposition is based on a low rank approximation of the covariance matrix. A covariance matrix  $C$  can always be factored as  $GG'$ , where  $G$  is an  $n \times n$  matrix, as it is positive semidefinite. This factorization can be found via Cholesky decomposition; see e.g. [Golub and Van Loan \(1996\)](#). Our goal, however, is to find a matrix  $\tilde{G}$  of size  $n \times p$ , with  $p \ll n$ , such that the difference  $C - \tilde{G}\tilde{G}'$  has norm less than a given value  $\eta$ . This can be achieved by incomplete Cholesky decomposition. This method has mainly been applied in statistical learning; see e.g. [Fine and Scheinberg \(2002\)](#) or [Bach and Jordan \(2005\)](#). Incomplete Cholesky decomposition differs from standard Cholesky decomposition in that all pivots that are below a certain threshold are simply skipped. If  $p$  is the number of non-skipped pivots, then we obtain a lower triangular matrix  $G$  with only  $p$  non-zero columns. Symmetric permutations of rows and columns are necessary during the factorization if we require the rank to be as small as possible ([Golub and Van Loan, 1996](#)). In that case, the stopping criterion involves the sum of remaining pivots.

#### 3.1. The algorithm

The algorithm involves picking one column of  $C$  at a time, and choosing the column to be added by greedily maximizing a lower bound on the reduction in the error of the approximation. It stops when the desired precision is reached.

After  $l$  steps, we have an approximation of the form  $\tilde{C}_l = \tilde{G}_l\tilde{G}_l'$ , where  $\tilde{G}_l$  is  $n \times l$ . The trace norm of the approximation error is

$$\begin{aligned} \|C - \tilde{G}_l\tilde{G}_l'\|_1 &= \text{tr}(C - \tilde{G}_l\tilde{G}_l') \\ &= \text{tr}(C) - \sum_{i=1}^l \|\tilde{G}[, i]\|_2^2. \end{aligned}$$

The gain between the iterations  $l - 1$  and  $l$  is then equal to  $\|\tilde{G}[, l]\|_2^2$ , that is the squared euclidean norm of the vector formed by the column  $l$  of  $G$ . Then the ranking of the  $n - l$  vectors that might be added in the following step should be done by computing  $\|\tilde{G}[, j]\|_2^2$  for all remaining columns  $j$ . As

this operation would involve all the terms of  $C - \tilde{G}_l \tilde{G}'_l$ , we rely instead on the following coarse lower bound:

$$\|G[, l]\|_2^2 > G[l, l]^2,$$

that is the sum of the squares of the elements of the column  $l$  is greater than the square of the diagonal element of that same column. Hence the choice of the column to add is made by comparing the diagonal elements of the remainder matrix. Each of these elements requires  $O(l)$  operations to compute. Moreover, the update of  $\tilde{G}_l$  has a cost of  $O(ln)$ , so the overall complexity is  $O(p^2n)$ .

The incomplete Cholesky method has many attractive features. Not only is its time complexity  $O(p^2n)$ , but also the only elements of  $C$  that are needed in the memory are the diagonal elements. Most of the other elements are never used and those that are needed can be computed on demand. The storage requirement is thus  $O(pn)$ . Also, the number  $p$  can be chosen online such that the approximation is as tight as desired.

### 3.2. Properties

From now on,  $G$  denotes the  $n \times p$  matrix resulting from the decomposition. The main property of the incomplete Cholesky decomposition is, by construction,

$$\|C - GG'\|_1 \leq \eta, \tag{6}$$

where  $\eta$  is set by the user. It can be useful to rescale the parameter in the following way:

$$\|C - GG'\|_1 \leq \eta^* \|C\|_1. \tag{7}$$

Then  $\eta^*$  can be interpreted as the proportion of variance that is discarded. In other words, the proportion  $1 - \eta^*$  of the variance is preserved in the approximation. Finally, it is worth noticing that  $\|C\|_1 = n$  when  $C$  is a correlation function.

To perform kriging, we have to invert  $C$ . Without the nugget effect,  $GG'$  is not invertible, but the following pseudoinverse can be used:

$$C^{-1} \approx G(G'G)^{-1}(G'G)^{-1}G'. \tag{8}$$

In the presence of a nugget effect,  $C$  is approximated by  $\sigma^2 I_n + GG'$ , where  $I_n$  is the  $n \times n$  identity matrix. A simple version of the Sherman–Morrison–Woodbury formula can be used to approximate its inverse:

$$(\sigma^2 I_n + GG')^{-1} = \frac{1}{\sigma^2} I_n - \frac{1}{\sigma^4} G \left( I_p + \frac{1}{\sigma^2} G'G \right)^{-1} G'. \tag{9}$$

The covariance function can contain a compactly supported component with a short range. In order to preserve the small scale behavior, the following approximation can be used:  $C \approx A + GG'$ , where  $A$  is sparse. The same formula as above can be used to invert the matrix:

$$(A + GG')^{-1} = A^{-1} - A^{-1}G(I_p + G'A^{-1}G)^{-1}G'A^{-1}. \tag{10}$$

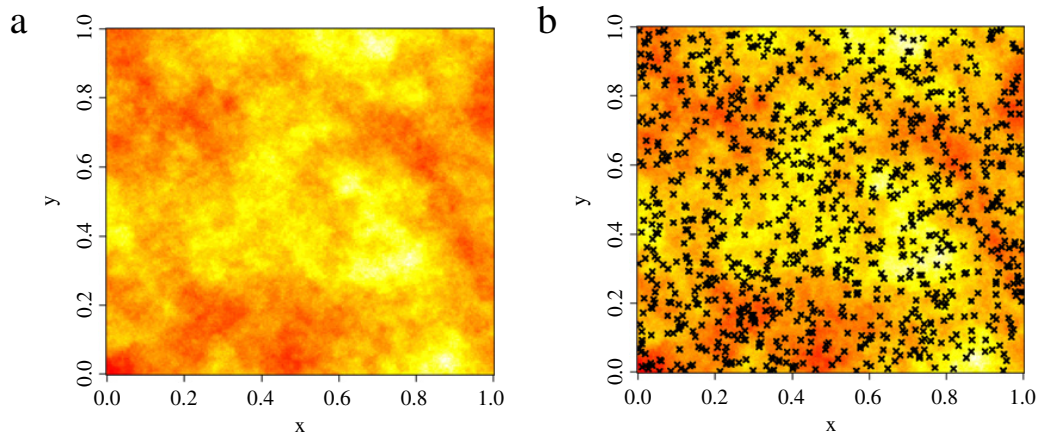
The formulas (8)–(10) are used instead of  $C^{-1}$  to compute the kriging estimate (1) and the associated prediction variance (2). Each inversion formula involves the inversion of the positive semidefinite  $G'G$ , that takes  $O(p^3)$  operations through standard Cholesky decomposition.

Finally, note that the inversion formulas used in (9) and (10) are the ones used for the low rank approximation and the combination respectively.

## 4. An application

In this section, we provide a comparison of numerical results obtained with the four following methods: standard kriging (no approximation), covariance tapering, the incomplete Cholesky method and a combination.

More precisely, we consider 1000 sample points from the simulation of a zero-mean Gaussian random function on  $[0, 1]^2$ , with several covariance models. For each method, we use the following



**Fig. 1.** Reference simulation *a* and initial sampling design *b*.

parameterization:

- Covariance tapering: an adapted taper in the sense of Furrer et al. (2006) is used, with a range  $\eta$ ,
- the incomplete Cholesky method: the precision used is  $1 - \eta$ , in terms of the proportion of variance (7). The pseudoinverse is used for the inversion (formula (8)),
- A combination (the incomplete Cholesky method + tapering): the error between the incomplete Cholesky decomposition, with precision  $1 - \eta$ , and the initial covariance is tapered by a taper with a range equal to  $\eta \times$  the range of the initial covariance.

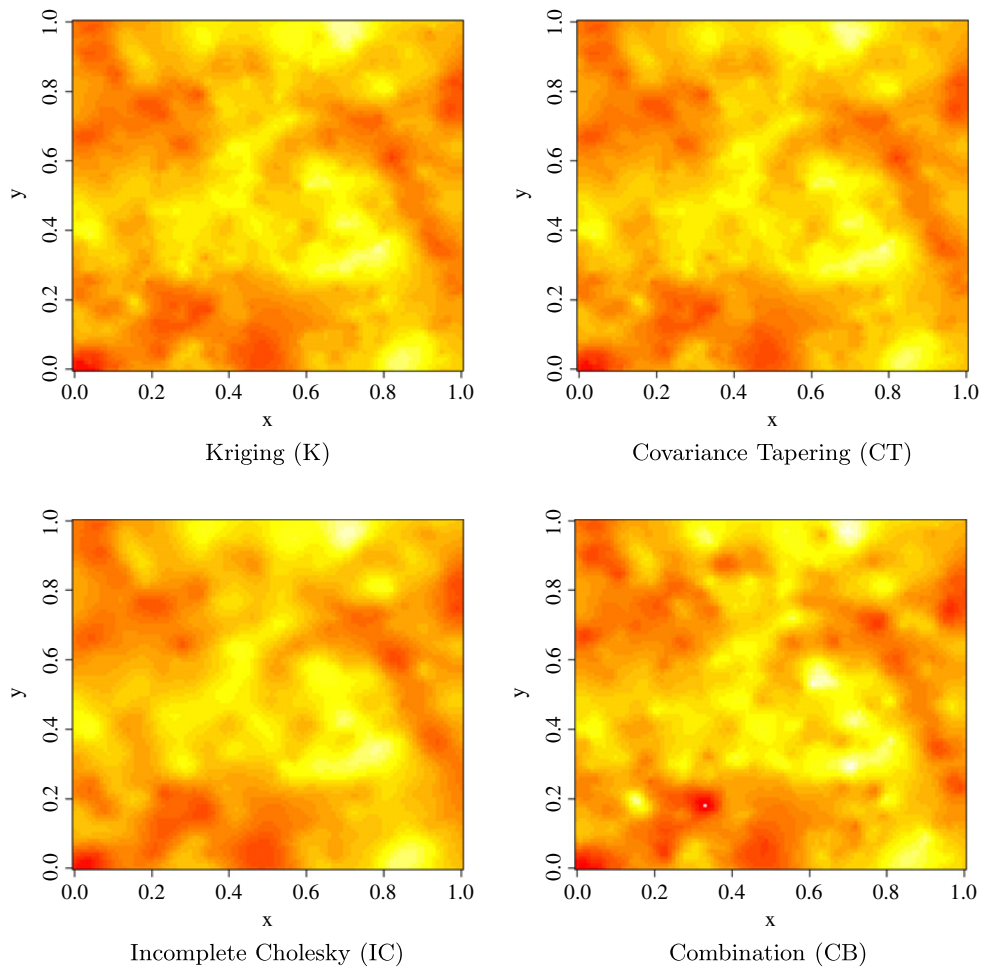
Two particular points are studied here: the sensitivity of the method to the sampling scheme and the sensitivity to the covariance model and the variations of its parameters.

#### 4.1. Sensitivity to the sampling scheme

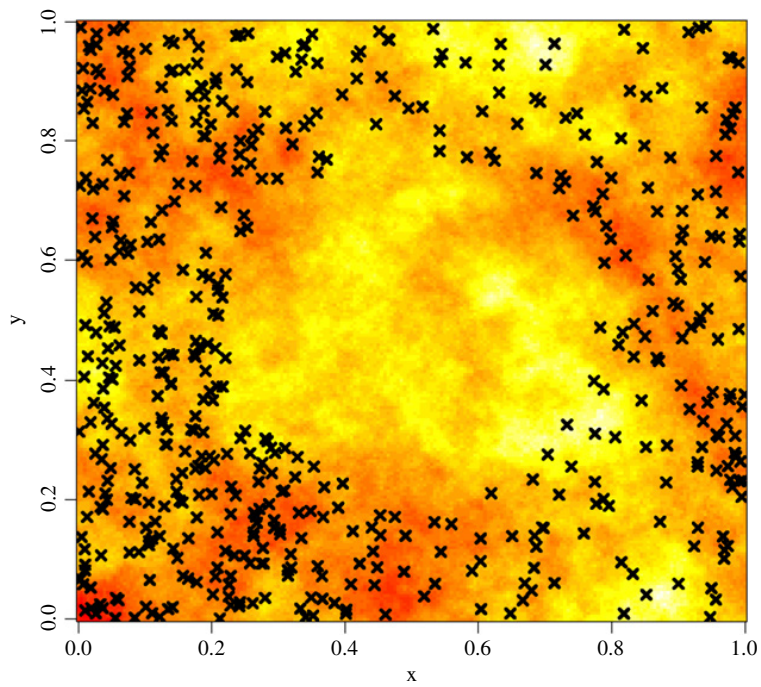
In this section, a spherical model with range 0.5 is considered. Fig. 1a represents a simulation of a zero-mean Gaussian random function with such a covariance model. Fig. 1b depicts the initial random uniform sampling design, that is made up of 1000 points.

All the kriging maps depicted in Figs. 2 and 4 were obtained with  $\eta = 0.1$ . For that value and the first sampling scheme, the sparse approximation generates a covariance matrix with 5.4% non-zero components, 138 columns in  $G$  for the incomplete Cholesky decomposition and an additional 0.88% non-zero components in the sparse matrix for the combination. Kriging is then performed with each of the four methods. The resulting maps are plotted in Fig. 2. The reference kriging map, obtained with standard kriging, is in the upper left corner. It can be seen that the covariance tapering reproduces the same patterns as the reference: it is very difficult to distinguish between the top two figures. The incomplete Cholesky method generates a smoother map, while reproducing the main patterns of the reference map. For the combination however, although it reproduces quite well the regularity of the map, it seems to be affected by artifacts: high values are too high, low values are too low. The empirical MSPE as a function of  $\eta$  computed from 100 similar simulations is plotted in Fig. 5a. The situation of Fig. 2 corresponds to the point with abscissa 0.1. The density of the matrix involved in the covariance tapering is 5.4%, 138 out of 750 components are used in the incomplete Cholesky method and an additional 0.88% density matrix is used in the combination. It can be seen that the MSPE obtained by each of the three methods are similar and close to the reference one. Next, we consider a different sampling design on the same realization. It is depicted in Fig. 3. It is made from the initial one, where the points within a disk of radius 0.28 around (0.5, 0.5) have been removed. Indeed, when studying natural phenomena, it is almost never possible to get a uniform sampling; holes may appear due to physical constraints like clouds for satellite data. Again, the kriging maps depicted below were obtained with  $\eta = 0.1$ . With the second sampling scheme, the sparse approximation gave a covariance matrix with 3.7% non-zero components, 121 columns in  $G$  for the incomplete Cholesky decomposition and an additional 1.1% non-zero components in the sparse matrix for the combination.

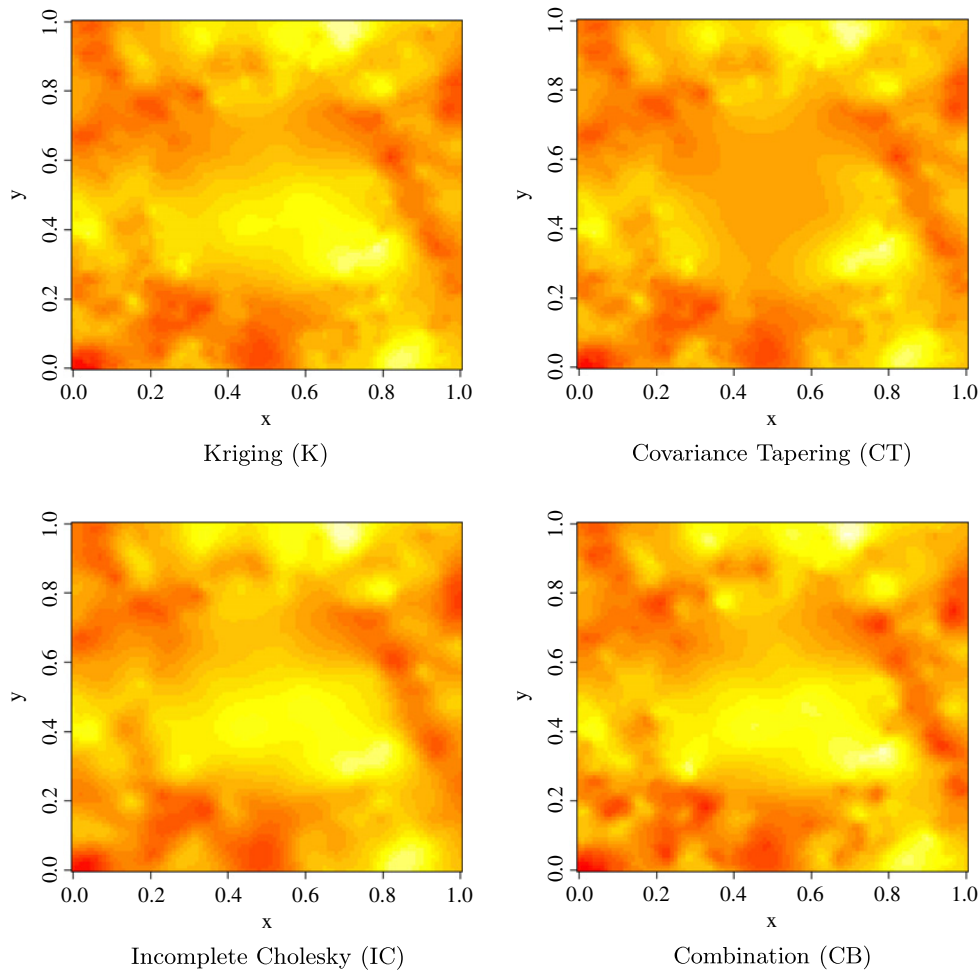




**Fig. 2.** Kriging results for each of the four methods for the first sampling scheme.



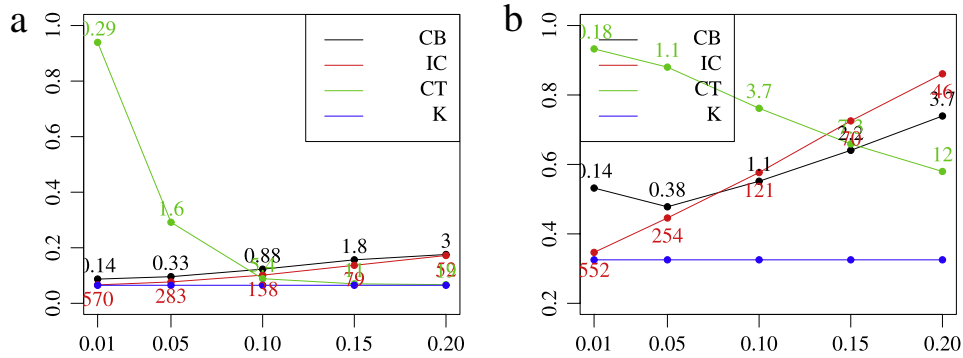
**Fig. 3.** Second sampling design.



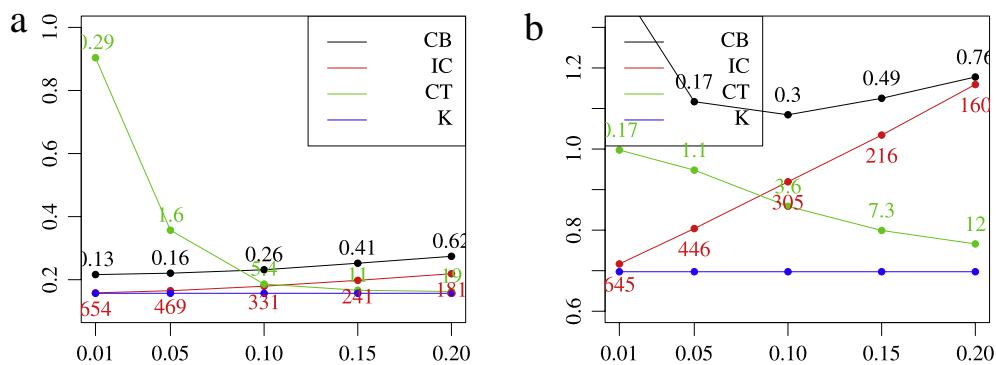
**Fig. 4.** Kriging results for each of the four methods for the second sampling scheme.

Kriging is performed by each of the four methods again, with this new sampling scheme, and the resulting maps are depicted in Fig. 4. The situation is now completely different: the tapering is unable to reproduce the large scale features of the kriging map in the area without sampling points, while the incomplete Cholesky method does reproduce these features and the combination does as well. Outside the *void* area, the maps produced are similar to the previous ones. The empirical MSPE plotted in Fig. 5b accounts for this situation: the difference between the value obtained by the tapering and the reference is almost double that obtained with the incomplete Cholesky method or the combination (0.44 versus 0.24). The previous figures showed the kriging maps obtained for  $\eta = 0.1$  and for only one realization. To offer a more thorough analysis, we now make  $\eta$  vary among (0.01, 0.05, 0.1, 0.15, 0.2) and for each value, we compute 100 realizations of the random function. For each simulation two samplings are realized: a random uniform sampling scheme of 1000 points and the associated *hole* sampling. 750 points are used for kriging while the remaining 250 are used to estimate the mean squared prediction error (MSPE) in the first case. In the second case, the MSPE is computed from the points within the hole. In both cases, the MSPE is averaged over the 100 realizations.

Fig. 5 shows the averaged MSPE over 100 simulations for each of the three algorithms when  $\eta$  varies. The numbers associated with each dot depend on the method used and vary as a function of  $\eta$ . For covariance tapering (CT), the values in green correspond to the density of the matrix, that is the percentage of non-zero values. For the incomplete Cholesky method (IC), in red, the values represent the number of components retained in the decomposition. For the combination, the number of components of the low rank term is the same as in IC, while the values in black are the densities of the sparse term. Finally, the blue horizontal line corresponds to the MSPE obtained with standard kriging.



**Fig. 5.** MSPE as a function of  $\eta$  for each of the four methods (plus the density of the sparse matrix for CB and CT, and the number of components for IC): uniform sampling *a*; hole sampling *b*.



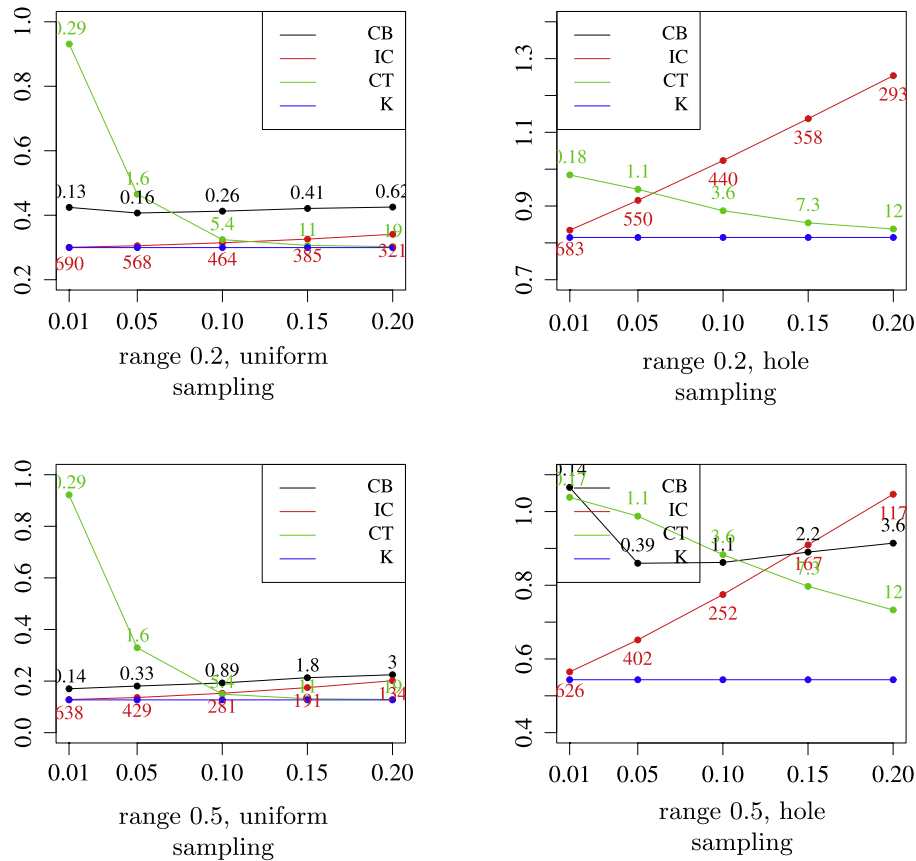
**Fig. 6.** MSPE as a function of  $\eta$  for each of the four methods (plus the density of the sparse matrix for CB and CT, and the number of components for IC): uniform sampling *a*; hole sampling *b*.

Graph *a* of Fig. 5 corresponds to the uniform sampling situation. It shows that the performances of the CT deteriorates quickly as  $\eta$  decreases. With  $\eta = 0.1$ , that corresponds to a density of 5.4% of the covariance matrix; the estimation of the MSPE is very close to the reference one however. In contrast, the difference between the estimated MSPE for K and that for IC and CB increases only slightly when  $\eta$  increases. In the same time, the number of components used in IC decreases drastically, from 570 to 52. It can be noticed that IC alone performs better than CB, whereas CB was meant to correct IC. This means that adding a sparse correction term to the incomplete Cholesky method does not seem to be relevant, at least in these settings. In the hole sampling case (graph *b*), CT does not perform well: the minimum MSPE obtained with  $\eta = 0.2$  is twice the reference value. Moreover, the density of the matrix is then 12% which is beyond the limits of sparsity. The MSPE values exhibited by CB are almost always greater than those of IC, except for large values of  $\eta$  in the hole case. With this sampling, CB exhibits performances better than those of IC for  $\eta \geq 0.1$ . For instance, for  $\eta = 0.1$ , CB allows one to improve slightly on IC. Nonetheless, the performances exhibited by each of the methods are farther from the reference in the hole sampling case, except for IC with 552 components which corresponds to an almost complete Cholesky case.

#### 4.2. Sensitivity to the covariance model

In this section, we study the sensitivity of the results obtained with each of the three approaches to the covariance model and parameters.

Fig. 6 depicts the MSPE in the case of a spherical covariance model with range 0.2. These results have to be compared to those of Fig. 5 where the range was 0.5, thus corresponding to a more regular covariance function. With the decrease of the range, CT gives slightly better results for both sampling schemes, as the initial covariance structure is less distorted in the operation. IC needs double the number of components to achieve results similar to those observed with a range of 0.5 with respect to the reference MSPE. The performances of CB also worsen when the range decreases, as the correction term only increases the estimated MSPE.

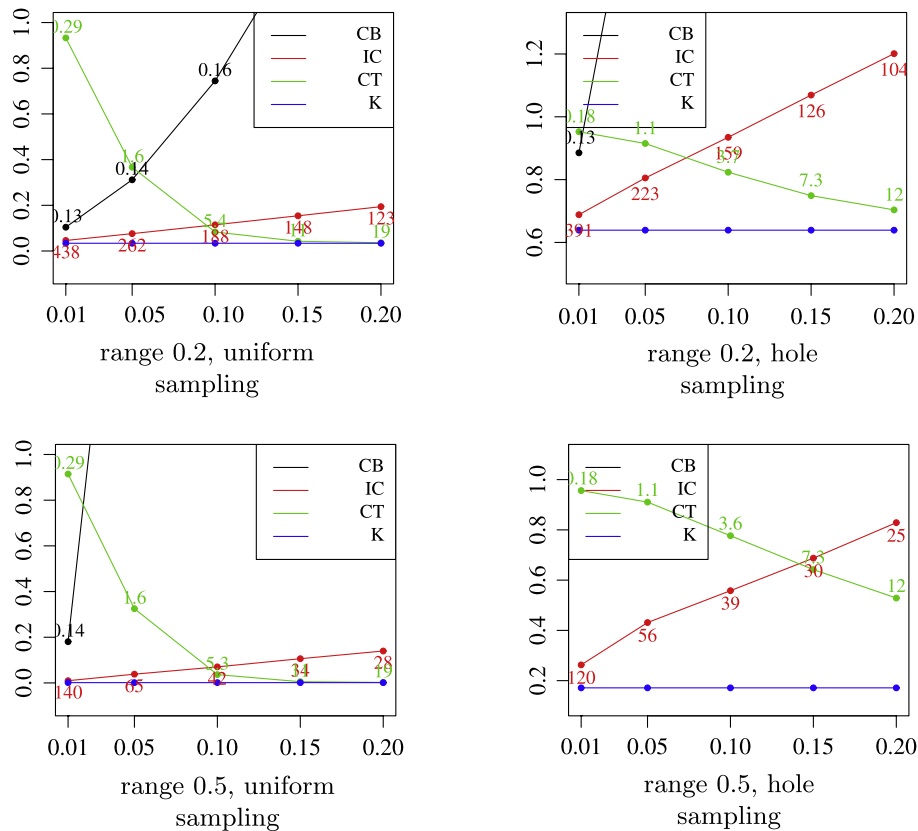


**Fig. 7.** MSPE as a function of  $\eta$  for each of the four methods (plus the density of the sparse matrix for CB and CT, and the number of components for IC); exponential covariance model.

Figs. 7–9 are obtained within the same settings as the previous ones, but with respectively the following three covariance models: exponential, Matérn with a smoothness parameter  $\nu = 2$ , and cubic (see e.g. Chilès and Delfiner (2012)). These models correspond to increasing degrees of smoothness of the related random function: non-differentiable for the exponential case, differentiable once for the Matérn case with  $\nu = 2$ , and differentiable twice for the cubic case, in the sense of the mean square differentiability. We will therefore study the sensitivity of the methods to the regularity of the covariance model. Both sampling schemes are used and crossed with two different ranges: 0.2 and 0.5.

We recognize more or less the same patterns as in the case of the spherical model: CB almost always gives worse results than IC, the tapering is always relatively bad in the hole sampling case, especially when the range is large, and the MSPE for IC is almost always linear in  $\eta$ . It is worth noticing that the number of components needed to represent a given proportion of the variance in IC decreases when the regularity of the covariance function increases: for a given covariance model and a fixed  $\eta$ , fewer components are needed for a 0.5 range than for 0.2 and far fewer components are needed for the cubic model than for the exponential at a given range. In particular, the number of components needed for the exponential covariance is very large whatever the range and may disqualify the method in this regard. In contrast, 32 components out of 750 are enough to represent 90% of the variance supported by the cubic covariance matrix with range 0.5 in the uniform case, and 42 for the Matérn model in the same situation.

To sum up, the results obtained have shown that there is clearly no best method overall that should be used in any situation. The sparse approximation (CT) is more appropriate for uniform sampling schemes combined with non-regular covariance models (short range and linear behavior at the origin) while the low rank approximation (IC) is more appropriate when the covariance model is smooth with a large range for any sampling scheme. On the other hand, the combination obtains very poor results in any situation. We believe that this is due to a numerical instability of the sparse term, when inverting it. Indeed, the low rank approximation is more or less accurate over the domain, depending on the



**Fig. 8.** MSPE as a function of  $\eta$  for each of the four methods (plus the density of the sparse matrix for CB and CT, and the number of components for IC); Matérn covariance model ( $\nu = 2$ , approximate range).

selected basis functions. In predictive process approximation for instance, the approximation is more accurate in the vicinity of the *knots* than elsewhere. Indeed, if  $x$  and  $y$  belong to the set of knots, then the approximation of  $C(x, y)$  is exact. In practice, the knots are not chosen at data points, but some of them may be close to one data point. Then the sparse approximation of the remainder of the difference between the true covariance matrix and its sparse approximation contains diagonal term values, and associated row and column term values, that are very close to zero. This makes the sparse matrix term unstable under inversion. The method may be more effective when a substantial nugget effect exists.

### 5. Discussion

When working on a grid, the incomplete Cholesky decomposition has the appealing property that it provides an approximation with a low number of components of the discretized random function. Let  $Z$  be a zero-mean Gaussian random function,  $\mathbf{x}$  the set of  $n$  grid nodes and  $C = C(\mathbf{x}, \mathbf{x})$  the covariance matrix of the vector  $Z(\mathbf{x})$ . We know that  $Z(\mathbf{x}) = C^{1/2}y$ , where  $C^{1/2}$  is a square root of  $C$ , for instance obtained by complete Cholesky decomposition, and  $y$  is a vector of  $n$  independent standard Gaussian random variables. Let us also have  $\tilde{Z}(\mathbf{x}) = G\tilde{y}$ , with  $G$  a  $n \times p$  matrix such that  $\|C - GG'\|_1 < \eta \|C\|_1$  and  $\tilde{y}$  a vector of  $p$  independent standard Gaussian random variables. If  $\tilde{y} = y_{\{1:p\}}$ , where  $y_{\{1:p\}}$  denotes the first  $p$  components of  $y$ , then  $\tilde{Z}(\mathbf{x})$  is an approximation of  $Z(\mathbf{x})$  that represents a proportion  $1 - \eta$  of the variance of  $Z$ . This approximated representation may be very useful for Bayesian inverse problems with Gaussian random function as a prior. Solving the inverse problem for  $\tilde{Z}(\mathbf{x})$  would give an approximate solution to the full problem.

Fig. 10 represents a realization of a standard Gaussian random function with spherical covariance of range 0.5 at different approximation levels. For each value of  $\eta$ , the approximation is computed by projecting the complete realization over the space spanned by the corresponding IC decomposition—in other words, on the space spanned by the corresponding columns of  $C$ . As  $\eta$  increases from 0 to 0.2, we can see the features of the realization fade away, from small scale to larger scale fluctuations,

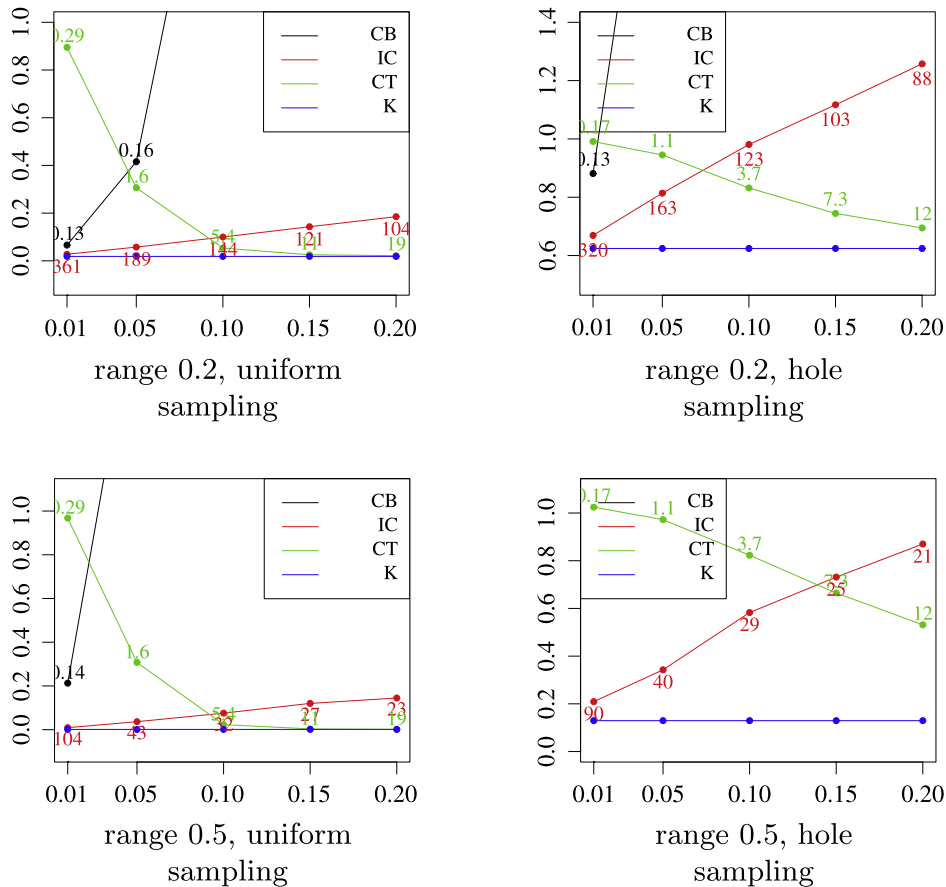


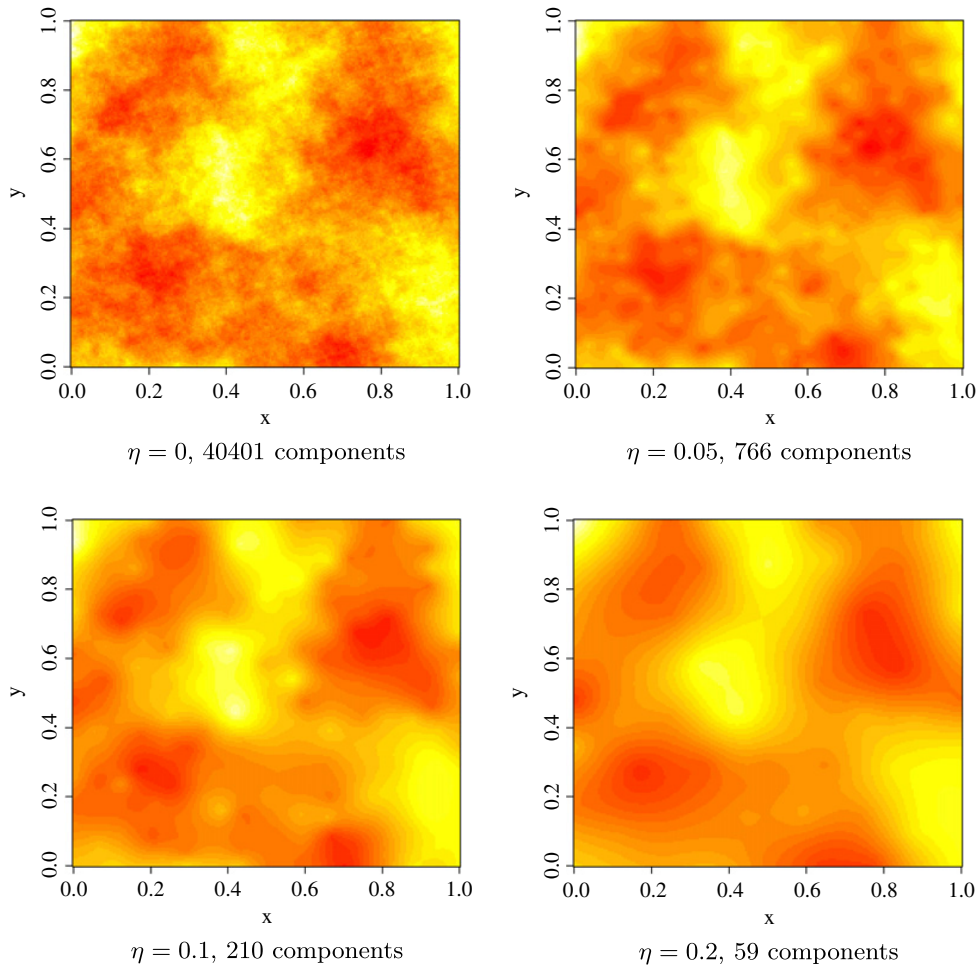
Fig. 9. MSPE as a function of  $\eta$  for each of the four methods (plus the density of the sparse matrix for CB and CT, and the number of components for IC); cubic covariance model.

while the number of components decreases from 40401 to 59. For some physical problems like fluid flow simulation, the large scale fluctuations may explain most of the response. Therefore, when solving inverse problems, the effort should focus on conditioning the most important components, which are far fewer than the total number of random variates needed for a complete realization. In this framework, the incomplete Cholesky decomposition can be an appealing dimension reduction technique, as it is much less computer resource demanding than others, like Karhunen–Loève expansion for instance; see Romary (2009).

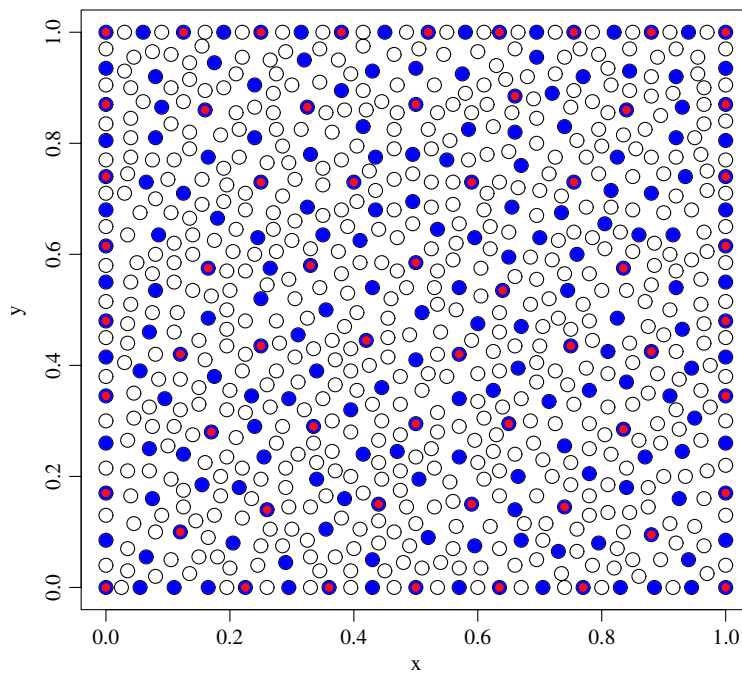
Finally, we represent in Fig. 11 the points of the grid corresponding to the pivots performed during the algorithm for different values of  $\eta$ , that is the columns selected during the algorithm. The black circles correspond to the 766 columns of  $C$  chosen among the 40401 possible ones to form the decomposition. The blue solid circles are the 210 points of the 90% approximation and the 59 smaller red disks correspond to the 80% approximation. We can see that they are uniformly spread over the  $[0, 1]^2$  square. Note that the sets of points are nested by construction. We recall that we approximate here a random function discretized on a  $201 \times 201$  grid. If we interpret this grid as an estimation grid for kriging, we could think of having an approximated kriging map when projecting the full one on the space supported by the  $p$  columns of  $C$  chosen in the IC decomposition. This is exactly what is done in the predictive process approach of Banerjee et al. (2008). Therefore, for a given estimation grid, the IC decomposition could be used to choose the number and location of the knots for the kriging map to achieve a given precision.

## 6. Conclusion

In this paper, we introduced the incomplete Cholesky decomposition algorithm, a method of numerical approximation for positive semidefinite matrices. It allows to build approximations of a



**Fig. 10.** Full (upper left corner) and approximate simulations (0.95% of the variance is preserved in the upper left corner, 0.9% in the lower left corner and 0.8% in the lower right corner).



**Fig. 11.** Knots selected by IC for different values of  $\eta$ .

covariance matrix whose accuracy is controlled by the proportion of the variance that one wants to retrieve. We introduced a way in which these approximations can be used for kriging, emphasizing its usefulness when the number of data is large.

We compared the results provided by this method to those from other popular methods of kriging for large datasets. We cannot conclude that there exists a best method overall: tapering performs better in the case of a dense, uniform sampling with a covariance that is not differentiable at the origin, while the incomplete Cholesky method gives accurate results with a very low number of components when the behavior at the origin of the covariance function is more regular. Depending on the regularity of the random function model and the sampling distribution, the user may choose between a sparse and a low rank approximation. The results obtained by the combination method were disappointing: the computed maps showed artifacts and the MSPE values were almost always greater than the incomplete Cholesky ones. This seems to be due to a numerical instability of the sparse term, when inverting it. Indeed, it admits by construction very small values where the low rank approximation is good. Adding a nugget effect may stabilize the results.

We have not studied how the number of components needed for a given accuracy scales with the number of data. According to Bach and Jordan (2003), in order to achieve a given required precision  $\eta$ , the rank  $p$  of an approximation to a Gaussian covariance matrix can be chosen as  $p = O(-\log(\eta))$ , that is, it does not depend on  $n$  ( $n$  needs to be large enough however). This is only valid for the Gaussian covariance, which has a very fast eigenvalue decay (Romary, 2009, see), but we can surmise that, for any covariance function, the number of components would stabilize for large  $n$ , as the estimation of the remaining sum of eigenvalues converges when  $n \rightarrow \infty$ .

We also showed that IC can be used to perform approximate simulations of Gaussian random functions at a low computational cost while drastically reducing the dimension of the input vector of random variates, which can be useful in inverse modeling. The algorithm can also be useful for selecting the knots in the predictive process approach (Banerjee et al., 2008) while controlling the error made when working on a given grid.

Finally, the main strength of IC with respect to existing methods is the quantification of the error made through the proportion of the variance. Another promising method that allows for quantifying the error made is by using hierarchical matrices (Saibaba and Kitanidis, 2012). Its application to kriging would consist of nested block sparse approximation of the covariance matrix so as to make its inversion possible by sparse complete Cholesky decomposition. This sparse approximation seems to be far more accurate than tapering and comes with a bound on the error.

## References

- Bach, F.R., Jordan, M.I., 2003. Kernel independent component analysis. *The Journal of Machine Learning Research* 3, 1–48.
- Bach, F.R., Jordan, M.I., 2005. Predictive low-rank decomposition for kernel methods. In: *Proceedings of the 22nd International Conference on Machine Learning, ICML '05*. ACM, New York, pp. 33–40. URL <http://doi.acm.org/10.1145/1102351.1102356>.
- Banerjee, S., Gelfand, A.E., Finley, A.O., Sang, H., 2008. Gaussian predictive process models for large spatial data sets. *Journal of the Royal Statistical Society. Series B* 70 (4), 825–848.
- Chilès, J.P., Delfiner, P., 2012. *Geostatistics, Modeling Spatial Uncertainty*, second ed. John Wiley & Sons, New York.
- Cressie, N., Johannesson, G., 2008. Fixed rank kriging for very large spatial data sets. *Journal of the Royal Statistical Society. Series B* 70, 209–226.
- Fine, S., Scheinberg, K., 2002. Efficient svm training using low-rank kernel representations. *Journal of Machine Learning Research* 2, 243–264. URL <http://dl.acm.org/citation.cfm?id=944790.944812>.
- Furrer, R., Genton, M.G., Nychka, D.W., 2006. Covariance tapering for interpolation of large spatial datasets. *Journal of Computational and Graphical Statistics* 15 (3), 502–523.
- Gelfand, A.E., Diggle, P.J., Fuentes, M., Guttorp, P. (Eds.), 2011. *Handbook of Spatial Statistics*. Chapman & Hall/CRC.
- Ghanem, R.G., Spanos, P.D., 1991. *Stochastic Finite Elements, A Spectral Approach*. Springer-Verlag.
- Golub, G., Van Loan, C., 1996. *Matrix Computations*. In: *Johns Hopkins Studies in the Mathematical Sciences*, Johns Hopkins University Press.
- Gribov, A., Krivoruchko, K., 2004. Geostatistical mapping with continuous moving neighbourhood. *Mathematical Geology* 36 (2), 267–281.
- Loève, M., 1955. *Probability Theory*. Van Nostrand, New York.
- Nychka, D., Wikle, C., Royle, J.A., 2002. Multiresolution models for nonstationary spatial covariance functions. *Statistical Modelling* 2, 315–331.
- Rivoirard, J., Romary, T., 2011. Continuity for kriging with moving neighborhood. *Mathematical geosciences* 43 (4), 469–481.
- Romary, T., 2009. Integrating production data under uncertainty by parallel interacting Markov chains on a reduced dimensional space. *Computational Geosciences* 13 (1), 103–122.



- Saibaba, A.K., Kitanidis, P.K., 2012. Efficient methods for large-scale linear inversion using a geostatistical approach. *Water Resources Research* 48.
- Sang, H., Huang, J., 2012. A full scale approximation of covariance functions for large spatial data sets. *Journal of the Royal Statistical Society. Series B* 74 (1), 111–132.
- Stein, M.L., 1993. A simple condition for asymptotic optimality of linear predictions of random fields. *Statistics & Probability Letters* 17, 399–404.
- Stein, M.L., 2008. A modeling approach for large spatial datasets. *Journal of the Korean Statistical Society* 37, 3–10.
- Sun, Y., Li, B., Genton, M., 2012. Geostatistics for large datasets. In: Porcu, E., Montero, J.-M., Schlather, M. (Eds.), *Advances and Challenges in Space–Time Modelling of Natural Events*. In: *Lecture Notes in Statistics*, Springer, pp. 55–77.

# Stochastic seismic tomography by interacting Markov chains

Alexis Bottero,<sup>1,2</sup> Alexandrine Gesret,<sup>1</sup> Thomas Romary,<sup>1</sup> Mark Noble<sup>1</sup>  
 and Christophe Maisons<sup>3</sup>

<sup>1</sup>MINES ParisTech, PSL Research University, Centre de Geosciences, 77 Fontainebleau, France. E-mail: alexis.bottero@gmail.com

<sup>2</sup>LMA, CNRS UPR 7051, Aix-Marseille University, Centrale Marseille, F-13453 Marseille Cedex 13, France

<sup>3</sup>Magnitude SAS, Centre Regain, Route de Marseille, 04 Sainte-Tulle, France

Accepted 2016 July 19. Received 2016 July 1; in original form 2016 March 11

## SUMMARY

Markov chain Monte Carlo sampling methods are widely used for non-linear Bayesian inversion where no analytical expression for the forward relation between data and model parameters is available. Contrary to the linear(ized) approaches, they naturally allow to evaluate the uncertainties on the model found. Nevertheless their use is problematic in high-dimensional model spaces especially when the computational cost of the forward problem is significant and/or the *a posteriori* distribution is multimodal. In this case, the chain can stay stuck in one of the modes and hence not provide an exhaustive sampling of the distribution of interest. We present here a still relatively unknown algorithm that allows interaction between several Markov chains at different temperatures. These interactions (based on importance resampling) ensure a robust sampling of any posterior distribution and thus provide a way to efficiently tackle complex fully non-linear inverse problems. The algorithm is easy to implement and is well adapted to run on parallel supercomputers. In this paper, the algorithm is first introduced and applied to a synthetic multimodal distribution in order to demonstrate its robustness and efficiency compared to a simulated annealing method. It is then applied in the framework of first arrival traveltimes seismic tomography on real data recorded in the context of hydraulic fracturing. To carry out this study a wavelet-based adaptive model parametrization has been used. This allows to integrate the *a priori* information provided by sonic logs and to reduce optimally the dimension of the problem.

**Key words:** Wavelet transform; Inverse theory; Tomography; Probability distributions.

## 1 INTRODUCTION

Tomographic inversion schemes aiming at reconstructing the subsurface structures from seismic traveltimes data are widely used (e.g. Rawlinson *et al.* 2010). The obtained wave propagation velocity distribution is usually a starting point for further seismological, structural or tectonic analysis. For a reliable and quantitative interpretation of the tomographic solution, an accurate velocity model with its associated uncertainties are required. Very often, the tomographic inverse problem is solved with iterative linearized approaches that minimize a misfit function. The misfit function usually measures the difference between observed and computed traveltimes as a function of the velocity model parameters. When the data-model relationship is highly non-linear, the misfit surface has a complex shape. It is most likely multimodal with a global maximum, representing the most likely solution, and a large number of secondary maxima, representing other possible solutions. The linearization makes the implicit assumption of a unique solution which is chosen thanks to a regularization procedure that reduces the solution non-uniqueness (Menke 2012). This is generally achieved by imposing the solution to be somehow similar or close to an initial *a priori* model. If the initial model is too far from the global minimum, the method may stop the inversion at any point where the misfit function does not change significantly. This can happen even for very simple tomographic inversion problem with only two layers (Weber 2000). In addition, the linearized approaches are not really adapted to provide reliable uncertainties. This is due either to possible multiple minima of the objective functions or to badly constrained area where many candidate models are all equally likely. Rawlinson *et al.* (2014) provide an extensive description of the existing methods to estimate uncertainties with local approaches and illustrate that a probabilistic sampling of the *a posteriori* distribution within a Bayesian framework (Tarantola 2005) is much more suitable.

In their pioneering work, Tarantola & Valette (1982) propose to sample the *a posteriori* probability density of the velocity model parameters. This posterior probability density function summarizes how the information brought by the seismic traveltimes data improves the *a priori* knowledge on the model parameters. Such a Bayesian approach is capable of dealing with highly non-linear problems. In the

Bayesian framework, either a single model maximizing the posterior distribution is sought, this is an optimization problem or the complete shape of the *a posteriori* distribution is inferred, this is a sampling problem. The sampling of the whole posterior distribution allows for an unbiased estimation of the velocity model uncertainties (Debski 2010) without any assumption on the shape of the misfit surface.

We focus here on Markov chain Monte Carlo (MCMC) methods introduced by Metropolis *et al.* (1953) for the universality of the method and the ease of its implementation. Monte Carlo type algorithms are useful to generate a large number of samples of statistically near-independent velocity models distributed according to the posterior distribution (Koren *et al.* 1991; Mosegaard & Tarantola 1995; Bodin & Sambridge 2009). In other words, their frequency of occurrence approximates the posterior probability density function. For example, the classical Metropolis–Hastings algorithm is known to converge towards the true posterior distribution (Robert & Casella 2013). However, the convergence can be quite slow. To overcome this drawback, during the first iterations of the algorithm, the true posterior distribution can be replaced by a tempered distribution. This corresponds to the simulated annealing (SA) algorithm which is commonly used in an optimization prospect where the final best velocity model is reached when the annealing temperature is close to zero (Salamon *et al.* 2002). Although particularly well suited for sampling and optimization (Robert & Casella 2013), two major impediments of classical MCMC methods exist. First, if the dimension of the problem is very large or if the forward problem is computationally expensive, the number of iterations needed for an adequate sampling of the posterior will never be reached. In addition, if the *a posteriori* distribution is multimodal, the chain may be stuck in one of the secondary modes. In order to sample very efficiently complex *a posteriori* distributions, it is possible to allow several Markov chains to interact at different temperatures, for example, according to the parallel tempering scheme (Geyer 1991). Recently Sambridge (2014) proposed to use parallel tempering for probabilistic sampling and multimodal optimization. Randomized swaps between any pair of adjacent temperature levels allow to significantly accelerate the convergence of the MCMC process. However in the parallel tempering algorithm exchanges can only occur between simultaneous states of the chains.

We go here a step forward by proposing an interacting MCMC (IMCMC) algorithm based on the population importance-resampling (PIR) MCMC sampler introduced by Andrieu *et al.* (2011) where exchanges are proposed taking advantage of the whole history of the chains. Romary (2009) and Romary (2010a) showed that this interaction scheme improves significantly the mixing properties and the convergence speed of the chains over other parallel MCMC schemes, including parallel tempering. This ensures better sampling and thus better results with less iterations for Bayesian inversion purpose. We first compare this IMCMC algorithm to a common SA algorithm for a highly multimodal function in order to test its robustness and efficiency. We then apply the IMCMC algorithm to invert a set of perforation shots recorded in the context of induced microseismicity. In this context, the calibrated velocity model has to be accurate since it is subsequently used for precise microseismic event locations that required for safety and exploitation optimization. In routine microseismic monitoring of hydraulic fracturing projects, sonic logs are used to construct a layered velocity model, which is then usually calibrated with the perforation shots from known locations (Eisner *et al.* 2009; Maxwell 2010). We propose an adaptive parametrization based on a optimally truncated decomposition of the sonic logs on a variable wavelet functions basis. This allows to use the sonic log as a prior information by describing a complex velocity model with a restricted number of wavelet coefficients. The wavelet type together with the coefficients become the model parameters of the inversion. We illustrate the efficiency and the robustness of our algorithm by applying it to the perforation shot data.

The paper is organized as follows. We first recall the formalism of the Bayesian inversion and briefly present the classical MCMC Metropolis–Hastings and SA algorithms. We then introduce our IMCMC algorithm and illustrate its performance for a highly multimodal distribution. We finally apply our approach to a real data set with an original adaptive model parametrization based on wavelet decomposition of sonic logs.

### 1.1 Mathematical setting, Bayesian context

The aim of traveltimes tomography is to infer the spatial structure  $\mathcal{S}$  of the seismic wave velocity field using observed traveltimes. It falls into the wide category of inverse problems. On the system  $\mathcal{S}$  we have a finite set of  $L$  imperfect indirect measurements  $\mathbf{d}^{\text{obs}} = (d_1^{\text{obs}}, \dots, d_L^{\text{obs}})$  belonging to the data space  $\mathcal{D} \subset \mathbb{R}^L$ , the measurement error being most of the time not perfectly known. The sought physical system  $\mathcal{S}$  is usually a continuous function of a general variable  $\mathbf{x}$  (space here)  $\mathcal{S}(\mathbf{x})$ . Practically, to reduce the dimension of the problem, this unknown system is approximated as a given function  $\mathcal{G}_b(\mathbf{x})$  that can be fully described by a finite set of  $N$  parameters  $\mathbf{b} = (b_1, \dots, b_N) \in \mathcal{M}$ .  $\mathcal{M} \subset \mathbb{R}^N$  is commonly called the model space

$$\mathcal{S}(\mathbf{x}) \approx \mathcal{G}_b(\mathbf{x}). \quad (1)$$

Knowing  $\mathcal{S}$  reduces then to find the best set of parameters  $\mathbf{b}$ . This approximation is called parametrization and can obviously affect the solution (e.g. Hartzell & Langer 1993; Köhn *et al.* 2012; Rawlinson *et al.* 2014). A good parametrization should extract the maximum information from the observed data, without introducing unjustified complexity to the model structure. The forward problem that predicts the synthetic measurements  $\mathbf{d}$  from a set of model parameters  $\mathbf{b}$  writes

$$\mathbf{d} = \mathbf{F}(\mathbf{b}) \quad (2)$$

where  $\mathbf{F}$  is the physical coupling that links the data  $\mathbf{d}$  to a given model  $\mathbf{b}$ . This function  $\mathbf{F}$ , generally non-linear, can only be imperfectly modeled. The inverse problem consists in finding the parameters  $\mathbf{b}$  for which  $\mathbf{d} = \mathbf{F}(\mathbf{b}) \approx \mathbf{d}^{\text{obs}}$ . In the following, we adopt the classical Bayesian framework (Tarantola 2005) in which all information is represented by probability distributions. This approach allows to transform the prior knowledge on model parameters into a posterior distribution while accounting properly for both the measurement and theoretical

uncertainties. For a given model  $\mathbf{b}$ , the forward problem is represented by the following  $N$ -dimensional probability distribution:

$$P(\mathbf{d}^{\text{obs}}|\mathbf{b}) \quad \text{probability of observing the data } \mathbf{d}^{\text{obs}} \text{ given the model } \mathbf{b} \quad (3)$$

commonly called the likelihood. It is a quantitative measure of the difference between the observed data  $\mathbf{d}^{\text{obs}}$  and the synthetic data  $\mathbf{d} \in \mathcal{D}$  computed from  $\mathbf{b} \in \mathcal{M}$ . This distribution depends on the stochastic noise associated to the observed data. It is common nevertheless to make the (simplifying) assumption that it has an uncorrelated Gaussian form:

$$P(\mathbf{d}^{\text{obs}}|\mathbf{b}) = e^{-\frac{\Delta(\mathbf{b})}{2}}, \quad (4)$$

in which  $\Delta(\mathbf{b})$  is a simple least-squares misfit function

$$\Delta(\mathbf{b}) = \left\| \frac{\mathbf{F}(\mathbf{b}) - \mathbf{d}^{\text{obs}}}{\sigma_d} \right\|^2 \quad (5)$$

where  $\sigma_d$  is the estimated standard deviation of the measurement errors that are assumed uncorrelated here. Some *a priori* information about the system is often available. This knowledge is expressed in terms of an *a priori*  $N$ -dimensional probability distribution on the values of the model parameters:

$$P(\mathbf{b}) \quad \textit{a priori} \text{ probability of the model } \mathbf{b} \quad (6)$$

The result of a non-linear inversion is not a unique value of  $\mathbf{b}$  but the following  $N$ -dimensional *a posteriori* probability distribution on the values of the parameters.

$$P(\mathbf{b}|\mathbf{d}^{\text{obs}}) \quad \textit{a posteriori} \text{ probability of the model } \mathbf{b} \text{ given the data } \mathbf{d}^{\text{obs}} \quad (7)$$

Bayes' rule (Bayes & Price 1763) allows to express explicitly this probability distribution:

$$P(\mathbf{b}|\mathbf{d}^{\text{obs}}) = \frac{P(\mathbf{d}^{\text{obs}}|\mathbf{b})P(\mathbf{b})}{P(\mathbf{d}^{\text{obs}})}. \quad (8)$$

Note that the term  $P(\mathbf{d}^{\text{obs}}) \equiv \int_{\mathcal{M}} P(\mathbf{d}^{\text{obs}}|\mathbf{b})P(\mathbf{b}) d\mathbf{b}$ , often called the probability of the data or the evidence (Sambridge *et al.* 2006), is generally intractable in high-dimensional problems but does not depend on any particular state of the system and can therefore be considered as a normalization constant. Hence, the problem to be solved amounts to evaluate

$$P(\mathbf{b}|\mathbf{d}^{\text{obs}}) \propto P(\mathbf{d}^{\text{obs}}|\mathbf{b})P(\mathbf{b}) = P(\mathbf{b}) \exp\left(-\frac{1}{2} \left\| \frac{\mathbf{F}(\mathbf{b}) - \mathbf{d}^{\text{obs}}}{\sigma_d} \right\|^2\right). \quad (9)$$

Knowing (9) exhaustively implies then to calculate the forward problem for all  $\mathbf{b} \in \mathcal{M} \subset \mathbb{R}^N$  or at least on a very large sample of models whose dimension makes the problem intractable in practice. MCMC methods are then used to sample these probability distributions in an efficient way, as they only involve computations of ratios of such probabilities.

## 1.2 Markov chain Monte Carlo (MCMC)

In the field of geophysical inversion, several methods have been proposed for solving (9). In a sampling purpose, one can cite the neighborhood algorithm (Sambridge 1999a,b) which seems to be limited to low-dimensional problems, or, in an optimization prospect, the genetic algorithm (based on Holland 1975) and the Kitanidis–Oliver algorithm (see Kitanidis 1995; Oliver 1996) whose speed of convergence is controversial. The efficiency for optimization of the last two has impelled some attempts of modifying them, so that they could also be used for sampling a probability distribution (see Sen & Stoffa 2013) but without success so far (Mosegaard & Sambridge 2002). We focus here on MCMC for the universality of the method and the ease of its implementation. Introduced by Metropolis *et al.* (1953), the aim of MCMC methods can be either optimization or sampling. The basic idea is to generate a random sequence of parameters  $\mathbf{b} \in \mathcal{M} \subset \mathbb{R}^N$

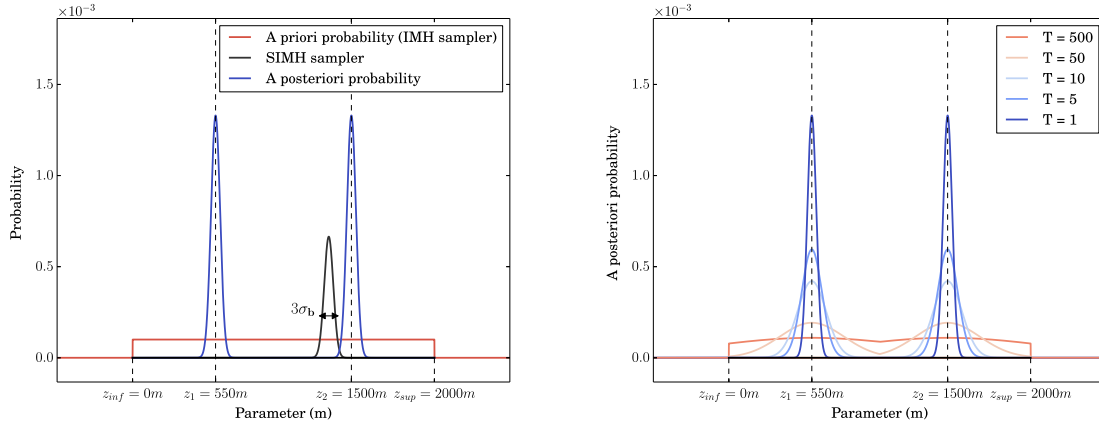
$$\mathbf{b}^0 \rightarrow \mathbf{b}^1 \rightarrow \mathbf{b}^2 \rightarrow \dots \rightarrow \mathbf{b}^i \rightarrow \dots \quad (10)$$

whose distribution  $\mathcal{B}$  equals asymptotically (and therefore approximately in finite time) the probability  $P(\mathbf{b}|\mathbf{d}^{\text{obs}})$ . The main characteristic of the chain is that every proposed state only depends on the previous one. The chain is said having converged when:

$$\mathcal{B} \approx P(\mathbf{b}|\mathbf{d}^{\text{obs}}). \quad (11)$$

Before the convergence is reached the chain is performing what is commonly called a burn-in period. In real problems, several methods exist to assess the convergence (e.g. Cowles & Carlin 1996; Brooks & Roberts 1998; Robert & Casella 2013) but obviously do not allow to rigorously ensure it. It is an assumption which has to be made. A set  $\kappa$  containing  $K$  samples computed after the convergence of the chain  $(\mathbf{b}^k)_{k \in \kappa} \in \mathcal{M}^K$  can then be used to estimate posterior expectations  $I_\kappa(h)$  of any function  $h(\mathbf{b})$ :

$$I_\kappa[h] \approx \int_{\mathcal{M}^K} P(\mathbf{b}|\mathbf{d}^{\text{obs}})h(\mathbf{b}) d\mathbf{b} \approx \frac{1}{K} \sum_{k \in \kappa} h(\mathbf{b}^k) \quad (12)$$



**Figure 1.** Left: example of a 1-D bimodal distribution with two samplers. The IMH sampler is not normalized. Right: effect of the temperature on this *a posteriori* example distribution  $P(\mathbf{b}|\mathbf{d}^{\text{obs}})$ .

and its associated variance:

$$\text{Var}_\kappa[h] = I_\kappa[(h - I_\kappa[h])^2] \approx \frac{1}{K} \sum_{k \in \kappa} (h(\mathbf{b}^k) - I_\kappa[h])^2. \quad (13)$$

For example,  $h(\mathbf{b}) = \mathbf{b}$  gives the mean posterior model and (13) an estimation of the associated uncertainties. Assuming the convergence is reached (11), the error on the value decreases as  $K^{-\frac{1}{2}}$  independently from the dimension of the space (Meyn & Tweedie 2012). However, if the Markov chain slowly moves in the model space, the states are highly correlated with each other and the chain is said to have poor mixing properties. In that case, the choice of the samples to be used  $(\mathbf{b}^k)_{k \in \kappa} \in \mathcal{M}^K$  is often made by taking off one sample out of several after the burn-in period to reduce the variance of the estimator  $I_\kappa[h]$ . As an illustration Bodin & Sambridge (2009) have used less than one sample out of 150 for their tomography of the Australian continent. The interest of this process called thinning is controversial and induces a considerable waste of resources. MacEachern & Berliner (1994) showed that more precise posterior estimates are obtained if the entire Markov chain is used.

### 1.2.1 MCMC sampling: the Metropolis–Hastings algorithm

The Metropolis–Hastings algorithm is an MCMC construction which can be tailored to sample from any target distribution, including posterior distributions in the Bayesian inversion context. As every state of a Markov chain only depends on the previous one, the chain behaviour depends on the rule chosen to go from one state to another. In the Metropolis–Hastings algorithm, a move is made in two steps. First, from the current state  $\mathbf{b}^i$  of the chain, a potential transition to a state  $\mathbf{b}^*$  is proposed following an instrumental distribution  $q(\mathbf{b}^i, \mathbf{b}^*)$  (probability of proposing a transition to  $\mathbf{b}^*$  from  $\mathbf{b}^i$ ), which has to be compatible with the prior. This distribution is called the sampler or the proposal distribution. The transition to  $\mathbf{b}^*$  is then accepted with a probability  $\alpha(\mathbf{b}^i, \mathbf{b}^*)$  defined as:

$$\alpha(\mathbf{b}^i, \mathbf{b}^*) = \begin{cases} \min\left(1, \frac{P(\mathbf{b}^*|\mathbf{d}^{\text{obs}})q(\mathbf{b}^i, \mathbf{b}^*)}{P(\mathbf{b}^i|\mathbf{d}^{\text{obs}})q(\mathbf{b}^*, \mathbf{b}^i)}\right) & \text{if } P(\mathbf{b}^i|\mathbf{d}^{\text{obs}})q(\mathbf{b}^i, \mathbf{b}^*) \neq 0, \\ 1 & \text{else.} \end{cases} \quad (14)$$

If the transition is accepted,  $\mathbf{b}^{i+1} = \mathbf{b}^*$ ; otherwise  $\mathbf{b}^*$  is rejected and the chain stays in the same state:  $\mathbf{b}^{i+1} = \mathbf{b}^i$ .

Several possibilities exist for the choice of the proposal distribution  $q$  (Robert & Casella 2013) and this choice strongly determines the inversion. For example, the Gibbs sampler is widely used for inverse problems (Mosegaard & Sambridge 2002) but the important number of calls to the forward problem limits its relevancy for seismic tomography applications. We here introduce two widely used proposal distributions that we have implemented:

(i) The independent sampler (IMH):  $q(\mathbf{b}^i, \mathbf{b}^*) = q(\mathbf{b}^*)$  where  $q$  is generally the *a priori* probability distribution (6) in Bayesian inversion. The proposed states are independent, they do not depend on the current one but are suggested according to the *a priori* information.

(ii) The symmetric increment random-walk sampler (SIMH):  $q(\mathbf{b}^i, \mathbf{b}^*) = q(|\mathbf{b}^i - \mathbf{b}^*|)$ . With this sampler, new states are proposed according to their distance to the current one.  $q(\mathbf{b}^*, \mathbf{b}^i) = q(\mathbf{b}^i, \mathbf{b}^*)$ , the acceptance probability then simplifies:

$$\alpha(\mathbf{b}^i, \mathbf{b}^*) = \begin{cases} \min\left(1, \frac{P(\mathbf{b}^*|\mathbf{d}^{\text{obs}})}{P(\mathbf{b}^i|\mathbf{d}^{\text{obs}})}\right) & \text{if } P(\mathbf{b}^i|\mathbf{d}^{\text{obs}}) \neq 0 \\ 1 & \text{else.} \end{cases} \quad (15)$$

We illustrate the two samplers in Fig. 1. For this 1-D bi-modal distribution example, the IMH sampler (red curve) equals the *a priori* distribution. It only depends on the state  $\mathbf{b}^*$ . The chain has thus good mixing properties since the successive states are not correlated (when the transition is accepted). In this sense, it is a global sampler that allows to explore coarsely the whole parameter space  $\mathcal{M}$  but commonly fails to identify the different modes (the different maxima) of the *a posteriori* distribution  $P(\mathbf{b}|\mathbf{d}^{\text{obs}})$ , especially in large dimensions. We also

represent in black in Fig. 1 an SIMH sampler  $q(\mathbf{b}^*|\mathbf{b}^i)$  based on a normal distribution  $\mathcal{N}(\mathbf{b}^i, \sigma_b^2)$  centred on the current state  $\mathbf{b}^i$  and with a standard deviation  $\sigma_b$ . The closer a state  $\mathbf{b}^*$  is from  $\mathbf{b}^i$ , the more likely it may be proposed. It is clear that for a small  $\sigma_b$  value (as shown here), the chain can spend a lot of iterations in the low probability area, it will then detect and permit a fine exploration of one of the modes. The SIMH is in this case a local sampler that will hardly leave a local mode. It does not give a wide overview of the distribution  $P(\mathbf{b}|\mathbf{d}^{\text{obs}})$ , the chain has low mixing properties as the successive states are strongly correlated. Conversely, if  $\sigma_b$  is too large, the sampler proposes large moves that are mostly rejected: the chain can eventually detect both modes but will probably not sample them correctly. For this sampler, an appropriate  $\sigma_b$  has to be chosen. It is a difficult tuning problem generally tackled by trial and error.

### 1.2.2 MCMC optimization: the simulated annealing algorithm (SA)

The SA has been introduced by Metropolis *et al.* (1953) and then generalized by Kirkpatrick *et al.* (1983) in the framework of optimization. It is a specialization of the previously introduced Metropolis–Hastings algorithm, the fundamental idea being to control the amplitude of the accepted perturbations thanks to a positive parameter  $T$  called the temperature. By considering the Gaussian likelihood (4) and defining the energy of the system as  $E(\mathbf{b}) \equiv \frac{\Delta(\mathbf{b})}{2} - \ln(P(\mathbf{b}))$ , we can rewrite the *a posteriori* distribution (8) as

$$P(\mathbf{b}|\mathbf{d}^{\text{obs}}) = \frac{e^{-E(\mathbf{b})}}{\int_{\mathcal{M}} e^{-E(\mathbf{b})} d\mathbf{b}} \quad (16)$$

and recognize a Gibbs–Boltzmann distribution of energy  $E$ . As for (9), we simply write:

$$P(\mathbf{b}|\mathbf{d}^{\text{obs}}) \propto e^{-E(\mathbf{b})}. \quad (17)$$

The SA consists in choosing a maximum temperature  $T_{\text{max}}$  and performing a classical MCMC sampling (for instance using an SIMH sampler) of the following deformed (or tempered) *a posteriori* Gibbs–Boltzmann distribution:

$$P_T(\mathbf{b}|\mathbf{d}^{\text{obs}}) \propto e^{-\frac{E(\mathbf{b})}{T}}. \quad (18)$$

The influence of the temperature parameter on a multimodal distribution is shown in Fig. 1 (right). It can be seen that at high temperature the *a posteriori* distribution gets close to a uniform distribution, the two modes are connected. Indeed at  $T \rightarrow +\infty$

$$P_T(\mathbf{b}|\mathbf{d}^{\text{obs}}) \xrightarrow{T \rightarrow +\infty} \text{uniform distribution}. \quad (19)$$

Decreasing the temperature parameter allows to detect and progressively reveal the modes. Obviously, at  $T = 1$ :

$$P_1(\mathbf{b}|\mathbf{d}^{\text{obs}}) = P(\mathbf{b}|\mathbf{d}^{\text{obs}}). \quad (20)$$

The temperature is slowly lowered from  $T_{\text{max}}$  to  $T = 0$  during the run. Close to  $T = 0$ , the distribution approximates a Dirac distribution at the global minimum for  $E(\mathbf{b})$  (if it is unique).

Using the energy the acceptance probability can be rewritten

$$\alpha(\mathbf{b}^i, \mathbf{b}^*) = \min\left(1, e^{\frac{E(\mathbf{b}^i) - E(\mathbf{b}^*)}{T}}\right). \quad (21)$$

Hence at  $T \rightarrow +\infty$ , every new model is accepted while at  $T \rightarrow 0$ , only the less energetic (the most probable) have a chance to be accepted. Since  $T$  decreases along the algorithm, the acceptance probability  $\alpha(\mathbf{b}^i, \mathbf{b}^*)$  varies with time and SA can be viewed as a non-homogeneous version of Metropolis–Hastings algorithm. Consequently, classical theoretical results on Markov chains do not apply for SA. Rigorously speaking, under some broad conditions, the algorithm converges to the global minimum of  $E(\mathbf{b})$  for a starting temperature high enough and a cooling in  $T \propto \frac{1}{\log(n)}$  (with  $n$  the iteration). This theoretical optimal cooling rate is intractable in practice and a geometrically decreasing sequence is generally used. The choice of the maximum temperature is done by considering some heuristic rules based on the acceptance rate (14) at  $T_{\text{max}}$  (Robert & Casella 2013).

Sampling a probability distribution with SA can be considered by stopping the cooling at  $T = 1$  and continuing a Metropolis–Hastings exploration. Nevertheless, for a multimodal posterior distribution, if the cooling has been made too fast and the chain is on a local minimum of energy far from the global minimum (in term of  $\sigma_b$  of the sampler), the chain will have an infinitely small probability to move toward and sample around the global minimum.

As just detailed, sampling correctly a complex posterior distribution and estimating the uncertainties may be problematic with classical Metropolis–Hastings or SA algorithms. Moreover, the number of iterations needed for a good sampling can be unreachable if the forward problem is too computationally expensive or the number of model parameters to retrieve too large. To cope with these impediments, we introduce in the next section a more elaborated algorithm.

## 2 INTERACTING MARKOV CHAINS

### 2.1 Theory and setting

#### 2.1.1 Presentation

We present here a still relatively unknown algorithm that brings a solution to the drawbacks of classical MCMC methods and seems therefore particularly suited for complex sampling problems such as Bayesian inversion. The basic idea is to run several chains at different temperatures and allow them to interact in order to sample efficiently an *a posteriori* distribution. The idea first appears in Geyer (1991). Its algorithm is called parallel tempering and has been originally applied in physicochemical simulations [see Earl & Deem (2005) and references therein]. Undoubtedly Monte Carlo algorithms are particularly adapted to parallel computing (see Rosenthal 2000; Hall *et al.* 2014). Nevertheless, even if parallel computing allows to explore a bigger part of the model space, the previously mentioned drawbacks of MCMC still exist: the procedure neither improves the convergence speed nor the mixing properties of the non-interacting chains. Most samples are still discarded either during burn-in or thinning and this constitutes a considerable waste of resources.

If running several non-interacting Markov chains in parallel is already widely used for inverse problems (e.g. Agostinetti *et al.* 2015), less studies have applied interacting Markov chains. This approach is also known under the names of exchange Monte Carlo or Metropolis coupled chains (Iba 2001). More recently, the statistical community paid attention to interacting Markov chains and proposed a theoretical background and new algorithms. In this approach, the chains at different temperatures can exchange information along their run. The term of these exchanges can be defined in several ways. For example, Sambridge (2014) recently used the basic parallel tempering algorithm where exchanges occur between simultaneous states of the chains. However, this approach had been compared with two other interaction schemes [equi-energy sampler Kou *et al.* (2006) and PIR of Andrieu *et al.* (2011)] in Romary (2009). These numerical results showed that the latter outperforms the other two for Bayesian inversion purpose. Indeed, it is based on a powerful interaction scheme making use of the whole history of the chains rather than the simultaneous states only. Our work is thus based on PIR interaction schemes which allows the low-temperature chains to access the history of the adjacent higher temperature ones. The first applications of the method in geosciences are Romary (2009, 2010a,b) where it has been used for reservoir characterization.

The previous discussions about MCMC and SA allows us to present easily this approach in the Bayesian inversion context. In a few words, several chains are run simultaneously, each one at a different temperature in the sense of SA. Instead of performing only Metropolis–Hastings updates, every chain can also jump to any past state of the chain at the next higher temperature with a probability defined by importance weights that will foster the low-energy states. The transition probability is then governed by eq. (26). Let us describe the approach in a more mathematical way.

#### 2.1.2 Mathematical setting

We recall that we aim at sampling from the posterior distribution (17).  $K$  Markov chains are running in parallel, each one at a different temperature  $T_l$  with  $l = 1, \dots, K$  in the sense of SA. The chain  $l$  is actually exploring the following tempered version of  $P(\mathbf{b}|\mathbf{d}^{\text{obs}})$ :

$$P_l(\mathbf{b}|\mathbf{d}^{\text{obs}}) \propto e^{-\frac{E(\mathbf{b})}{T_l}}. \quad (22)$$

We note  $q_l(\mathbf{b}, \mathbf{b}^*)$  the sampler of chain  $l$  (see Section 1.2.1). The temperature ladder verifies:

$$T_1 = 1 < T_2 < \dots < T_K = T_{\text{max}}. \quad (23)$$

In this approach, the chains at different temperatures can exchange information along their run. PIR allows importance sampling swaps between chains at adjacent temperatures. At each iteration, a swap move is proposed to any single chain (except the highest temperature one) with a probability  $\theta$ .

(i) If it is rejected, the chain performs a classical Metropolis–Hastings iteration as described in Section 1.2.1.

(ii) If it is accepted, an importance swap is proposed. It potentially allows the chain to move to any past state of the chain at the next higher temperature. This step is described below.

If we note  $\mathbf{b}^{(l,n)} = (b_1^{(l,n)}, b_2^{(l,n)}, \dots, b_N^{(l,n)}) \in \mathcal{M}$  the state number  $n$  generated by the chain  $l$  (at temperature  $T_l$ ), a state  $\mathbf{b}^{(l+1,*)}$  is picked among the past states of the adjacent higher temperature chain:

$$H_{n-1}^{(l+1)} \equiv \{\mathbf{b}^{(l+1,1)}, \mathbf{b}^{(l+1,2)}, \mathbf{b}^{(l+1,3)}, \dots, \mathbf{b}^{(l+1,n-1)}\}.$$

To each past state of that chain,  $\mathbf{b}^{(l+1,i)}$ ,  $i = 1, \dots, (n-1)$  is associated an importance weight:

$$r^{(l)}(\mathbf{b}^{(l+1,i)}) = \frac{P_l(\mathbf{b}^{(l+1,i)}|\mathbf{d}^{\text{obs}})}{P_{l+1}(\mathbf{b}^{(l+1,i)}|\mathbf{d}^{\text{obs}})} = \frac{e^{-\frac{E(\mathbf{b}^{(l+1,i)})}{T_l}}}{e^{-\frac{E(\mathbf{b}^{(l+1,i)})}{T_{l+1}}}} = e^{-E(\mathbf{b}^{(l+1,i)})\left(\frac{1}{T_l} - \frac{1}{T_{l+1}}\right)}. \quad (24)$$

which is the un-normalized ratio of the tempered distributions at  $T_l$  and  $T_{l+1}$ . Sampling a state  $\mathbf{b}^{(l+1,i^*)}$  of the chain  $l + 1$ , that is, virtually distributed according to  $P_{l+1}(\mathbf{b}^{(l+1,i^*)}|\mathbf{d}^{\text{obs}})$ , with respect to this weight allows this state to be considered as distributed according to  $P_{l+1}(\mathbf{b}^{(l+1,i^*)}|\mathbf{d}^{\text{obs}})$  following the principle of importance sampling (Robert & Casella 2013).

- The idea is pretty simple. Considering two probability density function,  $f$  and  $g$ , and a variable  $X$  that follows  $f$ . We can write:

$$E(h(X)) = \int h(x)f(x)dx = \int h(x)\frac{f(x)}{g(x)}g(x)dx. \quad (25)$$

Therefore, this expectation can be computed by Monte Carlo integration using random numbers following  $g$  and reweighted by the ratio  $\frac{f(x)}{g(x)}$ .

- This method gives more chance to the less energetic states of the chain  $l + 1$  to be proposed.
- The transition is then accepted or not with probability:

$$\beta(\mathbf{b}^{(l,n)}, \mathbf{b}^{(l+1,i^*)}) = \min\left(1, \frac{r^{(l)}(\mathbf{b}^{(l+1,i^*)})}{r^{(l)}(\mathbf{b}^{(l,n)})}\right) \quad (26)$$

$$= \min\left(1, \exp\left(\left(\frac{1}{T_l} - \frac{1}{T_{l+1}}\right)(E(\mathbf{b}^{(l,n)}) - E(\mathbf{b}^{(l+1,i^*)}))\right)\right). \quad (27)$$

If the energy of the proposed state  $\mathbf{b}^{(l+1,i^*)}$  is lower than that of  $\mathbf{b}^{(l,n)}$ , the exchange will be systematically accepted.

The efficiency of the algorithm relies on the type of sampler used by each chain. The highest temperature chain might use a global sampler (e.g. IMH) and thus explores coarsely the whole model space while the lower temperature chains use local samplers (e.g. SIMH) and thus are able to sample thoroughly the interesting regions discovered by the higher temperature chains. The standard deviation of the normal distribution of each SIMH sampler may also decrease with the temperature. We added a detailed flowchart of the algorithm in Appendix. As any MCMC-based inference method, the algorithm here proposed needs to be tuned. The parameters that have to be chosen are:

- The number  $K$  of chains.
- The sequence  $(T_l)_l$  of temperatures.
- The samplers at each temperature  $q_l(\mathbf{b}^*|\mathbf{b})$ .
- The probability  $\theta$  of proposing a swap.

The approach is purely empirical, based on numerical experiments. Romary (2010a) summarizes the existing guidelines and rules of thumb. Hence, for example, the number of chains to use should be the minimum number that ensures a good overlapping of the histograms of the tempered energies. For the choice of  $T_{\max}$  however it seems more suitable to us to run a single chain at temperature 1 sampling independently the prior distribution (6) during say,  $N = 500$  iterations:  $(\mathbf{b}^k)_{0 \leq k \leq 500}$ . The resulting energy distribution  $(E(\mathbf{b}^k))_{0 \leq k \leq 500}$  can then be used to choose  $T_{\max}$  in order to obtain a satisfactory transition acceptance rate (21) at the highest temperature: between 0.6 and 0.75 in practice. Knowing the number of chains and the maximum temperature, the temperature ladder can be built, for example, as a geometrical progression between 1 and  $T_{\max}$ .

### 2.1.3 Theoretical improvements

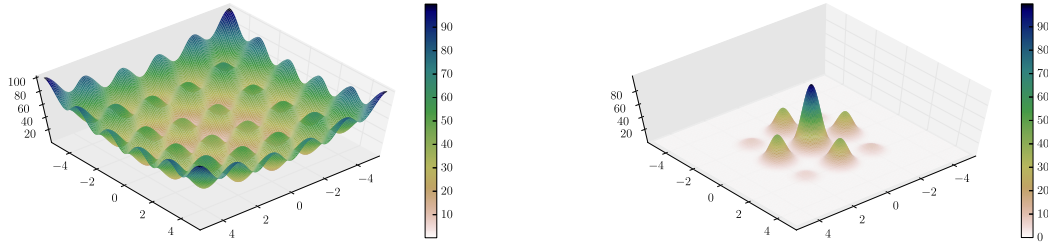
As we have seen, without exchange, the highest temperature chain would sample coarsely large volumes of the model space  $\mathcal{M}$  whereas low-temperature chains would precisely sample local regions but could become trapped in local energy minima. Allowing the lowest temperature chains to be regenerated in an interesting low energetic location previously sampled by the upper chains significantly enhances their mixing (Romary 2010a; Andrieu *et al.* 2011) and their speed of convergence toward the stationary regime, especially the lowest temperature one. Each chain can run physically in parallel, that is on different processors. However, if not working on a CPU cluster, running  $K$  chains rather than 1 during  $n$  iterations requires more computational effort. Nevertheless, the sampling will clearly be more efficient than running a single chain during  $nK$  iterations. Indeed, in practice, a single chain would not have been able to access to the whole region which is always within reach of the highest temperature chains during the parallel interacting sampling. Moreover, as shown in Romary (2010a), the states generated by parallel IMCMC are far less correlated than the ones generated by an SA or a classical Metropolis–Hastings algorithm thus providing less variable estimates. To estimate posterior expectations  $I[h]$  of a function  $h(\mathbf{b})$  (see eq. (12)) every sample generated at each temperature can also be used. Every state  $\mathbf{b}^{(l,n)}$  of the chain at temperature  $T_l$  just having to be weighted by the following importance weights:

$$\eta^{ln} = e^{-E(\mathbf{b}^{(l,n)})\left(1 - \frac{1}{T_l}\right)}. \quad (28)$$

Therefore, the estimate of  $I[h]$  after  $N$  iterations of  $K$  chains will take the form:

$$I[h] \approx \frac{1}{K} \sum_{l=1}^K \frac{\sum_{n=1}^N \eta^{ln} h(\mathbf{b}^{(l,n)})}{\sum_{n=1}^N \eta^{ln}}. \quad (29)$$





**Figure 2.** Left: 2-D Rastrigin function  $\mathbf{b} \mapsto \mathcal{R}(\mathbf{b})$ . Right: KDE of the *a posteriori* distribution sought  $P(\mathbf{b}|\mathbf{d}^{\text{obs}}) = P(\mathbf{b})e^{-\frac{1}{2}\left|\frac{\mathcal{R}(\mathbf{b})}{\sigma_d}\right|^2}$ .

**Table 1.** Features of the run for the interacting MCMC.

Number $K$ of chains	3		
Probability $\theta$ of proposing a swap	25 per cent		
Sequence $(T_l)_l$ of temperatures	Geometric sequence between $T_{\max} = 50$ and $T = 1$ :		
Samplers at each temperature	$T_0 = 1$	$T_1 = \sqrt{50}$	$T_{\max} = 50$
$q_l(\mathbf{b}^* \mathbf{b})$ :	SIMH	SIMH	IMH
$\sigma_{\mathbf{b}} = (\sigma_x, \sigma_y)$ :	$\mathcal{N}_2(\mathbf{b}, \mathbb{1}_2 \sigma_{\mathbf{b}}^2)$	$\mathcal{N}_2(\mathbf{b}, \mathbb{1}_2 \sigma_{\mathbf{b}}^2)$	$P(\mathbf{b}^*)$
	Linear decreasing functions of $T$ :		
	$\sigma_x = \sigma_y = \frac{5.12+5.12}{50}$	$\sigma_x = \sigma_y = \frac{5.12+5.12}{20.62}$	

With its estimated variance:

$$\text{Var}[h] \approx \frac{1}{K} \frac{\sum_{n=1}^N \eta^{ln} (h(\mathbf{b}^{(l,n)}) - I[h])^2}{\sum_{n=1}^N \eta^{ln}}. \tag{30}$$

Kou *et al.* (2006) have shown that using this weighted sum over the chains provides better estimates than using only the states generated by the lowest temperature chain (even if in practice the  $\eta^{ln}$  are very small).

### 2.2 Test on a highly multimodal distribution

In order to illustrate the efficiency and the robustness of the algorithm, we apply it to sample a 2-D multimodal distribution. The forward problem (2) is the following 2-D Rastrigin function  $\mathcal{R}$  depicted in Fig. 2 (left):

$$\mathcal{R} : \begin{cases} [-5.12, 5.12]^2 \longrightarrow \mathbb{R} \\ \mathbf{b} = (x, y) \longmapsto \mathcal{R}(\mathbf{b}) = \mathcal{R}(x, y) = 20 + (x^2 - 10 \cos(\pi x)) + (y^2 - 10 \cos(\pi y)). \end{cases}$$

It was first proposed by Torn & Zilinskas (1989) as a 2-D function and has been generalized for  $N$ -dimensions by Mühlenbein *et al.* (1991). It is a typical example of a non-linear multimodal function useful to test the performance of a sampling algorithm. It has a global minimum in  $(x, y) = (0, 0)$  and several local minima. For this test, following the notations of the previous section, a state  $\mathbf{b}$  is composed of two parameters  $x$  and  $y$  belonging to the model space  $\mathcal{M} = [-5.12, 5.12]^2$ . Our *a priori* knowledge on the values of these parameters is a uniform distribution over  $\mathcal{M}$  (every value of  $(x, y)$  has the same *a priori* probability):

$$\forall \mathbf{b} \in [-5.12, 5.12]^2 \quad P(\mathbf{b}) = \frac{1}{(5.12 + 5.12)^2}. \tag{31}$$

The data  $\mathbf{d}^{\text{obs}}$  are 0 with a significant standard deviation  $\sigma_d = \sigma_d = \sqrt{50}$ . Then, the energy  $E(\mathbf{b})$  of a state  $\mathbf{b}$  is

$$E(\mathbf{b}) = \frac{1}{2} \left| \frac{\mathcal{R}(\mathbf{b})}{\sigma_d} \right|^2 - \ln(P(\mathbf{b})) \tag{32}$$

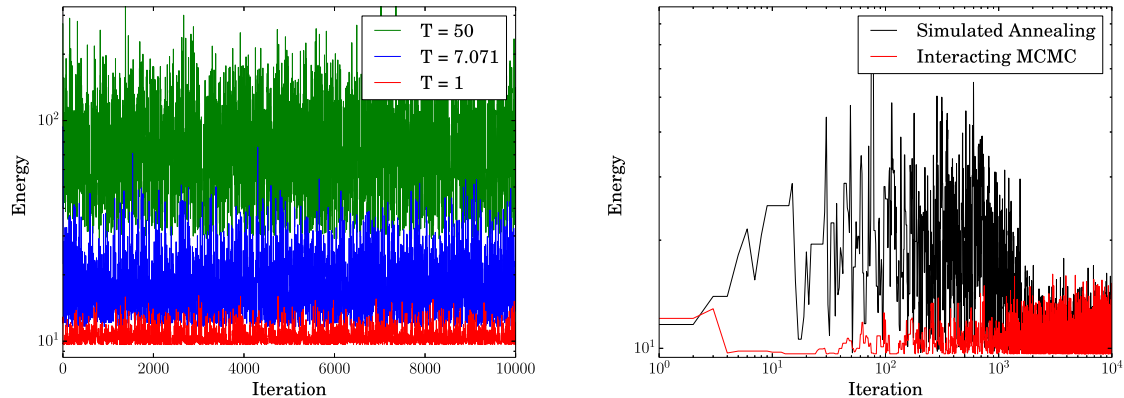
and the sought *a posteriori* distribution is

$$\mathbf{b} \mapsto P(\mathbf{b}|\mathbf{d}^{\text{obs}}) = e^{-E(\mathbf{b})} = P(\mathbf{b})e^{-\frac{1}{2}\left|\frac{\mathcal{R}(\mathbf{b})}{\sigma_d}\right|^2}. \tag{33}$$

We have sampled it perfectly by an accept/reject method using 100 000 samples and used a kernel density estimator (KDE, see Silverman 1986) to represent the resulting density. It is shown in Fig. 2 (right). We run two algorithms to sample this distribution: the IMCMC and an SA. Their features are shown in Tables 1 and 2. Having a look at these tables, it is important to note that the temperature sequences and the amplitudes of the proposed moves have been chosen in the same way for the two algorithms. In other words, when the SA chain is at the same

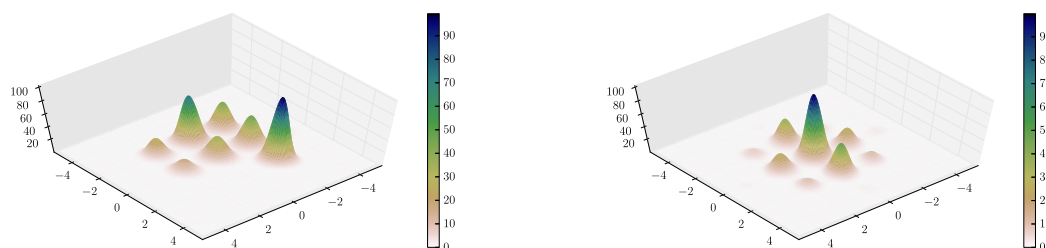
**Table 2.** Features of the run for the simulated annealing.

Evolution of $T$ with the iteration $n$	Geometric sequence. A decrease from $T_{\max} = 50$ to $T = 1$ is achieved in 2000 iterations the temperature being updated every 50 iterations: $T_n = 50^{-\frac{50}{2000}(\text{floor}[\frac{2000-n}{50}] + 1)}$
Sampler chosen $q(\mathbf{b}^* \mathbf{b})$ :	SIMH
$q_T(\mathbf{b}^* \mathbf{b})$ :	$\mathcal{N}_2(\mathbf{b}, \sigma_{\mathbf{b}}^2)$
$\sigma_{\mathbf{b}} = (\sigma_x, \sigma_y)$ :	Linear decrease between $\frac{5.12+5.12}{4}$ at $T_{\max}$ and $\frac{5.12+5.12}{50}$ at $T = 1$ Particularly: $\sigma_x(T = \sqrt{50}) = \frac{5.12+5.12}{20.62}$

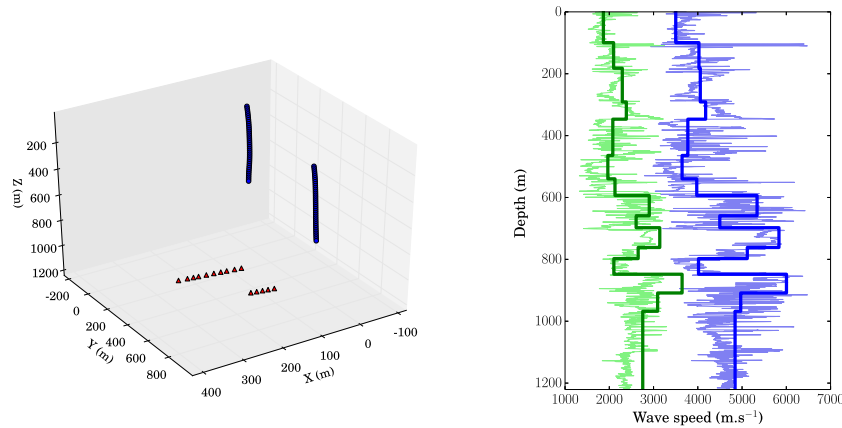


**Figure 3.** Left: energies of the interacting chains as a function of the iteration. Right: convergence in term of energy for the lowest temperature chain of interacting Markov chains algorithm (red) and the SA (black). Low-energy (high probability) regions are reached very quickly with interacting Markov chains algorithm.

temperature as one of the interacting chains they behave alike (they can propose the same moves). The only difference being their ability to interact or not. The three interacting Markov chains ran during  $N = 10\,000$  iterations whereas the SA ran during 30 000 iterations after the 2000 burn-in iterations. The energies of states  $(x, y)$  generated by the IMCMC are shown in Fig. 3 (left). We observe a good overlapping of the energy domain explored by the three chains ensuring a satisfactory exchange rate between the chains. Fig. 3 (right) is a comparison of the energy evolution for the 10 000 first samples of each algorithm, it demonstrates clearly the higher speed of convergence of the IMCMC algorithm compared to the SA. No burn-in is required with IMCMC, whereas more than 1000 iterations are needed by SA to converge. The posterior distributions simulated by each algorithm are then constructed from the generated samples using a KDE. The results shown in Fig. 4 have to be compared to the theoretical density drawn in Fig. 2 (right). At  $T = 1$ , the standard deviation of the sampler  $\sigma_x = \sigma_y = \frac{5.12+5.12}{50}$  gives a chance to the chain to go from one mode to another. Nevertheless, the SA algorithm ends stuck in one or several modes and does not succeed in identifying the central mode as the principal one (Fig. 4, left). The results presented here are the best we managed to obtain in term of sampling with the SA method (it has detected seven of the nine modes) and a standard deviation of the sampler at  $T = 1$  similar as the one used by the lowest temperature chain of the IMCMC algorithm. It seems important to highlight the high variability of the results obtained by SA as the chain ends stuck in different modes for each different run. The IMCMC detected the nine principal modes (Fig. 4, right) in 100 per cent of the runs performed, even with far less iterations (down to 2000). The consequences are huge in term of computational cost. Most important, the estimations of the uncertainties on the values of the parameters  $\mathbf{b} = (x, y)$  would be much more reliable with the IMCMC. The algorithm gives thus a robust way to evaluate the uncertainties associated with the model parameters and it is interesting because of its speed of convergence and its stability.



**Figure 4.** Left: KDE of the posterior distribution simulated by simulated annealing. Right: KDE of the posterior distribution simulated by the interacting Markov chains. To be compared with Fig. 2.



**Figure 5.** This figure illustrates the *a priori* knowledge on the area studied. Left: geometrical setting for the first arrival times recorded. The circles represent the receivers and the triangles are the sources. Right: light blue and green curves are, respectively,  $P$ - and  $S$ -wave velocity profiles (sonic logs) acquired on the site. These curves are noted  $V_{\log}^P(z)$  and  $V_{\log}^S(z)$  in the text. Dark blue and green curves are, respectively,  $V_P$  and  $V_S$  vertical profiles of the 1-D reference calibrated velocity model.

### 3 REAL DATA CASE STUDY

In this last section, we present the results of our inversion approach applied to a real data set recorded in the context of hydraulic fracturing. We have two types of measurements available: a sonic log which informs on the wave propagation velocity of the medium at a given frequency and some traveltimes from perforation shots. We first build the prior model from the sonic logs. We then update this prior knowledge on the model parameters by estimating the posterior distribution of the model parameters knowing the traveltimes data.

#### 3.1 Setting of the problem

The acquisition geometry is represented in Fig. 5 (left) in a relative Cartesian coordinate system. The monitoring network consisted of two wireline arrays of 40 3C-receivers that were deployed in two different vertical wells (circles in Fig. 5, left). The depth range of these arrays is between about 50 and 750 m. Fifteen perforation shots performed along two horizontal wells (triangles in Fig. 5, left) at a depth of 1050 m have been considered. For each of the 15 perforation shots, almost all 80 first  $P$ -wave and approximately 40  $S$ -wave arrivals could be picked. These  $P$  and  $S$  first arrival traveltimes constitute our data set for the inversion. Following the notations introduced in Section 1, we write:

$$\mathbf{d}^{\text{obs}} = (d_p^1, \dots, d_p^{15 \times 80}, d_s^1, \dots, d_s^{15 \times 40}) \in \mathcal{D} = \mathbb{R}^{\times 15 \times (80+40)}. \quad (34)$$

Over the entire data set, we estimate the average  $P$ - and  $S$ -wave picking uncertainties to be around 2.5 and 5 ms, respectively. Compressive and shear velocity logs represented in Fig. 5 (right, light curves) have also been acquired in one of the wells using a sonic wireline tool that emits a signal with a central frequency around 3 kHz. They will be noted  $V_{\log}^P(z)$  and  $V_{\log}^S(z)$ , respectively. The measurements have been made every half-foot. These logs are obtained at a very high frequency and, assuming that the medium is dispersive, probably do not represent the subsurface as seen by the seismic waves in the frequency band of the microseismic events. Indeed disagreements between seismic data and sonic log data are often recognized (Sheriff & Geldart 1995). However, they can be considered as *a priori* information for an inversion. The sonic logs are generally used to build a layered velocity model, which is then calibrated with a tomography using perforation shots from known locations (e.g. Bardainne & Gaucher 2010). In this example, a velocity model was derived from the logs and from *a priori* geological information (subhorizontal layers). The resulting 1-D calibrated velocity model consisted of a blocky horizontal model with 15 layers (see dark curves in Figs 5, right).  $P$  velocities range from 3500 to 6000  $\text{ms}^{-1}$  and  $V_P/V_S$  ratio varies from 1.6 to 1.95.

For this example, a layered geometrical parametrization is not necessarily appropriate: how many layers do we need to describe the subsurface? Is it physically relevant to work with layers and not with a smoother representation? This parametrization is by itself a strong *a priori* assumption since the number of layers is fixed. Most important, this representation is not optimum in term of dimension since  $2N$  parameters ( $V_P$  and  $V_S$  in each layer) are needed to represent  $N$  static horizontal layers. If the interface depths are also considered as unknowns,  $3N - 1$  parameters are required. In a Bayesian context, the computational cost depends on the number of model parameters to retrieve, the dimension of the problem should thus be just sufficient for the resolution needed. We propose here to use a parametrization based on an optimally truncated decomposition of the velocity logs on wavelet functions basis. This allows to describe complex velocity models with a restricted number of parameters.

### 3.2 A self-parametrizing approach using wavelet theory

#### 3.2.1 The wavelet decomposition, a multiscale analysis tool

Since the 1980s, wavelet theory has developed rapidly and has become an important tool for multiscale signal analysis (see Burrus *et al.* 1997; Mallat 1999). It has been applied successfully in many different fields, such as image compression and processing, optimum signal smoothing or speech analysis. It has also been used several times for model parametrization [Chiao & Kuo (2001), Bennett & Malinverno (2005), Loris *et al.* (2007), Tikhotsky & Achauer (2008), Simons *et al.* (2011) or more recently Hawkins & Sambridge (2015)]. The original goal of wavelet theory is to build orthonormal basis of the Hilbert space  $L^2(\mathbb{R})$  of square integrable functions of  $\mathbb{R}$ , expressed as:

$$\{\psi_{jk}(x) : (j, k) \in \mathbb{Z}^2\} = \{2^{\frac{j}{2}} \psi(2^j x - k) : (j, k) \in \mathbb{Z}^2\} \quad (35)$$

that is, composed of dyadic translations and dilations of a unique function, preferably regular and localized, called the mother wavelet  $\psi$ . This basis is orthonormal:

$$\forall j, k, l, m \quad \langle \psi_{jk}, \psi_{lm} \rangle = \delta_{jl} \delta_{km} \quad (36)$$

where  $\delta_{jl}$  is the Kronecker delta and  $\langle f, g \rangle$  is the standard inner product  $\langle f, g \rangle = \int_{-\infty}^{\infty} f(x) \overline{g(x)} dx$  of  $L^2(\mathbb{R})$ . Every function  $h \in L^2(\mathbb{R})$  can be expanded in a unique way in this basis as:

$$h(x) = \sum_{j=-\infty}^{\infty} \sum_{k=-\infty}^{\infty} a_{jk} \psi_{jk}(x) \quad (37)$$

with convergence of the series understood to be convergent in norm. Such a representation of a function  $h$  is known as a wavelet series. In the discrete case  $h = (h(x_j))_{j=1, \dots, N} \in \mathbb{R}^N$  can be fully described by  $N$  coefficients:

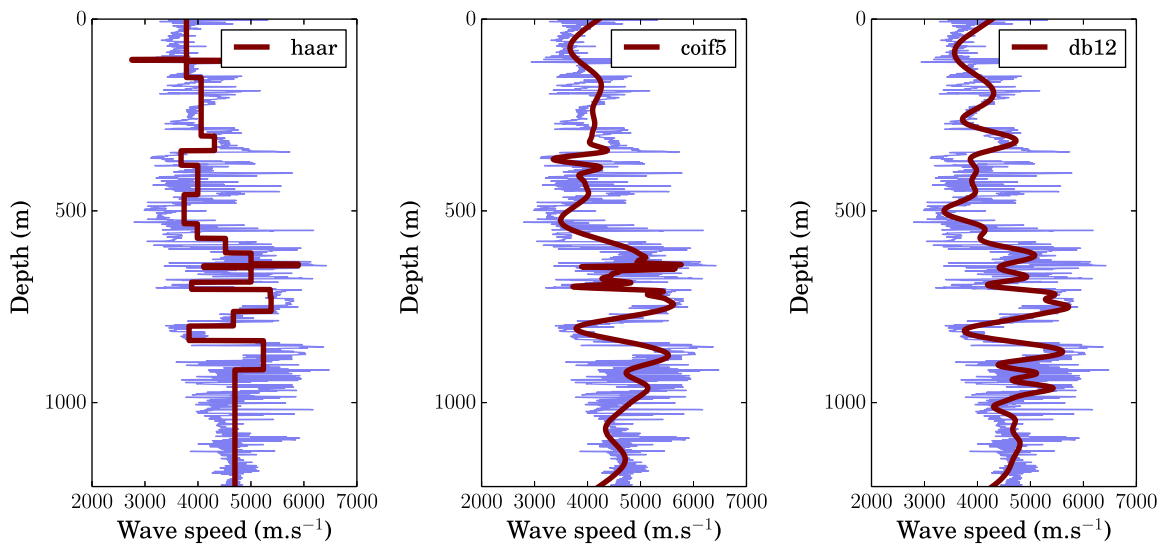
$$h(x_j) = \sum_{i=1}^N a_i \psi_i(x_j). \quad (38)$$

The first coefficients describe the global shape of the function while the last ones characterize its small-scale details. From (38), by keeping the  $M$  most energetic coefficients (i.e. the biggest  $|a_i|^2$ ) and setting the other coefficients to 0, we construct the following approximation  $h_M$  of  $h$ :

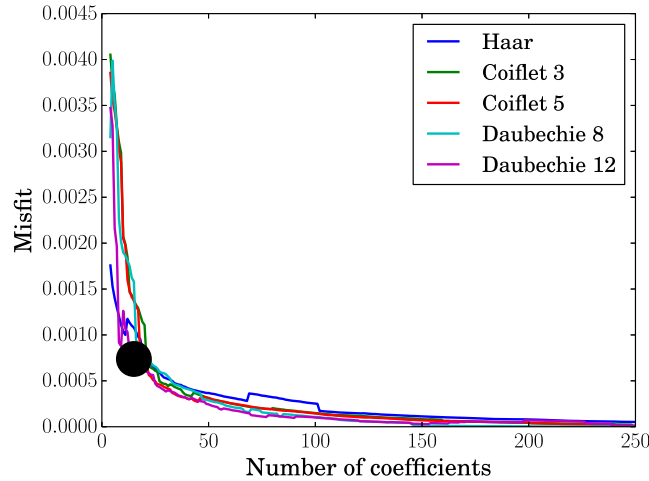
$$h_M(x_j) = \sum_{i=1}^M c_{i'} \psi_{i'}(x_j) \quad (39)$$

where  $i'$  is the index of the  $i$ th most energetic coefficient (see Jansen 2012). It is interesting to note (and very convenient) that  $h_M$  is an optimal approximation in the mean-square sense. Moreover, this thresholding does not cause any phase shift nor ringing as the Gibbs phenomenon for filtering in frequency domain.

Fig. 6 shows examples of filtering of the  $P$ -wave velocity log (blue curve) using three different wavelets and 15 coefficients (red curves). The mother wavelet determines the global appearance of the profile and the first coefficients mainly control its lowest frequency variations.



**Figure 6.** Left: filtering of the  $P$ -wave velocity log using a Haar wavelet and 15 coefficients. Middle: filtering of the  $P$ -wave velocity log using a Coiflet wavelet with five vanishing moments and 15 coefficients. Right: filtering of the  $P$ -wave velocity log using a Daubechie wavelet with 12 vanishing moments and 15 coefficients. This illustrates the global aspect of the profiles based on the different wavelets. Coiflet 3 is very similar to Coiflet 5 and Daubechie 8 is similar to Daubechie 12.



**Figure 7.** Least-squares misfit in term of traveltimes between the raw  $P$ -wave velocity  $\log V_{\log}^P(z)$  and the filtered logs  $V_{\log, \psi}^{P, M}(z)$  as a function of the number  $M$  of coefficients kept and for the set of mother wavelets  $\psi$  considered in this paper. The black dot spots the selected configuration, namely: 15 coefficients.

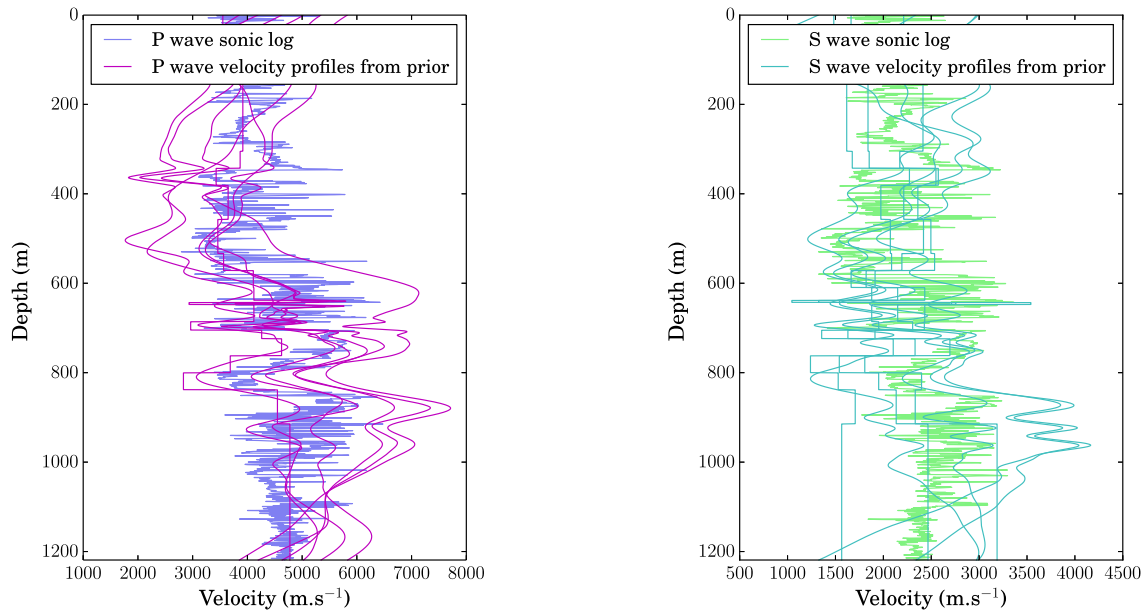
Rather complex features of the velocity log are described with 15 coefficients. With a geometrical-layered parametrization, 15 parameters represent only eight layers, a very small number compared to the layer number obtained with the Haar wavelet parametrization (Fig. 6, left).

### 3.2.2 The parametrization chosen

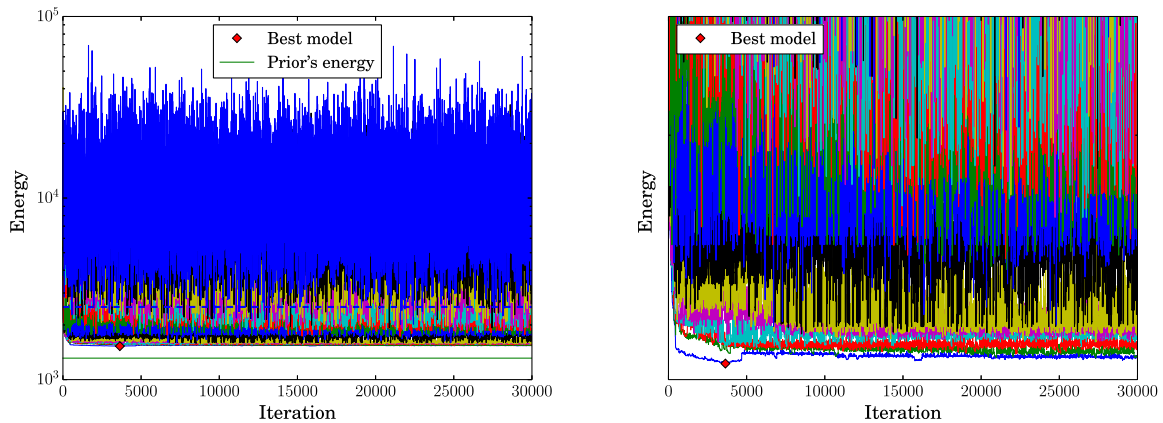
The velocity profile parametrization is based on the filtering described above. Indeed, a velocity profile can be constructed from the choice of a mother wavelet  $\psi$  and a set  $\mathbf{c}$  of coefficients  $c_i$  through the relation (39). The idea is to use the mother wavelet's type and its coefficients as the parameters  $\mathbf{b}$  of the inversion. Letting the wavelet type be part of the inversion is a similar approach to the Bayesian Model averaging (Hoeting *et al.* 1999). The mother wavelets belong to a discrete space  $\{\psi_1, \psi_2, \psi_3, \dots\}$ . The number of coefficients to keep have still to be determined. We adopt the following approach: from the raw  $P$ -wave velocity  $\log V_{\log}^P(z)$  represented in light colours in Fig. 5 (right), we generate synthetic first arrival times  $\mathbf{d}_0$  with the eikonal solver of Noble *et al.* (2014) for an arbitrary shot located near the bottom of the profile (at  $z = 900$  m) and recorded at synthetic receivers located close to the surface ( $z = 50$  m).  $V_{\log}^P(z)$  is then iteratively filtered with a truncated wavelet decomposition. We note  $V_{\log, M}^P(z)$  the filtered  $P$ -wave sonic log reconstructed from  $M$  coefficients. For each truncated profile, first arrival times  $\mathbf{d}_M$  are computed with the eikonal solver and a misfit  $\Delta_M = \|\mathbf{d}_0 - \mathbf{d}_M\|^2$  is calculated. In Fig. 7, we represent this misfit with respect to the number of coefficients kept to describe the velocity model for five different mother wavelets. The best trade-off between a minimum number of coefficients and a resolution sufficient to explain the synthetic data  $\mathbf{d}_0$  seems to be a filtering with  $M \approx 15$  coefficients regardless of the mother wavelet type. Hence, for the inversion, a state of the Markov chain will contain the wavelet type plus 30 parameters: 15 for the  $P$ -wave velocity and 15 for  $S$ -wave velocity. We note:  $\mathbf{b} = (\psi, c_1^P, \dots, c_{15}^P, c_1^S, \dots, c_{15}^S) \equiv (\psi, c_1, \dots, c_{30})$  and  $\mathcal{M} \subset \mathbb{N} \times \mathbb{R}^{30}$ . The *a priori* knowledge on the values of the parameters  $\mathbf{b}$  can be represented by a 31-D probability distribution:  $P(\mathbf{b})$ . For this problem we choose, *a priori*, five different wavelet types: Haar wavelet, Daubechie wavelets with 8 and 12 vanishing moments and Coiflet wavelets with 3 and 5 vanishing moments, these mother wavelets having, *a priori*, the same chance to be drawn. These different wavelets provide velocity profiles with distinct aspect. The reader can make up his mind by looking at Fig. 6 which illustrates the appearance of the profiles based on the different wavelets. Coiflet 3 is very similar to Coiflet 5 and Daubechie 8 is similar to Daubechie 12. We also suppose a uniform *a priori* probability distribution on each coefficient  $c_j$ , that is, for each wavelet type  $\psi_i$  and each coefficient  $c_j$  we define an interval  $[\inf_{\psi_i}(c_j), \sup_{\psi_i}(c_j)]$  on which  $c_j$  can take any value with the same probability. Out of this domain any value has a probability 0. Hence, the *a priori* probability is:

$$P(\mathbf{b}) = \begin{cases} \prod_{i=1}^5 \prod_{j=1}^{30} \frac{1}{\sup_{\psi_i}(c_j) - \inf_{\psi_i}(c_j)} & \text{if } \forall k \quad c_k \in [\inf_{\psi}(c_k), \sup_{\psi}(c_k)] \\ 0 & \text{else.} \end{cases} \quad (40)$$

We note  $V_{\log, \psi}^{P, 15}(z)$  and  $V_{\log, \psi}^{S, 15}(z)$  the  $P$ - and  $S$ -wave velocity logs filtered by truncated wavelet decomposition using wavelet  $\psi$  and 15 coefficients. Three examples of filtering are shown in Fig. 6. Filtered sonic logs are fully described by their 15 wavelet coefficients that we will note:  $(l_{1, \psi}^P, \dots, l_{15, \psi}^P)$  and  $(l_{1, \psi}^S, \dots, l_{15, \psi}^S)$ . In order to generate models not too far from  $V_{\log, \psi}^{P, 15}(z)$  and  $V_{\log, \psi}^{S, 15}(z)$ , we have chosen to sample around the  $l_{i, \psi}^P$  and  $l_{i, \psi}^S$ . Hence, we have chosen  $\inf_{\psi_i}(c_j^P) = 0.6l_{j, \psi_i}^P$ ,  $\sup_{\psi_i}(c_j^P) = 1.4l_{j, \psi_i}^P$ ,  $\inf_{\psi_i}(c_j^S) = 0.6l_{j, \psi_i}^S$  and  $\sup_{\psi_i}(c_j^S) = 1.4l_{j, \psi_i}^S$ . To summarize, the parameters of the inversion are perturbed wavelet coefficients of the filtered logs. This strategy allows us to include the *a priori* information brought by the seismic logs while keeping a relatively wide model space. A sample of  $P$ - and  $S$ -wave velocity profiles generated with this prior can be seen in Fig. 8 together with the  $P$ - and  $S$ -wave measured velocity logs.



**Figure 8.** Samples of  $P$ - and  $S$ -wave velocity profiles from prior shown together with the sonic logs. It illustrates the diversity of profiles reachable *a priori*.



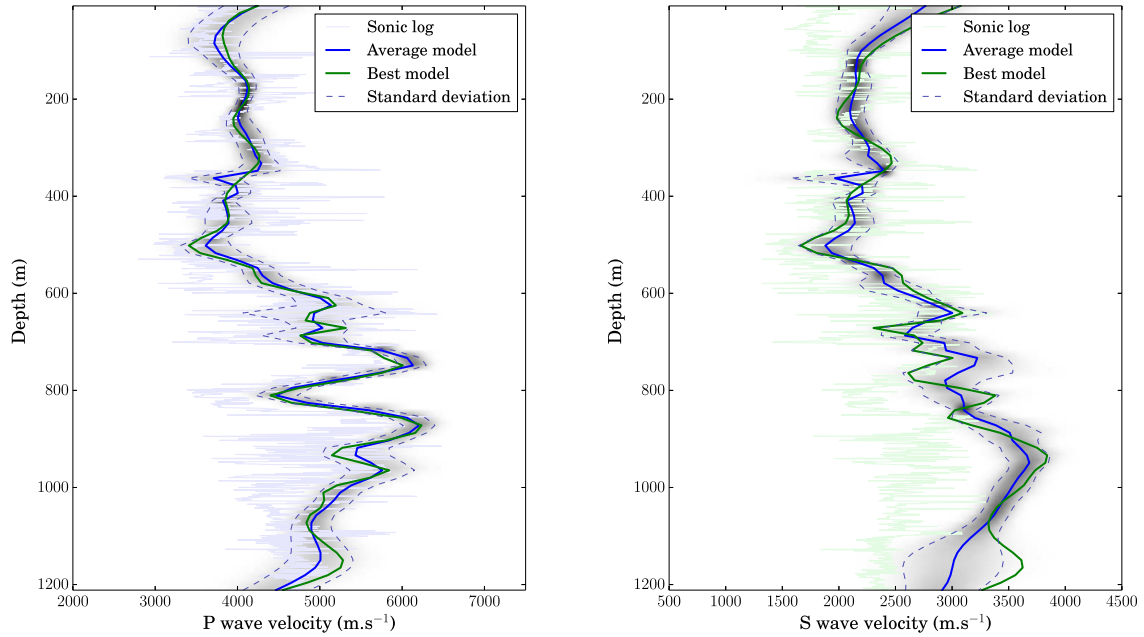
**Figure 9.** Energies of the interacting Markov chains as a function of the iteration for the real case study. The red diamond spots the best model generated. The dashed blue line represents the threshold below which the models generated are acceptable (one standard deviation of difference between synthetic and real data). The right figure is a zoom on the lowest temperature chains. These figures illustrate the efficiency of the algorithm especially showing the large number of acceptable models sampled from the first iterations.

### 3.3 Results and discussion

For this traveltimes tomography, the forward problem is the 3-D computation of the theoretical traveltimes  $\mathbf{d}$  that a configuration of the velocity model produces. The velocity model is discretized on a rectangular grid ( $n_x = 35 \times n_y = 35 \times n_z = 80$ ) and a new eikonal solver (Noble *et al.* 2014) is used to calculate, for each perforation shot (triangles in Fig. 5, left), the traveltimes at the receiver positions (circles in Fig. 5, left). The forward problem is the most resource demanding part of the algorithm. A parallel network of 30 CPUs (Intel®Xeon®CPU E5-2407 @ 2.20 GHz) has been used, at each iteration, to run the eikonal solver for both  $P$  and  $S$  waves for the 15 shots. As seen in Section 1 (eqs (8) and (16)) an energy is computed afterward from a least-squares misfit between  $\mathbf{d}$  and  $\mathbf{d}^{\text{obs}}$ . Fifteen interacting Markov chains ran during  $N = 30\,000$  iterations. This number ensured a good overlapping of the histograms of the tempered energies and allows them to satisfactorily exchange their states. The whole tomography lasted  $\sim 8$  hr. We represent the energies of the different chains with respect to the iteration number in Fig. 9. In that figure, the dashed blue line represents the threshold below which the models generated are acceptable (one standard deviation of difference between synthetic and real data). It illustrates the high number of acceptable models found. It has to be noted that the best model tested has been found rather quickly but not reached again during the simulation indicating that, providing the algorithm has finally converged, it probably corresponds to a thin mode which did not need to be sampled to give a good overview of the *a posteriori* probability distribution. The algorithm converges in less than 5000 iterations for the features indicated in Table 3. We first estimate the  $P$  and  $S$  best models (green curves in Fig. 10) and average models (dark blue curves in Fig. 10) applying the formula (29). In this case, the obtained profiles seem coherent with the sonic logs. In order to estimate the velocity model uncertainties, we represent marginal posterior probability density function for  $V_P$  and  $V_S$  velocity profiles. The resulting density plots are shown in Fig. 10 with darker colours representing higher probability.

**Table 3.** Features of the run for the interacting MCMC applied to the real case study. We note  $\sigma_{\mathbf{e}} = (\sigma_{c_1}, \dots, \sigma_{c_{30}})$  and  $\sigma_{\mathbf{e}}^2 = (\sigma_{c_1}^2, \dots, \sigma_{c_{30}}^2)$ .

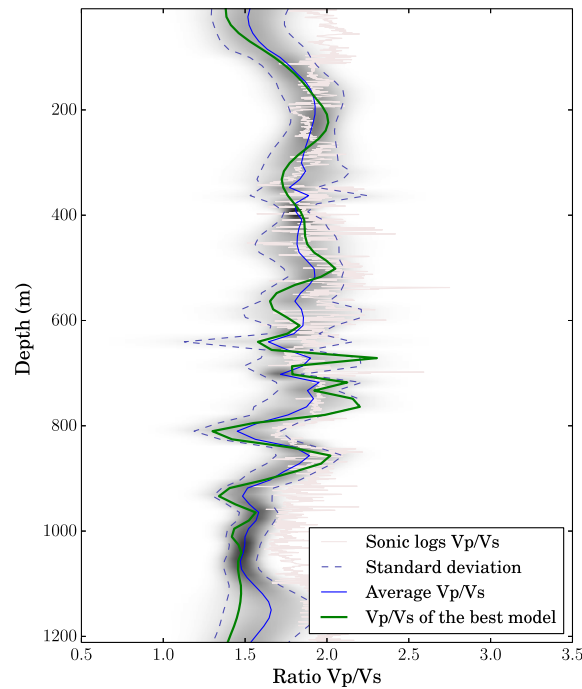
Number $K$ of chains	15
Probability $\theta$ of proposing a swap	25 per cent
Sequence $(T_k)_k$ of temperatures	Geometric sequence between $T_{14} = T_{\max} = 10\,000$ and $T_0 = 1$ : $\forall k \in \{0, 1, \dots, 14\} \quad T_k = T_{\max}^{\frac{k}{14}}$
Number of parameters changed at each iteration (including wavelet type)	1 if $T = 1$ 31 if $T = T_{\max}$ 4 otherwise
Samplers at each temperature	SIMH if $T \neq T_{\max}$ IMH otherwise
$q_l(\mathbf{b}^*   \mathbf{b})$	$\mathcal{N}_{30}(\mathbf{c}, \mathbb{I}_{30} \sigma_{\mathbf{e}}^2)$ if $T \neq T_{\max}$ $P(\mathbf{b}^*)$ otherwise
$\forall j \quad \sigma_{c_j}$	If $T \neq T_{\max} \quad \sigma_{c_j} = \frac{\inf_{\psi_i(c_j)} - \inf_{\psi_l(c_j)}}{5}$



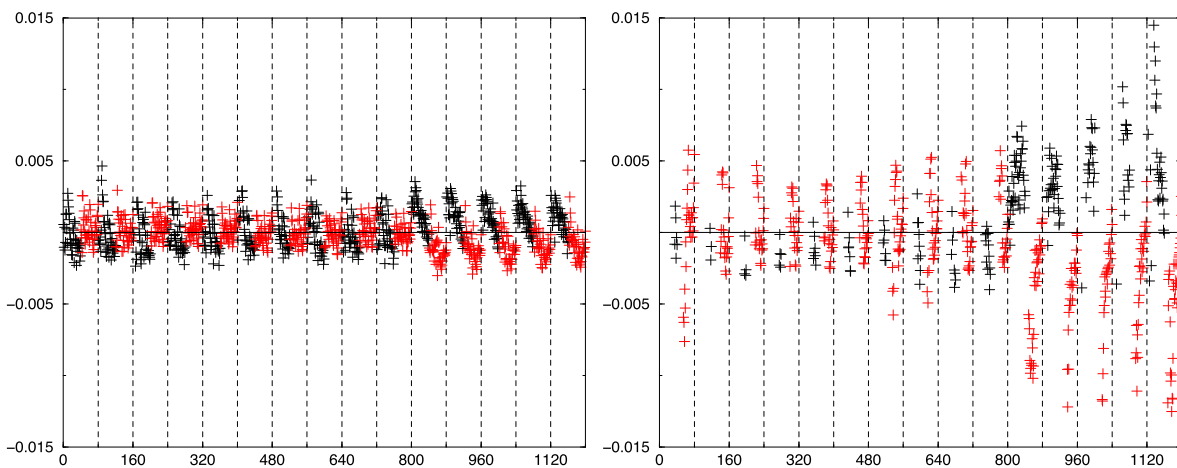
**Figure 10.** Left: results of the tomography for  $P$ -wave velocities. The figure shows the average model (blue) and *a posteriori* standard deviation (dashed blue)—calculated with the formulae (29) and (30)—together with the best model (green) and the measured sonic log (bright blue). The resulting density plots are also shown with darker colours representing higher probability. Right: same for  $S$ -wave velocities. The most valuable information obtained are the standard deviations which tell us about the knowledge on the area obtained by the inversion.

We also estimate the standard deviation (dashed blue lines in Fig. 10) from all the samples generated by the chains. This figure illustrates that between 200 and 750 m (in particular for  $S$  waves), the velocity model is well constrained and thus reliable. This is in agreement with the ray coverage linked to the source and receiver positions since both receiver arrays cover this depth range. We also represent average  $V_P/V_S$  and associated standard deviation in Fig. 11, this allows us to check the validity of our approach since the obtained ratios for the average model are in an acceptable range, between about 1.5 and 2. Differences between the confidence intervals and the sonic logs between 750 and 1000 m are fairly notable. This discrepancy seems higher for the  $S$  velocities. A very plausible explanation in this acquisition context could be the presence of mycolour dispersive and/or very strong anisotropic layers not accounted for in our isotropic eikonal solver.

We finally compare the residuals obtained with the reference calibrated velocity model (dark curves in Fig. 5, right) and with the new average model (blue curves in Fig. 10). Black crosses in Fig. 12 represent residuals at the receivers of the first antenna (at  $Y = -200$  m), red crosses at the receivers of the second antenna (at  $Y = 400$  m) and the 15 shots are separated by the vertical dashed lines. For this reference velocity model, most of the residuals are smaller than the standard deviation of the picking uncertainty (0.0025 s for  $P$  waves and 0.005 s for  $S$  waves). However, the residuals are not at random for the mycolour last five shots since they are distributed with almost all the positive for the first antenna and almost all the negative for the second antenna. This particular distribution is confirmed by the  $P$  (Fig. 14) and  $S$  (Fig. 15) residual histograms sorted by antenna and by shot line (upper figures). This specific distribution is not retrieved with the average model based on a wavelet parametrization since the  $P$  and  $S$  residuals for the last five shots are centred on zero for both antenna (Figs 13–15). The best model gives similar residuals as those obtained with the average model with a good distribution and most of the residuals smaller than the standard deviations of the picking uncertainties. This comparison illustrates the gain in accuracy provided by the wavelet-based parametrization compared to the layer-based parametrization with much better residual distribution. Moreover, the classical parametrization with layers is dependent of the operator who has to determine an initial velocity model from the sonic logs. With the wavelet parametrization, no operator is involved and the parametrization is only driven by the sonic log data.



**Figure 11.** In green, the ratio  $V_P(z)/V_S(z)$  for the best model and in blue, the average  $V_P(z)/V_S(z)$ . The standard deviation is drawn as two dashed blue lines and the ratio  $V_{\log}^P(z)/V_{\log}^S(z)$  for the sonic logs is drawn in bright red. Darker colours of the density plot represent higher probability. The obtained ratios for the average model are in an acceptable range, between about 1.5 and 2.



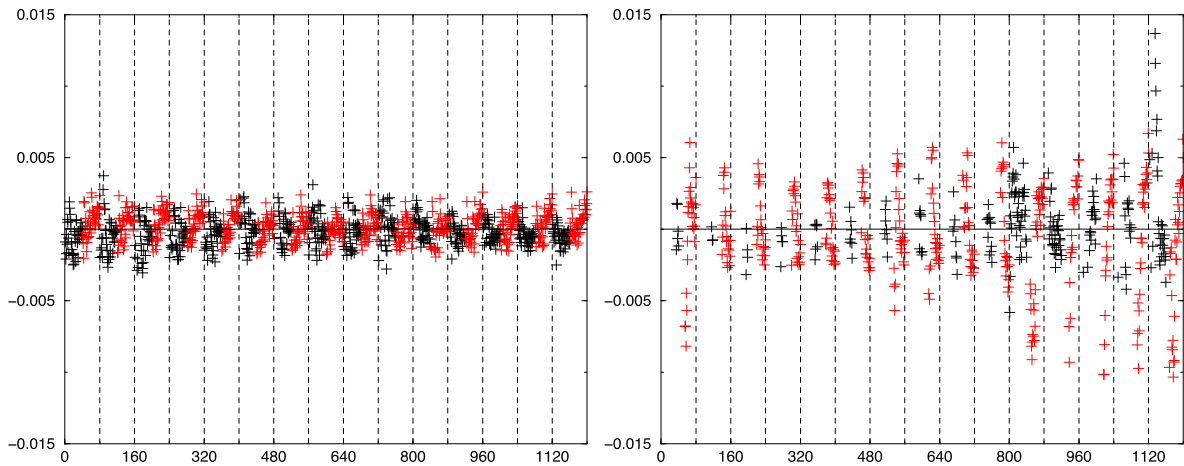
**Figure 12.** Left:  $P$ -wave arrival time residuals for the reference calibrated velocity model of Fig. 5 (right, dark blue curve). The 15 perforation shots are separated by vertical dashed lines. Black crosses represent residuals for the first receiver antenna and red crosses for the second receiver antenna. Right: same for  $S$ -wave arrival times. Residuals are not at random for the last five shots since they are almost all positive (respectively, negative) for the first (respectively, second) antenna.

This application of our new efficient Bayesian tomography algorithm to real data with an adaptive model parametrization based on wavelet coefficients is very convincing. Indeed, this first attempt gives, with only 30 parameters, elaborated final  $P$  and  $S$  velocity models with reliable associated uncertainty estimations indicating regions where the velocities are well constrained and thus reliable and also the regions where the velocities are poorly constrained.

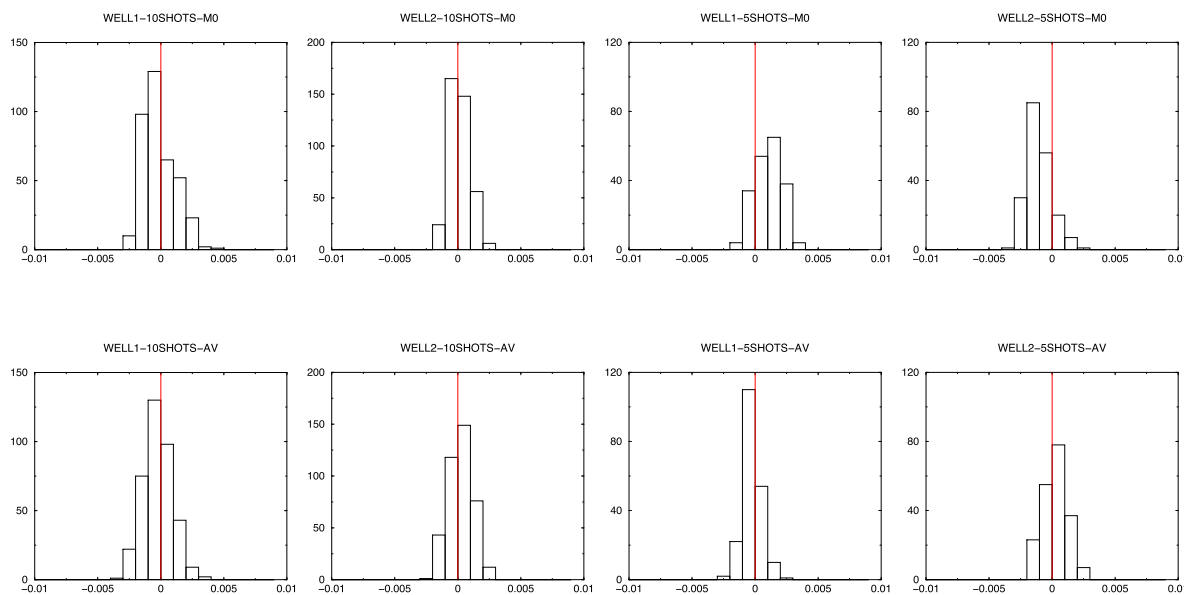
#### 4 CONCLUSIONS

In this paper, we have described and applied a still relatively unknown algorithm for Bayesian inversion. The main idea is to sample the *a posteriori* distribution using several interacting Markov chains evolving at different scales. The exchanges between the chains allow a precise sampling of the high probability zones of the model space while avoiding the chains becoming stuck in a local probability maxima. The algorithm has first been tested on a synthetic multimodal 2-D distribution where it has shown its robustness and efficiency for sampling compared to an SA approach. No burn-in period was required to be compared to more than 1000 iterations needed by SA to converge. It





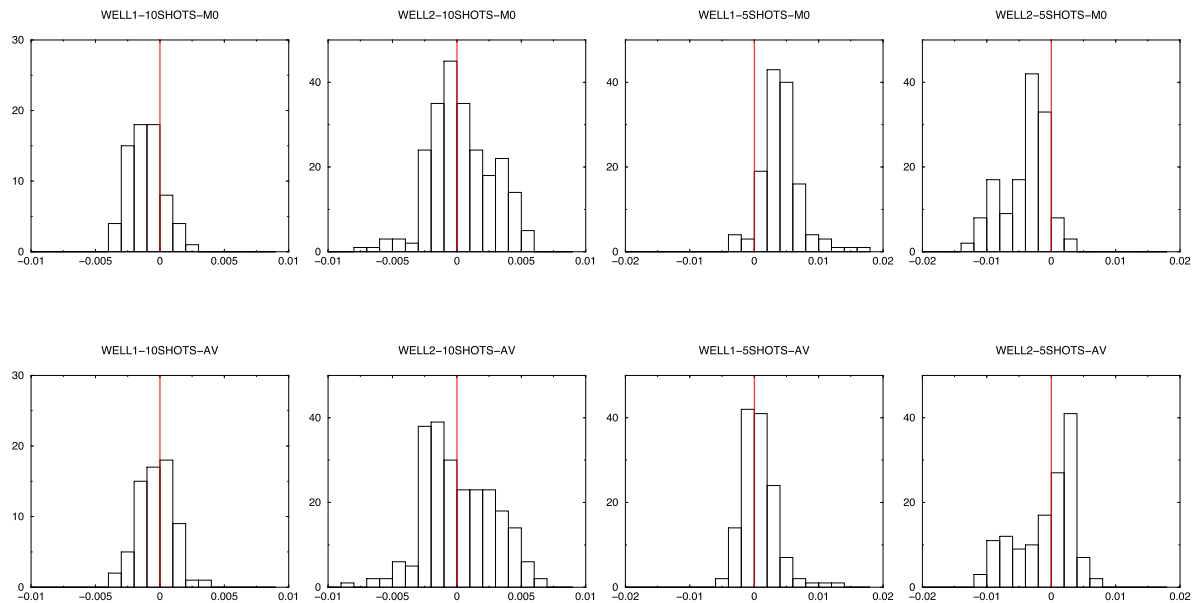
**Figure 13.** Residuals for the new average model (blue curve in Fig. 10). Information is plotted in the same way as Fig. 12. Good distribution of residuals that are centred on 0 for the last five shots.



**Figure 14.** Top figures: histograms of the  $P$ -wave arrival time residuals for the reference calibrated velocity model of Fig. 5 (right, dark blue curve). From left to right: histogram for the 10 shots (at  $Y = 100$  m) recorded at the first receiver antenna (at  $Y = -200$  m), for the 10 shots (at  $Y = 100$  m) recorded at the second receiver antenna (at  $Y = 400$  m), for the five shots (at  $Y = 400$  m) recorded at the first receiver antenna (at  $Y = -200$  m) and for the five shots (at  $Y = 400$  m) recorded at the second receiver antenna (at  $Y = 400$  m). Bottom: same for the new average model (blue curve in Fig. 10).

has then been implemented for a 3-D seismic first arrival traveltimes tomography on a real data example in an efficient way. Agostinetti *et al.* (2015) have recently presented a 3-D probabilistic local earthquake tomography with a Voronoi cell parametrization. They report very long computation times with several independent Markov chains. We could expect a great decrease in running time by using interacting instead of independent MCMC. Moreover, in the examples presented in this paper, only the computation of the forward problem has been performed on several processors. All the chains have run together on a single CPU and the algorithm already gave good results in a very efficient way. The evolution of each chain could also be parallelized on a CPU cluster, enhancing even more its performances. Communications between processors would be limited since only one energy value has to be sent by each processor at each iteration.

On the real-data example typical of hydraulic fracturing context, instead of using a geometrical parametrization with a layered velocity model built from the sonic logs, we have developed an adaptive wavelet-based model parametrization. This allowed us to reduce the dimension of the problem while including properly on the inversion the *a priori* information provided by sonic logs, without involving any operator. Our stochastic tomography of perforation shots provided reliable estimates of the uncertainties. This is important, in particular in the microseismic context where the final velocity model is further used to locate seismic events. With a standard linearized inversion method or global optimization approach, the errors on the resulting velocity model cannot be properly taken into account in the seismic event location. On the contrary, the samples of velocity model provided by our stochastic approach could be used to propagate velocity model uncertainties to the seismic event location uncertainties (Gesret *et al.* 2015). The 1-D velocity model assumption seems valid for this case study but for more complex media, such as generally faced in seismic refraction, the parametrization methodology should be extended using 2-D or



**Figure 15.** Histograms of the  $S$ -wave arrival time residuals. Information is plotted in the same way as Fig. 14.

3-D wavelet bases. We noted that anisotropy could explain the discrepancy between well logs and tomography results below 750 m, it would thus be interesting to do a future inversion including anisotropic parameters.

In this paper, the number of wavelet coefficients has been fixed. Future development should be made on the algorithm to allow transdimensional inversions such as proposed by Malinverno (2002) or Bodin *et al.* (2012) with our adaptive wavelet-based parametrization. For example, the reversible jump algorithm of Green (1995) could be adapted to multichain sampling. The number of wavelet coefficients to be kept would hence become part of the inversion, leading to a resolution adapted to the observed data. The algorithm would then adapt to the complexity of the medium and become an even more efficient approach to tomography.

Our software package is collaborative and available open-source at: <https://github.com/bottero/IMCMCrun>, it includes all the tools necessary to reproduce the results presented in this paper.

## ACKNOWLEDGEMENTS

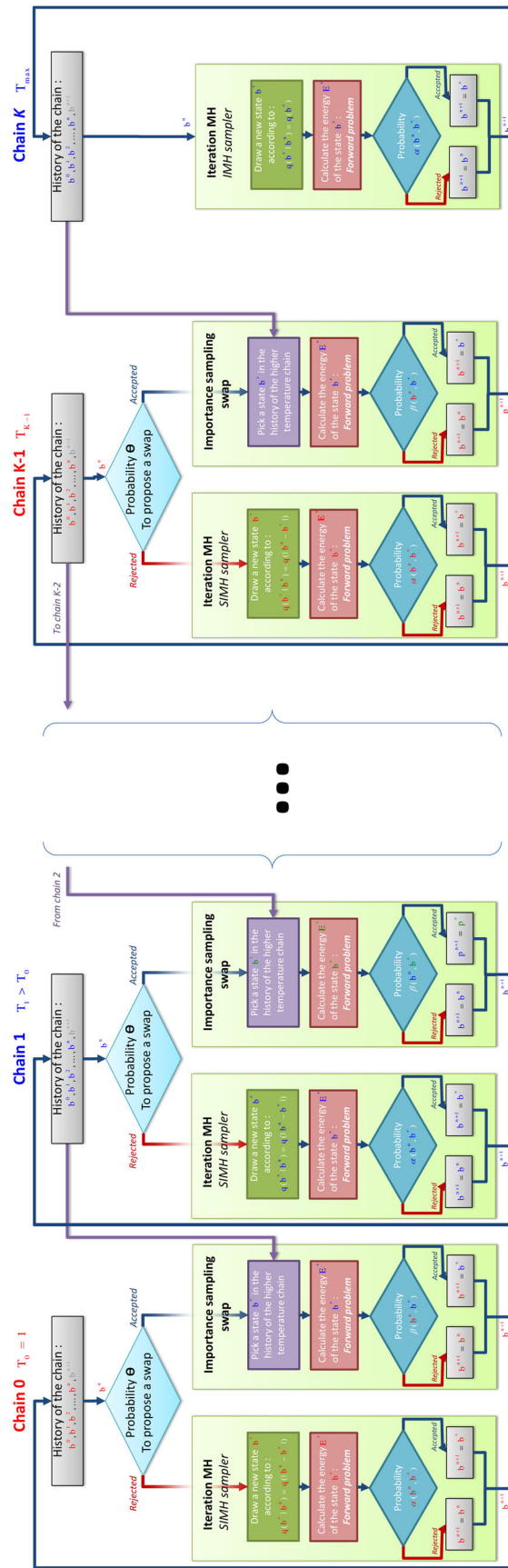
We are grateful to N. Desassis and F. Ors (Mines ParisTech, Centre de Géosciences, Fontainebleau, France) for fruitful discussions. The internship of AB has been made possible by IPGP and ENS Cachan, France. We also thank Michael Afanasiev and an anonymous reviewer for their useful comments that improved the manuscript.

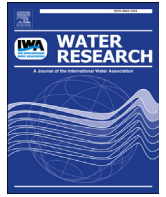
## REFERENCES

- Agostinetti, N.P., Giacomuzzi, G. & Malinverno, A., 2015. Local three-dimensional earthquake tomography by trans-dimensional Monte Carlo sampling, *Geophys. J. Int.*, **201**(3), 1598–1617.
- Andrieu, C., Jasra, A., Doucet, A. & Del Moral, P., 2011. On nonlinear Markov Chain Monte Carlo, *Bernoulli*, **17**(3), 987–1014.
- Bardainne, T. & Gaucher, E., 2010. Constrained tomography of realistic velocity models in microseismic monitoring using calibration shots, *Geophys. Prospect.*, **58**(5), 739–753.
- Bayes, M. & Price, M., 1763. An essay towards solving a problem in the doctrine of chances. By the late rev. Mr. Bayes, FRS communicated by Mr. Price, in a letter to John Canton, AMFRS, *Phil. Trans.*, **53**, 370–418.
- Bennett, N.N. & Malinverno, A., 2005. Fast model updates using wavelets, *Multiscale Model. Simul.*, **3**(1), 106–130.
- Bodin, T. & Sambridge, M., 2009. Seismic tomography with the reversible jump algorithm, *Geophys. J. Int.*, **178**(3), 1411–1436.
- Bodin, T., Sambridge, M., Rawlinson, N. & Arroucau, P., 2012. Transdimensional tomography with unknown data noise, *Geophys. J. Int.*, **189**(3), 1536–1556.
- Brooks, S.P. & Roberts, G.O., 1998. Convergence assessment techniques for Markov Chain Monte Carlo, *Stat. Comput.*, **8**(4), 319–335.
- Burrus, C.S., Gopinath, R.A. & Guo, H., 1997. *Introduction to Wavelets and Wavelet Transforms: A Primer*, Prentice Hall, 268 pp.
- Chiao, L.-Y. & Kuo, B.-Y., 2001. Multiscale seismic tomography, *Geophys. J. Int.*, **145**, 517–527.
- Cowles, M.K. & Carlin, B.P., 1996. Markov Chain Monte Carlo convergence diagnostics: a comparative review, *J. Am. Stat. Assoc.*, **91**(434), 883–904.
- Debski, W., 2010. Seismic tomography by Monte Carlo sampling, *Pure appl. Geophys.*, **167**(1–2), 131–152.
- Earl, D.J. & Deem, M.W., 2005. Parallel tempering: theory, applications, and new perspectives, *Phys. Chem. Chem. Phys.*, **7**(23), 3910–3916.
- Eisner, L., Duncan, P.M., Heigl, W.M. & Keller, W.R., 2009. Uncertainties in passive seismic monitoring, *Leading Edge*, **28**(6), 648–655.
- Gesret, A., Desassis, N., Noble, M., Romary, T. & Maisons, C., 2015. Propagation of the velocity model uncertainties to the seismic event location, *Geophys. J. Int.*, **200**(1), 52–66.
- Geyer, C.J., 1991. Markov chain Monte Carlo maximum likelihood, in *Computer Science and Statistics: Proceedings of 23rd Symposium on the Interface*, pp. 156–163, ed. Keramigas, E.M., Interface Foundation, Fairfax Station, VA.
- Green, P.J., 1995. Reversible jump Markov Chain Monte Carlo computation and Bayesian model determination, *Biometrika*, **82**(4), 711–732.
- Hall, C., Ji, W. & Blaisten-Barojas, E., 2014. The metropolis Monte Carlo method with cuda enabled graphic processing units, *J. Comput. Phys.*, **258**, 871–879.

- Hartzell, S. & Langer, C., 1993. Importance of model parameterization in finite fault inversions: application to the 1974 Mw 8.0 Peru earthquake, *J. geophys. Res.*, **98**(B12), 22 123–22 134.
- Hawkins, R. & Sambridge, M., 2015. Geophysical imaging using trans-dimensional trees, *Geophys. J. Int.*, **203**(2), 972–1000.
- Hoeting, J.A., Madigan, D., Raftery, A.E. & Volinsky, C.T., 1999. Bayesian model averaging: a tutorial, *Stat. Sci.*, **14**(4), 382–417.
- Holland, J.H., 1975. *Adaptation in Natural and Artificial System: An Introduction with Application to Biology, Control and Artificial Intelligence*, University of Michigan Press.
- Iba, Y., 2001. Extended ensemble Monte Carlo, *Int. J. Mod. Phys. C*, **12**(05), 623–656.
- Jansen, M., 2012. *Noise Reduction by Wavelet Thresholding*, Vol. 161, Springer Science & Business Media.
- Kirkpatrick, S., Gelatt, C.D. & Vecchi, M.P., 1983. Optimization by simulated annealing, *Science*, **220**(4598), 671–680.
- Kitanidis, P.K., 1995. Quasi-linear geostatistical theory for inverting, *Water Resour. Res.*, **31**(10), 2411–2419.
- Köhn, D., De Nil, D., Kurzman, A., Przebindowska, A. & Bohlen, T., 2012. On the influence of model parameterization in elastic full waveform tomography, *Geophys. J. Int.*, **191**(1), 325–345.
- Koren, Z., Mosegaard, K., Landa, E., Thore, P. & Tarantola, A., 1991. Monte Carlo estimation and resolution analysis of seismic background velocities, *J. geophys. Res.*, **96**(B12), 20 289–20 299.
- Kou, S., Zhou, Q. & Wong, W.H., 2006. Discussion paper equi-energy sampler with applications in statistical inference and statistical mechanics, *Ann. Stat.*, **34**(4), 1581–1619.
- Loris, I., Nolet, G., Daubechies, I. & Dahlen, F., 2007. Tomographic inversion using  $\ell_1$ -norm regularization of wavelet coefficients, *Geophys. J. Int.*, **170**(1), 359–370.
- MacEachern, S.N. & Berliner, L.M., 1994. Subsampling the gibbs sampler, *Am. Stat.*, **48**(3), 188–190.
- Malinverno, A., 2002. Parsimonious Bayesian Markov Chain Monte Carlo inversion in a nonlinear geophysical problem, *Geophys. J. Int.*, **151**(3), 675–688.
- Mallat, S., 1999. *A Wavelet Tour of Signal Processing*, Academic Press.
- Maxwell, S., 2010. Microseismic: growth born from success, *Leading Edge*, **29**(3), 338–343.
- Menke, W., 2012. *Geophysical Data Analysis: Discrete Inverse Theory*, Academic Press.
- Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H. & Teller, E., 1953. Equation of state calculations by fast computing machines, *J. Chem. Phys.*, **21**(6), 1087–1092.
- Meyn, S.P. & Tweedie, R.L., 2012. *Markov Chains and Stochastic Stability*, Springer Science & Business Media.
- Mosegaard, K. & Sambridge, M., 2002. Monte Carlo analysis of inverse problems, *Inverse Probl.*, **18**(3), R29–R54.
- Mosegaard, K. & Tarantola, A., 1995. Monte Carlo sampling of solutions to inverse problems, *J. geophys. Res.*, **100**(B7), 12 431–12 447.
- Mühlenbein, H., Schomisch, M. & Born, J., 1991. The parallel genetic algorithm as function optimizer, *Parallel Comput.*, **17**(6), 619–632.
- Noble, M., Gesret, A. & Belayouni, N., 2014. Accurate 3-d finite difference computation of traveltimes in strongly heterogeneous media, *Geophys. J. Int.*, **199**(3), 1572–1585.
- Oliver, D.S., 1996. On conditional simulation to inaccurate data, *Math. Geol.*, **28**(6), 811–817.
- Rawlinson, N., Pozgay, S. & Fishwick, S., 2010. Seismic tomography: a window into deep earth, *Phys. Earth Planet. Inter.*, **178**(3), 101–135.
- Rawlinson, N., Fichtner, A., Sambridge, M. & Young, M.K., 2014. Chapter one—seismic tomography and the assessment of uncertainty, *Adv. Geophys.*, **55**, 1–76.
- Robert, C. & Casella, G., 2013. *Monte Carlo Statistical Methods*, Springer Science & Business Media.
- Romary, T., 2009. Integrating production data under uncertainty by parallel interacting Markov Chains on a reduced dimensional space, *Comput. Geosci.*, **13**(1), 103–122.
- Romary, T., 2010a. Bayesian inversion by parallel interacting Markov Chains, *Inverse Probl. Sci. Eng.*, **18**(1), 111–130.
- Romary, T., 2010b. History matching of approximated lithofacies models under uncertainty, *Comput. Geosci.*, **14**(2), 343–355.
- Rosenthal, J.S., 2000. Parallel computing and monte carlo algorithms, *Far East J. Theor. Stat.*, **4**(2), 207–236.
- Salamon, P., Sibani, P. & Frost, R., 2002. *Facts, Conjectures, and Improvements for Simulated Annealing*, SIAM.
- Sambridge, M., 1999a. Geophysical inversion with a neighbourhood algorithm—I. Searching a parameter space, *Geophys. J. Int.*, **138**(2), 479–494.
- Sambridge, M., 1999b. Geophysical inversion with a neighbourhood algorithm—II. Appraising the ensemble, *Geophys. J. Int.*, **138**(3), 727–746.
- Sambridge, M., 2014. A parallel tempering algorithm for probabilistic sampling and multimodal optimization, *Geophys. J. Int.*, **196**, 357–374.
- Sambridge, M., Gallagher, K., Jackson, A. & Rickwood, P., 2006. Trans-dimensional inverse problems, model comparison and the evidence, *Geophys. J. Int.*, **167**(2), 528–542.
- Sen, M.K. & Stoffa, P.L., 2013. *Global Optimization Methods in Geophysical Inversion*, Cambridge Univ. Press.
- Sheriff, R.E. & Geldart, L.P., 1995. *Exploration Seismology*, Cambridge Univ. Press.
- Silverman, B.W., 1986. *Density Estimation for Statistics and Data Analysis*, Vol. 26, CRC press.
- Simons, F.J. *et al.*, 2011. Solving or resolving global tomographic models with spherical wavelets, and the scale and sparsity of seismic heterogeneity, *Geophys. J. Int.*, **187**(2), 969–988.
- Tarantola, A., 2005. *Inverse Problem Theory and Methods for Model Parameter Estimation*, SIAM, Philadelphia, Pennsylvania, USA.
- Tarantola, A. & Valette, B., 1982. Generalized nonlinear inverse problems solved using the least squares criterion, *Rev. Geophys. Space Phys.*, **20**(2), 219–232.
- Tikhotsky, S. & Achauer, U., 2008. Inversion of controlled-source seismic tomography and gravity data with the self-adaptive wavelet parametrization of velocities and interfaces, *Geophys. J. Int.*, **172**(2), 619–630.
- Torn, A. & Zilinskas, A., 1989. *Global Optimization*, Springer-Verlag.
- Weber, Z., 2000. Seismic traveltime tomography: a simulated annealing approach, *Phys. Earth Planet. Inter.*, **119**(1), 149–159.

APPENDIX: FLOWCHART OF THE INTERACTING MCMC ALGORITHM PRESENTED





# Oxygen data assimilation for estimating micro-organism communities' parameters in river systems

Shuaitao Wang<sup>\*</sup>, Nicolas Flipo, Thomas Romary

Geosciences and Geoengineering Department, MINES ParisTech, PSL University, 35 Rue Saint-Honoré, 77300, Fontainebleau, France



## ARTICLE INFO

### Article history:

Received 30 April 2019

Received in revised form

30 July 2019

Accepted 21 August 2019

Available online 26 August 2019

### Keywords:

Data assimilation

Dissolved oxygen

Parameter estimation

Particle filter

ProSe-PA

## ABSTRACT

The coupling of high frequency data of water quality with physically based models of river systems is of great interest for the management of urban socio-ecosystems. One approach to exploit high frequency data is data assimilation which has received an increasing attention in the field of hydrology, but not for water quality modeling so far. We present here a first implementation of a particle filtering algorithm into a community-centered hydro-biogeochemical model to assimilate high frequency dissolved oxygen data and to estimate metabolism parameters in the Seine River system. The procedure is designed based on the results of a former sensitivity analysis of the model (Wang et al., 2018) that allows for the identification of the twelve most sensible parameters all over the year. Those parameters are both physical and related to micro-organisms (reaeration coefficient, photosynthetic parameters, growth rates, respiration rates and optimal temperature). The performances of the approach are assessed on a synthetic case study that mimics 66 km of the Seine River. Virtual dissolved oxygen data are generated using time varying parameters. This paper aims at retrieving the predefined parameters by assimilating those data. The simulated dissolved oxygen concentrations match the reference concentrations. The identification of the parameters depends on the hydrological and trophic contexts and more surprisingly on the thermal state of the river. The physical, bacterial and phytoplanktonic parameters can be retrieved properly, leading to the differentiation of two successive algal blooms by comparing the estimated posterior distribution of the optimal temperature for phytoplankton growth. Finally, photosynthetic parameters' distributions following circadian cycles during algal blooms are discussed.

© 2019 Elsevier Ltd. All rights reserved.

## 1. Introduction

The coupling of high frequency data of water quality with physically based model of river systems is of great interest for the management of urban socio-ecosystems. One approach to exploit high frequency data is data assimilation which combines observations and short-range forecasts to estimate the distribution of the true state of a process (Wikle and Berliner, 2007). Assimilating high frequency data allows for the identification of the multiple sources of model uncertainty relative to parameters, model structure, forcing data (e.g. temperature, wind speed and solar radiation) and observations (Evensen, 2003), which limit the validation and application of water quality models (Beven, 1989; Polus et al., 2011).

A lot of data assimilation techniques exist in literature such as the variational methods (Sasaki, 1955, 1958), the Kalman filter (KF (Kalman, 1960)), the extended Kalman filter (EKF (Beck, 1987)), the ensemble Kalman filter (EnKF (Evensen, 1994)), and the particle filter (PF (Doucet et al., 2001)). A review of these methods can be found in the literature (Wikle and Berliner, 2007; Cappe et al., 2007; Särkkä, 2013). The variational methods minimizing a cost function have been widely used in numerical weather prediction (Courtier et al., 1994, 1998; Gauthier et al., 2007; Kleist and Ide, 2015; Yuçel et al., 2015). The major drawbacks of variational methods are the numerical implementation complexity and the computational cost. The KF handles only linear models with Gaussian errors, which are not consistent with the hydrologic and water quality models.

To extend to nonlinear models, the EKF uses the linearized formula tangent to the forward model, which is an approximation and would be very costly to implement for high-dimensional systems (Evensen, 2003). The EnKF uses Monte Carlo samples to

<sup>\*</sup> Corresponding author.

E-mail addresses: [shuaitao.wang@mines-paristech.fr](mailto:shuaitao.wang@mines-paristech.fr), [shuaitaowang@outlook.com](mailto:shuaitaowang@outlook.com) (S. Wang), [nicolas.flipo@mines-paristech.fr](mailto:nicolas.flipo@mines-paristech.fr) (N. Flipo), [thomas.romary@mines-paristech.fr](mailto:thomas.romary@mines-paristech.fr) (T. Romary).

approximate the forecast distribution and then applies linear update formulas to obtain the posterior distribution (Evenesen, 2003), while the particle filter based methods estimate the forecast and posterior distributions using discrete probability densities (Arulampalam et al., 2002) obtained via Bayes' theorem (Bayes, 1763). Since the EnKF and particle filter resolve the two major drawbacks related to the use of EKF and handle highly nonlinear models, the EnKF and particle filter based methods have become the most commonly used sequential data assimilation techniques in hydrologic modeling for state-parameter estimation (Moradkhani et al., 2005; Weerts and El Serafy, 2006; Andreadis et al., 2007; Salamon and Feyen, 2009; Plaza et al., 2012; DeChant and Moradkhani, 2012; Vrugt et al., 2013; Shi et al., 2014; Abbaszadeh et al., 2018).

In biogeochemical oceanography, the efficiency of the ensemble-based Kalman filters has been assessed on ocean models for ocean biogeochemical state and parameter estimation (Simon and Bertino, 2012; Simon et al., 2012; Gharamti et al., 2017; Yu et al., 2018). Due to the lack of data, the difficulties in representing complex hydro-ecosystems and the computational cost, there are still relatively few applications about the state-parameter estimation by assimilating high-frequency dissolved oxygen (DO) concentrations in river systems. Pastres et al. (2003) have applied the EKF to update three parameters of a simple DO-chlorophyll model in the lagoon of Venice. Mao et al. (2009) used EKF to forecast algal blooms and dissolved oxygen dynamics in a marine ecosystem. The recent studies focus particularly on forecasting algal bloom dynamics using EnKF in river system (Kim et al., 2014) or in lakes (Huang et al., 2013; Page et al., 2018). Xue et al. (2012) applied also EnKF in the Massachusetts Bay to design the optimal monitoring sites for DO measurements.

However, as stated by Wikle and Berliner (2007, p. 10–11), “We assume that the forecast distribution can be characterized by its first two moments (or, equivalently, that it is Gaussian with mean and (estimated) variance/covariance matrix). ... However, in nonlinear cases, since Gaussianity cannot hold for all time, the EnKF must yield biased samples and estimates, even for unlimited sample sizes”. The recent investigations concluded also that the assumption of a Gaussian error structure for the forecast distribution may not be realistic for hydrologic systems and phytoplankton dynamics and the authors suggested testing the particle filter based methods to overcome this problem (Plaza et al., 2012; Pasetto et al., 2012; DeChant and Moradkhani, 2012; Huang et al., 2013). The feasibility of Bayesian inference with a Particle Markov Chain Monte Carlo algorithm is tested on a simple predator-prey model in ecological research (Kattwinkel and Reichert, 2017).

In this paper, we present the ProSE-PA model, which consists in a first implementation of a particle filtering algorithm into the community-centered hydro-biogeochemical model ProSE (Even et al., 1998, 2007b; Flipo et al., 2004; Vilmin et al., 2015b) in order to assimilate 15 min-DO data and estimate metabolisms' parameters in the Seine River system. The procedure is designed from the results of a sensitivity analysis of the biogeochemical module, C-RIVE, of the ProSE model that identified the twelve most sensible parameters of the model (Wang et al., 2018). Those parameters are both physical (water re-aeration by fluvial navigation) and physiological (for instance growth rate of heterotrophic bacteria and photosynthetic parameters of phytoplankton). The performances of ProSE-PA are assessed on a synthetic case study that mimics 66 km of the Seine River and generates virtual “observation” data of DO. The objectives of this study are to retrieve the predefined parameters used to generate the “observation” data and to distinguish two successive algal blooms by identifying different physiological properties.

The manuscript is organized as follows. The section 2 presents

the ProSE model and the study area, followed by the ProSE-PA approach including the mathematical formulations of the particle filter, the resampling algorithm as well as the full numerical algorithm. The considered parameters, the “observation” data and the input data of ProSE model are described in section 2.3. The numerical settings and the computational cost are given in section 3.1. Then, we evaluate the simulated DO concentrations by four statistical criteria (section 3.2) and show the identification of metabolisms' parameters in frame of state classification (section 3.3). The real time parameter identification is discussed in section 4.1, followed by the identifiability of two successive algal blooms by comparing the different physiological properties. The photosynthetic parameters' distributions following circadian cycles are described in section 4.3. The sensitivity of ProSE-PA performances to the observation error is discussed in the section 4.4. To finish, a brief conclusion is given in section 5.

## 2. Material and methods

### 2.1. Synthetic case study

#### 2.1.1. The ProSE model

The ProSE model (Even et al., 1998, 2007b; Flipo et al., 2004; Vilmin et al., 2015b), couples three libraries (a hydrodynamic library, a transport library and a biogeochemical library), that simulate together the hydro-biogeochemical functioning of a river system. The hydraulic module solves the 1D shallow water equations with a finite volume scheme. The transport module simulates the advection and dispersion of both particulate and dissolved substances, including also water re-oxygenation by overflowing over hydraulic works. The biogeochemical library is based on the RIVE conceptual model which is a community-centered model (Billen et al., 1994; Garnier et al., 1995) (<https://www.fire.upmc.fr/river>). The cycles of carbon, nutrients and dissolved oxygen are simulated in both water column and sediment layer. The compound exchanges between water column and benthic layers have been successively developed (Even et al., 2004, 2007a; Flipo et al., 2004, 2007; Vilmin et al., 2015b). The ProSE model has been well validated and largely applied in the Seine River system (Even et al., 1998, 2004, 2007a; Polus et al., 2011; Raimonet et al., 2015; Vilmin et al., 2015a, b, 2016, 2018).

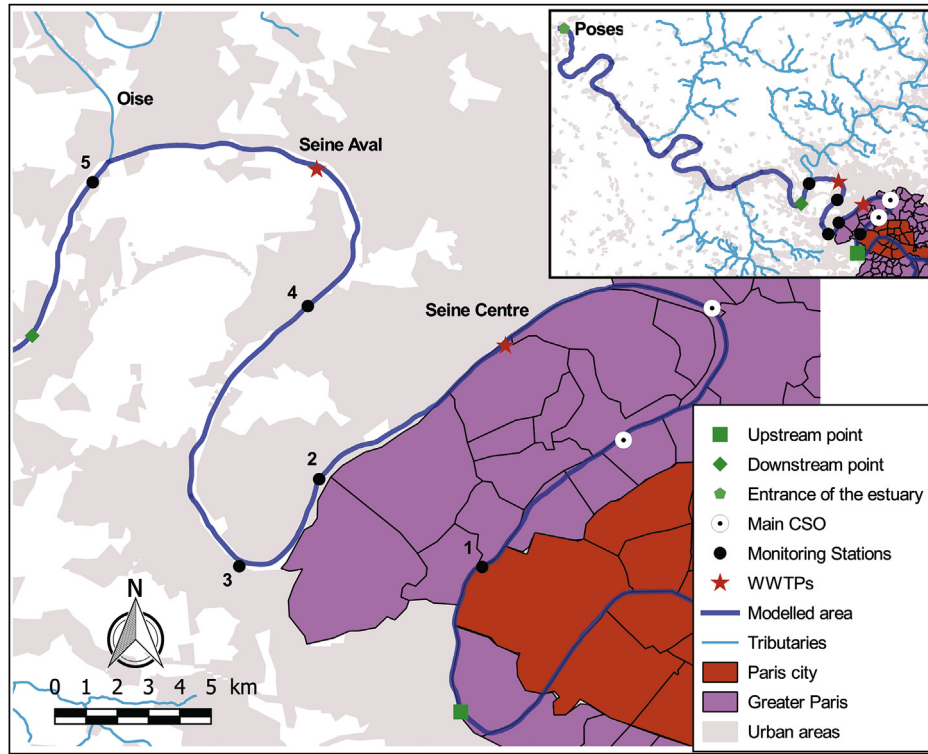
#### 2.1.2. Study area

The study area is located downstream Paris city and consists of 66 km of the Seine River (Fig. 1). The bathymetry data of the channel is extracted from the full 220 km Seine model used in recent studies (Vilmin et al., 2016, 2018). This area is highly impacted by human activities including two Waste Water Treatment Plants (WWTPs, Seine Aval and Seine Centre) and two major Combined Sewer Overflows (CSOs). The Seine Aval (SAV) is the largest WWTP of Europe, which treats the effluents of over 6.5 million equivalent inhabitants (Rocher et al., 2011). During rain events, CSOs discharge large amount of suspended solids, organic matters and nutrients into the Seine River (Even et al., 2007b). Assimilating DO is therefore of great interest for the decision makers in this area.

### 2.2. Data assimilation framework using particle filtering, the ProSE-PA approach

#### 2.2.1. State-space model

To begin with, we represent the ProSE model as a state-space model (Kalman, 1960). A state-space model uses inputs, state variables, outputs to describe the evolution of a system over time. In our case, it uses three equations (Eq. (1), (2), (3)). Let  $\mathbf{y}$  represents



**Fig. 1.** The description of study area and monitoring sites. Full Seine model is shown at upper-right corner; Monitoring stations from upstream to downstream: 1. Suresnes, 2. Chatou, 3. Bougival, 4. Sartrouville, 5. Andresy.

the state variable vector of the system in terms of physics (DO concentrations in this study) and  $\mathbf{x}$  be the vector of model parameters. The state variable and the model parameters are both modelled as random variables  $\mathbf{Y}$ ,  $\mathbf{X}$  characterized by their probability distribution functions (pdf).  $\mathbf{x}$  and  $\mathbf{y}$  denote the realizations of random variables  $\mathbf{X}$ ,  $\mathbf{Y}$ . The observation vector  $\mathbf{y}^*$  is a realization of the random variable  $\mathbf{Y}^*$ . The state-space model explains the temporal evolution of the system by assuming that it behaves as a Markov process (Markov, 1906). This means that the values at  $t$  depend on the values at  $t - 1$  only:

$$\mathbf{x}_t = \mathbf{x}_{t-1} + \boldsymbol{\eta}_t \quad (1)$$

$$\mathbf{y}_t = M(\mathbf{y}_{t-1}, \boldsymbol{\mu}_t, \mathbf{x}_t) + \boldsymbol{\nu}_t \quad (2)$$

$$\mathbf{y}_t^* = h(\mathbf{y}_t) + \boldsymbol{\varepsilon}_t \quad (3)$$

The transition of model parameter  $\mathbf{x}$  from  $t - 1$  to  $t$  is described by a Gaussian perturbation ( $\boldsymbol{\eta}_t$ , Eq. (1)). In the above equations,  $M$  is the forward model (in our case ProSE) and  $h$  is the observation operator relating the forecasted state variable ( $\mathbf{y}_t$ ) to the observation ( $\mathbf{y}_t^*$ ). In other words,  $h$  denotes the selection of model cells where observations are available.  $\mathbf{y}_{t-1}$  stands for the posterior state variable at previous time step ( $t - 1$ ) or equivalently for the prior state variable at current time step ( $t$ ). The symbols  $\boldsymbol{\mu}_t$  and  $\mathbf{x}_t$  represent respectively the prior forcing data (e.g. temperature, wind speed and solar radiation) and the prior model parameters at time step  $t$ . The variables  $\boldsymbol{\nu}_t$  and  $\boldsymbol{\varepsilon}_t$  characterize the unknown model and measurement errors respectively. The ProSE model aims at conserving the mass balance. The model does it with accuracy, the relative errors being in the order  $10^{-5}$  or  $10^{-6}$ . This is far below observation errors. Therefore, no model errors are considered ( $\boldsymbol{\nu}_t = 0$ ) in our case.

Note that the nomenclature we use here is not canonical. In dynamical system literature, the parameters would have been called the “states” because they are the hidden dynamic variables that govern the system’s behaviour. In forward modelling and especially hydrology, the state variables designate the variable that are computed by the model, while the parameters represents the time varying functional inputs of the numerical model. We consider the latter as the reader is more likely to be accustomed to this definition. Please note that, with this definition, the input fluxes at the boundary of the model are not called parameters, but boundary conditions.

### 2.2.2. Bayesian inference

We seek to estimate the posterior pdf of both the random variable  $\mathbf{Y}$  and the parameter  $\mathbf{X}$  conditionally to the observations  $\mathbf{Y}^*$ . To simplify the writing, we define the state vector  $\mathbf{Z} = [\mathbf{Y}^T, \mathbf{X}^T]^T$  and its realizations  $\mathbf{z} = [\mathbf{y}^T, \mathbf{x}^T]^T$ . Therefore, our goal can be reformulated in finding the conditional pdf of  $\mathbf{Z}$  knowing  $\mathbf{y}^*$  noted  $f_{\mathbf{Z}}(\mathbf{z}|\mathbf{y}^*)$ . Through Bayes’ theorem (Bayes, 1763), the posterior pdf  $f_{\mathbf{Z}}(\mathbf{z}|\mathbf{y}^*)$  can be deduced from the product of the prior pdf ( $f_{\mathbf{Z}}(\mathbf{z})$ ) and the likelihood distribution ( $f_{\mathbf{Y}^*}(\mathbf{y}^*|\mathbf{z})$ ), up to a normalizing constant independent from  $\mathbf{Z}$ :

$$f_{\mathbf{Z}}(\mathbf{z}|\mathbf{y}^*) \propto f_{\mathbf{Y}^*}(\mathbf{y}^*|\mathbf{z})f_{\mathbf{Z}}(\mathbf{z}) \quad (4)$$

The prior pdf  $f_{\mathbf{Z}}(\mathbf{z})$  gives the prior knowledge on  $\mathbf{z}$  before the measurements are taken. The likelihood  $f_{\mathbf{Y}^*}(\mathbf{y}^*|\mathbf{z})$  describes the pdf of the observation data given the model predictions.

### 2.2.3. Sequential form

The formulation (4) does not relate to time. To introduce a sequential form for data assimilation, we define the temporal trajectories of the random variables  $\mathbf{Z}$  and  $\mathbf{Y}^*$ ,  $\mathbf{Z}_{1:t} \equiv \{\mathbf{Z}_1, \dots, \mathbf{Z}_t\}$  and  $\mathbf{Y}_{1:t}^* \equiv \{\mathbf{Y}_1^*, \dots, \mathbf{Y}_t^*\}$ , discretized over time by sequences of state ( $\mathbf{Z}_i, \mathbf{Y}_i^*$ ,

$i = 1, \dots, t$ ), as well as their realizations  $\mathbf{z}_{1:t}, \mathbf{y}_{1:t}^*$ . The posterior pdf of the trajectory  $\mathbf{Z}_{1:t}$  given observation  $\mathbf{y}_{1:t}^*$  writes  $f_{\mathbf{Z}_{1:t}}(\mathbf{z}_{1:t}|\mathbf{y}_{1:t}^*)$ .

Next, we can write the posterior pdf  $f_{\mathbf{Z}_{1:t}}(\mathbf{z}_{1:t}|\mathbf{y}_{1:t}^*)$  under a recursive form (Arulampalam et al., 2002; Doucet et al., 2001; Wikle and Berliner, 2007; Särkkä, 2013) owing to the Markovian nature of the process. We simplify the notations rewriting a pdf  $f_{\mathbf{Z}}(\mathbf{z})$  into  $f(\mathbf{z})$ .

$$f(\mathbf{z}_{1:t}|\mathbf{y}_{1:t}^*) \propto f(\mathbf{y}_t^*|\mathbf{z}_{1:t}, \mathbf{y}_{1:t-1}^*)f(\mathbf{z}_{1:t}|\mathbf{y}_{1:t-1}^*) \quad (5)$$

$$= f(\mathbf{y}_t^*|\mathbf{z}_t)f(\mathbf{z}_t|\mathbf{z}_{1:t-1}, \mathbf{y}_{1:t-1}^*)f(\mathbf{z}_{1:t-1}|\mathbf{y}_{1:t-1}^*) \quad (6)$$

$$= f(\mathbf{y}_t^*|\mathbf{z}_t)f(\mathbf{z}_t|\mathbf{z}_{t-1})f(\mathbf{z}_{1:t-1}|\mathbf{y}_{1:t-1}^*), \quad (7)$$

variables knowing all the past observations as well as the current one. In this paper, we note the filtering posterior pdf at time  $t$  as  $f(\mathbf{z}_t|\mathbf{y}_{1:t}^*)$  which can be approximated as,

$$f(\mathbf{z}_t|\mathbf{y}_{1:t}^*) \approx \sum_{i=1}^N \hat{\omega}_t^i \delta \mathbf{z}_t^i \quad (10)$$

where  $\delta(\cdot)$  is a Dirac delta function and  $N$  denotes the number of particles. That is the filtering distribution is approximated by a discrete distribution, whose probability mass function is defined by the normalised importance weights and charges the particles.

If we assume that the observation error is Gaussian, then the likelihood probability of each particle can be calculated using the pdf of the multivariate normal distribution:

$$\ln L(\mathbf{y}_t^*|\mathbf{z}_t^i) = -\frac{m}{2} \ln(2\pi) - \frac{1}{2} \ln(|\Sigma|) - \frac{1}{2} (\mathbf{y}_t^* - h(\mathbf{y}_t^i))^T \Sigma^{-1} (\mathbf{y}_t^* - h(\mathbf{y}_t^i)) \quad (11)$$

$$f(\mathbf{y}_t^*|\mathbf{z}_t^i) = \frac{L(\mathbf{y}_t^*|\mathbf{z}_t^i)}{\sum_{i=1}^N L(\mathbf{y}_t^*|\mathbf{z}_t^i)}$$

where we use first the Bayes' theorem to get equation (5), then the Bayes' theorem and the Markov property for equation (6) and the Markov property to obtain the sequential relation (7). We propagate  $\mathbf{z}_t$  using the evolution equations (1) and (2) in order to model the distribution  $f(\mathbf{z}_t|\mathbf{z}_{t-1})$ . Note in case of a deterministic formulation ( $(\nu_t, \eta_t) = (0, 0)$  in equations (1) and (2)),  $f(\mathbf{z}_t|\mathbf{z}_{t-1})$  is a Dirac measure. Consequently, we can access the posterior pdf  $f(\mathbf{z}_{1:t}|\mathbf{y}_{1:t}^*)$  of the trajectory  $\mathbf{Z}_{1:t}$  by updating it at each time step. This will be done through the use of sampling techniques.

#### 2.2.4. The particle filter

The particle filter aims at approximating the posterior pdf  $f(\mathbf{z}_{1:t}|\mathbf{y}_{1:t}^*)$  of the trajectory  $\mathbf{Z}_{1:t}$  knowing  $\mathbf{y}_{1:t}^*$  by a set of particles (simulations) associated with weights ( $\omega_t$ ). It is generally impossible to sample from  $f(\mathbf{z}_{1:t}|\mathbf{y}_{1:t}^*)$  directly. To address this problem, sequential importance sampling has been suggested (Doucet et al., 2000; Liu, 2001). For the ease of reading, the principle of the importance sampling and the definition of the importance weights are detailed in B. We give here the weight update formula for each particle directly, which is a consequence of the sequential decomposition of the posterior of the trajectory obtained in (Eq. (7)).

$$\omega_t^i = f(\mathbf{y}_t^*|\mathbf{z}_t^i) \omega_{t-1}^i \quad (8)$$

$$\hat{\omega}_t^i = \frac{\omega_t^i}{\sum \omega_t^i} \quad (9)$$

where  $f(\mathbf{y}_t^*|\mathbf{z}_t^i)$  denotes the likelihood function, which quantifies how  $\mathbf{y}_t^*$  is likely to be observed given  $\mathbf{z}_t^i$  at time  $t$ .  $\omega_t^i$  and  $\omega_{t-1}^i$  stand for the posterior and prior weights at time  $t$ . In other words, the posterior weight at time  $t-1$  serves as a prior weight at time  $t$ . We also define the normalised importance weight,  $\hat{\omega}_t^i$ , associated with the particle  $i$ .

In practice, it is usually not necessary to estimate the posterior pdf  $f(\mathbf{z}_{1:t}|\mathbf{y}_{1:t}^*)$  of the trajectory  $\mathbf{Z}_{1:t}$  knowing  $\mathbf{y}_{1:t}^*$ . We are interested here in the marginal distribution of  $f(\mathbf{z}_{1:t}|\mathbf{y}_{1:t}^*)$ , called filtering distribution  $f(\mathbf{z}_t|\mathbf{y}_{1:t}^*)$ . It represents the distribution of the state

where  $m$  is the number of observation sites and  $\Sigma$  is the error covariance matrix of the observations.

#### 2.2.5. Resampling algorithm

A common problem when using particle filtering is the degeneracy phenomena. This occurs when almost all the particles have zero or very small importance weights. Only a few particles have significant importance weights. Therefore, the discrete probability densities cannot represent the posterior pdf of a given state adequately. A resampling procedure reduces the degeneracy effect. The basic idea of the resampling is to discard particles that have a small weight and to duplicate particles with a large weight. Generally, it is not necessary to perform resampling at every time step, but only when necessary. A way to monitor the need of resampling is to estimate the "effective" number of particles by the effective sample size ( $N_{eff}$ ). The  $N_{eff}$  cannot be evaluated exactly but it can be estimated it by (Kong et al., 1994; Doucet et al., 2000):

$$\widehat{N}_{eff} = \frac{1}{\sum_{i=1}^N (\hat{\omega}_t^i)^2} \quad (12)$$

When  $N_{eff}$  is below a threshold ( $N_{thres} = \alpha \cdot N$ ) predefined by the user, resampling is performed. After the resampling step, all weights are reset to  $1/N$ . As the particles having a important weight may be copied many times, which results in sample impoverishment problem. To maintain the diversity of the ensemble, a random perturbation is added to the parameters' value after the resampling step (eq. (13)).

$$\mathbf{x}_{t+1}^i = \mathbf{x}_{t, resampling}^i + \boldsymbol{\eta}_t^i \quad \boldsymbol{\eta}_t^i \sim N(0, (s \cdot \Phi)^2) \quad (13)$$

where  $s$  is a percent perturbation predefined by the user (0.1 in this study) and  $\Phi$  denotes the parameter range.

The resampling technique used in this work is referred to as systematic resampling. The procedure of systematic resampling is thoroughly described in the literature (Kitagawa, 1996; Moradkhani et al., 2005; Li et al., 2015).



### 2.2.6. Numerical algorithm

The coupling of the particle filtering algorithm with ProSE model is called ProSE-PA. PA stands for Parallel computing and data Assimilation. A full description of the ProSE-PA approach is given below (Fig. 2).

### 2.3. Description of the synthetic case study

#### 2.3.1. Parameters considered and virtual "observation" data

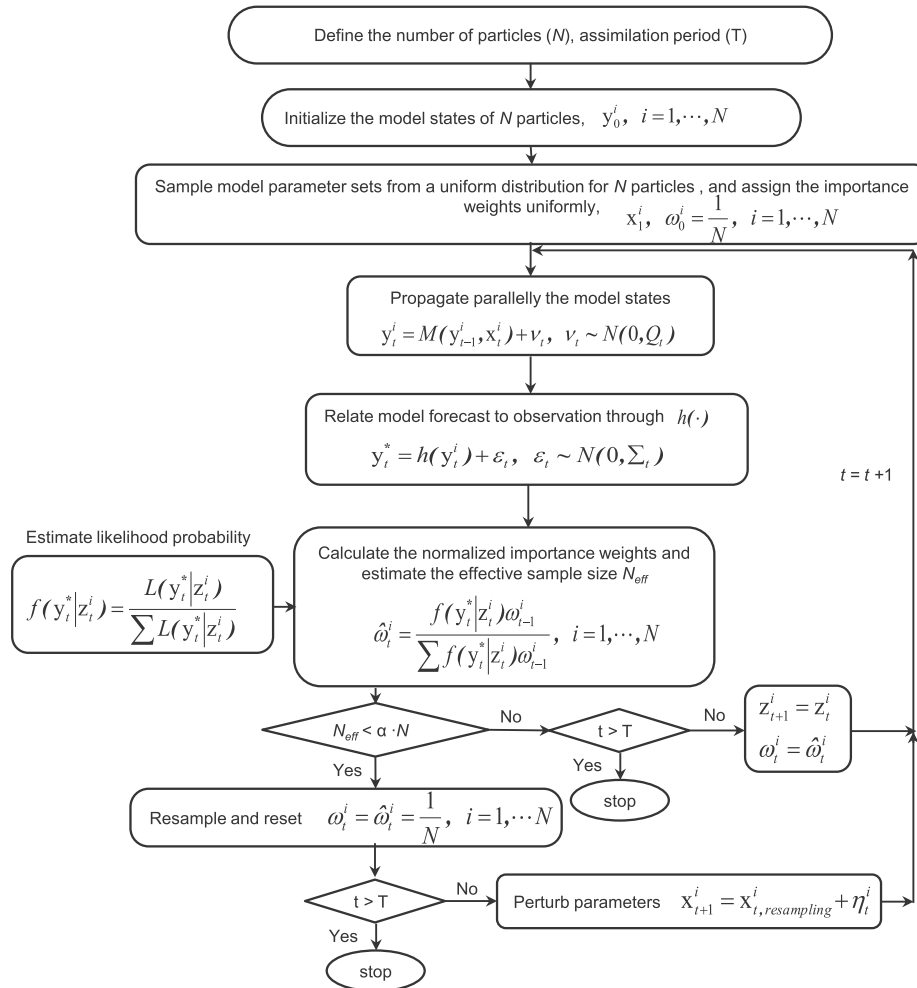
Twelve parameters of the ProSE model were identified in different hydrological and trophic contexts in a previous work (Wang et al., 2018). Those parameters are both physical and physiological (Table 1).

Five monitoring stations are located in the study area (Suresnes, Chatou, Bougival, Sartrouville, and Andresy, from upstream to downstream respectively, Fig. 1). The assimilation period is the year 2011 when algal blooms occurred in March and July in the Seine River (Vilmin et al., 2016). The two successive algal blooms are characterized by different physiological properties, especially the optimal temperature for growth of phytoplankton (Vilmin, 2014). A high-frequency DO (dissolved oxygen) dataset ( $\mathbf{y}^{ref}$ , every 15 min), which corresponds to the frequency of the real measurement, is generated using predefined parameters (Table 1). Three phytoplanktonic parameters are modified manually at day 139 in order to represent the two different algal blooms ( $P_{max,pp}$ ,  $R_{m,pp}$  and  $T_{opt,pp}$ ,

see Table 1 for parameter definition). In this case study, the "observation" data are obtained adding a Gaussian error to the reference dataset with a mean of zero and a standard deviation of  $0.01 \times \mathbf{y}^{ref}$  ( $\mathbf{y}_t^* = \mathbf{y}_t^{ref} + \varepsilon_t$ ,  $\varepsilon_t \sim N(0, (0.01 \times \mathbf{y}_t^{ref})^2)$ ). A standard deviation of  $0.01 \times \mathbf{y}^{ref}$  ensures that the 95% observation errors are less than  $0.02 \times \mathbf{y}^{ref}$ , which is coherent with the sensors used in the Seine River system (Garnier et al., 2019). The reference dataset and predefined parameters are assumed to be true and are assessed via data assimilation. In this study, no model errors are considered ( $\nu_t = 0$ , see Eq. (2)). The major aims of this work are to retrieve the reference DO concentrations and the dominant predefined parameters identified by Wang et al. (2018) over time.

#### 2.3.2. Input data for ProSE model

Apart from hydraulic (time varying river discharge and CSOs data), geometric data (river channels), the ProSE model also requires time varying concentrations of the biogeochemical variables (micro-organism biomass, dissolved oxygen, nutrients, organic matters, suspended solids) entering the system and the meteorological data (solar irradiance, water temperature and wind speed). These data are used as boundary conditions of the ProSE model. The quantification of uncertainties on boundary conditions is an ambitious topic that extent far beyond the scope of this paper and will require the study of real systems. Here we only investigate uncertainties on parameters.



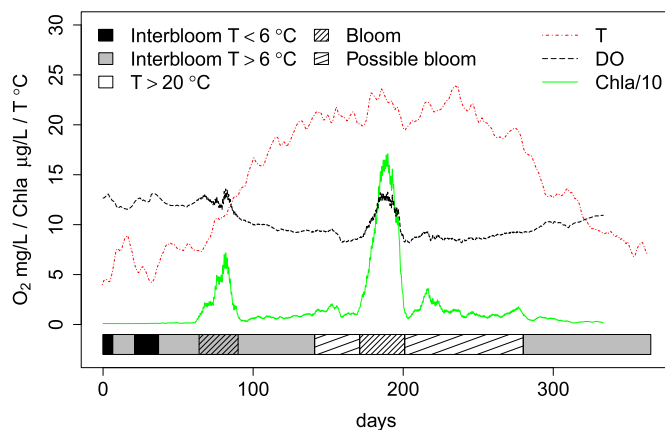
**Fig. 2.** The flowchart of ProSE-PA approach. The state  $\mathbf{z}^i$  denotes  $i$ th particle or a realization of random variable  $\mathbf{Z} = [\mathbf{Y}^T, \mathbf{X}^T]^T$ .  $\mathbf{y}^i$  and  $\mathbf{x}^i$  signify model state and parameter set of particle  $i$ .  $Q_t$  and  $\Sigma_t$  correspond to the model error covariance matrix and the observation error covariance matrix respectively. In this work, no model errors are considered,  $\nu_t = 0$ .

**Table 1**  
Reference parameters considered in data assimilation (Parameter range from Wang et al. (2018)).

Parameters	Description	Range	Reference March	Reference July	Unit
<b>Physical parameters</b>					
$\eta_{water}$	Light extinction coefficient for pure water	[0.2, 0.8]	0.32	0.32	[m-1]
$K_{navig}$	Reaeration coefficient related to the navigation	[0, 0.05]	0.025	0.025	[m.h-1]
<b>Bacterial parameters</b>					
$\mu_{max,hb}$	Maximum growth rate of bacteria	[0.01, 0.13]	0.04	0.04	[h-1]
$mort_{hb}$	Maximum mortality rate of bacteria	[0.01, 0.08]	0.02	0.02	[h-1]
$T_{opt,hb}$	Optimal temperature for bacterial growth	[10, 35]	25	25	[°C]
$Y_{hb}$	Bacterial growth yield	[0.03, 0.5]	0.25	0.25	[-]
<b>Phytoplanktonic parameters</b>					
$\alpha_{pp}$	Photosynthetic capacity	[0.0003, 0.0018]	0.0012	0.0012	[m2.s.μE-1.h-1]
$\eta_{chla,pp}$	Light extinction coefficient by algal self-shading	[0.006, 0.054]	0.02	0.02	[L.μgchla-1.m-1]
$Chla/C_{pp}$	Ratio of chlorophyll <i>a</i> to carbon	[50, 7.69]	28.57	28.57	[μgchla.mgC-1]
$P_{max,pp}$	Maximum photosynthesis rate	[0.09, 0.546]	0.2	0.4	[h-1]
$R_{m,pp}$	Respiration of maintenance	[0.001, 0.021]	0.002	0.01	[h-1]
$T_{opt,pp}$	Optimal temperature for growth of phytoplankton	[10, 37]	13	23	[°C]

### 2.3.3. Qualitative description of the simulation period: state classification

According to the former sensitivity analysis of the biogeochemical module, C-RIVE, of the ProSE model (Wang et al., 2018), we can classify the year 2011 into different periods (Fig. 3). The polygons at the bottom show this classification (Fig. 3). The water temperature increases from black to white (color gradient) and the line shaded polygons represent algal bloom periods ( $C_{chla} > 10 \mu\text{g/L}$ ). Wang et al. (2018) concluded that at low temperatures (Interbloom  $T < 6^\circ\text{C}$ , black polygons), the river system is controlled by the reaeration process ( $K_{navig}$ ). At moderate and high water temperatures, the maximum growth rate of bacteria ( $\mu_{max,hb}$ ) is most sensitive to variation of DO concentrations when no algal bloom occurs (Interbloom  $T > 6^\circ\text{C}$ , gray polygons). Once algal bloom occurs, phytoplanktonic parameters related to respiration ( $R_{m,pp}$ ) and to photosynthesis ( $\alpha_{pp}$ ,  $P_{max,pp}$  etc.) are dominant. A supplementary sensitivity analysis by Sobol method (Sobol, 1993) during algal bloom shows that the optimal temperature ( $T_{opt,pp}$ ) plays an important role on the growth of phytoplankton and that the sensitivity of  $R_{m,pp}$  (respiration of maintenance) depends on the water temperature (Fig. A1). When the water temperature exceeds  $20^\circ\text{C}$ , the total sensitivity index of  $R_{m,pp}$  decreases dramatically (Fig. A1).



**Fig. 3.** State classification of synthetic case study over time, concluded by Wang et al. (2018). The red line (dash-dotted) represents water temperature; The black line (dashed) denotes reference DO concentrations at the Bougival station; The green line (solid) shows the evolution of chla concentrations. The polygons at the bottom correspond to the different periods. The water temperature increases from black to white. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

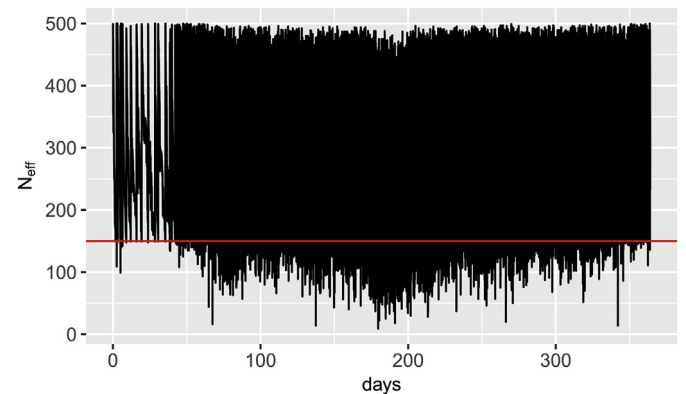
### 2.4. Numerical setting and computational cost

According to a test of the number of particles (not shown here), we select 500 particles in this work which is sufficient to i) match the observed oxygen concentrations and ii) identify the posterior pdfs of sensible parameters. The OpenMP Application Programming Interface (API) is implemented in the code to simulate the 500 particles in parallel. For a 1 year-simulation period (365 days) at a 15-min time step, the computation takes 1.13 days with 20 processors (Intel(R) Xeon(R) CPU E5-2640 v4 @ 2.40 GHz). A resampling threshold of  $\alpha = 0.3$  is chosen, corresponding to a minimum effective sample size of 150.

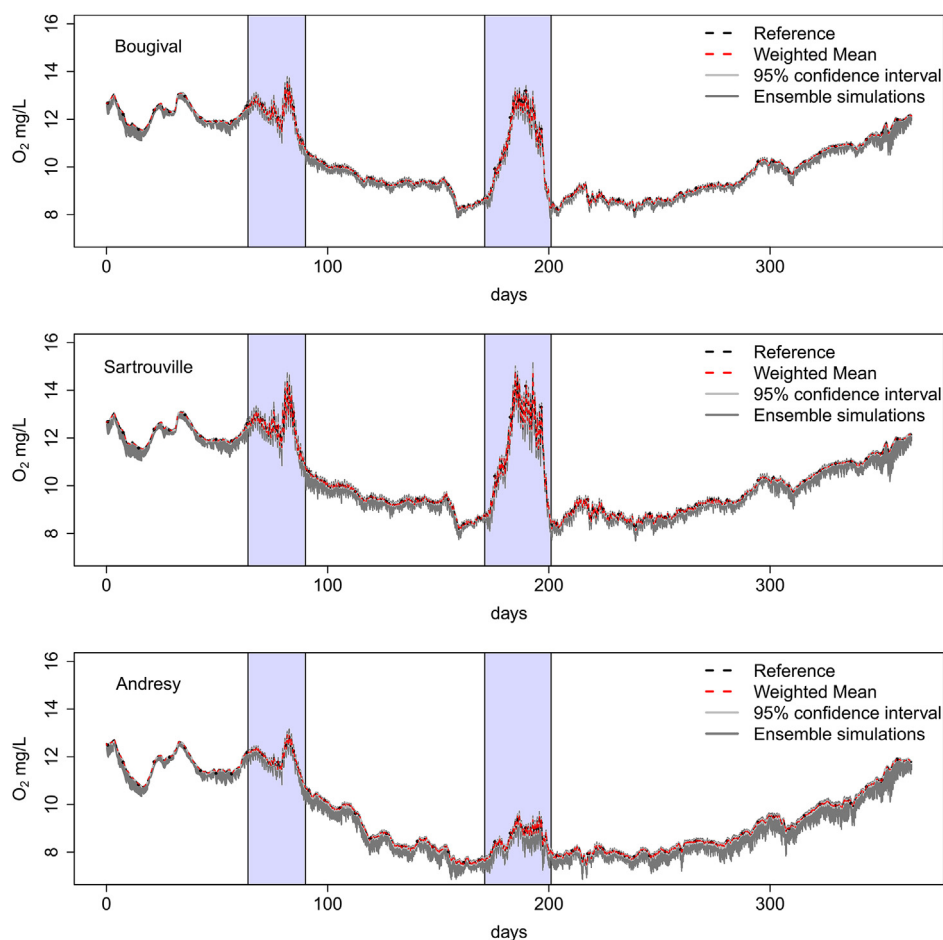
## 3. Results: oxygen simulation and parameter identification

### 3.1. Effective sample size for resampling procedure

Although the effective sample size ( $N_{eff}$ ) reduce fast after two months, synchronously with the start of the first algal bloom (day 64), the parameter perturbation allows the restoration of particle diversity (Fig. 4,  $N_{eff} > 450$  after resampling), which indicates a proper exploration of the parameter space by the algorithm. This is achieved by selecting a configuration of the Gaussian random walk (Pearson, 1905) through its standard deviation. After several trials (not shown here), the standard deviation of the random walk has been set to 0.1 times the parameters' range (eq. (13)).



**Fig. 4.** The evolution of effective sample size during assimilation. Red line corresponds to the minimum effective sample size 150. Frequent resamplings ensure a proper searching of the parameter space. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)



**Fig. 5.** Simulated DO concentrations at Bougival, Sartrouville and Andresy stations. The dashed red lines denote ensemble weighted means. The reference DO data are represented by dashed black lines. To clarify the illustration, weekly reference DO concentrations are shown by black points. The gray areas correspond to the ensemble simulations (dark gray areas) and the 95 percentile confidence intervals (light gray areas). The two algal bloom periods are shown by blue polygons. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

### 3.2. Almost perfect DO simulation with ProSE-PA

Fig. 5 shows simulated DO concentrations at three downstream monitoring stations. Those stations are selected because they are less sensitive to input boundary conditions and therefore permit to evaluate the conceptual structure of the code and the conceptualization of processes. The model performances are evaluated by four statistical criteria (RMSE, Root-Mean-Square-Error; MAE, Maximum Absolute Error; NSE, Nash-Sutcliffe Efficiency; R, Correlation coefficient).

The ensemble weighted average DO concentrations at all stations match the reference DO concentrations. Two algal blooms are well retrieved. The first starts at 64th day and the second around 171st day (Fig. 5). The maximum RMSE and MAE values between ensemble weighted mean and reference DO data are obtained at the Suresnes station with an error of 0.035 mgO<sub>2</sub>/L and an error of 0.321 mgO<sub>2</sub>/L respectively (Table 2). Furthermore, all NSEs and correlation coefficients (R) are close to 1, which signifies a perfect match of simulated DO concentration to the reference DO data. In addition, the 95 percentile confidence intervals are very narrow, which means a perfect simulation also.

### 3.3. Parameter identification in the frame of the state classification

In this section, the identification of physical, bacterial and

**Table 2**

Statistical criteria of the weighted average DO concentrations of 500 particles/simulations at five monitoring sites.

Stations	RMSE	NSE	MAE	r
Suresnes	0.0345	0.9995	0.3209	0.9998
Chatou	0.0173	0.9999	0.1347	0.9999
Bougival	0.0192	0.9998	0.1627	0.9999
Sartrouville	0.0201	0.9998	0.1765	0.9999
Andresy	0.0163	0.9998	0.1350	0.9999

RMSE: Root-Mean-Square-Error.

NSE: Nash-Sutcliffe Efficiency.

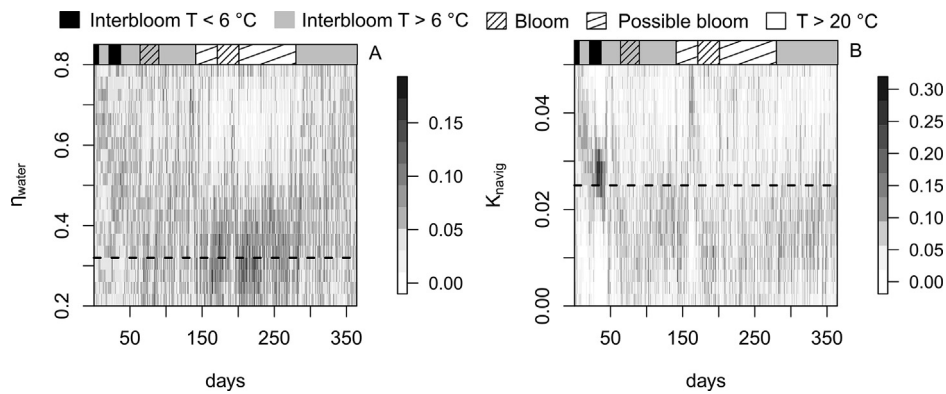
MAE: Maximum Absolute Error.

r: Correlation coefficient.

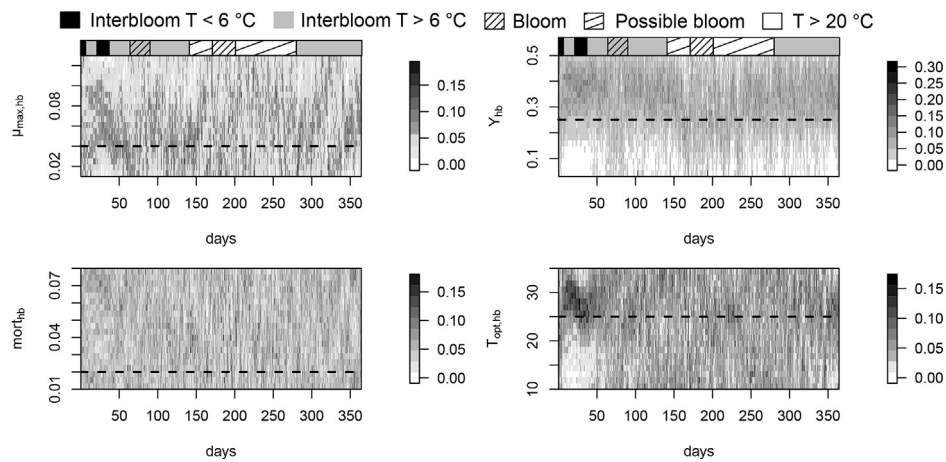
phytoplanktonic parameters is shown in plots displaying normalised weights over time (Figs. 6–8) for each parameter of interest. The daily normalised weights over time permits to represent the evolution of the posterior pdf of each parameter. Each parameter range is divided in 20 intervals. The sum of daily normalised weights in each interval which approximates the posterior pdf are shown by a image plot. The dashed black line represents the pre-defined value of parameter used to generate “observation” data.

#### 3.3.1. Physical parameters

Two physical parameters are assimilated in this paper, the light



**Fig. 6.** Daily normalised weights of physical parameters: light extinction coefficient for pure water ( $\eta_{water}$ ) and reaeration coefficient due to navigation ( $K_{navig}$ ). The dashed line represents the predefined reference value. Line shaded polygons correspond to algal bloom periods and water temperature increases from black to white (color gradient). See Fig. 3 for detailed state classification.



**Fig. 7.** The normalised importance weights of bacterial parameters. See Table 1 for parameter definition. The dashed line represents the predefined reference value. Line shaded polygons correspond to algal bloom periods and water temperature increases from black to white (color gradient). See Fig. 3 for detailed classification.

extinction coefficient for pure water ( $\eta_{water}$ ) and the reaeration coefficient related to the navigation ( $K_{navig}$ ). Albeit light extinction by clear water ( $\eta_{water}$ ) is a physical parameter, it is an important control for phytoplankton growth (Wang et al., 2018). It is not very well determined most of the time, except during algal blooms when the posterior pdfs focus around the reference value of 0.32  $m^{-1}$  (see the line shaded polygons in Fig. 6A).  $K_{navig}$  controls the reaeration process and is the most sensitive parameter to variation of DO in winter when the water temperature is below 6 °C (Wang et al., 2018). In this period (see the second black polygon in Fig. 6B), the particle filter achieves a very fine identification of its pdf. Seeing other periods of the year, its effect remains negligible and logically it remains unidentified.

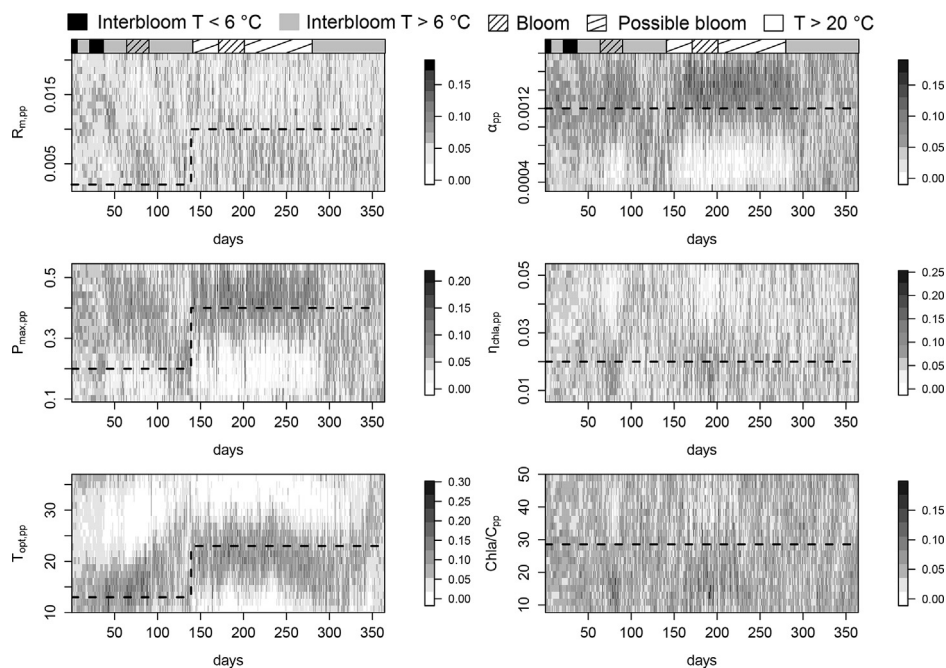
### 3.3.2. Bacterial parameters

In the previous sensitivity analysis (Wang et al., 2018), the maximum growth rate of bacteria ( $\mu_{max,hb}$ ) is the first ranked parameter out of algal bloom with a moderate water temperature ( $T > 6$  °C) and the bacterial growth yield ( $Y_{hb}$ ) is identified as the second sensitive parameter for inter algal bloom periods (both low and moderate water temperatures).  $\mu_{max,hb}$  can be retrieved well for several periods. These periods match globally the gray polygons which correspond to inter algal bloom periods with a water temperature  $T > 6$  °C (Fig. 7).  $Y_{hb}$  is slightly overestimated during the simulation, except for the algal blooms periods (see the line shaded

polygons in Fig. 7). The optimal temperature for bacterial growth ( $T_{opt,hb}$ ) can be determined at the beginning of simulation. However, the maximal mortality rate of bacteria ( $mort_{hb}$ ) spreads uniformly over the parameter range during the data assimilation period.

### 3.3.3. Phytoplanktonic parameters

It can be observed that the respiration of maintenance ( $R_{m,pp}$ ), the photosynthetic capacity ( $\alpha_{pp}$ ), the light extinction coefficient by algal self-shading ( $\eta_{chla,pp}$ ), and the optimal temperature for growth of phytoplankton ( $T_{opt,pp}$ ) are well estimated during the first algal bloom (see the gray line shaded polygon in Fig. 8), while the maximum photosynthesis rate ( $P_{max,pp}$ ) is overestimated (Fig. 8). Similarly, almost all the phytoplanktonic parameters can be retrieved during the second algal bloom (see the clear line shaded polygons in Fig. 8), but not for  $R_{m,pp}$  and  $Chla/C_{pp}$ . The ratio of chlorophyll *a* to carbon ( $Chla/C_{pp}$ ) is unidentified during the simulation, because  $Chla/C_{pp}$  has little effect on the variation of DO concentrations (Wang et al., 2018). The particle filter captures the change of phytoplanktonic properties after 139 day for  $P_{max,pp}$  and  $T_{opt,pp}$ . During algal blooms, five among seven parameters including  $\eta_{water}$  can be retrieved. Even though the two other parameters are not well estimated, we obtain satisfying simulated DO concentrations during algal blooms (Fig. 5). Although  $\alpha_{pp}$ ,  $\eta_{water}$ ,  $\eta_{chla,pp}$  and  $T_{opt,pp}$ , are always identified by the algorithm during blooms, this is



**Fig. 8.** The normalised importance weights of phytoplanktonic parameters. See Table 1 for parameter definition. The dashed line represents the predefined reference value. Line shaded polygons correspond to algal bloom periods and water temperature increases from black to white (color gradient). See Fig. 3 for detailed classification.

not the case for  $R_{m,pp}$  and  $P_{max,pp}$  which are supposed to be important during algal bloom (Wang et al., 2018). This is further discussed in the next section.

## 4. Discussion

### 4.1. Specifications for real time parameter identification

As shown by the results, the identification of the different parameters is not always ensured during the data assimilation period. This result can be explained by the parameters' sensitivity in contrasted hydrological and trophic contexts. One parameter that is dominant on DO concentration can be well estimated, while the other insensitive parameters have a relatively large uncertainty. In this section, we discussed the real time parameter identification.

#### 4.1.1. Parameters identified between algal blooms

Out of algal blooms, the identified parameters are  $\mu_{max,hb}$  and  $K_{navig}$ , which is coherent with the previous sensitivity analysis (Wang et al., 2018). At low temperatures ( $T < 6^\circ\text{C}$ ), bacterial and phytoplanktonic activities are limited, the reaeration ( $K_{navig}$ ) controls DO concentration within water column. The river system is governed by physical processes. The daily posterior pdfs of  $K_{navig}$  at low temperature period (day 21–38) are shown (Fig. 9A). Although the pdfs of  $K_{navig}$  from day 21 to day 28 have more dispersion than those for day 29–38, their modes are centered on the reference value (Fig. 9A). However, the water temperature is below  $6^\circ\text{C}$  at the beginning of the simulation (days 0–6),  $K_{navig}$  is not very well estimated (see the first black polygon in Fig. 6B). This can be explained by the fact that the prior pdf of  $K_{navig}$  is uniform at the start of the simulation and this period is too short.

When the water temperature increases ( $T > 6^\circ\text{C}$ ), the heterotrophic bacteria develops and degrades organic matters.  $\mu_{max,hb}$  becomes the most important parameter governing the DO concentration in river system (Wang et al., 2018). The identification of  $\mu_{max,hb}$  in those periods can be expected (see the gray polygons in Fig. 7). The pdfs of  $\mu_{max,hb}$  are displayed for day 45–62 when the

water temperature increases (Fig. 9B). It can be clearly seen that the modes of the pdfs of  $\mu_{max,hb}$  move towards the reference value (Fig. 9B). The particle filter captures the change from physical control ( $K_{navig}$ ) to bacterial control ( $\mu_{max,hb}$ ).

Next, the mortality of phytoplankton at the end of algal blooms supplies organic matters for bacterial growth.  $\mu_{max,hb}$  is thus well estimated after algal blooms (see the gray polygons day 91–140 and day 280–300 in Fig. 7). This indicates that the algorithm adapts the evolution of the regime of river system from autotroph to heterotroph. However,  $\mu_{max,hb}$  is not well estimated for day 300–325, even if this period corresponds to inter algal blooms. The correlation between the water temperature and the DO concentration in this period (Fig. 3) signifies that the DO concentration is controlled by water temperature and then colsed to saturation. The reaeration and micro-organism activities are not important in this period. Therefore, no parameter can be identified in this period.

#### 4.1.2. Parameters identified during algal blooms

The maximum photosynthesis rate ( $P_{max,pp}$ ) is third ranked parameter during algal blooms in the sensitivity analysis (Wang et al., 2018). However,  $P_{max,pp}$  is overestimated with a stable posterior pdf during the first algal bloom (see line shaded polygon day 64–90 in Fig. 8). The particles with  $P_{max,pp}$  close to  $0.2\text{ h}^{-1}$  (predefined value) are under weighted. That means the compensation of  $P_{max,pp}$  with the other parameters exist. The parameter interaction of  $P_{max,pp}$  has been shown by the difference between the total sensitivity index and the first order sensitivity index (Wang et al., 2018). Nonetheless, the DO concentrations have been well estimated.

The former sensitivity analysis which used almost constant water temperatures didn't highlight the sensitivity of  $T_{opt,pp}$ . The supplementary result shows that the sensitivity of  $T_{opt,pp}$  depends on the water temperature (Fig. A1). Therefore, the water temperature is a crucial factor for algal blooms when  $T_{opt,pp}$  is defined. This is the reason why  $T_{opt,pp}$  can be identified before algal blooms, but not after algal blooms (Fig. 8).

Contrarily to the first algal bloom, the identification of  $R_{m,pp}$  fails

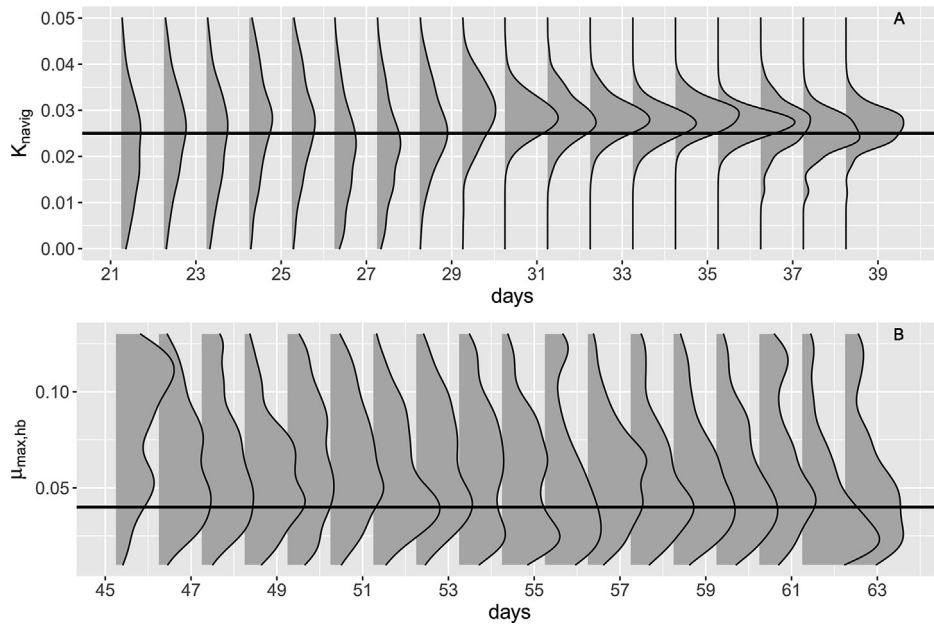


Fig. 9. The posterior pdf of  $K_{navig}$  during low temperature period (A) and the posterior pdf of  $\mu_{max,hb}$  with a moderate temperature (B); Black lines represent the predefined values.

for the second algal bloom (Fig. 8). The supplementary sensitivity analysis shows that when the water temperature exceeds 20 °C, the total sensitivity index of  $R_{m,pp}$  reduces dramatically (Fig. A1). The water temperatures are over 21 °C during the second bloom (Fig. 3). Therefore,  $R_{m,pp}$  is insensitive during the second algal bloom and remains unidentified. The other identified parameters match the sensitivity analysis ( $\alpha_{pp}$ ,  $\eta_{water}$ ,  $\eta_{chla,pp}$ ).

In conclusion, the identification of the parameters depends on not only the hydrological and trophic contexts but also on the thermal state of the river system. The detailed identification of the parameters is resumed in Fig. 10.

4.2. Possibility to identify different phytoplanktonic communities

In this paper, we define two different phytoplanktonic properties to distinguish blooms in March and in July ( $R_{m,pp}$ ,  $P_{max,pp}$  and  $T_{opt,pp}$ ). The particle filter detects well the change of phytoplanktonic properties (Fig. 8), especially the optimal temperature for the growth of phytoplankton ( $T_{opt,pp}$ ). The posterior pdfs of  $T_{opt,pp}$  are

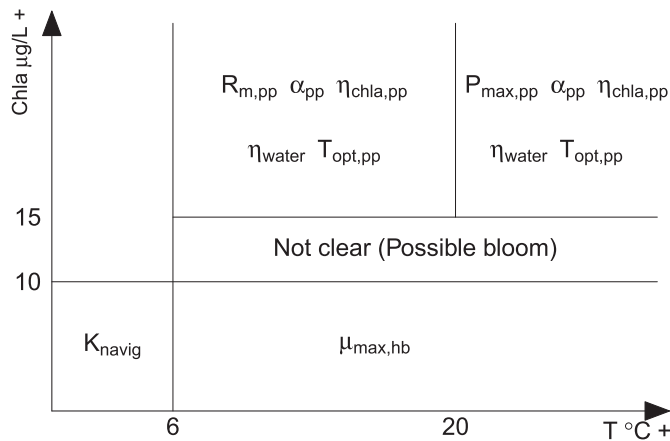


Fig. 10. The identification of parameters in different trophic ( $C_{chla}$ ) and thermal ( $T$  °C) contexts of river system. See Table 1 for parameter definition.

stable during the two algal blooms (Fig. 11). It is therefore possible to differentiate phytoplanktonic communities in real time by comparing the posterior distributions of  $T_{opt,pp}$ . The different optimal temperatures that have been used for modeling the phytoplanktonic communities during spring and summer are the ones determined in the Loire and Seine rivers (Descy et al., 2012; Garnier et al., 1995; Vilmin, 2014).

4.3. Circadian rhythm

At night, the photosynthetic parameters are absolutely insensitive and thus should remain unidentified. Therefore, a circadian rhythm (day and night) can be expected for the photosynthetic parameters. To study the circadian rhythm of photosynthetic parameters, the posterior pdfs of  $P_{max,pp}$ ,  $\alpha_{pp}$  and  $T_{opt,pp}$  are displayed every 3 h during days 184–186 (Fig. 12). It can be clearly noticed that the posterior pdfs of the three parameters have more dispersion at night (00:00–6:00) than those during the day (9:00–18:00). Their modes don't match the reference values at night, while the three parameters are well estimated during the day (Fig. 12). The ProSe-PA approach retrieves well the circadian rhythm of the photosynthetic parameters' sensitivity.

4.4. Performances' sensitivity to the observation error

The proof of concept of the particle filter has been achieved assuming a relative error on observations of 1%. In this section, we assess the impact of the relative error on both ProSe-PA ability to simulate oxygen concentrations and to identify parameter posterior pdfs. To this aim, various relative observation errors are tested from 1% to 10%. The Kling-Gupta Efficiencies (KGE) (Gupta et al., 2009; Kling et al., 2012) are calculated to evaluate the model performances. KGE is based on the decomposition of the mean squared error and NSE performance criteria (Gupta et al., 2009). KGEs range from -Inf to 1. Essentially, the closer to 1, the more accurate the model is.

The results show that the model performance decreases slightly with the increase of the observation relative error. Nevertheless, ProSe-PA retrieves in average the reference DO concentrations for

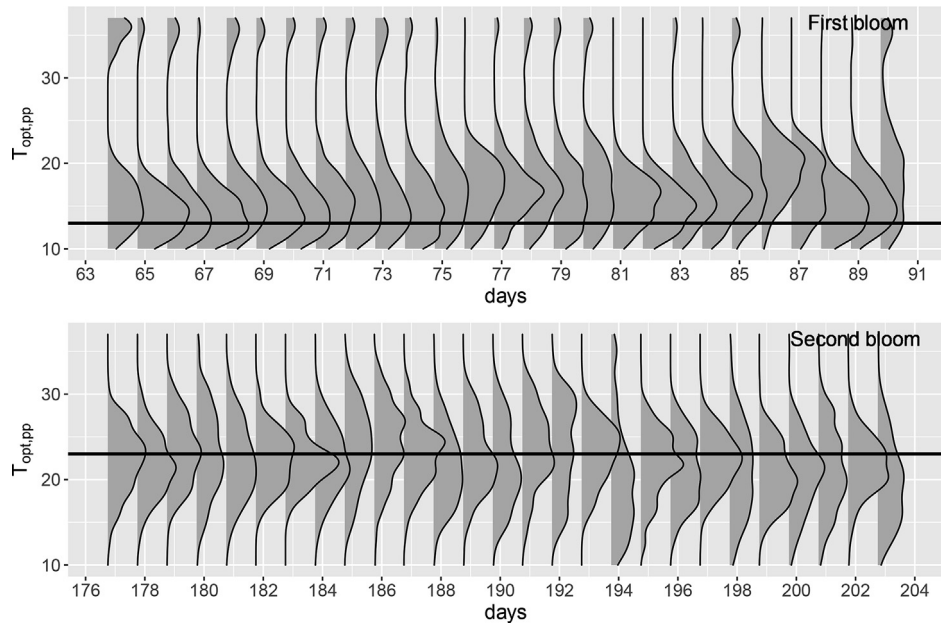


Fig. 11. The posterior distributions of  $T_{opt,pp}$  during the first and the second algal bloom at 6 p.m. Black line represents the predefined value.

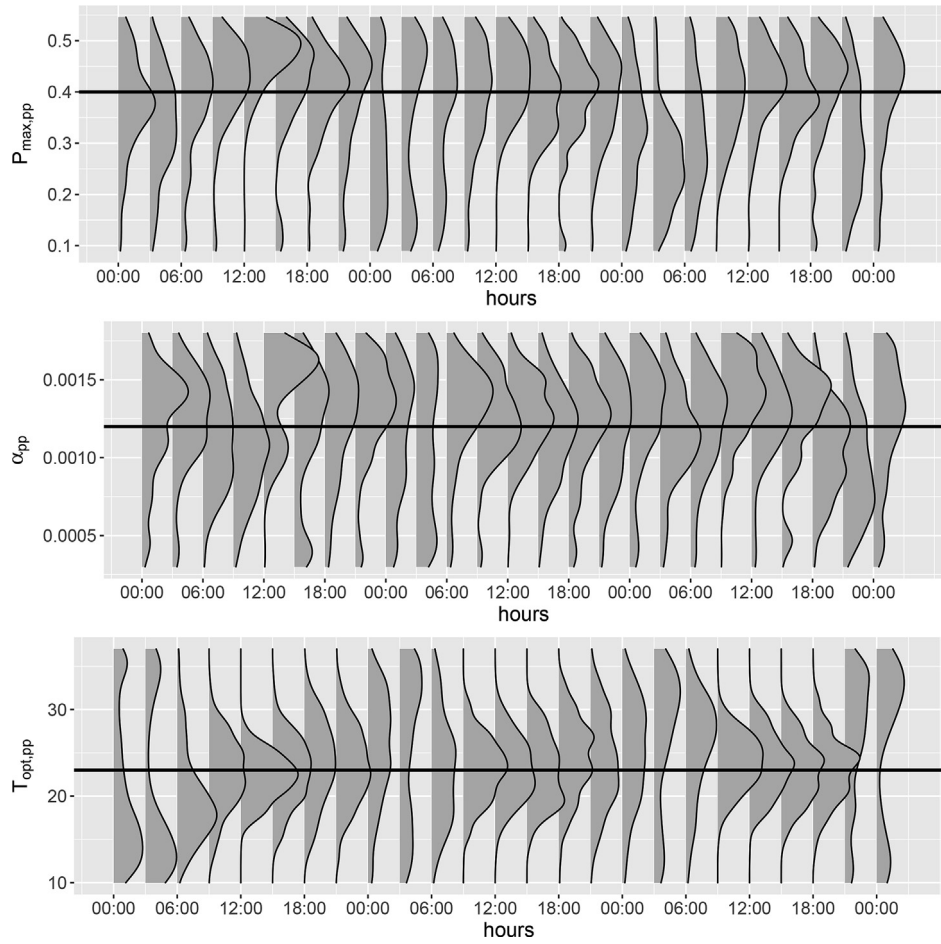


Fig. 12. The posterior distributions of  $P_{max,pp}$ ,  $\alpha_{pp}$  and  $T_{opt,pp}$  during 184–186 days.

**Table 3**  
Thresholds of the relative observation error for the parameter identifiability (1%–10%).

Parameters	Thresholds	Parameters	Thresholds	Parameters	Thresholds
$R_{m,pp}$	10% (T < 20 °C)	$\eta_{water}$	10%	$Y_{hb}$	< 7.5%
$P_{max,pp}$	10%	Chla / $C_{pp}$	Not stable	$mort_{hb}$	Not stable
$\alpha_{pp}$	10%	$T_{opt,pp}$	10%	$T_{opt,hb}$	< 5%
$\eta_{chla,pp}$	10%	$\mu_{max,hb}$	< 5%	$K_{navig}$	< 5%

10%: The parameters are always identified for the relative observation errors tested. See the Table 1 for the parameters' definition and units.

all observation relative errors tested. KGEs are always larger than 0.96 (not shown here) whatever the relative error. The identifiability of the bacterial parameters ( $\mu_{max,hb}$ ,  $Y_{hb}$  and  $T_{opt,pp}$ ) and the reaeration coefficient related to navigation ( $K_{navig}$ ) is not ensured when the observation relative error exceeds 5% while ProSE-PA is able to capture phytoplanktonic properties for all tested relative errors (1%–10%, Table 3), except for the ratio of chlorophyll *a* to carbon (Chla /  $C_{pp}$ ) which has little influence on the variation of DO concentrations (Wang et al., 2018).

## 5. Conclusions

In this work, we present a first implementation of particle filter into a hydro-biogeochemical model for metabolism's parameter estimation. The assimilation of a 15-min "observation" DO data is realized in the Seine River system on a synthetic case study.

- It can be concluded that the particle filter is an efficient method for the biogeochemical data assimilation and for the metabolism's parameter estimation in urban river systems.
- The ProSE-PA approach is capable to retrieve perfectly the reference DO concentrations at all stations.
- The identification of the metabolism parameters depends on the hydrological and trophic contexts and more surprisingly on the thermal state of the river system.
- The data assimilation method adapts to the trophic state's change of the Seine River system and the circadian cycle of photosynthetic parameters is well captured by the ProSE-PA approach.
- It is possible to distinguish phytoplanktonic species by identifying the different physiological properties via data assimilation.
- The ProSE-PA model is operational and transferable to waste water manager for assessing the impact of their practices now. However, it is necessary to quantify input uncertainties (forcing data, reject water of WWTPs and CSOs etc.), which are not considered in this paper.

## Software availability

A virtual machine including the ProSE-PA0.30 executable and a tutorial based on the synthetic dataset will be accessible online. If one wants to access the source code of the ProSE-PA model, please contact Dr. Nicolas Flipo. The software license is currently under discussion.

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Acknowledgements

This work is a contribution to the PIREN-SEINE research program, part of the french Long Term Socio-Ecological Research (LTSER) site "Zone Atelier Seine". We thank the editor and the three anonymous reviewers for their fruitful remarks which helped improve the overall quality of the paper.

## Appendix A. Sobol sensitivity analysis of C-RIVE module during algal bloom at different temperatures

The sensitivity analysis of C-RIVE model during algal blooms is realized every two degrees from 2 °C to 26 °C (Fig. A1). The results show that the optimal temperature for the growth of phytoplankton ( $T_{opt,pp}$ ) is the most influential parameter at low temperatures (< 6 °C). When the water temperature exceeds 20 °C, the importance of the respiration of maintenance ( $R_{m,pp}$ ) reduces drastically. For more detailed information about the sensitivity analysis of C-RIVE model, one can consult Wang et al. (2018).

## Appendix B. Importance sampling and definition of the importance weights in particle filtering

We suppose here that we want to simulate a random variable  $\mathbf{U}$  which may have a complex distribution. The importance sampling method uses an instrumental distribution called the importance distribution  $\pi(\cdot)$  from which we know how to draw samples (Särkkä, 2013). The importance sampling relies on the following formula:

$$\int f_{\mathbf{U}}(\mathbf{u}) d\mathbf{u} = \int \left[ \frac{f_{\mathbf{U}}(\mathbf{u})}{\pi_{\mathbf{U}}(\mathbf{u})} \right] \pi_{\mathbf{U}}(\mathbf{u}) d\mathbf{u} \quad (\text{B1})$$

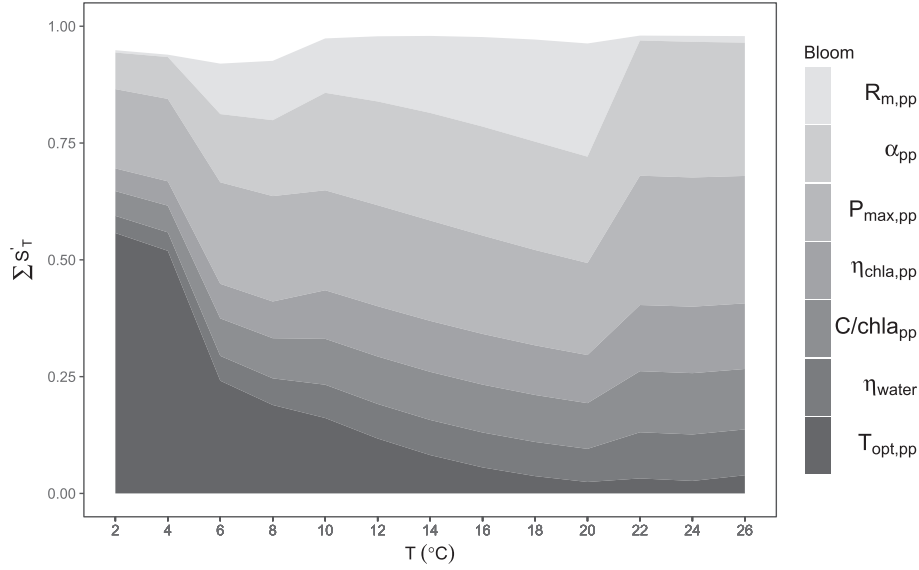
where  $\pi_{\mathbf{U}}(\mathbf{u})$  is a proposal pdf for the random variable  $\mathbf{U}$  from which we can draw samples and  $f_{\mathbf{U}}(\mathbf{u})$  is the posterior pdf of given state  $\mathbf{U}$  from which we cannot directly or efficiently draw samples. The equation (B.1) transforms the expectation of  $\mathbf{U}$  into the expectation of the term  $\left[ \frac{f_{\mathbf{U}}(\mathbf{u})}{\pi_{\mathbf{U}}(\mathbf{u})} \right]$ . We can perform Monte Carlo approximation to estimate the expectation of  $g(\mathbf{U})$ , denoted  $\mathbf{E}(g(\mathbf{U}))$ , for any function  $g$ :

$$\begin{aligned} \mathbf{E}(g(\mathbf{U})) &= \int \left[ g(\mathbf{u}) \frac{f_{\mathbf{U}}(\mathbf{u})}{\pi_{\mathbf{U}}(\mathbf{u})} \right] \pi_{\mathbf{U}}(\mathbf{u}) d\mathbf{u} \\ &\approx \frac{1}{N} \sum_{i=1}^N g(\mathbf{u}^i) \frac{f_{\mathbf{U}}(\mathbf{u}^i)}{\pi_{\mathbf{U}}(\mathbf{u}^i)} \\ \omega^i &= \frac{1}{N} \frac{f_{\mathbf{U}}(\mathbf{u}^i)}{\pi_{\mathbf{U}}(\mathbf{u}^i)} \end{aligned} \quad (\text{B2})$$

where  $\mathbf{u}^i$  ( $i = 1, \dots, N$ ) is *i*th sample (particle) drawn from the importance distribution  $\pi_{\mathbf{U}}(\mathbf{u})$  and  $\omega^i$  denotes the weight



associated with  $i$ th sample (particle). With particular choices of  $g$ , we can hence retrieve the expectation of  $\mathbf{U}$ , its variance, quantiles, etc.



**Fig. A1.** : The evolution of the total sensitivity indices (normalised to 100%) with temperature for Bloom condition (From bottom to top:  $T_{opt,pp}$ ,  $\eta_{water}$ ,  $C/Chla_{pp}$ ,  $P_{max,pp}$ ,  $\alpha_{pp}$ ,  $R_{m,pp}$ ). See Table 1 for parameter definition. For more detailed information about sensitivity analysis, one can consult Wang et al. (2018).

We can use importance sampling to address the problem of sampling directly from the posterior pdf  $f(\mathbf{z}_{1:t}|\mathbf{y}_{1:t}^*)$  of the trajectory  $\mathbf{Z}_{1:t}$ . Let the importance distribution,  $\pi(\mathbf{z}_{1:t}|\mathbf{y}_{1:t}^*)$ , from equation (7) we can compute importance weights,

$$\omega_t = \frac{f(\mathbf{z}_{1:t}|\mathbf{y}_{1:t}^*)}{\pi(\mathbf{z}_{1:t}|\mathbf{y}_{1:t}^*)} \propto \frac{f(\mathbf{y}_t^*|\mathbf{z}_t)f(\mathbf{z}_t|\mathbf{z}_{t-1})f(\mathbf{z}_{1:t-1}|\mathbf{y}_{1:t-1}^*)}{\pi(\mathbf{z}_{1:t}|\mathbf{y}_{1:t}^*)} \quad (\text{B3})$$

Then, assuming the importance distribution  $\pi(\mathbf{z}_{1:t}|\mathbf{y}_{1:t}^*)$  to be Markovian, that is

$$\begin{aligned} \pi(\mathbf{z}_{1:t}|\mathbf{y}_{1:t}^*) &= \pi(\mathbf{z}_t|\mathbf{z}_{1:t-1}, \mathbf{y}_{1:t}^*)\pi(\mathbf{z}_{1:t-1}|\mathbf{y}_{1:t-1}^*) \\ &= \pi(\mathbf{z}_t|\mathbf{z}_{t-1}, \mathbf{y}_t^*)\pi(\mathbf{z}_{1:t-1}|\mathbf{y}_{1:t-1}^*), \end{aligned} \quad (\text{B4})$$

then equation (B.3) becomes,

$$\begin{aligned} \omega_t &\propto \frac{f(\mathbf{y}_t^*|\mathbf{z}_t)f(\mathbf{z}_t|\mathbf{z}_{t-1})f(\mathbf{z}_{1:t-1}|\mathbf{y}_{1:t-1}^*)}{\pi(\mathbf{z}_t|\mathbf{z}_{t-1}, \mathbf{y}_t^*)\pi(\mathbf{z}_{1:t-1}|\mathbf{y}_{1:t-1}^*)} \\ &\propto \frac{f(\mathbf{y}_t^*|\mathbf{z}_t)f(\mathbf{z}_t|\mathbf{z}_{t-1})}{\pi(\mathbf{z}_t|\mathbf{z}_{t-1}, \mathbf{y}_t^*)}\omega_{t-1} \end{aligned} \quad (\text{B5})$$

In other words, we derive an update formula for the importance weights in the same form as the update formula for the posterior pdf (Eq. (7)). If we can draw  $N$  samples (particles) from the importance distribution  $\pi(\mathbf{z}_t|\mathbf{y}_{1:t}^*)$ ,

$$\mathbf{z}_t^i \sim \pi(\mathbf{z}_t|\mathbf{y}_{1:t}^*) \quad i = 1, \dots, N \quad (\text{B6})$$

a weight update formula for each particle can be written as follows,

$$\begin{aligned} \omega_t^i &\propto \frac{f(\mathbf{y}_t^*|\mathbf{z}_t^i)f(\mathbf{z}_t^i|\mathbf{z}_{t-1}^i)}{\pi(\mathbf{z}_t^i|\mathbf{z}_{t-1}^i, \mathbf{y}_t^*)}\omega_{t-1}^i \\ \omega_t^i &= \frac{\omega_t^i}{\sum \omega_t^i} \end{aligned} \quad (\text{B7})$$

where we define the normalised importance weight ( $\hat{\omega}_t^i$ ) associated with the particle  $i$ . Normalizing the weights allows to evaluate the problem of the normalizing constant in the Bayesian representation of the posterior ((5)). Equation (B.7) shows that we only need to store the state  $\mathbf{z}_t$  at time  $t$ . In addition, it is usually not necessary to estimate the full pdf  $f(\mathbf{z}_{1:t}|\mathbf{y}_{1:t}^*)$  in practice. In this work, we are interested in the marginal distribution  $f(\mathbf{z}_t|\mathbf{y}_{1:t}^*)$  at time  $t$ , called filtering distribution. The filtering posterior pdf  $f(\mathbf{z}_t|\mathbf{y}_{1:t}^*)$  at time  $t$  can be approximated as,

$$f(\mathbf{z}_t|\mathbf{y}_{1:t}^*) \approx \sum_{i=1}^N \hat{\omega}_t^i \delta(\mathbf{z}_t - \mathbf{z}_t^i) \quad (\text{B.8})$$

where  $\delta(\cdot)$  is a Dirac delta function.

The performance of the above algorithm depends on the quality of the importance distribution  $\pi(\cdot)$ . Typically, the optimal importance distribution is (Doucet et al., 2001; Särkkä, 2013),

$$\pi(\mathbf{z}_t|\mathbf{z}_{t-1}, \mathbf{y}_t^*) = f(\mathbf{z}_t|\mathbf{z}_{t-1}, \mathbf{y}_t^*) \quad (\text{B9})$$

In practice, however we generally do not know the distribution  $f(\mathbf{z}_t|\mathbf{z}_{t-1}, \mathbf{y}_t^*)$  and we rather propagate  $\mathbf{z}_t$  using the evolution equations (1) and (2), which terms to draw samples from  $f(\mathbf{z}_t|\mathbf{z}_{t-1})$ . In that case, the equations (B.5) and (B.7) simplifies into

$$\omega_t \propto f(\mathbf{y}_t^*|\mathbf{z}_t)\omega_{t-1} \quad (\text{B10})$$

## References

- Abbaszadeh, P., Moradkhani, H., Yan, H., 2018. Enhancing hydrologic data assimilation by evolutionary Particle Filter and Markov chain Monte Carlo. *Adv. Water Resour.* 111, 192–204.
- Andreadis, K., Clark, E., Lettermaier, D., Alsdorf, D., 2007. Prospects for river discharge and depth estimation through assimilation of swath-altimetry into a raster-based hydrodynamics model. *Geophys. Res. Lett.* 34, L10403.
- Arulampalam, M.S., Maskell, S., Gordon, N., Clapp, T., 2002. A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking. *IEEE Trans. Signal Process.* 50 (2), 174–188.
- Bayes, T., 1763. An essay towards solving a problem in the doctrine of chances. *Philos. Trans. R. Soc. Lond.* 53, 370–418.
- Beck, M., 1987. Water quality modelling : a review of the analysis of uncertainty. *Water Resour. Res.* 23 (8), 1393–1442.
- Beven, K., 1989. Changing ideas in hydrology. The case of physically-based model. *J. Hydrol.* 105, 157–172.
- Billen, G., Garnier, J., Hanset, P., 1994. Modelling phytoplankton development in whole drainage networks: the RIVERSTRAHLER Model applied to the Seine river system. *Hydrobiologia* 289, 119–137.
- Cappe, O., Godsill, S.J., Moulines, E., 2007. An overview of existing methods and recent advances in sequential Monte Carlo. *Proc. IEEE* 95 (5), 899–924.
- Courtier, P., Andersson, E., Heckley, W., Vasiljevic, D., Hamrud, M., Hollingsworth, A., Rabier, F., Fisher, M., Pailleux, J., 1998. The ECMWF implementation of three-dimensional variational assimilation (3D-Var). I: Formulation. *Q. J. R. Meteorol. Soc.* 124 (550), 1783–1807.
- Courtier, P., Thépaut, J.N., Hollingsworth, A., 1994. A strategy for operational implementation of 4D-Var, using an incremental approach. *Q. J. R. Meteorol. Soc.* 120 (519), 1367–1387.
- DeChant, C.M., Moradkhani, H., 2012. Examining the effectiveness and robustness of sequential data assimilation methods for quantification of uncertainty in hydrologic forecasting. *Water Resour. Res.* 48 (4).
- Descy, J.-P., Leita, M., Everbecq, E., Smitz, J.-S., Deliége, J.-F., 2012. Phytoplankton of the River Loire, France: a biodiversity and modelling study. *J. Plankton Res.* 34 (2), 120–135.
- Doucet, A., de Freitas, N., Gordon, N., 2001. *Sequential Monte Carlo Methods in Practice*. Springer.
- Doucet, A., Godsill, S., Andrieu, C., 2000. On sequential Monte Carlo sampling methods for bayesian filtering. *Stat. Comput.* 10 (3), 197–208.
- Even, S., Bacq, N., Ruelland, D., Billen, G., Garnier, J., Poulin, M., Théry, S., Blanc, S., 2007a. New tools for modelling water quality of hydrosystems: an application in the Seine River basin in the frame of the Water Framework Directive. *Sci. Total Environ.* 375 (1–3), 274–291.
- Even, S., Mouchel, J.M., Servais, P., Flipo, N., Poulin, M., Blanc, S., Chabanel, M., Paffoni, C., 2007b. Modeling the impacts of combined sewer overflows on the river seine water quality. *Sci. Total Environ.* 375 (1–3), 140–151.
- Even, S., Poulin, M., Garnier, J., Billen, G., Servais, P., Chesterikoff, A., Coste, M., 1998. River ecosystem modelling: application of the ProSe model to the Seine river (France). *Hydrobiologia* 373, 27–37.
- Even, S., Poulin, M., Mouchel, J.M., Seidl, M., Servais, P., 2004. Modelling oxygen deficits in the Seine river downstream of combined sewer overflows. *Ecol. Model.* 173, 177–196.
- Evensen, G., 1994. Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics. *J. Geophys. Res.: Oceans* 99 (C5), 10143–10162.
- Evensen, G., Nov 2003. The Ensemble Kalman Filter: theoretical formulation and practical implementation. *Ocean Dyn.* 53 (4), 343–367.
- Flipo, N., Even, S., Poulin, M., Tusseau-Vuillemin, M.H., Améziane, T., Dauta, A., 2004. Biogeochemical modelling at the river scale: plankton and periphyton dynamics - grand Morin case study. *France. Ecol. Model.* 176, 333–347.
- Flipo, N., Rabouille, C., Poulin, M., Even, S., Tusseau-Vuillemin, M., Lalande, M., 2007. Primary production in headwater streams of the Seine basin: the Grand Morin case study. *Sci. Total Environ.* 375, 98–109.
- Garnier, J., Billen, G., Coste, M., 1995. Seasonal succession of diatoms and chlorophyceae in the drainage network of the river Seine: observations and modelling. *Limnol. Oceanogr.* 40 (4), 750–765.
- Garnier, J., Marescaux, A., Guillon, S., Vilmin, L., Rocher, V., Billen, G., Thieu, V., Silvestre, M., Passy, P., Raimonet, M., Groleau, A., Théry, S., Tallec, G., Flipo, N., 2019. The Seine River Basin. *Handbook of Environmental Chemistry*. Springer. Ch. Ecological functioning of the Seine River: from long-term modelling approaches to high-frequency data analysis, p. (in press).
- Gauthier, P., Tanguay, M., Laroche, S., Pellerin, S., Morneau, J., 2007. Extension of 3DVAR to 4DVAR: implementation of 4DVAR at the meteorological service of Canada. *Mon. Weather Rev.* 135 (6), 2339–2354.
- Gharanti, M., Tjiputra, J., Bethke, I., Samuelsen, A., Skjelvan, I., Bentsen, M., Bertino, L., 2017. Ensemble data assimilation for ocean biogeochemical state and parameter estimation at different sites. *Ocean Model.* 112, 65–89.
- Gupta, H.V., Kling, H., Yilmaz, K.K., Martinez, G.F., 2009. Decomposition of the mean squared error and nse performance criteria: implications for improving hydrological modelling. *J. Hydrol.* 377 (1), 80–91.
- Huang, J., Gao, J., Liu, J., Zhang, Y., 2013. State and parameter update of a hydrodynamic-phytoplankton model using ensemble Kalman filter. *Ecol. Model.* 263, 81–91.
- Kalman, R.E., 1960. A new approach to linear filtering and prediction problems. *J. Basic Eng.* 82 (1), 35–45.
- Kattwinkel, M., Reichert, P., 2017. Bayesian parameter inference for individual-based models using a Particle Markov chain Monte Carlo method. *Environ. Model. Softw.* 87, 110–119.
- Kim, K., Park, M., Min, J., Ryu, I., Kang, M., Park, L.J., 2014. Simulation of algal bloom dynamics in a river with the ensemble Kalman filter. *J. Hydrol.* 519, 2810–2821.
- Kitagawa, G., 1996. Monte Carlo filter and smoother for non-Gaussian nonlinear state space models. *J. Comput. Graph. Stat.* 5 (1), 1–25.
- Kleist, D.T., Ide, K., 2015. An OSSE-based evaluation of hybrid variational-ensemble data assimilation for the NCEP GFS. Part II: 4D-EnVar and hybrid variants. *Mon. Weather Rev.* 143 (2), 452–470.
- Kling, H., Fuchs, M., Paulin, M., 2012. Runoff conditions in the upper danube basin under an ensemble of climate change scenarios. *J. Hydrol.* 424–425, 264–277.
- Kong, A., Liu, J.S., Wong, W.H., 1994. Sequential imputations and Bayesian missing data problems. *J. Am. Stat. Assoc.* 89 (425), 278–288.
- Li, T., Villarrubia, G., Sun, S., Corchado, J.M., Bajo, J., 2015. Resampling methods for particle filtering: identical distribution, a new method, and comparable study. *Front. Inf. Technol. Elect. Eng.* 16 (11), 969–984.
- Liu, J.S., 2001. *Monte Carlo Strategies in Scientific Computing*. Springer.
- Mao, J., Lee, J.H., Choi, K., 2009. The extended Kalman filter for forecast of algal bloom dynamics. *Water Res.* 43 (17), 4214–4224.
- Markov, A.A., 1906. Extension of the Law of Large Numbers to Dependent Quantities (In Russian). *Izvestiia Fiz.-Matem. Obsch.*, vol. 15. Kazan Univ., pp. 135–156 (2nd Ser.).
- Moradkhani, H., Hsu, K., Gupta, H., Sorooshian, S., 2005. Uncertainty assessment of hydrologic model states and parameters: sequential data assimilation using the particle filter. *Water Resour. Res.* 41 (5).
- Page, T., Smith, P.J., Beven, K.J., Jones, I.D., Elliott, J.A., Maberly, S.C., Mackay, E.B., Ville, M.D., Feuchtmayr, H., 2018. Adaptive forecasting of phytoplankton communities. *Water Res.* 134, 74–85.
- Pasetto, D., Camporese, M., Putti, M., 2012. Ensemble Kalman filter versus particle filter for a physically-based coupled surface/subsurface model. *Adv. Water Resour.* 47, 1–13.
- Pastres, R., Ciavatta, S., Solidoro, C., 2003. The extended Kalman Filter (EKF) as a tool for the assimilation of high frequency water quality data. *Ecol. Model.* 170 (2), 227–235.
- Pearson, K., 1905. The problem of the random walk. *Nature* 72 (1865), 294.
- Plaza, D.A., De Keyser, R., De Lannoy, G.J.M., Giustarini, L., Matgen, P., Pauwels, V.R.N., 2012. The importance of parameter resampling for soil moisture data assimilation into hydrologic models using the particle filter. *Hydrol. Earth Syst. Sci.* 16 (2), 375–390.
- Polus, E., Flipo, N., de Fouquet, C., Poulin, M., 2011. Geostatistics for assessing the efficiency of distributed physically-based water quality model. Application to nitrates in the Seine River. *Hydrol. Process.* 25 (2), 217–233.
- Raimonet, M., Vilmin, L., Flipo, N., Rocher, V., Laverman, A., 2015. Modelling the fate of nitrite in an urbanized river using experimentally obtained nitrifier growth parameters. *Water Res.* 73, 373–387.
- Rocher, V., Garcia-Gonzalez, E., Paffoni, C., Thomas, W., 2011. La production de nitrates lors de la dénitrification des eaux usées: un sujet sensible et complexe ! *Eau, Ind., Les. Nuisances* 344, 80–83.
- Salamon, P., Feyen, L., 2009. Assessing parameter, precipitation, and predictive uncertainty in a distributed hydrological model using sequential data assimilation with the particle filter. *J. Hydrol.* 376, 428–442.
- Särkkä, S., 2013. *Bayesian Filtering and Smoothing*. Cambridge University Press.
- Sasaki, Y., 1955. A fundamental study of the numerical prediction based on the variational principle. *Journal of the Meteorological Society of Japan* 33 (6), 262–275. Ser. II.
- Sasaki, Y., 1958. An objective analysis based on the variational method. *Math. Model. Civ. Eng.* 36 (3), 77–88. Ser. II.
- Shi, Y., Davis, K.J., Zhang, F., Duffy, C.J., Yu, X., 2014. Parameter estimation of a physically based land surface hydrologic model using the ensemble Kalman filter: a synthetic experiment. *Water Resour. Res.* 50 (1), 706–724.
- Simon, E., Bertino, L., 2012. Gaussian anamorphosis extension of the denkf for combined state parameter estimation: application to a 1d ocean ecosystem model. *J. Mar. Syst.* 89 (1), 1–18.
- Simon, E., Samuelsen, A., Bertino, L., Dumont, D., 2012. Estimation of positive sum-to-one constrained zooplankton grazing preferences with the denkf: a twin experiment. *Ocean Sci.* 8 (4), 587–602. <https://www.ocean-sci.net/8/587/2012/>.
- Sobol, I., 1993. Sensitivity estimates for on linear mathematical models. *Math. Model. Civ. Eng.* 1.
- Vilmin, L., 2014. *Modélisation du fonctionnement biogéochimique de la seine de l'agglomération parisienne à l'estuaire à différentes échelles temporelles*. Ph.D. thesis. MINES ParisTech.
- Vilmin, L., Aissa-Grouz, N., Garnier, J., Billen, G., Mouchel, J.M., Poulin, M., Flipo, N., 2015a. Impact of hydro-sedimentary processes on the dynamics of soluble reactive phosphorus in the Seine River. *Biogeochemistry* 122, 229–251.
- Vilmin, L., Flipo, N., de Fouquet, C., Poulin, M., 2015b. Pluri-annual sediment budget in a navigated river system: the Seine River (France). *Sci. Total Environ.* 502, 48–59.
- Vilmin, L., Flipo, N., Escoffier, N., Groleau, A., 2018. Estimation of the water quality of a large urbanized river as defined by the european WFD: what is the optimal sampling frequency? *Environ. Sci. Pollut. Control Ser.* 25 (24), 23485–23501.
- Vilmin, L., Flipo, N., Escoffier, N., Rocher, V., Groleau, A., 2016. Carbon fate in a large temperate human-impacted river system: focus on benthic dynamics. *Glob. Biogeochem. Cycles* 30 (7), 1086–1104.

- Vrugt, J.A., ter Braak, C.J., Diks, C.G., Schoups, G., 2013. Hydrologic data assimilation using particle Markov chain Monte Carlo simulation: theory, concepts and applications. *Adv. Water Resour.* 51, 457–478.
- Wang, S., Flipo, N., Romary, T., 2018. Time-dependent global sensitivity analysis of the C-RIVE biogeochemical model in contrasted hydrological and trophic contexts. *Water Res.* 144, 341–355.
- Weerts, A.H., El Serafy, G.Y.H., 2006. Particle filtering and ensemble Kalman filtering for state updating with hydrological conceptual rainfall-runoff models. *Water Resour. Res.* 42 (9).
- Wikle, C.K., Berliner, L.M., 2007. A Bayesian tutorial for data assimilation. *Phys. D Nonlinear Phenom.* 230 (1), 1–16.
- Xue, P., Chen, C., Beardsley, R.C., 2012. Observing system simulation experiments of dissolved oxygen monitoring in Massachusetts Bay. *J. Geophys. Res.: Oceans* 117 (C5).
- Yu, L., Fennel, K., Bertino, L., Gharamti, M.E., Thompson, K.R., 2018. Insights on multivariate updates of physical and biogeochemical ocean variables using an ensemble kalman filter and an idealized model of upwelling. *Ocean Model.* 126, 13–28.
- Yucel, I., Onen, A., Yilmaz, K., Gochis, D., 2015. Calibration and evaluation of a flood forecasting system: utility of numerical weather prediction model, data assimilation and satellite-based rainfall. *J. Hydrol.* 523, 49–66.