UNIVERSITÉ TOULOUSE III PAUL SABATIER

INSTITUT DE MATHÉMATIQUES DE TOULOUSE

MÉMOIRE D'HABILITATION À DIRIGER DES RECHERCHES

Discipline : **Mathématiques**

Spécialité : **Statistiques**

Présentée et soutenue publiquement

le 21 janvier 2009 à Toulouse par

Bertrand IOOSS

Ingénieur-chercheur au CEA

Direction de l'Énergie Nucléaire

Centre de Cadarache

Sujet du mémoire

CONTRIBUTIONS AU TRAITEMENT DES INCERTITUDES EN MODÉLISATION NUMÉRIQUE :

PROPAGATION D'ONDES EN MILIEU ALÉATOIRE ET ANALYSE STATISTIQUE D'EXPÉRIENCES SIMULÉES

Jury

M. Anestis ANTONIADIS	Université Grenoble I	Rapporteur
M. Mark ASCH	Université de Picardie	Rapporteur
M. Jean-Marc AZAÏS	Université Toulouse III	Président
M. Fabrice GAMBOA	Université Toulouse III	Examinateur
M. Michel SCHMITT	École des Mines de Paris	Examinateur
M. Michael L. STEIN	University of Chicago, USA	Rapporteur
M. Stefano TARANTOLA	Joint Research Centre, Ispra, Italie	Examinateur

Résumé

Contributions au traitement des incertitudes en modélisation numérique : propagation d'ondes en milieu aléatoire et analyse d'expériences simulées

Le présent document constitue mon mémoire d'habilitation à diriger des recherches. Il retrace mon activité scientifique de ces douze dernières années, depuis ma thèse jusqu'aux travaux réalisés en tant qu'ingénieur-chercheur du CEA Cadarache. Les deux chapitres qui structurent ce document correspondent à deux domaines de recherche relativement différents mais se référant tous les deux au traitement des incertitudes dans des problèmes d'ingénierie. Le premier chapitre établit une synthèse de mes travaux sur la propagation d'ondes hautes fréquences en milieu aléatoire. Il concerne plus spécifiquement l'étude des fluctuations statistiques des temps de trajet des ondes acoustiques en milieu aléatoire et/ou turbulent. Les nouveaux résultats obtenus concernent principalement l'introduction de l'anisotropie statistique des champs de vitesse lors de la dérivation des expressions des moments des temps en fonction de ceux du champ de vitesse des ondes. Ces travaux ont été essentiellement portés par des besoins en géophysique (exploration pétrolière et sismologie). Le second chapitre aborde le domaine de l'utilisation des techniques probabilistes pour prendre en compte les incertitudes des variables d'entrée d'un modèle numérique. Les principales applications que j'évoque dans ce chapitre relèvent du domaine de l'ingénierie nucléaire qui offre une grande variété de problématiques d'incertitude à traiter. Tout d'abord, une synthèse assez complète est réalisée sur les méthodes statistiques d'analyse de sensibilité et d'exploration globale de modèles numériques. La construction et l'exploitation d'un métamodèle (fonction mathématique peu coûteuse se substituant à un code de calcul coûteux) sont ensuite illustrées par mes travaux sur le modèle processus gaussien (krigeage). Deux thématiques complémentaires sont finalement abordées : l'estimation de quantiles élevés de réponses de codes de calcul et l'analyse de codes de calcul stochastiques. Une conclusion met en perspective ces travaux dans le contexte plus général de la simulation numérique et de l'utilisation de modèles prédictifs dans l'industrie.

Abstract

Contributions to the uncertainty management in numerical modelisation : wave propagation in random media and analysis of computer experiments

The present document constitutes my habilitation thesis report. It recalls my scientific activity of the twelve last years, since my PhD thesis until the works completed as a research engineer at CEA Cadarache. The two main chapters of this document correspond to two different research fields both referring to the uncertainty treatment in engineering problems. The first chapter establishes a synthesis of my work on high frequency wave propagation in random medium. It more specifically relates to the study of the statistical fluctuations of acoustic wave traveltimes in random and/or turbulent media. The new results mainly concern the introduction of the velocity field statistical anisotropy in the analytical expressions of the traveltime statistical moments according to those of the velocity field. This work was primarily carried by requirements in geophysics (oil exploration and seismology). The second chapter is concerned by the probabilistic techniques to study the effect of input variables uncertainties in numerical models. My main applications in this chapter relate to the nuclear engineering domain which offers a large variety of uncertainty problems to be treated. First of all, a complete synthesis is carried out on the statistical methods of sensitivity analysis and global exploration of numerical models. The construction and the use of a metamodel (inexpensive mathematical function replacing an expensive computer code) are then illustrated by my work on the Gaussian process model (kriging). Two additional topics are finally approached : the high quantile estimation of a computer code output and the analysis of stochastic computer codes. We conclude this memory with some perspectives about the numerical simulation and the use of predictive models in industry. This context is extremely positive for future researches and application developments.

Remerciements

Entre Toulouse, Solaize, Paris, La Havane, Caracas, Medellin, Pékin et Oulan Bator, Fabrice Gamboa a réussi l'exploit de trouver le temps de superviser mon mémoire. Même s'il l'a lu couché dans sa yourte, en mangeant du fromage de lait de yak fermenté en trinquant avec Jean-Claude autour de quelques verres de vodka mongole, je lui suis infiniment reconnaissant de ses critiques et conseils. Merci aussi pour ton dynamisme communicatif, ton ouverture d'esprit exceptionnelle et surtout pour avoir accepté spontanément de soutenir un dossier quelque peu particulier. L'un de mes souhaits dans le futur serait que l'on puisse travailler ensemble autour de thématiques de recherche ambitieuses.

Merci à Anestis Antoniadis d'avoir rapporté cette habilitation, et d'en être en quelque sorte un peu responsable (mais pas coupable) en étant le premier a m'avoir donné un avis (positif) sur la recevabilité de mon dossier. Ce coup de pouce m'a décidé à me lancer dans l'aventure. Je tiens donc à te remercier pour ton intérêt pour ce sujet, pour ton extrême compétence, mais surtout pour ta générosité. J'exprime également ma reconnaissance à Mark Asch qui a accepté de rapporter ce mémoire. Lors de notre première rencontre à Cadarache en 2003, j'avais bien perçu ton intérêt pour les problèmes issus de l'industrie. Les appréciations que tu as portées sur mon travail m'ont vraiment fait plaisir. Furthermore, I am glad to thank Michael Stein as a reviewer of this thesis. Your criticisms about my works are of great value for me and I hope to meet you soon. Je tiens également à remercier Jean-Marc Azaïs, grand promoteur de collaborations universités-industries, qui m'a permis de soutenir au sein de l'école doctorale de Toulouse et qui a accepté de présider mon jury. Je remercie Michel Schmitt qui m'aura finalement honoré de sa présence au sein des jurys de mes deux soutenances : thèse et habilitation. Cette présence permet d'assurer une certaine continuité entre mes travaux récents et mes travaux de jeunesse initiés par les chercheurs de l'École des Mines de Paris. Enfin, je remercie chaleureusement Stefano Tarantola d'avoir accepté de s'intéresser à mes travaux. Le Sue qualità umane e la sua experienza nel dominio dell'analisi di sensibilità sono un esempio per me (merci `reverso.net`).

Je souhaite à présent remercier tout particulièrement mon ex sans qui ce projet d'habilitation n'aurait pas vu le jour, j'ai nommé Nicolas Devictor, ex-chef du laboratoire où j'officie. Le positionnement de notre équipe au sein du CEA, les études et les sujets de R&D que tu m'as proposés ont créé un contexte extrêmement favorable pour mon travail. Cette habilitation provient, pour une grande part, de ton management motivant, ouvert aux initiatives et incitatif. Je n'oublie pas non plus ce que je dois à ma collègue de bureau, Nadia Pérot, qui m'a apporté des avis constructifs et pertinents dès que je la sollicitais, mais surtout un soutien moral sans faille accompagné de petits macarons bien délicieux. Notre binôme, issu de six ans de vie commune (sans un heurt), prouve que dans la vie professionnelle $\mathbb{P}(1+1>2)>0$ et force probablement l'admiration de tous. Et puis, last but not least, je m'incline devant le senior (je devrais dire seigneur) de l'équipe, Michel Marquès, dont la gentillesse légendaire n'a d'égale que sa modestie.

Le soutien, parfois malicieux (spécial dédicace à Bernard qui me remet à ma place à chaque ascension ventousienne), de tous les membres du Laboratoire de Conduite et de Fiabilité des Réacteurs (LCFR) du CEA Cadarache m'a été indispensable et je les en remercie grandement. Plus particulièrement, les remarques constructives sur le mémoire et le soutien logistique de Frédéric Bertrand, chef du LCFR, ont été très appréciables. Je remercie aussi Jean-Claude Garnier, chef du Service d'Etudes des Systèmes Innovants, et Alain Porracchia, chef du Département d'Étude des Réacteurs, pour leur appui dans mes démarches. Ah, en passant, il ne faut pas que j'oublie de citer les anciens de la fneube 2 (Obog, Dave, Anne-Cé, Jé, Lo, Fred, Ben et son chien, . . .), sinon ils ne voudront plus de moi à leur table. Sans vous, j'aurai probablement sombrer en grande dépression post-nucléaire à l'automne 2003.

Le contenu de ce mémoire ne serait pas le même sans l'apport de toutes les personnes avec lesquelles j'ai travaillé de manière étroite et qui m'ont permis d'avancer sur de nombreux problèmes et de me lancer dans de nouvelles voies de recherche. Je citerai d'abord les petits jeunots, qui m'ont bien aidé et avec qui j'ai bien rigolé, que sont Vincent Feuillard, Claire Cannamela et Gilles Pujol, et ensuite tous les autres : Josselin Garnier, Roger Phan Tan Luu, Julien Jacques, Béatrice Laurent, François Van Dorpe, Elena Volkova, Michel Jullien, Mustafa Touati, Yann Samuelides, David Geraets, Philippe Blanc-Benon sans oublier tous les stagiaires que j'ai tenté d'encadrer. Je porte un toast plus soutenu à deux autres gamins : Amandine Marrel et Mathieu Ribatet. Certains des travaux présentés dans ce mémoire vous doivent énormément et votre disponibilité est un modèle du genre. A mon avis et vu mon envie, on ne tardera pas à retravailler rapidement ensemble. Je souligne également l'influence d'Alain Galli et de Professeur Lulu, sans qui rien ne serait arrivé.

L'aspect générique et transverse de mon domaine de recherche m'a permis de côtoyer des ingénieurs et chercheurs extrêmement compétents et passionnants. Je remercie tout d'abord les collègues avec qui j'ai collaboré sur le traitement des incertitudes à la Direction de l'Energie Nucléaire du CEA : Fabrice Gaudier, Agnès de Crecy (et son inséparable Pascal), Cyrille de Saint Jean (et son inséparable Gilles) . . . , et surtout Jean-Marc Martinez qui m'a permis de développer ma pugnacité. Je félicite aussi les membres du groupe de travail "Incertitudes" de l'IMdR (Etienne de Rocquigny, Fabien Mangeant, Erick Herbin, Nicolas Fischer, Yann Richet, Guennadi Andrianov, Olivier Vasseur, Eric Chojnacki, j'en oublie ?) qui m'ont soutenu de manière exemplaire lors de récentes défaillances gastriques. J'applaudis enfin les membres du bureau du GdR MASCOT-NUM (Les 3 F, Luc, Jean-Claude, Hervé et consors) qui me permettent de participer à de belles actions communes.

Pour finir, parlons de la VIE EN DEHORS DU BOULOT, même si j'admet que parfois j'ai tendance à l'oublier (*mea culpa*, je sens que je vais me rattraper). Alors je lance un cri du coeur à mes amis et aux monuments niçois qui étaient déjà dans mes remerciements de thèse il y a tout juste 10 ans : Dume, Housse, Titou, Stéph, Alain, leurs conjointes, la pissaladière, la socca, l'OGCN et les Dum Dum Boys. En parlant de music à tendance noisy-punk, Fatal Error ayant laissé la place à Porcheline, merci à mes deux musiciens cadarachiens, Arno et Jief, qui me permettent de m'exprimer, parfois violemment, mais toujours dans un micro ou en frappant/poussant/tirant sur une belle corde métallique qui ne demande que ça. En parlant de musiciens, ça me fait penser à Vincent, et donc à Fabienne, qui nous ont été d'un grand secours dans la dernière ligne droite : encore merci pour le baby-sitting prolongé. Et puis, je rend hommage à mon paternel qui est probablement la source de ma motivation première. J'ai une pensée permanente pour Momo, pour mes deux zozios, Orlane et Emeric, sans qui rien n'aurait de sens et pour ma douce, SoΦ qui me supporte avec tant de courage.

Table des matières

Table des figures

Liste des travaux

MONOGRAPHIE (en préparation)

[I1] B. Iooss, H. Monod, G. Pujol and C. Storlie. *Sensitivity Analysis with R.* En préparation.

ARTICLES de revues à comité de lecture

[Ia1] B. Iooss. Seismic reflection traveltimes in two-dimensional statistically anisotropic random media. *Geophysical Journal International,* 135 :999-1010, 1998.

[Ia2] M. Touati, B. Iooss and A. Galli. Quantitative control of migration : a geostatistical attempt. *Mathematical Geology,* 31 :277-295, 1999.

[Ia3] B. Iooss, Ph. Blanc-Benon and C. Lhuillier. Statistical moments of travel times at second order in isotropic and anisotropic random media. *Waves in Random Media,* 10 :381-394, 2000.

[Ia4] B. Iooss, C. Lhuillier and H. Jeanneau. Numerical simulation of transit-time ultrasonic flowmeters : uncertainties due to fluid turbulence. *Ultrasonics,* 40 :1009-1015, 2002.

[Ia5] B. Iooss, D. Geraets, T. Mukerji, Y. Samuelides, M. Touati and A. Galli. Inferring the statistical distribution of velocity heterogeneities by statistical traveltime tomography. *Geophysics,* 68(5) :1714-1730, 2003.

[Ia6] B. Iooss, F. Van Dorpe and N. Devictor. Response surfaces and sensitivity analyses for an environmental model of dose calculations. *Reliability Engineering and System Safety,* 91 :1241-1251, 2006.

[Ia7] F. Van Dorpe, B. Iooss, V. Semenov, O. Sorokovikova, A. Fokin and Y. Margerit. Atmospheric transfer modeling with 3D Lagrangian dispersion codes compared with SF6 tracer experiments at regional scale. *Science and Technology of Nuclear Installations*, Volume 2007, Article ID 30863, 13 pages, doi :10.1155/2007/30863, 2007.

[Ia8] E. Volkova, B. Iooss and F. Van Dorpe. Global sensitivity analysis for a numerical model of radionuclide migration from the "RRC Kurchatov Institute" radwaste disposal site. *Stochastic Environmental Research and Risk Assessment*, 22 :17-31, 2008.

[Ia9] A. Marrel, B. Iooss, F. Van Dorpe and E. Volkova. An efficient methodology for modeling complex computer codes with Gaussian processes. *Computational Statistics and Data Analysis,* 52 :4731-4744, 2008.

[Ia10] G. Noguere, D. Bernard, C. De Saint Jean, B. Iooss, F. Gunsing, K. Kobayashi, S. Mughabghab and P. Siegler. Assessment and propagation of the ^{237}Np nuclear data uncertainties in integral calculations by Monte Carlo techniques. *Nuclear Science and Engineering,* 160 :108-122, 2008.

[Ia11] C. Cannamela, J. Garnier and B. Iooss. Controlled stratification for quantile estimation. *Annals of Applied Statistics*, 2 :1554-1580, 2008.

[Ia13] B. Iooss and M. Ribatet. Global sensitivity analysis of computer models with functional inputs. *Reliability Engineering and System Safety,* in press, 2009.

[Ia14] A. Marrel, B. Iooss, B. Laurent and O. Roustant. Calculations of Sobol indices for the Gaussian process metamodel. *Reliability Engineering and System Safety,* 94 :742-751, 2009.

[Ia15] C. De Saint Jean, G. Noguere, B. Habert and B. Iooss. A Monte Carlo approach of nuclear model parameters uncertainties propagation. *Nuclear Science and Engineering,* 161 : 363-370, 2009.

Articles de revues soumis

[Ia12] B. Iooss, M. Ribatet and A. Marrel. Global sensitivity analysis of stochastic computer models with generalized additive models. *Technometrics*, submitted, 2006.

CONFÉRENCES

Actes de conférences avec comité de sélection

[Ic1] B. Iooss. Caractérisation probabiliste de réflecteurs en sismique réflexion. Journées de Juin, École des Mines de Paris, Centre de Géostatistique, Fontainebleau, France, juin 1998. *Les Cahiers de Géostatistique,* 6 : 61-73, École des Mines de Paris, 1998.

[Ic2] B. Iooss and Y. Samuelides. Inversion of velocity statistical parameters from traveltimes. *Proceedings of 16th International Congress on Acoustics and 135th Meeting of the Acoustical Society of America,* 2319-2320, Seattle, USA, juin 1998.

[Ic3] B. Iooss, A. Galli and M. Touati. Velocity correlation function estimation from seismic reflection traveltimes. *68th SEG expanded abstract,* 1724-1727, New-Orleans, USA, septembre 1998.

[Ic4] B. Iooss and A. Galli. Statistical tomography for seismic reflection data. 6th International Geostatistics Congress (Geostats' 2000), Cape Town, South Africa, avril 2000. *Geostats' 2000*, Kleingeld W. & Krige D. (eds). [CD-ROM]. s.l. : Geostatistical Association of Southern Africa, 2000.

[Ic5] B. Iooss, N. Devictor and F. Van Dorpe. Response surfaces and sensitivity analyses for an environmental model of dose calculations. K.M. Hanson and F.M. Hemez (eds). *Proceedings of 4th International Conference on Sensitivity Analysis of Model Output,* 260-269, Santa Fe, New Mexico, USA, mars 2004. Los Alamos National Laboratory, 2005.

[Ic6] B. Iooss and M. Ribatet. Analyse de sensibilité globale de modèles numériques à paramètres incontrôlables. *Actes des XXXVIIIièmes Journées de Statistique,* Clamart, France, juin 2006.

[Ic7] C. De Saint Jean, G. Noguere and B. Iooss. Sensitivity and uncertainty studies of average cross section parameters with Monte-Carlo sampling. *Proceedings of PHYSOR-2006,* Vancouver, Canada, septembre 2006.

[Ic8] G. Noguere, C. De Saint Jean, B. Iooss, P. Schilleebeeckx and P. Siegler. Production of multigroup data covariance in the resonance range by Monte-Carlo calculations. *Proceedings of PHYSOR-2006,* Vancouver, Canada, septembre 2006.

[Ic9] M. Petelet, O. Asserin, B. Iooss and A. Loredo. Echantillonnage LHS des propriétés matériau des aciers pour l'analyse de sensibilité globale en simulation numérique du soudage. *MATERIAUX 2006,* Dijon, France, novembre 2006.

[Ic10] B. Iooss and M. Ribatet. Global sensitivity analysis of computer models with functional inputs. *5th International Conference on Sensitivity Analysis of Model Output,* Budapest, Hungary, juin 2007.

[Ic11] A. Marrel, B. Iooss and O. Roustant. Analytical calculations of Sobol indices for the Gaussian process metamodel. *5th Intern. Conf. on Sensitivity Analysis of Model Output,* Budapest, Hungary, juin 2007.

[Ic12] M. Petelet, O. Asserin and B. Iooss. Application of global sensitivity analysis method in welding simulation. *5th International Conference on Sensitivity Analysis of Model Output,* Budapest, Hungary, juin 2007.

[Ic13] M. Petelet, O. Asserin, B. Iooss and A. Loredo. Quantification des effets des propriétés matériau sur le résultat du calcul en simulation numérique du soudage. *18ème Congrès Français de Mécanique,* Grenoble, France, août 2007.

[Ic14] B. Iooss, L. Boussouf, A. Marrel and V. Feuillard. Numerical study of algorithms for metamodel construction and validation. S.Martorell, C. Guedes Soares and J. Barnett (eds). *Safety, Reliability and Risk Analysis - Proceedings of the ESREL 2008 Conference,* 2135-2141, CRC Press, Valencia, Espagne, septembre 2008.

[Ic15] B. Auder and B. Iooss. Global sensitivity analysis based on entropy. S.Martorell, C. Guedes Soares and J. Barnett (eds). *Safety, Reliability and Risk Analysis - Proceedings of the ESREL 2008 Conference,* 2107-2115, CRC Press, Valencia, Espagne, septembre 2008.

[Ic16] G. Lorenzo, P. Zanocco, M. Giménez, M. Marquès, B. Iooss, R. Bolado Lavin, F. Pierro, G. Galassi, F. D'Auria and L. Burgazzi. Reliability assessment of the thermal hydraulic phenomena related to a CAREM-like passive RHR system. S.Martorell, C. Guedes Soares and J. Barnett (eds). *Safety, Reliability and Risk Analysis - Proceedings of the ESREL 2008 Conference,* 2899-2907, CRC Press, Valencia, Espagne, septembre 2008.

[Ic17] N. Jeannée, Y. Desnoyers, F. Lamadie and B. Iooss. Geostatistical sampling optimization of contaminated premises. *DEM 2008 - Decommissionning challenges : an industrial reality ?,* Avignon, France, septembre 2008.

[Ic18] N. Pérot and B. Iooss. Quelques problématiques d'échantillonnage statistique pour le démantèlement d'installations nucléaires. *Conférence $\lambda\mu$16,* Avignon, France, octobre 2008.

[Ic19] B. Iooss, M. Marquès, F. Gaudier, B. Spindler and B. Tourniaire. Uncertainty assessments in severe accident scenarios using the URANIE software. *35rd ESReDA Seminar on "Uncertainty in Industrial Practice - Generic best practices in uncertainty treatment",* Marseille, France, novembre 2008.

Abstracts de conférences avec comité de sélection

[Ic20] B. Iooss, P. Bazin, A. de Crecy, J. Garnier, C. Cannamela and R. Phan-Tan-Luu. Quantile estimation via the use of a metamodel : application on a nuclear safety computer code. *VI Colloquium Chemiometricum Mediterraneum,* Saint Maximin La Sainte Baume, France, septembre 2007.

[Ic21] A. Marrel, B. Laurent and B. Iooss. Utilisation des processus gaussiens pour l'analyse de sensibilité, mise en œuvre sur une sortie fonctionnelle d'un code de calcul. *Joint Meeting of the Statistical Society of Canada and the Société Française de Statistique,* Ottawa, Canada, mai 2008.

MÉMOIRES de SOUTENANCE

[Im1] B. Iooss. *Mise en oeuvre d'estimateurs de la volatilité dans des modèles de sauts purs.* Rapport de stage de D.E.A., Université Paris VII, France, 1995.

[Im2] B. Iooss. *Tomographie Statistique en Sismique Réflexion : Estimation d'un Modèle de Vitesse Stochastique.* Thèse de l'École des Mines de Paris, 1998.

RAPPORTS INTERNES

[Ir1] B. Iooss. *Approche probabiliste du traitement de données sismiques (champ de vitesse et réflecteur aléatoires).* Rapport N-27/96/G, École des Mines de Paris, Centre de Géostatistique, Fontainebleau, France, 1996.

[Ir2] B. Iooss. *Revue sur la propagation d'onde acoustique en milieu turbulent.* Note Technique CEA DRN/DER/SSAE/LSMR 99/0051, CEA Cadarache, France, 1999.

[Ir3] B. Iooss et C. Lhuillier. *Influence de la turbulence et des gradients sur les mesures de débit par ultrasons.* Note Technique CEA DRN/DER/SSAE/LSMN 00/0029, CEA Cadarache, France, 2000.

[Ir4] B. Iooss. *PROUST_G Version 2.0 : Manuel d'utilisation du logiciel de la propagation ultrasonore en milieu turbulent.* Note Technique CEA DRN/DER/SSAE/LSMN 00/0037, CEA Cadarache, France, 2000.

[Ir5] B. Iooss. *Misfit statistics and data decimation in seismic traveltime tomography.* In *KIM Annual Report 2001,* pp 89-106, Institut Français du Pétrole, Rueil-Malmaison, France, 2002.

[Ir6] B. Iooss et F. Van Dorpe. *Analyses de sensibilité du code de calcul d'impact dosimétrique GASCON.* Note Technique CEA DEN/CAD/DER/STR/LCFR 2002/0032 et DEN/CAD/DED/SAMRA 02/053, CEA Cadarache, France, 2002.

[Ir7] B. Iooss. *Analyse statistique de la base de données de traction de boulons des caissons de générateurs de vapeur de Phénix.* Note Technique CEA DEN/CAD/DER/STR/LCFR 2003/0008, CEA Cadarache, France, 2003.

[Ir8] B. Iooss et S. Campos. *Différentiation automatique du module MARGARET V3.1.* Note Technique CEA DEN/CAD/DER/STR/LCFR 2003/0011, CEA Cadarache, France, 2003.

[Ir9] S. Ndao, A. Bouloré et B. Iooss. *Rapport de stage - Simplification du modèle de densification "MOGADOR-DENS V0".* Note Technique CEA DEN/CAD/DEC/SESC/LSC 03-030, CEA Cadarache, France, 2003.

[Ir10] B. Iooss et F. Van Dorpe. *Etude de la sensibilité aux méthodes d'interpolation des champs de température et de vent du logiciel MINERVE.* Note Technique CEA/DEN/CAD/DER/SESI/LCFR/NT DO 3 19/02/04, CEA Cadarache, France, 2004.

[Ir11] P-M. Pair et B. Iooss. *Rapport de stage - Construction de surfaces de réponse non linéaires : Etude comparative de nouvelles méthodes de régression.* Note Technique CEA/DEN/CAD/DER/SESI/LCFR/NT DO 21 17/06/04, CEA Cadarache, France, 2004.

[Ir12] N. Devictor, M. Marquès, N. Pérot et B. Iooss. *Description of methods for uncertainty and sensitivity analysis in support of Level 2 PSA.* Note Technique CEA/DEN/CAD/DER/SESI/LCFR/NT DO 38 17/11/04, CEA Cadarache, France, 2005.

[Ir13] B. Iooss. *Analyse d'incertitudes et de sensibilité du code METEOR (crayon UO_2).* Note Technique CEA/DEN/CAD/DER/SESI/LCFR/NT DO 5 16/02/05, CEA Cadarache, France, 2005.

[Ir14] E. Volkova, F. Van Dorpe et B. Iooss. *Modélisation du transport de ^{90}Sr en milieu poreux saturé et analyse de sensibilité du modèle : application sur un site de stockage temporaire de déchets radioactifs (CRR Kurchatov Institute, Russie).* Note Technique CEA/DEN/CAD/DTN/SMTM/LMTE 2005/63, CEA Cadarache, France, 2005.

[Ir15] A. Marrel et B. Iooss. *Rapport de stage - Modélisation des codes de calcul dans le cadre des processus gaussiens.* Note Technique CEA/DEN/CAD/DER/SESI/LCFR/NT DO 16 04/10/05, CEA Cadarache, France, 2005.

[Ir16] M. Ribatet et B. Iooss. *Rapport de stage - Modélisation de la moyenne et de la dispersion : une approche par les GLM.* Note Technique CEA/DEN/CAD/DER/SESI/LCFR/NT DO 20 29/11/05, CEA Cadarache, France, 2005.

[Ir17] P. Heyraud, B. Iooss, A. Bouloré et C. De Bellis. *Gestion des incertitudes dans la base de données CRACO.* Note Technique CEA/DEN/CAD/DEC/SESC/LSC 05-040, CEA Cadarache, France, 2006.

[Ir18] B. Iooss et F. Van Dorpe. *Analyse de sensibilité du code d'impact radiologique MIRAGE.* Note Technique CEA/DEN/CAD/DER/SESI/LCFR/NT DO 5 01/03/06, CEA Cadarache, France, 2006.

[Ir19] B. Iooss. *Manuel utilisateur du logiciel SSURFER V1.2 : programmes en R d'analyses d'incertitudes, de sensibilités, et de construction de surfaces de réponse.* Note Technique CEA/DEN/CAD/DER/SESI/LCFR/NT DO 6 08/03/06, CEA Cadarache, France, 2006.

[Ir20] B. Iooss et N. Pérot. *Représentativité d'un échantillon de faible taille.* Note CEA/DEN/CAD/DER/SESI/LCFR DO 71 13/11/06, CEA Cadarache, France, 2006.

[Ir21] B. Iooss, P. Bazin, A. de Crecy, C. Cannamela et J. Garnier. *Compte rendu du projet CEMRACS 2006 sur l'estimation de quantiles de codes de calcul.* Note Technique CEA/DEN/CAD/DER/SESI/LCFR/NT DO 22 29/11/06, CEA Cadarache, France, 2006.

[Ir22] B. Auder et B. Iooss. *Analyse de sensibilité globale basée sur l'entropie.* Note Technique CEA/DEN/CAD/DER/SESI/LCFR/NT DO 04 12/03/07, CEA Cadarache, France, 2007.

[Ir23] M. Marquès, B. Iooss et B. Spindler. *Guide utilisateur LEONAR Version 1.* Note Technique CEA/DEN/CAD/DER/SESI/LCFR/NT DO 07 20/04/07, CEA Cadarache, France, 2007.

[Ir24] B. Tourniaire, B. Spindler, B. Iooss et M. Marquès. *LEONAR V1 - Modélisation des phénomènes hors cuve et méthodes statistiques et probabilistes.* Note Technique CEA DEN/DTN/SE2T/LPTM/2007-207, CEA Grenoble, France, 2007.

[Ir25] B. Spindler, B. Tourniaire, B. Iooss et M. Marquès. *LEONAR V1 - Note de "validation" : modules physiques, méthodes statistiques, illustration d'un calcul couplé.* Note Technique CEA DEN/DTN/SE2T/LPTM/2007-208, CEA Grenoble, France, 2007.

[Ir26] B. Iooss et N. Pérot. *Quelques outils statistiques pour étudier la représentativité d'un échantillon de données de faible taille.* Note Technique CEA/DEN/CAD/DER/SESI/LCFR/NT DO 13 20/09/07, CEA Cadarache, France, 2007.

[Ir27] V. Bakirdjian et B. Iooss. *Rapport de stage : Etude géostatistique pour la cartographie de cellules contaminées.* Note Technique CEA/DEN/CAD/DER/SESI/LCFR/NT DO 14 24/09/07, CEA Cadarache, France, 2007.

[Ir28] B. Iooss, E. Volkova et A. Marrel. *Analyse de sensibilité d'un modèle hydrogéologique dépendant de simulations géostatistiques du champ de perméabilité.* Note Technique CEA/DEN/CAD/DER/SESI/LCFR/NT DO 15 02/10/07, CEA Cadarache, France, 2007.

[Ir29] B. Iooss. *Spécifications fonctionnelles pour URANIE liées aux besoins du logiciel LEONAR : tests d'ajustement statistique.* Note Technique CEA/DEN/CAD/DER/SESI/LCFR/NT DO 20 23/11/07, CEA Cadarache, France, 2007.

[Ir30] N. Pérot et B. Iooss. *Echantillonnage pour le démantèlement : norme AFNOR 8550 de 1994. Démonstration et résultats de simulation.* Note Technique CEA/DEN/CAD/DER/SESI/LCFR/NT DO 22 17/12/07, CEA Cadarache, France, 2007.

[Ir31] B. Spindler, B. Tourniaire, J-M. Seiler, G. Ratel, B. Iooss et M. Marquès. *LEONAR V2 - Modélisation des phénomènes physiques et méthodes statistiques et probabilistes.* Note Technique CEA DEN/DTN/SE2T/LPTM/2008-259, CEA Grenoble, France, 2008.

[Ir32] B. Iooss, M. Marquès et B. Spindler. *Guide utilisateur LEONAR Version 2.* Note Technique CEA/DEN/CAD/DER/SESI/LCFR/NT DO 3 04/03/08, CEA Cadarache, France, 2008.

[Ir33] B. Iooss. *Global sensitivity analysis methods with spatially-dependent inputs.* Note Technique CEA/DEN/CAD/DER/SESI/LCFR/NT DO 4 06/03/08, CEA Cadarache, France, 2008.

[Ir34] B. Iooss et A. Marrel. *Performing sensitivity analysis of cpu time consuming models using metamodels.* Note Technique CEA/DEN/CAD/DER/SESI/LCFR/NT DO 7 21/03/08, CEA Cadarache, France, 2008.

[Ir35] B. Iooss et N. Devictor. *Presentation of performance assessment results by alternative approaches.* Note Technique CEA/DEN/CAD/DER/SESI/LCFR/NT DO 6 21/03/08, CEA Cadarache, France, 2008.

[Ir36] N. Pérot et B. Iooss. *Estimation par régression multi-linéaire sur la réalisation expérimentale de la puissance maximum déposée (P_{max}) et de l'énergie pour différents types de pulses dans le réacteur CABRI.* Note Technique CEA/DEN/CAD/DER/SESI/LCFR/NT DO 9 16/04/08, CEA Cadarache, France, 2008.

[Ir37] B. Iooss. *Spécifications fonctionnelles pour URANIE : analyses de sensibilité basées sur des tests statistiques.* Note Technique CEA/DEN/CAD/DER/SESI/LCFR/NT DO 13 28/08/08, CEA Cadarache, France, 2008.

[Ir38] B. Iooss, M. Marquès et B. Spindler. *Guide utilisateur de LEONAR Version 2.1-0.* Note Technique CEA/DEN/CAD/DER/SESI/LCFR/NT DO 14 14/10/08, CEA Cadarache, France, 2008.

[Ir39] B. Tourniaire, B. Spindler, J-M. Seiler, G. Ratel, B. Iooss et M. Marquès. *LEONAR V2.1 - Modélisation des phénomènes physiques et méthodes statistiques et probabilistes.* Note Technique CEA DEN/DTN/SE2T/LPTM/2008-300/a, CEA Grenoble, France, 2008.

Chapitre 1

Introduction générale

La simulation numérique consiste à reproduire un ou des phénomènes physiques à l'aide de modèles constitués d'équations mathématiques et résolus par des calculs sur ordinateur. Cette révolution scientifique, datant d'une soixantaine d'années, a conduit à de nombreuses avancées : meilleure compréhension de la physique, prédiction de phénomènes, aide à la planification d'expériences réelles, remplacement d'expériences réelles par des simulations, outils de réalité virtuelle, ... Malheureusement, dans certaines situations, la méconnaissance de la physique et de la réalité limite son utilisation. De manière schématique, on peut distinguer trois sources d'imprécision majeure :

1. les imprécisions dues à une modélisation théorique simplifiée ou erronée de la réalité (*e.g.* phénomène physique ou chimique non pris en compte, suppression de la variabilité spatiale ou temporelle d'un paramètre du modèle) que l'on appelle communément incertitudes de modèle ;
2. les imprécisions dues à la résolution numérique des équations mathématiques (*e.g.* schéma numérique, critère de convergence), appelées incertitudes numériques ;
3. les imprécisions sur les paramètres et données d'entrée du phénomène physique (*e.g.* vitesse du vent, porosité du sous-sol), qui se classent en incertitudes épistémiques (aléas dus à une méconnaissance) et en incertitudes stochastiques (aléas intrinsèques à la nature des variables). Ces entrées doivent pourtant être spécifiées pour réaliser un calcul du modèle numérique.

A partir des années 1980, les problématiques de prise en compte et de réduction des incertitudes ont été identifiées par les industriels comme des enjeux majeurs de leurs processus de décision s'appuyant sur des modèles prédictifs. En France, dès le début des années 1990, de grands projets de gestion des incertitudes ont été lancés, notamment dans les industries pétrolière et métallurgique. De part ma formation universitaire en mathématiques appliquées, statistique et probabilités, mon intérêt pour les problèmes industriels liés à la modélisation aléatoire, et mon arrivée dans le monde de la R&D en 1995, mes travaux de recherche se sont naturellement tournés vers ce contexte global du traitement des incertitudes en simulation.

Durant mes six premières années de recherche, motivées (et financées) principalement par les problématiques d'exploration pétrolière, j'ai travaillé sur le premier type d'imprécision, les incertitudes de modèles, dans le cadre de la modélisation de la propagation d'ondes acoustiques. Les incertitudes considérées concernaient la non prise en compte, dans l'interprétation des signaux acoustiques, de la variabilité spatiale des champs de vitesse des ondes d'échelle supérieure à la longueur d'onde. Durant les six années suivantes, supporté par certains programmes de la Direction de l'Énergie Nucléaire du Commissariat à l'Énergie Atomique (simulation numérique, maîtrise des risques, sûreté des réacteurs, gestion du combustible, ...), je me suis attaché au troisième type d'imprécisions, concernant les données d'entrée d'un modèle de simulation numérique. Je me suis notamment efforcé d'utiliser et de développer des méthodes statistiques rigoureuses de traitement des incertitudes, et ce dans un cadre générique (indépendamment d'un modèle et d'une problématique physique). Ce type d'études, de plus en plus appliqué au contexte industriel, a produit des outils extrèmement utiles aux ingénieurs qui développent, valident ou utilisent des modèles prédictifs.

Intégrée au sein d'une équipe dédiée à l'étude des incertitudes du Commissariat à l'Énergie Atomique de Cadarache (CEA Cadarache), cette dernière activité s'est avérée passionnante, du fait notamment de la diversité des domaines d'applications concernés et de la multiplicité des outils statistiques utilisés. L'un des *credo* de notre équipe porte notamment sur la résolution de problèmes atypiques, *i.e.* de problèmes d'ingénieurs difficiles : questions imprécises et/ou mal posées, modèle numérique non robuste, peu voire pas de données, code numérique coûteux ou peu maniable, nombre d'entrées excessivement important, ... La résolution de ces problèmes passent souvent par l'utilisation d'outils statistiques avancés, issus de travaux de recherche plus ou moins récents du monde académique. Cette approche a suscité, de notre part, une forte volonté d'échanges afin d'intéresser des chercheurs universitaires à nos problèmes. L'une de nos réussites est que cette démarche, menée avec d'autres organismes de recherche industrielle (comme EDF R&D, IFP, EADS, ONERA, ...), a porté ces fruits. En effet, de nombreuses initiatives de collaboration sur cette thématique ont récemment vu le jour, dont certains résultats tangibles sont présentés aux §3.1, §3.4 et au chapitre 4.

Ma demande d'habilitation à diriger des recherches entre dans le cadre de ces échanges entre équipes de R&D industrielles et équipes universitaires. Il est souhaitable que les échanges se fassent à double sens, les universitaires s'intéressant aux problématiques industrielles et les ingénieurs de recherche s'investissant dans des activités de nature plus académique (publications, formation, encadrements de doctorants, ...). Ma demande est par ailleurs motivée par le souhait du CEA de disposer en interne de chercheurs habilités à diriger des recherches. Elle est également renforcée par notre volonté de pérenniser l'activité "Traitement générique des incertitudes" à la Direction de l'Energie Nucléaire du CEA. Bien entendu, d'autres motivations importantes sont plus personnelles, avec notamment le désir de présenter dans un cadre global tous les travaux de recherche que j'ai réalisés durant mes douze premières années d'exercice. Ceci constitue l'objet de ce mémoire et me permet de les commenter et les critiquer de manière plus aboutie que dans des publications scientifiques. Malgré mon souci de synthèse, il est apparu cependant inévitable de séparer ce mémoire en deux parties, distinguant mes périodes "pétrolière" et "nucléaire".

Le chapitre suivant est consacré à mes travaux sur la caractérisation stochastique des champs de vitesse dans la modélisation de la propagation d'ondes. En effet, durant ma thèse, mon post-doctorat et mon activité à l'Institut Français du Pétrole (IFP), je me suis intéressé à l'étude des moments des temps de trajet des ondes en fonction des moments des champs de vitesse dans lesquels elles se propagent. Les années 1990 ont vu un effort de recherche notable consacré à ce sujet en France, ce qui m'a permis de collaborer avec des chercheurs issus d'horizons divers : stagiaires, thésards, post-doctorants, ingénieurs de recherche et chercheurs universitaires. Nos travaux, qui ont donné lieu à quelques avancées théoriques et méthodologiques, sont à classer dans les domaines des mathématiques appliquées et de leurs implications pour la physique. Nous avons notamment apporté une vision légèrement différente de celle des physiciens : approche géostatistique pour la modélisation des milieux aléatoires, interrogation systématique du domaine de validité des approximations physiques, étude fine et validation systématique de ces approximations à l'aide de simulations numériques lourdes, ... Durant ces recherches, les domaines d'application ont été variés (sismique et hydraulique).

Le troisième chapitre de ce mémoire concerne la problématique du traitement des incertitudes en simulation et de l'analyse statistique des réponses des codes numériques. La première section passe en revue les différentes méthodes d'analyse de sensibilité de modèles. La deuxième section présente mes travaux sur l'utilisation des processus gaussiens comme métamodèle, modèle se substituant à un code de calcul lorsque celui-ci est trop coûteux pour être "exploré" à l'aide d'outils statistiques. Les deux dernières sections présentent des sujets plus originaux sur lesquels je me suis concentré récemment : l'estimation de quantiles élevés de sorties de codes de calcul et le traitement des sorties des codes de calcul stochastiques, *i.e.* de modèles numériques qui contiennent un aléa incontrôlable. Ce dernier thème de recherche me permet d'introduire le problème de la prise en compte des variables fonctionnelles en analyse d'incertitude de modèles, sujet sur lequel je lance actuellement quelques actions de recherche.

Le dernier chapitre dresse un bilan et des perspectives de ces travaux. Les annexes sont constituées de mon curriculum vitae détaillé (intégrant une synthèse courte de tous mes travaux de recherche, responsabilités et encadrements scientifiques) et des résumés de toutes mes publications à revue.

English version - General introduction

Numerical simulation consists in reproducing one or several physical phenomena with a numerical model. Generally speaking, a numerical model is built on mathematical equations and solved by computer calculations. This scientific revolution, beginning sixty years ago, led to numerous contributions :
- better understanding of the physics,
- prediction of phenomena,
- help to the design of real experiments,
- replacement of real experiments by simulations,
- etc.

Unfortunately, in many situations, the physics misunderstanding limits its use. In a simplistic way, one can distinguish three sources of major imprecision :

1. inaccuracies due to a simplified or erroneous theoretical modelling of the reality (e.g. physical or chemical phenomenon not taken into account, deletion of the spatial or temporal variability of a model parameter) that is commonly called model uncertainty ;
2. inaccuracies due to the numerical resolution of the mathematical equations (e.g. numerical scheme, convergence criterion), called numerical uncertainty ;
3. inaccuracies about the input parameters and input data of the physical phenomenon (e.g. wind speed, groundwater porosity), which are classified in epistemic uncertainties (insufficient knowledge alea) and in stochastic uncertainties (intrinsic alea). Nevertheless, these input variables have to be specified in order to perform one calculation of the numerical model.

From 1980s, uncertainty problems were identified in industry as major stakes. This kind of problems occurs in the industrial decision processes based on predictive models. In France, from the 1990s, big projects of uncertainty management were launched in the petroleum and metallurgical industries for example. I began my research activities in 1995. At this time, I had studied mainly applied mathematics and more precisely probability and statistics. I was highly interested by the random modelling in industrial problems. So that, my activities naturally turned to the uncertainty management context in simulation models.

During my first six research years, motivated (and financed) mainly by oil exploration problems, I worked on the first imprecision type (model uncertainties). My topic concerned the acoustical wave propagation modelling. The associated uncertainties were provoked by the deletion, in the interpretation of the acoustical signals, of the wave velocity spatial variability of scale larger than the wavelength. During the next six years, I worked on the third imprecision type (input data and parameter uncertainties) for the research projects of the French Nuclear Energy Division. For example, these projects are related to numerical simulation, risk control, nuclear reactor safety and fuel management. I tried to use and develop rigorous statistical methods in a generic framework, i.e. independently of a model and of a physical problem. This kind of studies produces useful tools to engineers who develop, validate or use predictive models. The interest about these problems grows more and more in the industrial context.

Inside a research team dedicated to the uncertainty studies of the Commissariat à l'Énergie Atomique de Cadarache (CEA Cadarache), this last activity turned out fascinating. It was mainly due to the application domains variety and the involved statistical tools multiplicity. One of our activity concerns the resolution of atypical and difficult engineering problems :

– imprecise or ill-posed questions,
– non robust numerical model,
– little or no input data to quantify parameter uncertainties,
– cpu time expensive numerical model,
– excessively large number of random inputs,
– etc.

Resolution of these problems often needs the use of advanced statistical tools. This approach has necessitated to increase our contacts with academic researchers. These initiatives have been led with other industrial research institutes (as EDF R&D, IFP, EADS, ONERA, ...). They have provided some successful results. Indeed, numerous collaborations on this research theme were recently born. Some results of these collaborations are mentioned in §3.1, §3.4 and in chapter 4.

My *habilitation* demand enters within these exchanges between industrial and academic research teams. These exchanges are made with a double meaning : the academics provide interests about the industrial problems and the research engineers are involved in more academic activities (as publications, training, PhD student supervision). Besides, I was motivated by my research institute which needs habilited researchers. This habilitation will also allow to reinforce the activity "Uncertainty management" in the Nuclear Energy Division. Naturally, other motivations are more personal. For example, I wanted to present in a global context all my research works. That allows me to comment and to criticize these results in a single report. This can be done in a more succeeded way than in any scientific publications. However, in spite of this synthesis effort, it is inevitable to divide this report in two parts, in order to distinguish my "oil" and "nuclear" periods.

The following chapter is dedicated to my works on the velocity stochastic characterization in the wave propagation modelling. Indeed, this was my research subject during my PhD thesis, my postdoctoral position and my activity at the French Institute of Petroleum (IFP). It concerns the study of the wave traveltimes statistical moments according to the statistical moments of the velocity field in which the waves propagate. The 1990s saw a considerable research effort dedicated to this subject in France. So that, I have collaborated with various researchers as trainees, PhD students, industrial research engineers and academic researchers. Our works have produced some theoretical and methodological contributions in the applied mathematics domain and their implications for the physics. We notably brought a slightly different view from that of the physicists : geostatistical approach for the random media modelling, interrogations about approximation validity domains, systematic validation of the estimates by huge numerical simulations, ... I have also applied these researches on various application domains, mainly exploration seismics and hydraulics.

The uncertainty management problem in numerical simulation is the topic of the third chapter. More precisely, I have studied some statistical approaches to analyze numerical code outputs. The first section reviews the sensitivity analysis of model output methods. The second section presents my works on the Gaussian process model (kriging) as a metamodel. A metamodel is a simple mathematical model substituting itself for a computer code. A metamodel is useful when the complete numerical model is too time consuming to be "investigated" using statistical tools. The last two sections present more original subjects on which I recently concentrated. The first one relates to the high order quantile estimation problem of a computer code output. The second one is concerned by the statistical analysis of stochastic computer codes, i.e. numerical models which contain an uncontrollable alea source. This latter problem allows me to introduce the problem of the functional variables in uncertainty analysis. It is one of my recent interest topic.

The last chapter draws up balance sheet and perspectives of these works. Appendices are constituted by my detailed curriculum vitae and all my journal publication summaries. My curriculum vitae integrates a short synthesis of all my research works, responsibilities and scientific supervisions.

Chapitre 2

Modélisation stochastique des incertitudes de vitesse en propagation d'ondes

2.1 Introduction

Les phénomènes de propagation d'ondes interviennent dans une grande quantité de domaines scientifiques et peuvent être parfois associés à des enjeux industriels importants (exploration pétrolière, télécommunications, électronique, contrôle non destructif). Lorsque le milieu traversé par les ondes est hétérogène, inconnu et trop complexe pour être modélisé de manière déterministe, il est courant de le modéliser par un champ aléatoire. Les caractéristiques du milieu (*e.g.* raideurs élastiques, densité, température, vitesses du vent, ...) en chaque point sont donc des variables aléatoires qui sont corrélées spatialement. Dans ce cadre, le champ d'onde devient lui-même un processus aléatoire et on le caractérise par sa moyenne, sa covariance, voire ses moments d'ordre plus élevé.

2.1.1 État de l'art

Introduite par des astrophysiciens (Chandrasekhar [38]), la théorie de la propagation d'ondes en milieu aléatoire s'est fortement développée dans les années cinquante et soixante grâce à l'école physicienne russe (Chernov [41], Barabanenkov et al. [14]), puis a été considérablement enrichie par les apports des mathématiciens (Keller [115], Frisch [70], Asch et al. [9], Papanicolaou [164], Fouque et al. [67]). Son développement a été porté par la multiplicité de ses domaines d'application, parmi lesquels on peut citer l'électromagnétisme (Ishimaru [105]), la radiophysique (Tatarskii [211], Rytov et al. [181]), l'acoustique atmosphérique (Ostashev [163]), l'acoustique marine (Uscinski [218], Flatté et al. [65]), l'optique et les télécommunications (Andrews & Phillips [5]), la surveillance des réacteurs nucléaires (Fiorina [64]), la sismologie (Sato & Fehler [193]), la sismique pétrolière (Jannaud [108], Iooss et al. [97]), la physique des plasmas, l'hydrodynamique, l'imagerie médicale, ... Dans chacun de ces domaines d'application, les paramètres d'hétérogénéité ou de turbulence dépendent de la nature et de l'environnement (les conditions aux limites) du milieu de propagation. Par exemple, la sismologie s'intéresse aux constantes élastiques et à la densité des roches. En océanographie, la température, la salinité et la pression de l'eau modifient la célérité des ondes alors que dans l'atmosphère, ce sont la température et l'humidité de l'air qui sont concernées. Dans ce chapitre, ces paramètres spatialement hétérogènes seront modélisés par des champs aléatoires continus (Cressie [44]).

Mathématiquement, les méthodes utilisées reposent dans leur grande majorité sur la même idée de base : on identifie un paramètre d'échelle ε, puis, on utilise soit une méthode de perturbation (si ε est petit) pour obtenir une équation simplifiée que l'on moyenne ensuite, soit une méthode d'analyse stochastique qui consiste à moyenner l'équation stochastique d'abord et résoudre les équations des moments ensuite. Le choix de ε dépend notamment de la taille l_ε des hétérogénéités. Différents rapports de cette taille à la longueur d'onde dominante λ_0 de l'onde considérée conduisent à trois régimes différents pour la physique de la propagation :

- ◇ le cadre des basses fréquences est celui où la taille caractéristique des hétérogénéités est nettement plus petite que la longueur d'onde ($l_\varepsilon \ll \lambda_0$). Il n'est pas possible d'étudier les interactions de l'onde avec chacune des hétérogénéités et le milieu de propagation est alors remplacé par un milieu fictif homogène appelé milieu effectif. La difficulté consiste à obtenir les propriétés du milieu effectif par moyennage des propriétés des constituants du milieu initial. Dans le cas 1D (milieu stratifié), des approximations de type diffusion (dans la théorie des processus stochastiques) permettent d'obtenir des résultats théoriques à la limite $l_\varepsilon \to 0$ (cf. Fouque et al. [67] pour l'ouvrage le plus récent sur ce sujet) ;

- quand les hétérogénéités sont de taille comparable à la longueur d'onde (cadre des moyennes fréquences, $l_\varepsilon \sim \lambda_0$), des phénomènes très complexes peuvent se produire : diffractions multiples, rétro-diffractions, atténuation, résonance, localisation, ... C'est le cas le plus difficile scientifiquement car les hétérogénéités ne sont pas assez petites pour pouvoir être lissées (comme dans la théorie du milieu effectif), et pas assez grandes pour pouvoir être considérées constantes sur la distance d'une longueur d'onde (comme dans l'optique géométrique). La théorie du transfert radiatif fournit des équations de transport de l'énergie de l'onde et donne des indications sur le comportement moyen d'une onde dans un milieu hétérogène (Ishimaru [105], Sato & Fehler [193]). Elle permet aussi d'expliquer le phénomène de la localisation : à cause des diffractions multiples et arrières dans une zone fortement stratifiée, l'énergie d'une onde décroît exponentiellement, et peut rester emprisonnée à l'intérieur d'une zone du milieu de propagation (Papanicolaou [164], Fouque et al. [67]) ;

- quand la taille de hétérogénéités est supérieure à la longueur d'onde ($l_\varepsilon \gg \lambda_0$), ce sont les phénomènes de transmission et de diffraction faible qui gouvernent la propagation. Le champ d'onde est faiblement déformé par chaque hétérogénéité et, en se plaçant en régime périodique (onde monochromatique), les notions d'amplitude et de phase sont conservées. On se restreint alors souvent à l'étude des fluctuations des amplitude et temps de trajet de l'onde, quantités qui sont facilement mesurées sur les sismogrammes. Dans ce régime, et notamment dans son cas limite qui est l'optique géométrique, les résultats théoriques sont nombreux pour des milieux aléatoires pouvant être relativement complexes (statistiquement anisotrope, localement stationnaire, quasi-stationnaire, cf §2.2.2). Par exemple, les deux premiers moments des amplitudes et des temps de trajet des ondes peuvent être exprimés en fonction des deux premiers moments du champ de vitesse des ondes (Chernov [41], Tatarskii [211], Ishimaru [105], Rytov et al. [181], Ostashev [163]). C'est dans ce cadre hautes fréquences que mes travaux se sont inscrits, du fait principalement de leurs liens avec l'étude des temps de trajet des ondes.

2.1.2 Contributions

Mes travaux de recherche se sont intéressés à la théorie de la propagation d'onde en milieu aléatoire d'un point de vue pratique, pour aborder successivement deux problèmes industriels, en exploration sismique pétrolière et en contrôle non destructif par ultrasons, détaillés ci-dessous.

1. La sismique réflexion est employée en prospection pétrolière depuis plusieurs décennies pour donner une image de la structure d'un bassin sédimentaire. Elle consiste à créer artificiellement dans le sous-sol (resp. en mer) un ébranlement (resp. une onde de compression) et à enregistrer les réponses du milieu en différents récepteurs situés en surface (Fig. 2.1). On obtient une bonne couverture du sous-sol en répétant et en déplaçant le dispositif le long du terrain examiné. A partir des signaux sismiques et de la vitesse de propagation des ondes, on tente de reconstituer une carte du sous-sol. Pour ce faire, une étape préalable d'inversion du champ de vitesse sismique à partir des temps de trajet des ondes est nécessaire. Il s'agit là d'un problème inverse mal posé (Tarantola [208]) qui peut être traité par des méthodes de régularisation (processus nommé tomographie).

   Malheureusement, les méthodes sismiques utilisées en pratique ont des limites au niveau de la

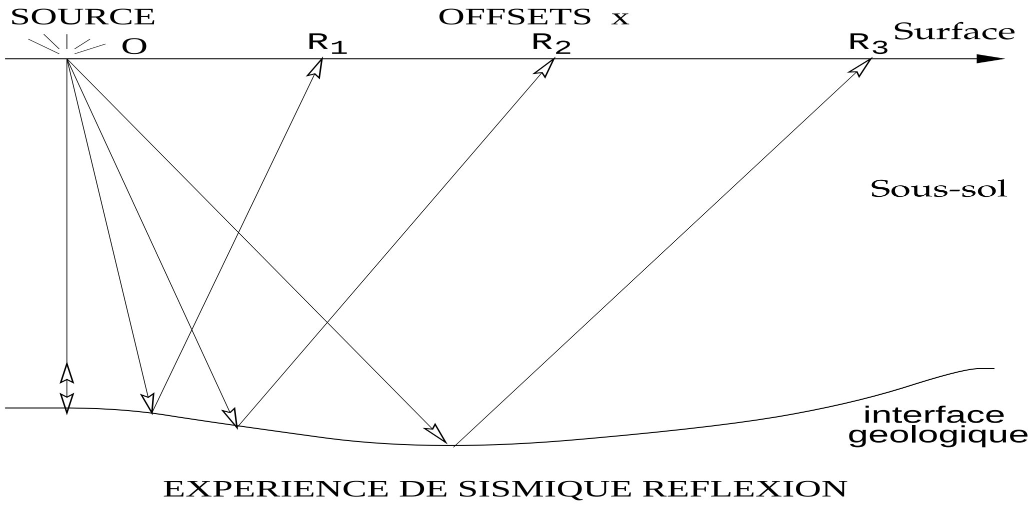


FIG. 2.1 – Expérience de sismique réflexion 2D. L'offset x est la distance entre la source O et le récepteur R.

résolution spatiale du champ de vitesse qu'elles sont susceptibles de caractériser (Thore & Juliard [212], Williamson & Worthington [229]). Basées sur l'approximation de l'équation d'ondes par l'optique géométrique, ces méthodes déterministes interprètent mal les événements provoqués par de trop petites hétérogénéités de vitesse qui ont un impact non négligeable sur les temps d'arrivée. Au début des années 1990, Matheron [148] s'est intéressé à ce sujet de recherche en modélisant de manière probabiliste ces hétérogénéités de vitesse. Au sein du Centre de Géostatistique de l'École des Mines de Paris, Alain Galli a poursuivi ces travaux en dirigeant successivement trois thèses sur le sujet (de 1992 à 2002). Mes travaux de thèse (Iooss [92]), qui ont fait suite à ceux d'une première thèse (Touati [214]), ont contribué à fournir une procédure pour quantifier statistiquement les hétérogénéités de vitesse sismique (cf. §2.4). Ces résultats de recherche, encore non exploités de manière industrielle en exploration pétrolière, ont été poursuivis par quelques géophysiciens et semblent déboucher sur des applications (Kaslilar et al. [114]). Au cours de ces travaux, j'ai également obtenu de nouveaux résultats sur les conditions de validité des différentes approximations (Rytov, parabolique, optique géométrique) qui permettent d'exprimer analytiquement les temps de trajet des ondes (et leurs moments) en fonction de perturbations de vitesse statistiquement anisotropes (cf. §2.3).

2. La deuxième problématique industrielle que j'ai abordée concernait les mesures de débit dans les tuyauteries industrielles (de liquide ou de gaz) qui sont couramment réalisées à l'aide de techniques ultrasonores. C'est le cas par exemple sur certains circuits des réacteurs nucléaires à eau pressurisée. Les techniques ultrasonores de débitmétrie présentent l'avantage d'être non destructives, non intrusives, faciles à utiliser et possèdent un temps de réponse extrêmement court (Lynnworth [138]). Certaines sont basées sur la différence de temps de trajet entre deux ondes ultrasonores propagées dans le sens et à contresens du flux (Fig. 2.2). Si le poids des incertitudes prépondérantes est connu, celles liées aux profils (thermiques et cinématiques) et aux turbulences (thermiques et cinématiques) du milieu le sont beaucoup moins.

   En collaboration avec Philippe Blanc-Benon[1] et Christian Lhuillier[2], ce sujet a donné lieu à un travail amont sur l'analyse des fluctuations statistiques des temps de trajet et au développement de méthodes de simulation de la propagation d'ondes en milieu en mouvement. Le résultat le plus marquant resulte de la démonstration analytique de la non linéarité de la variance des temps au deuxième ordre, observée dans des simulations numériques précédemment à mes travaux et qui a été confirmée expérimentalement depuis (cf. §2.3). Les outils numériques développés ont permis ensuite d'évaluer les incertitudes (dues aux turbulences thermique et cinématique de l'eau) sur

[1]Laboratoire de Mécanique des Fluides et d'Acoustique, CNRS/École Centrale de Lyon
[2]CEA Cadarache, Direction de l'Énergie Nucléaire

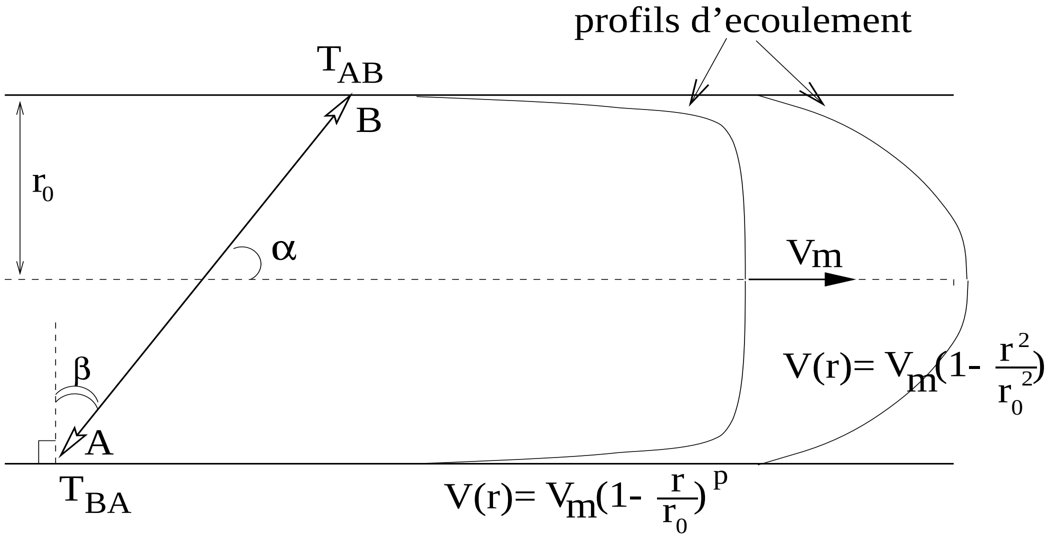


FIG. 2.2 – Principe du débitmètre à ultrasons. Profils d'écoulement en régime turbulent et laminaire.

l'estimation du débit par les débitmètres ultrasonores à temps de transit. Dans ce mémoire, l'application à la débitmétrie ne sera pas évoquée. On notera cependant la publication récente de Franchini et al. [68] qui apporte une solution analytique à l'un des problèmes sur les incertitudes en débitmétrie que j'ai résolu par simulation numérique dans Iooss et al. [98].

La section suivante présente le type de modélisation choisie pour les hétérogénéités du milieu de propagation, à savoir des champs aléatoires caractérisés par leur structure de covariance. La troisième section décrit ma contribution à la théorie de la propagation d'ondes en milieu aléatoire, qui concerne essentiellement l'étude, par méthodes perturbatives, des temps de trajet en milieu aléatoire anisotrope. La quatrième section traite de ma contribution sur l'inversion de la covariance des temps de trajet (nommée "tomographie statistique"). Finalement, à partir d'une synthèse de mes travaux en acoustique et en sismique, une conclusion permet de mettre en évidence un certain nombre d'axes de recherche ouverts.

2.2 Modélisation aléatoire des milieux hétérogènes et turbulents

Dans cette section, les choix de modélisation pour les milieux aléatoires sont présentés.

2.2.1 Milieux stationnaires anisotropes

En physique, un champ aléatoire $\varepsilon(\boldsymbol{r}) \in \mathbb{R}$ (où $\boldsymbol{r} \in \mathbb{R}^3$) est souvent caractérisé en première approximation par sa moyenne et sa covariance (*e.g.* Rytov et al. [180]). C'est également la base de la géostatistique linéaire qui consiste à caractériser un milieu uniquement à l'aide de ses deux premiers moments statistiques (Chilès & Delfiner [42]).

En pratique, on fait souvent des hypothèses d'invariance des moments par translation. Par exemple, on peut utiliser l'hypothèse de stationnarité d'ordre deux : la moyenne et les corrélations sont invariantes par translation. On suppose alors que la structure spatiale de ε est décrite par sa fonction de covariance $C_\varepsilon : \boldsymbol{r} \in \mathbb{R}^3 \to C_\varepsilon(\boldsymbol{r}) \in \mathbb{R}$:

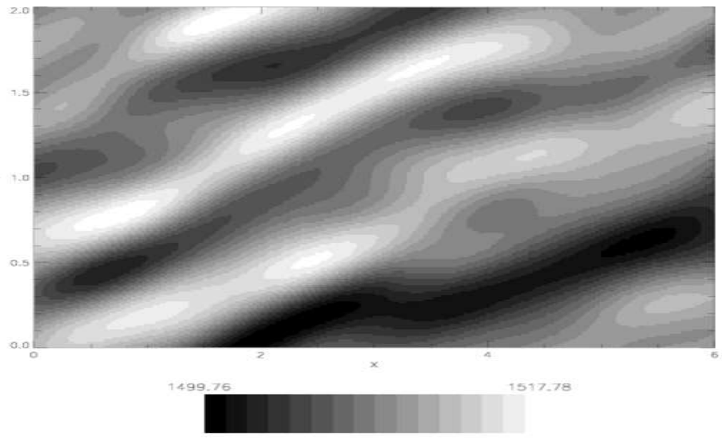
$$\mathrm{Cov}[\varepsilon(\boldsymbol{r}_1), \varepsilon(\boldsymbol{r}_2)] = C_\varepsilon(\boldsymbol{r}_1 - \boldsymbol{r}_2) = \sigma_\varepsilon^2 N(\boldsymbol{r}_1 - \boldsymbol{r}_2) \ , \ \forall (\boldsymbol{r}_1, \boldsymbol{r}_2) \in \mathbb{R}^3 \times \mathbb{R}^3 \ , \tag{2.1}$$

où $N : \boldsymbol{r} \in \mathbb{R}^3 \to N(\boldsymbol{r}) \in \mathbb{R}$ est la fonction de covariance normalisée de ε et σ_ϵ^2 est la variance de ε (le palier en géostatistique), qui donne l'amplitude typique des fluctuations du champ aléatoire. En physique, des fluctuations spatiales stationnaires sont souvent appelées fluctuations homogènes, le terme "stationnaire" étant réservé à la dépendance temporelle (Yaglom [231]).

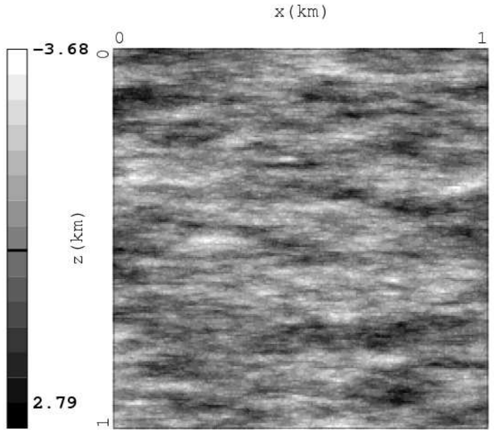
Soit $\boldsymbol{r} = (x, y, z) \in \mathbb{R}^3$ où (x, y) sont les coordonnées horizontales et z est la coordonnée verticale. En nous plaçant dans le cadre de l'hypothèse d'anisotropie géométrique (Chilès & Delfiner [42]), qui signifie que les lignes d'iso-covariance sont des ellipses concentriques, $N(\cdot)$ se définit par :

$$N(\boldsymbol{r}) = C_0\left(\sqrt{\frac{x^2}{a_x^2} + \frac{y^2}{a_y^2} + \frac{z^2}{a_z^2}}\right) , \tag{2.2}$$

où $C_0 : r \in \mathbb{R} \to C_0(r) \in \mathbb{R}$ est appelée covariance standardisée, $a_x \in \mathbb{R}_+^*$, $a_y \in \mathbb{R}_+^*$ et $a_z \in \mathbb{R}_+^*$ sont respectivement les longueurs de corrélation (portées en géostatistique) horizontale, azimuthale et verticale. Ces longueurs de corrélation représentent les tailles caractéristiques des hétérogénéités du champ aléatoire (Matheron [147]). Il est possible, sans difficulté supplémentaire, d'étendre le modèle (2.2) à des stratifications inclinées en introduisant un angle azimuthal et un angle polaire (Wackernagel [224]). Des exemples de milieux aléatoires géométriquement anisotropes sont donnés en Figure 2.3. On distingue bien les différentes hétérogénéités en forme de lentilles, ce qui corrobore notre raisonnement sur les tailles des hétérogénéités vues par les ondes et les longueurs de corrélations dans les différentes directions spatiales. Pour prendre en compte des ondes d'angle d'incidence quelconque, Samuelides & Mukerji [190], Iooss [92], Iooss et al. [94] et Kravtsov et al. [127] utilisent notamment à la place de (a_x, a_y, a_z) les longueurs de corrélation parallèle $l_\parallel$ et transverses $l_{\perp 1}$ et $l_{\perp 2}$ à la direction principale de propagation de l'onde.



(a)

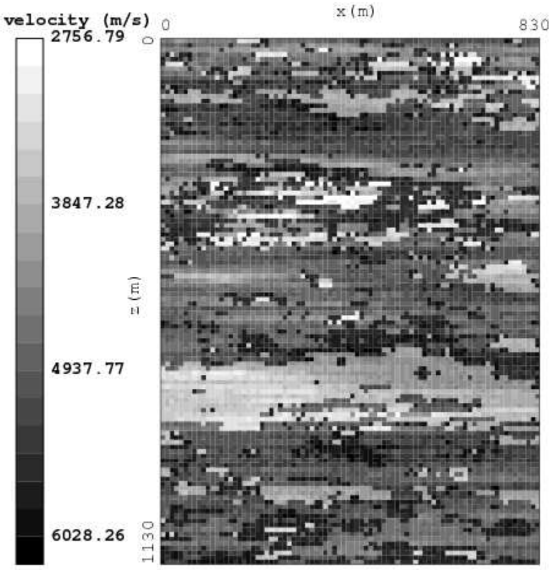


(b)

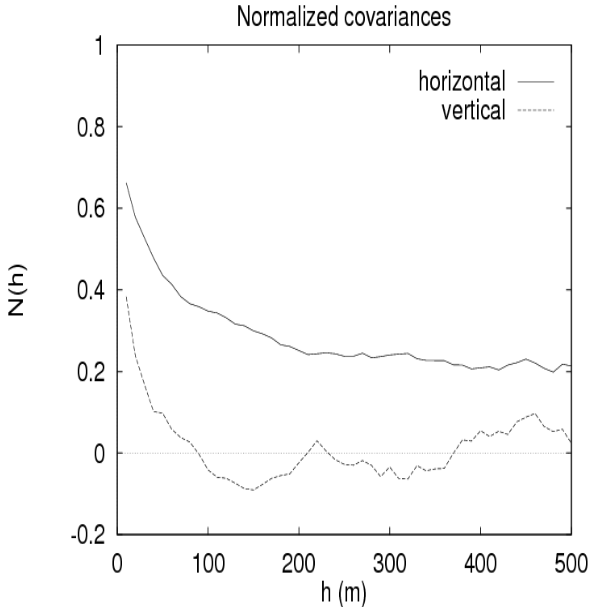
FIG. 2.3 – Exemples de champs aléatoires 2D géométriquement anisotropes. (a) Champ de vitesse d'ondes acoustiques dans l'eau (moyenne $c_0 = 1509$ m/s, covariance gaussienne $C_0(h) = \exp(-h^2)$, $h \in \mathbb{R}$, $\sigma_\varepsilon = 4$ m/s, $a_x = 1$ m, $a_z = 0.25$ m, pendage d'anisotropie $\theta = 11.5^o$, unités des axes en mètres). (b) Perturbations de vitesse des ondes sismiques (covariance exponentielle $C_0(h) = \exp(-|h|)$ de variance unité, $a_x = 0.1$ km, $a_z = 0.02$ km).

Cette prise en compte de l'anisotropie a été longtemps oubliée dans la théorie de la propagation d'ondes en milieu aléatoire. Elle a enfin été considérée dans les ouvrages de Flatté et al. [65] et de Rytov et al. [181] en distinguant les coordonnées spatiales dans les covariances et dans les densités spectrales (mais sans utiliser la notation de la covariance standardisée). Cette notation apparaît plus explicitement dans Munk & Zachariasen [155] (covariance gaussienne uniquement), Jannaud [109] et Kon [124] pour modéliser respectivement les stratifications océanographiques, géologiques et atmosphériques. Mes recherches en sismique pétrolière, où des bassins sédimentaires (donc des milieux fortement stratifiés) sont caractérisés, m'ont incité à rentrer dans ce cadre plus réaliste que le cadre isotrope. C'est l'une des contributions originales de mes travaux (Iooss [92], Iooss et al. [94]) : l'introduction de la covariance standardisée dans les équations, notamment pour la dérivation et l'étude numérique des domaines de validité de certaines approximations en milieu aléatoire anisotrope (cf §2.3.1 et §2.3.2). Depuis quelques années, d'autres auteurs, notamment en sismologie, utilisent ce modèle d'anisotropie géométrique (parfois appelé “anisomérie” ou “anisométrie”) aussi bien pour des travaux théoriques que numériques (cf. Klimeš [121], Kravtsov et al. [127, 126], Margerin [140], Saito [183]).

Bien entendu, le modèle d'anisotropie géométrique (2.2), qui est le modèle anisotrope le plus simple qui soit, est parfois peu plausible dans les applications. En géologie, des hypothèses de type anisotropie zonale sont plus réalistes pour les milieux fortement stratifiés (Chilès & Delfiner [42]). L'anisotropie zonale introduit notamment des paliers (*i.e.* variances) différents suivant les directions spatiales, en superposant une covariance isotrope et une covariance possédant une anisotropie géométrique très forte. La Figure 2.4 (a) présente un champ de vitesse sismique réaliste, construit à partir de données de puits[3] complétées par des techniques de simulations géostatistiques de lithofaciès[4]. La covariance déduite de ce modèle présente une nette anisotropie zonale avec effet de pépite[5] (Fig. 2.4 (b)).



(a)



(b)

FIG. 2.4 – Exemple de réservoir pétrolier réaliste (issu de Iooss et al. [97]). (a) Champ de vitesse sismique (avec anisotropie zonale). (b) Covariances expérimentales normalisées dans les directions horizontale et verticale.

L'anisotropie géométrique présente néanmoins le grand avantage de permettre l'obtention de résultats explicites pour les moments des temps de trajet (cf chap. 2.3), en passant aux covariances ou densités spectrales standardisées lors de la simplification des intégrales. De plus, dans Iooss et al. [97], est présentée une application dans laquelle les équations obtenues sous anisotropie géométrique sont néanmoins robustes par rapport à cette hypothèse.

Touati [214] et Touati et al. [215] préfèrent utiliser une fonction de covariance factorisée (aussi appelée covariance séparable) :

$$N(\boldsymbol{r}) = N(x, y, z) = N_1\left(\frac{x}{a_x}\right) N_2\left(\frac{y}{a_y}\right) N_3\left(\frac{z}{a_z}\right) \ , \ \boldsymbol{r} = (x, y, z) \in \mathbb{R}^3 \ , \tag{2.3}$$

où $N_1(\cdot)$, $N_2(\cdot)$ et $N_3(\cdot)$ sont des fonctions de covariance monodimensionnelles, $a_x \in \mathbb{R}_+^*$, $a_y \in \mathbb{R}_+^*$ et $a_z \in \mathbb{R}_+^*$ sont respectivement les longueurs de corrélation horizontale, azimuthale et verticale. Ce type de covariance ne rentre pas dans le cadre des modèles géométriquement anisotropes mais permet aussi de simplifier les intégrations. Ce modèle est cependant peu réaliste pour la modélisation de structures physiques, par exemple en géostatistique et en modélisation de la turbulence.

[3]mesures acoustiques fines réalisées par des sondes situés dans des puits forés verticalement
[4]faciès d'une couche sédimentaire pour ce qui est des minéraux qui la composent
[5]discontinuité à l'origine traduisant une erreur de mesure ou la présence d'une microstructure

2.2.2 Milieux non stationnaires

L'hypothèse de stationnarité d'ordre deux (Eq. (2.1)) est une hypothèse également peu réaliste en pratique. Même s'il est possible de superposer un macro-modèle au champ aléatoire stationnaire pour lui imposer certaines tendances globales, cette hypothèse amène des contraintes parfois peu souhaitables sur le champ aléatoire (*e.g.* l'homoscédasticité). Il est aisé de relâcher cette hypothèse en utilisant l'hypothèse d'acroissements stationnaires (avec linéarité de l'espérance des accroissements), plus connue sous le nom d'hypothèse intrinsèque en géostatistique et de stationnarité locale en physique. Le champ aléatoire est alors caractérisé par les deux premiers moments de ces accroissements. La moyenne des accroissements (la dérive) est une fonction linéaire $m_\varepsilon : \boldsymbol{h} \in \mathbb{R}^3 \to m_\varepsilon(\boldsymbol{h}) \in \mathbb{R}$:

$$m_\varepsilon(\boldsymbol{h}) = \mathbb{E}[\varepsilon(\boldsymbol{r}+\boldsymbol{h}) - \varepsilon(\boldsymbol{r})] = \boldsymbol{\alpha}^t \boldsymbol{h} \ , \ \forall (\boldsymbol{r}, \boldsymbol{h}) \in \mathbb{R}^3 \times \mathbb{R}^3 \ , \tag{2.4}$$

et avec $\boldsymbol{\alpha} \in \mathbb{R}^3$.

Pour la covariance des accroissements, il est facile de montrer qu'elle peut s'exprimer à l'aide de la fonction "variance d'accroissements" (cf. par exemple Iooss [92]). Cette dernière fonction est donc utilisée pour définir le moment d'ordre deux des accroissements. La demi-variance des accroissements $\gamma_\varepsilon : \boldsymbol{h} \in \mathbb{R}^3 \to \gamma_\varepsilon(\boldsymbol{h}) \in \mathbb{R}$, fonction appelée variogramme en géostatistique (Matheron [147]) et fonction de structure en physique (Yaglom [231]), ne dépend pas des points d'appui mais seulement de leur différence :

$$\gamma_\varepsilon(\boldsymbol{h}) = \frac{1}{2}\mathrm{Var}[\varepsilon(\boldsymbol{r}+\boldsymbol{h}) - \varepsilon(\boldsymbol{r})] \ , \ \forall (\boldsymbol{r}, \boldsymbol{h}) \in \mathbb{R}^3 \times \mathbb{R}^3 \ . \tag{2.5}$$

Ainsi, la théorie de la propagation d'onde en milieu aléatoire s'applique si le champ de vitesse est intrinsèque dans la direction de propagation et stationnaire perpendiculairement (Tatarskii [211]). Une telle généralisation permet de prendre en compte les variations linéaires de la moyenne et de la variance de la vitesse dans la direction de propagation.

Un autre modèle de non stationnarité plus général, nommé quasi-stationnarité, est utilisé par Rytov et al. [180, 181] et a été repris récemment par Kravtsov et al [127, 126] :

$$\mathrm{Cov}[\varepsilon(\boldsymbol{r}_1), \varepsilon(\boldsymbol{r}_2)] = \sigma_\varepsilon^2(\boldsymbol{r}_+) K(\boldsymbol{r}_1 - \boldsymbol{r}_2, \boldsymbol{r}_+) \ , \ \forall (\boldsymbol{r}_1, \boldsymbol{r}_2) \in \mathbb{R}^3 \times \mathbb{R}^3 \ , \tag{2.6}$$

où $\boldsymbol{r}_+ = (\boldsymbol{r}_1 + \boldsymbol{r}_2)/2$ est le vecteur situant le centre de gravité de $\boldsymbol{r}_1$ et $\boldsymbol{r}_2$, $\sigma_\varepsilon^2(\cdot)$ est la variance de ε (fonction non constante) et $K(\boldsymbol{r}_1 - \boldsymbol{r}_2, \cdot)$ est une fonction de covariance normalisée qui vaut un quand $\boldsymbol{r}_1 = \boldsymbol{r}_2$. Ce modèle de fluctuations quasi-stationnaires (appelées localement stationnaires en statistique) permet à la variance et aux longueurs de corrélation d'évoluer lentement dans une direction. La théorie de la propagation d'ondes en milieu aléatoire peut ainsi être développée dans ce type de milieux, $\boldsymbol{r}_+$ étant la direction principale de propagation de l'onde. Ceci présente, par exemple, un intérêt en sismologie et en océanographie où la structure des hétérogénéités peut varier en fonction de la profondeur.

2.2.3 Modélisation stochastique de la turbulence

Les deux sections précédentes ont traité d'un champ aléatoire scalaire représentant dans notre contexte les fluctuations de la célérité des ondes. En milieu immobile (typiquement solide) et pour des ondes acoustiques, seul ce champ aléatoire scalaire est introduit dans l'équation d'ondes. Dans des milieux fluides, la célérité du son dépend de la température résultant de la turbulence thermique et une relation directe relie ces deux champs scalaires (célérité et température). En milieu en mouvement, un autre paramètre intervient dans l'équation d'ondes : la vitesse du fluide résultant de la turbulence cinématique. Il s'agit maintenant d'un champ aléatoire vectoriel $\boldsymbol{v}(\boldsymbol{r}) = (v_1(\boldsymbol{r}), v_2(\boldsymbol{r}), v_3(\boldsymbol{r})) \in \mathbb{R}^3$ (où $\boldsymbol{r} \in \mathbb{R}^3$).

Un champ aléatoire vectoriel stationnaire (resp. intrinsèque) est défini par ses fonctions de covariance $C_{ij}(\cdot)$ (resp. variogrammes $\gamma_{ij}(\cdot)$) directes et croisées ($i, j = 1..3$) :

$$C_{ij}(\boldsymbol{h}) = \mathrm{Cov}[v_i(\boldsymbol{r}), v_j(\boldsymbol{r}+\boldsymbol{h})] \ , \ \forall (\boldsymbol{r}, \boldsymbol{h}) \in \mathbb{R}^3 \times \mathbb{R}^3 \ . \tag{2.7}$$

La théorie de Kolmogorov (dont la première version date de 1941) permet de décrire en termes statistiques la structure d'une turbulence cinématique. Elle est basée sur une hypothèse de cascade d'énergie : quand le nombre critique de Reynolds est dépassé, des instabilités locales se créent, de tailles plus faibles que l'échelle caractéristique et indépendantes de l'écoulement géniteur. Puis, pendant le temps de retournement d'un tourbillon, celui-ci perd une fraction de son énergie cinétique pour former un tourbillon de taille inférieure (Frisch [71]). La cascade se poursuit jusqu'à une taille d'hétérogénéité où l'énergie se dissipe sous forme de chaleur par dissipation visqueuse. Cette échelle inférieure limite l_0, appelée échelle de dissipation de Kolmogorov, dépend donc de la viscosité cinématique du fluide. En supposant que la turbulence est intrinsèque et isotrope, Kolmogorov obtient la forme du spectre d'énergie des tourbillons à l'intérieur du domaine de transfert d'énergie $[2\pi/L_0, 2\pi/l_0]$ où $L_0 \in \mathbb{R}_+^*$ et $l_0 \in \mathbb{R}_+^*$ sont les échelles d'injection et de dissipation d'énergie (Lesieur [132], Frisch [71]).

Ces arguments physiques permettent de déduire les structures de covariance adaptées à la caractérisation stochastique des milieux turbulents que l'on va utiliser dans la section suivante. On utilise usuellement la densité spectrale $\Phi : \boldsymbol{k} \in \mathbb{R}^3 \to \Phi(\boldsymbol{k}) \in \mathbb{R}$, transformée de Fourier de la covariance $C(\cdot)$:

$$\Phi(\boldsymbol{k}) = \frac{1}{(2\pi)^3} \iiint_{\mathbb{R}^3} C(\boldsymbol{r}) \exp(-i\boldsymbol{k}.\boldsymbol{r})\, d\boldsymbol{r} \ , \ \forall \boldsymbol{k} \in \mathbb{R}^3 \ . \tag{2.8}$$

La Figure 2.5 donne un exemple d'écoulement turbulent 2D dont les fluctuations de vitesse isotropes ont une densité spectrale de Kolmogorov : $\Phi_{v_1}(\boldsymbol{k}) = \Phi_{v_2}(\boldsymbol{k}) = \alpha \|\boldsymbol{k}\|^{-11/3}$ où $\alpha \in \mathbb{R}_+^*$, $v_1(\cdot)$ et $v_2(\cdot)$ sont les composantes horizontale et verticale (supposées indépendantes) des fluctuations de vitesse. Bien entendu, ce spectre est simpliste et de nombreux travaux basées sur des hypothèses plus réalistes ont été réalisés depuis la détermination de ce spectre (modélisant par exemple les phénomènes d'intermittence, cf. Frisch [71]).

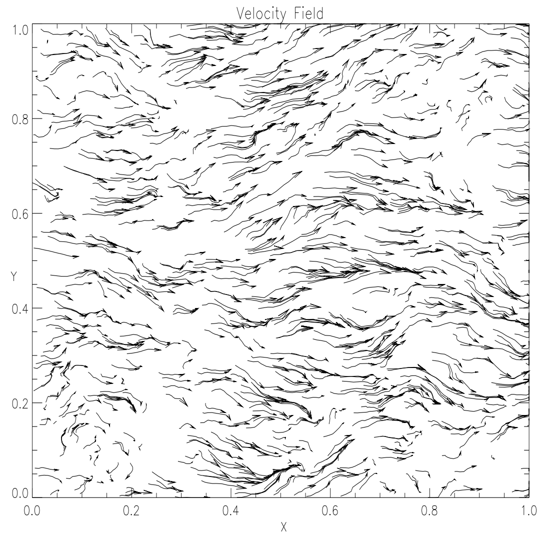


FIG. 2.5 – Exemple d'écoulement fluide bidimensionnel turbulent de vitesse moyenne $\boldsymbol{v}_0 = (10$ m/s, 0 m/s). Perturbations de vitesse à spectre de Kolmogorov pour chaque composante de la vitesse : $L_0 = 0.2$ m, $l_0 = 0.002$ m, écart types $\sigma_{v_1} = \sigma_{v_2} = 5$ m/s. Les unités des axes sont en mètres.

Remarque 2.2.1 Dans la suite, on se limite à l'hypothèse de milieu figé, qui est valide pour un faible nombre de Mach moyen (rapport de la vitesse moyenne du fluide sur la célérité moyenne des ondes acoustiques). Cette hypothèse est largement respectée dans les applications industrielles que j'ai étudiées (en hydraulique, cf. Iooss et al. [98]). En dehors de cette hypothèse, il faudrait prendre en compte en plus l'évolution temporelle du milieu lors de la propagation de l'onde. Utiliser des modèles statistiques évoluant avec le temps est un problème encore ouvert pour la propagation d'ondes en milieu turbulent en mouvement. En géostatistique, de nombreux travaux sont d'ores et déjà disponibles sur la simulation de champs aléatoires spatio-temporels (Kyriakidis & Journel [130]).

2.2.4 Quelques familles de fonctions de covariance et de variogrammes

Dans cette section, on présente trois grandes familles de modèles paramétriques de covariance ou de variogramme utilisées dans les domaines abordés dans mes travaux (géostatistique, géophysique, mécanique des fluides, modélisation de codes numériques). Les ouvrages de Chilès & Delfiner [42], Cressie [44], Abrahamsen [2] dressent une liste plus complète de toutes les fonctions de covariance avec leurs avantages et leurs inconvénients. On note $d \in \mathbb{N}^*$ la dimension du support du champ aléatoire ($d = 2$ ou 3 en propagation d'ondes mais éventuellement supérieure en modélisation des réponses de codes de calcul).

- ⋆ Soient $\boldsymbol{r} = (r_1, \ldots, r_d) \in \mathbb{R}^d$ et $(a_i)_{i=1..d}$ les longueurs de corrélation du champ aléatoire ($a_i \in \mathbb{R}_+^*$), le modèle exponentiel généralisé s'écrit

$$N(\boldsymbol{r}) = \prod_{i=1}^{d} \exp\left(-\frac{|r_i|}{a_i}\right)^{p_i} \;,\; \text{avec} \;\; 0 < p_i \leq 2 \;\forall\, i = 1 \ldots d \;, \tag{2.9}$$

où $(p_i)_{i=1..d}$ sont les paramètres puissances du modèle. C'est une covariance factorisée qui n'est pas standardisable (pas d'anisotropie géométrique). Ce modèle possède un grand nombre de degrés de liberté (deux fois plus de paramètres à ajuster qu'il y a de dimensions), ce qui explique l'intérêt qu'il suscite dans la modélisation des réponses de codes de calcul. De plus, il a la propriété de se factoriser, ce qui permet de simplifier les intégrations multidimensionnelles. Pour modéliser des milieux aléatoires, les paramètres puissance sont généralement pris égaux entre eux : $p_1 = \ldots = p_d = p$ (Diggle & Ribero [58]).

Si $p = 2$, on obtient une covariance de forme gaussienne, infiniment dérivable à l'origine, qui modélise des milieux extrêmement réguliers et lisses (cf. Fig. 2.3 (a) pour une réalisation d'un milieu à covariance gaussienne). De par sa maniabilité dans les expressions analytiques, cette fonction est l'une des plus populaires en physique et en géostatistique. Des arguments théoriques (matrice de covariance particulièrement mal conditionnée) déconseillent cependant son utilisation (Stein [203]). Si $p = 1$, on obtient une covariance de forme exponentielle, également très populaire dans les applications (cf. Fig. 2.3 (b) pour une réalisation d'un milieu à covariance exponentielle).

- ⋆ Le modèle de Matèrn (appelée covariance K-Bessel en géostatistique) s'écrit (Matèrn [146])

$$C_0(h) = \frac{2^{1-p}}{\Gamma(p)} |h|^p K_p(|h|) \;,\; \forall h \in \mathbb{R} \;, \tag{2.10}$$

où $p \in \mathbb{R}_+$ est le paramètre puissance du modèle, $\Gamma(\cdot)$ est la fonction gamma et $K_p(\cdot)$ est la fonction de Bessel modifiée du second type et d'ordre p. Ce modèle modélise des milieux géométriquement anisotropes (Eq. (2.10)), mais il possède également une formulation factorisée (Santner et al. [191]). Le cas $p = 0.5$ correspond à une covariance exponentielle, le cas $p = 0$ correspond à un milieu dit de type fractal (car il possède des propriétés d'auto-similarité) alors que la covariance gaussienne est obtenue quand $p \to \infty$. Le modèle de Matèrn est particulièrement recommandé, notamment par Stein [203], car il permet de modéliser des champs aléatoires avec différents degrés de régularité. La covariance (2.10) est $\lceil p \rceil - 1$ fois différentiable, où $\lceil . \rceil$ correspond à la fonction "partie entière supérieure".

En physique, la covariance de Matèrn est appelée covariance de von Karman et provient de considérations basées sur la théorie de Kolmogorov qui donne une forme analytique au spectre d'énergie dans le domaine des nombres d'onde (cf. §2.2.3). Le variogramme standardisé du modèle de Kolmogorov s'écrit ($\gamma_0 : r \in \mathbb{R} \to \gamma_0(r) \in \mathbb{R}$) :

$$\left\{ \begin{array}{rcl} \gamma_0(0) & = & 0 \;, \\ \gamma_0(h) & = & \alpha |h|^{\frac{2}{3}} \;,\; \forall h \in \mathbb{R} \text{ tel que } l_0 \ll |h| \ll L_0 \;, \end{array} \right. \tag{2.11}$$

et avec $\alpha \in \mathbb{R}_+^*$. Ce variogramme modélise correctement les structures fines de la turbulence dans l'intervalle $[2\pi/L_0, 2\pi/l_0]$. Pour éviter la singularité du spectre de Kolmogorov dans les basses fréquences et intégrer une modélisation correcte des grosses structures, le spectre de von Karman prend en compte une fréquence de coupure. Il correspond au modèle de Matèrn (Eq. (2.10)) avec $p = 1/3$ en 3D et $p = 5/6$ en 2D. Ce modèle est fréquemment utilisé pour modéliser les milieux cinématiquement turbulents. Pour généraliser ce modèle à la prise en compte des hautes fréquences (petites hétérogénéités), il est possible d'introduire dans son spectre un filtre gaussien et ainsi obtenir le modèle de von Karman modifié. La fonction de covariance associée se nomme covariance de Kummer car elle fait intervenir la fonction hypergéométrique de confluence de Kummer (Andrews & Phillips [5]).

⋆ Le modèle puissance n'est pas stationnaire mais intrinsèque. Son variogramme standardisé s'écrit (Chilès & Delfiner [42])

$$\gamma_0(h) = |h|^p \ , \ \forall h \in \mathbb{R} \ , \tag{2.12}$$

où $0 < p < 2$ est le paramètre puissance du modèle. En fait, ce variogramme correspond à celui d'un mouvement brownien fractionnaire. Ce modèle a des propriétés d'auto-similarité car on ne peut pas y associer d'échelles caractéristiques du phénomène. En effet, si $a \in \mathbb{R}^*$, $\gamma_0(h/a) = |a|^{-p}\gamma_0(h)$. Il est donc invariant sous un changement d'échelle d'observation. Le cas limite $p = 0$ correspond à l'effet de pépite pur (aucune corrélation) où le processus est un bruit blanc gaussien. Pour $p = 1$, on obtient le variogramme du mouvement brownien standard ($\gamma_0(h) = |h|$).

2.3 Propagation d'ondes acoustiques hautes fréquences en milieu aléatoire

Différentes méthodologies peuvent être mises en œuvre dans l'étude d'un phénomène physique gouverné par des équations mathématiques : une approche théorique, une approche numérique et une approche expérimentale. L'idéal est de pouvoir confronter les résultats de ces trois approches. En propagation d'ondes en milieu aléatoire, l'approche théorique consiste à formuler les moments statistiques du champ d'onde en fonction de ceux du milieu de propagation. Cela est possible dans les cas d'école, *e.g.* ondes monochromatiques, espace libre sans obstacle ni réflecteur, ondes planes ou sphériques, champs moyens sans effet. La technique numérique, quant-à elle, consiste à simuler à l'aide d'un code de calcul la propagation des ondes à l'intérieur de différentes réalisations (issues de simulations) du milieu turbulent (défini par ses deux premiers moments statistiques). Le moyennage final (pour un grand nombre de simulations) des champs d'onde obtenus permet de calculer leurs caractéristiques statistiques, sans qu'il y ait de relation explicite avec les moments du champ turbulent. Finalement, l'approche expérimentale consiste à propager, de manière naturelle ou artificielle, une onde de source connue dans un milieu hétérogène dont on connaît la structure de covariance (ou le spectre), et à enregistrer le champ d'onde transmis à l'aide d'un dispositif d'acquisition adéquat.

Mes recherches en propagation d'ondes en milieu aléatoire ont été guidées de manière assez complémentaire par les trois approches suivantes :

▷ d'un point de vue théorique, j'ai opté pour une approche "physique" des problèmes, en m'appuyant sur une bibliographie fournie (principalement celle de Chernov, Tatarskii, Rytov, Kravtsov et Ostashev). Une telle approche permet d'avoir une bonne intuition des phénomènes que l'on néglige lors des simplifications d'équations. Pour plus de rigueur sur les aspects probabilistes et mathématiques du sujet, on peut se référer par exemple à Hoffman [88], Frisch [70] ou Bal [13]. J'ai obtenu quelques résultats théoriques nouveaux, notamment avec Philippe Blanc-Benon et Christian Lhuillier. L'un de ces résultats (non linéarité de la variance des temps en fonction de la distance de propagation) a été par la suite reproduit de manière expérimentale par Andreeva & Durgin [4] (cf. §2.3.4) ;

- ▷ mes travaux en modélisation numérique se sont principalement appuyés sur la méthode de résolution de l'équation d'onde acoustique par différences finies, méthode de référence pour valider et étudier la robustesse de résultats théoriques. Cette méthode de simulation, particulièrement lourde à mettre en œuvre, a été introduite par Frankel & Clayton [69] pour la propagation d'ondes en milieu aléatoire et je l'ai systématiquement utilisée pour valider mes développements. À l'aide de la simulation numérique, j'ai aussi pu résoudre une problématique industrielle posée par EDF et liée à la débitmétrie ultrasonore (cf. Fig. 2.2, Iooss et al. [98]).

- ▷ au niveau expérimental, en utilisant des données d'exploration sismique pétrolière, j'ai pu confirmer dans Iooss [92] et Iooss et al. [97] l'un des phénomènes prédit par Touati [214], à savoir la décroissance de la variance des temps de trajet avec l'offset (distance source-récepteur, cf. Fig. 2.1).

Dans cette section, on note $X \in \mathbb{R}^+$ la distance de propagation de l'onde étudiée, et $l_\parallel \in \mathbb{R}_+^*$ et $l_\perp \in \mathbb{R}_+^*$ les longueurs de corrélation (des fluctuations de vitesse de l'onde) parallèle et transverse à la direction de propagation de l'onde. Pour pouvoir effectuer des statistiques sur un champ d'onde propagé, il faut qu'il ait traversé beaucoup d'hétérogénéités, ce qui se note en terme d'approximation physique par

$$l_\parallel \ll X \; . \tag{2.13}$$

Pour simplifier, on se limite dans la suite aux milieux 3D stationnaires et transversalement isotropes (où $l_\perp = l_{\perp 1} = l_{\perp 2}$). Les résultats que l'on montre dans les sections suivantes peuvent être étendus sans réelles difficultés aux milieux anisotropes où $l_{\perp 1} \neq l_{\perp 2}$, aux milieux intrinqèques et même quasi-stationnaires.

2.3.1 Approximation de Rytov parabolique

Mes travaux se sont concentrés sur l'étude de la propagation des ondes acoustiques. L'équation de Helmholtz s'écrit en espace libre (hors de la zone source), dans un milieu 3D immobile et en régime périodique (ondes monochromatiques de longueur d'onde $\lambda_0 = 2\pi/k_0$ avec $k_0 \in \mathbb{R}_+^*$ appelé nombre d'onde) de la manière suivante :

$$\Delta u(\boldsymbol{r}) + k_0^2[1 + \varepsilon(\boldsymbol{r})]u(\boldsymbol{r}) = 0 \; , \tag{2.14}$$

où $u : \boldsymbol{r} \in \mathbb{R}^3 \to u(\boldsymbol{r}) \in \mathbb{Z}$ est le champ de l'onde monochromatique (défini par une amplitude et une phase) et $\varepsilon(\boldsymbol{r}) \in \mathbb{R}$ est une perturbation de la lenteur (inverse de la vitesse) au carré :

$$\frac{1}{c^2(\boldsymbol{r})} = \frac{1}{c_0^2}[1 + \varepsilon(\boldsymbol{r})] \; , \tag{2.15}$$

avec $c(\boldsymbol{r}) \in \mathbb{R}_+^*$ le champ de vitesse de l'onde, $c_0 \in \mathbb{R}_+^*$ une constante et $\mathbb{E}[\varepsilon(\boldsymbol{r})] = 0$. L'équation (2.14) est une équation aux dérivées partielles linéaire de type elliptique, dont l'un des coefficients est une fonction aléatoire. Elle appartient donc à la classe des équations linéaires stochastiques. L'équation de Helmholtz n'est pas soluble analytiquement et de multiples approximations ont donc été développées pour la résoudre, parmi lesquelles les méthodes intégrales (Kirchhoff), les méthodes de perturbation (Born, Rytov), l'optique géométrique et les méthodes numériques. Dans mes travaux, je me suis principalement intéressé aux méthodes hautes fréquences (approximation parabolique, approximation de Rytov et optique géométrique) qui permettent de définir la notion de temps de trajet, quantité utilisée dans mes applications industrielles.

On se place dans le régime des petites perturbations de lenteur (hétérogénéités "faibles") :

$$\sigma_\varepsilon = o(1) \; , \tag{2.16}$$

où $\sigma_\varepsilon \in \mathbb{R}_+$ est l'écart type de ε. En physique, la condition (2.16) s'écrit $\sigma_\varepsilon \ll 1$. Il est possible de simplifier l'équation (2.14) quand on s'intéresse uniquement au champ d'onde primaire, en approximant le front d'onde sphérique par un front parabolique. Cette approximation, nommée approximation parabolique, revient à ne considérer que des diffractions vers l'avant et à l'intérieur d'un petit cône autour de la direction principale de propagation à laquelle on s'intéresse. Dans la littérature, elle est aussi connue sous le nom de la méthode de l'équation parabolique (Tappert [207]). En raisonnant en milieu aléatoire anisotrope, j'ai montré :

Théorème 2.3.1 *Si les conditions*

$$\frac{\lambda_0}{l_\perp} = O\left(\frac{\lambda_0}{l_\parallel}\right) \ , \ \frac{\lambda_0}{l_\perp} = o\left(\left(\frac{\lambda_0}{X}\right)^{\frac{1}{4}}\right) \quad \textit{et} \quad X\sigma_b = o(1) \tag{2.17}$$

sont respectées (avec $\sigma_b \in \mathbb{R}_+$ section effective de rétrodiffraction, fonction de λ_0 et $C_\varepsilon(\cdot)$), alors, quand $\frac{\lambda_0}{l_\perp} \to 0$, la limite de $u^*(\boldsymbol{r})\ \forall \boldsymbol{r} \in \mathbb{R}^3$ solution de l'équation (2.14), si elle existe, tend vers $u^{**}(\boldsymbol{r})$ solution de l'équation (2.18) :

$$2ik_0\partial_\parallel[u(\boldsymbol{r})\exp(-ik_0X)] + \Delta_\perp[u(\boldsymbol{r})\exp(-ik_0X)] + k_0^2\varepsilon(\boldsymbol{r})[u(\boldsymbol{r})\exp(-ik_0X)] = 0 \ . \tag{2.18}$$

Les opérateurs $\partial_\parallel$ et $\Delta_\perp$ représentent respectivement la dérivée partielle parallèle et le laplacien transverse à la direction de propagation (le repère est centré sur l'axe de propagation principal de l'onde).

Les conditions asymptotiques du théorème ($\lambda_0 \ll l_\perp$ et $\lambda_0 \ll l_\parallel$) illustrent le fait que l'on se place dans un cadre haute fréquence : les hétérogénéités sont grandes devant la longueur d'onde. La deuxième condition de (2.17), qui se traduit par $\sqrt{\lambda_0 X} \ll \frac{l_\perp^2}{\lambda_0}$, correspond à l'approximation du front sphérique par le front parabolique alors que la troisième revient à négliger les diffractions arrières (condition aui a moins d'influence par rapport aux autres). Tatarskii [211] et Rytov et al. [181] ont donné la deuxième condition de validité dans un milieu isotrope : $\sqrt{\lambda_0 X} \ll \frac{l_\varepsilon^2}{\lambda_0}$ où l_ε est la longueur de corrélation des fluctuations de vitesse. Ma contribution dans Iooss [92, 91] a été de l'exprimer en milieu anisotrope (en faisant apparaître la longueur de corrélation transverse $l_\perp$). J'ai également confirmé la dépendance de l'approximation parabolique à cette condition grâce à des simulations numériques (méthode des différences finies sur l'équation d'onde).

Remarque 2.3.1 Si le milieu est en mouvement, l'équation parabolique contient des termes supplémentaires faisant intervenir les composantes du champ de vitesse vectoriel $\boldsymbol{v}(\boldsymbol{r}) \in \mathbb{R}^3$ du milieu. Ostashev [163] montre que l'on peut se ramener à l'équation (2.18) en remplaçant la perturbation ε par une perturbation effective $\varepsilon_{\mathit{eff}} = \varepsilon + 2v_\parallel/c_0$, où $v_\parallel$ est la composante de $\boldsymbol{v}$ parallèle à la direction de propagation principale de l'onde. La condition nécessaire à cette approximation est la suivante :

$$\left(\frac{l_\perp}{\lambda_0}\right)^2 \frac{\sigma_{v_\perp}^2}{c_0^2} = o(1) \ , \tag{2.19}$$

où $l_\perp$ est la longueur de corrélation transverse des hétérogénéités de vitesse et $\sigma_{v_\perp}^2$ est la variance de la composante transverse à la direction de propagation du champ de vitesse du milieu. Cette condition permet de négliger l'influence des composantes transverses de la vitesse. Tous les résultats que l'on montre dans la suite de ce mémoire concernent des milieux immobiles (avec une covariance $C_\varepsilon(\cdot)$), mais pourront être appliqués directement aux milieux en mouvement (avec une covariance effective) si la condition (2.19) est valide.

Pour faire apparaître le temps d'arrivée de l'onde à partir des équations (2.14) et (2.18), on représente le champ d'onde au moyen de l'exponentielle complexe $\Psi : \boldsymbol{r} \in \mathbb{R}^3 \to \Psi(\boldsymbol{r}) \in \mathbb{Z}^+$ (Chernov [41]) :

$$u(\boldsymbol{r}) = \exp[\Psi(\boldsymbol{r})] \ , \ \text{où} \ \ \Psi(\boldsymbol{r}) = \log[A(\boldsymbol{r})] + ik_0c_0T(\boldsymbol{r}) \ , \tag{2.20}$$

avec $A : \boldsymbol{r} \in \mathbb{R}^3 \to A(\boldsymbol{r}) \in \mathbb{R}_+^*$ et $T : \boldsymbol{r} \in \mathbb{R}^3 \to T(\boldsymbol{r}) \in \mathbb{R}^+$ l'amplitude et le temps d'arrivée de l'onde. Cette nouvelle représentation permet notamment de découpler le terme en $\varepsilon(\boldsymbol{r})u(\boldsymbol{r})$ de l'équation (2.18), ce qui permettra d'obtenir des équations exprimant les moments de $u(\boldsymbol{r})$.

L'approximation de Rytov (aussi appelée méthode des perturbations lisses) consiste alors à supposer que le champ d'onde est faiblement déformé par les hétérogénéités qu'il rencontre, de telle sorte que l'on puisse négliger les termes d'ordre supérieur à un dans la série asymptotique de Ψ (en puissances de ε). Au premier ordre, on obtient une solution approchée explicite de l'équation d'onde parabolique (2.18) (Rytov et al. [181]) :

Théorème 2.3.2 *Si la condition*

$$\mathbb{E}[(\lambda_0 \nabla_\perp \Psi)^2] = o(\sigma_\varepsilon) \tag{2.21}$$

est respectée, alors, quand $\sigma_\varepsilon \to 0$, la limite de $u^{**}(\boldsymbol{r})$ $\forall \boldsymbol{r} \in \mathbb{R}^3$ solution de l'équation (2.18), si elle existe, tend vers $u_1(\boldsymbol{r})$ dont l'expression est

$$\log[u_1(\boldsymbol{r})] = \log[u_0(\boldsymbol{r})] - \frac{k_0^2}{4\pi} \int_0^X \iint_{-\infty}^{\infty} \frac{1}{(X-x')\eta(x')} \exp\left[\frac{ik_0}{2} \frac{(\eta(x')\boldsymbol{y} - \boldsymbol{y}')^2}{(X-x')\eta(x')}\right] \varepsilon(x', \boldsymbol{y}')\, d\boldsymbol{y}' dx' \ , \tag{2.22}$$

où $\eta(x) = 1$ et $u_0(\boldsymbol{r}) = e^{ik_0\|\boldsymbol{r}\|}$ pour une onde plane et $\eta(x) = \dfrac{x}{X}$ et $u_0(\boldsymbol{r}) = \dfrac{e^{ik_0\|\boldsymbol{r}\|}}{4\pi\|\boldsymbol{r}\|}$ pour une onde sphérique.

L'approximation (2.22), nommée approximation de Rytov, est indispensable pour pouvoir travailler avec les temps d'arrivée. En effet, sa non validité implique que le champ d'onde a subi une forte atténuation (Bailly et al. [12]) et qu'il est alors impossible de relever les temps de trajet de l'onde sur le signal devenu incohérent (Samuelides [189]). Ceci a été confirmé par simulations numériques dans Iooss [92]. La condition (2.21) signifie que le champ d'onde est faiblement déformé sur des distances de l'ordre de la longueur d'onde. Un domaine de validité plus explicite que $\mathbb{E}[(\lambda_0 \nabla_\perp \Psi)^2] \ll \sigma_\varepsilon$ a été donné en milieu isotrope (Shapiro et al. [198], Samuelides [189]) : $\sigma_\varepsilon^2 X l_\varepsilon \ll \lambda_0^2$ où l_ε est la longueur de corrélation des fluctuations de vitesse. Grâce à des arguments heuristiques, j'ai étendu cette condition de validité pour les milieux anisotropes dans Iooss [92] en la confirmant par simulations numériques (méthode des différences finies sur l'équation d'onde). L'heuristique que j'ai proposée est la suivante :

Heuristique 2.3.1 *En milieu anisotrope, la condition (2.21) de validité de l'approximation de Rytov est équivalente à la condition*

$$\frac{\sigma_\varepsilon^2 X l_\parallel}{\lambda_0^2} = o(1) \ . \tag{2.23}$$

Cette condition (2.23) a été démontrée récemment par Saito [183].

L'approximation de Rytov parabolique (*i.e.* approximation de Rytov sur l'équation d'onde parabolique) consiste donc à considérer des ondes faiblement diffractées vers l'avant de la propagation. Ce domaine de validité est pleinement compatible avec les applications en sismique pétrolière où les ordres de grandeur sont cohérents avec ce régime (longueur d'onde décamétrique, hétérogénéités de taille hectométrique, distance de propagation d'ordre kilométrique).

2.3.2 Optique géométrique

Dans mes travaux, je me suis également intéressé à un cadre plus restrictif que l'approximation de Rytov parabolique, également plausible dans de nombreuses applications, et qui permet d'aller plus loin au niveau analytique. C'est le domaine de l'optique géométrique, cadre très haute fréquence de la propagation des ondes dans lequel plus aucun phénomène de diffraction n'est pris en compte, valide sous les conditions

$$\frac{\lambda_0}{l_\perp} = O\left(\frac{\lambda_0}{l_\parallel}\right) = o(1) \ . \tag{2.24}$$

Le développement dans l'équation de Helmholtz (2.14) de l'amplitude de l'onde (cf. Eq. (2.20)) sous la forme d'une série de Taylor proportionnelle à la longueur d'onde (expansion de Debye, cf. Kravtsov [126] pour les milieux immobiles et Ostashev [163] pour les milieux en mouvement) conduit à un système d'équations indépendantes de la fréquence. La première équation de ce système nous intéresse plus particulièrement car ne faisant intervenir que le temps de trajet de l'onde :

$$[\nabla T(\boldsymbol{r})]^2 = \frac{1}{c_0^2}[1+\varepsilon(\boldsymbol{r})] \ . \tag{2.25}$$

Cette équation est connue sous le nom d'"eikonale" et décrit le principe de Fermat dans un milieu hétérogène. Elle est valide sous l'hypothèse d'absence de diffraction, *i.e.* quand la taille transverse de la zone de Fresnel du front d'onde est petite devant la taille transverse des hétérogénéités (Kravtsov [126]). Cette condition s'écrit en milieu aléatoire anisotrope (Samuelides & Mukerji [190])

$$\frac{\sqrt{\lambda_0 X}}{l_\perp} = o(1) \ , \tag{2.26}$$

et a été confirmée dans Iooss [92, 91] par simulations numériques (différences finies sur l'équation d'onde). Par rapport à la condition haute fréquence (2.24), la condition (2.26) se reécrit

$$\frac{\lambda_0}{l_\perp} = o\left(\frac{l_\perp}{X}\right) \ . \tag{2.27}$$

L'équation de l'eikonale (2.25) n'est pas analytiquement soluble. Des résolutions numériques très puissantes ont été développées, notamment celle basée sur la théorie des rais qui permet de visualiser les trajectoires de l'onde qui transportent son énergie. Les techniques basées sur le tracé de rais sont beaucoup moins coûteuses en temps de calcul que la résolution de l'équation d'onde complète. Par contre, la théorie des rais n'est plus valide lors de phénomènes singuliers tels que les caustiques (zones de convergence des rais) et les zones d'ombre (zones d'absence de rais) (Kravtsov [126]). Il a été démontré dans Kulkarny & White [129] et White [227] que les caustiques apparaissent à des distances de propagation de l'ordre de $l_\varepsilon \sigma_\varepsilon^{-2/3}$ en milieu isotrope (où l_ε représente la taille des hétérogénéités). Ceci a été observé numériquement par de nombreux auteurs, parmi lesquels Blanc-Benon et al. [23, 25] et Samuelides & Mukerji [190]. En milieu aléatoire anisotrope, Samuelides & Mukerji [190] supposent que c'est la taille transverse des hétérogénéités $l_\perp$ qui joue un rôle dans la condition de validité de la théorie des rais.

Pour résoudre analytiquement l'équation non linéaire (2.25), une méthode de perturbation est classiquement utilisée en développant T en série asymptotique en puissances de ε. À l'approximation au second ordre, on obtient (Boyse & Keller [27]) :

$$T(\boldsymbol{r}) = \frac{X}{c_0} + \frac{1}{2c_0}\int_0^X \varepsilon\left(\frac{u\boldsymbol{r}}{r}\right) du - \frac{1}{4c_0}\int_0^X \left[\int_0^u \frac{v}{u}\nabla_\perp \varepsilon\left(\frac{v\boldsymbol{r}}{r}\right) dv\right]^2 du \ . \tag{2.28}$$

Usuellement, seuls les deux premiers termes de cette équation sont utilisés (approximation au premier ordre). Quelques auteurs se sont cependant aussi intéressés aux effets des termes du deuxième ordre. J'ai pu expliciter dans Iooss et al. [94] la condition de validité de cette approximation en milieu anisotrope, en montrant :

Théorème 2.3.3 *Sous la condition*

$$\sigma_\varepsilon = o\left(\frac{l_\perp^2}{X^{\frac{3}{2}} l_\parallel^{\frac{1}{2}}}\right) \ , \tag{2.29}$$

alors $T(\boldsymbol{r}) \overset{\sigma_\varepsilon \to 0}{\longrightarrow} T^*(\boldsymbol{r})\ \forall \boldsymbol{r} \in \mathbb{R}^3$ *avec* $T^*(\boldsymbol{r})$ *s'exprimant par*

$$T^*(\boldsymbol{r}) = \frac{X}{c_0} + \frac{1}{2c_0}\int_0^X \varepsilon(\frac{u\boldsymbol{r}}{r})\, du \ . \tag{2.30}$$

Cette approximation au premier ordre est reliée à celle de la théorie des rais (apparition des caustiques). On voit que la condition (2.29), qui équivaut à une limite sur la distance de proopagation de l'onde

$$X \ll \left(\frac{l_\perp^4}{l_\parallel}\right)^{1/3} \sigma_\varepsilon^{-2/3} , \tag{2.31}$$

est plus complexe qu'une simple substitution de l_ε par $l_\perp$ dans la condition $X \ll \alpha l_\varepsilon \sigma_\varepsilon^{-2/3}$ qui avait été obtenue en milieu isotrope.

2.3.3 Moyenne des temps de trajet au second ordre : le "velocity shift"

Que ce soit dans l'approximation de Rytov parabolique (Eq. (2.22)) ou dans l'optique géométrique (Eq. (2.28)), l'espérance du terme au premier ordre est nulle car $\mathbb{E}[\varepsilon(\boldsymbol{r})] = 0$. Pour obtenir une approximation fine de la moyenne du temps de trajet, il faut donc calculer l'espérance du terme au deuxième ordre. Cet effet du deuxième ordre sur les temps de trajet n'a pas été évoqué dans la littérature jusqu'à l'utilisation des techniques de tracé de rais et de simulation numérique en sismologie dans les années 1980 (Müller et al. [154], Petersen [168]). Par la suite, les géophysiciens l'ont principalement caractérisé en terme de *velocity shift*, c'est-à-dire de déviation de la vitesse effective d'un milieu (*i.e.* la vitesse vue par l'onde) par rapport à la vitesse moyenne statistique. À l'approximation du deuxième ordre, la vitesse effective v_{eff} d'un milieu est définie par

$$v_{\text{eff}} = \mathbb{E}\left[\frac{X}{T(\boldsymbol{r})}\right] = \mathbb{E}\left(\frac{X}{X/c_0 + T_1(\boldsymbol{r}) + T_2(\boldsymbol{r})}\right) = c_0\left\{1 - \frac{c_0}{X}\mathbb{E}[T_2(\boldsymbol{r})] + O\left(\left(\frac{c_0}{X}T_1\right)^2\right)\right\} , \tag{2.32}$$

où $T(\boldsymbol{r})$ est le temps d'arrivée d'une onde de distance de propagation X, $T_1(\boldsymbol{r})$ et $T_2(\boldsymbol{r})$ sont les termes du premier ordre et second ordre dans le développement asymptotique de $T(\boldsymbol{r})$ (deuxième et troisième termes de l'équation (2.28)). Le *velocity shift* s'explique physiquement par le fait que l'onde choisit préférentiellement les trajectoires les plus rapides pour joindre une source à un récepteur (principe de Fermat).

Dans l'optique géométrique, Roth et al. [177], Mukerji et al. [153] et Boyse & Keller [27] se sont intéressés à l'estimation du *velocity shift* en milieu isotrope. Samuelides & Mukerji [190] ont abordé ce problème en milieu anisotrope à covariance gaussienne. Iooss [92] et Iooss et al. [94] ont étendu ce résultat aux milieux géométriquement anisotropes en introduisant la formulation de la covariance standardisée. La propriété que j'ai démontrée est la suivante :

Propriété 2.3.1 *En milieu aléatoire 3D, géométriquement anisotrope et transversalement isotrope, l'espérance des temps de trajet à l'ordre deux vaut*

$$\mathbb{E}[T_2(\boldsymbol{r})] = \frac{1}{\eta}\frac{\sigma_\varepsilon^2}{4c_0}X^2\frac{l_\parallel}{l_\perp^2}\int_0^\infty \frac{C_0'(u)}{u}\,du , \tag{2.33}$$

où $\eta = 1$ pour une onde plane et $\eta = 3$ pour une onde sphérique.

Dans les modèles de covariance classiques (strictement décroissants), l'intégrale sur C_0' est négative et donc $\mathbb{E}[T_2(\boldsymbol{r})]$ est négative. Ainsi, on confirme que $v_{\text{eff}} > v_0$, la vitesse du milieu effectif est supérieure à la vitesse moyenne du milieu réel. D'autre part, tant que l'optique géométrique est valide, le *velocity shift* augmente linéairement avec la distance de propagation X et dépend de $l_\parallel/l_\perp^2$: il augmente avec la taille longitudinale des hétérogénéités et diminue avec leurs tailles transverses.

En dehors de l'optique géométrique, dans le cadre plus général de l'approximation de Rytov parabolique, le calcul du *velocity shift* est à notre connaissance assez récent : Shapiro et al. [198] pour les milieux isotropes, Samuelides [189] pour les milieux anisotropes gaussiens et Saito [183] pour les milieux géométriquement anisotropes. Après une augmentation linéaire du *velocity shift* en fonction de X dans l'optique géométrique, son évolution sature dans l'approximation parabolique, puis il décroît très lentement vers zéro. Samuelides [189], Iooss [92] et Iooss & Samuelides [103] présentent des études détaillées du comportement du *velocity shift* en milieu anisotrope à l'aide de simulations numériques basées sur la méthode des différences finies sur l'équation d'onde. Ces travaux ont ainsi permis d'observer le phénomène de saturation du *velocity shift*.

2.3.4 Variance des temps de trajet au second ordre en optique géométrique

Dans l'optique géométrique, l'expression de la variance des temps de trajet au premier ordre est bien connue depuis l'ouvrage de Chernov [41] (pour une onde plane en milieu isotrope). Rytov et al. [181] ont étendu ce résultat aux milieux anisotropes. Avec la fonction de covariance standardisée, donc en milieu géométriquement anisotrope, on obtient l'expression (Touati [214])

$$\mathrm{Var}[T_1(\boldsymbol{r})] = \frac{\sigma_\varepsilon^2}{2c_0^2} X l_\parallel \int_0^\infty C_0(u)\, du \,. \tag{2.34}$$

Cette formule, appelée approximation de Chernov, montre une augmentation linéaire de la variance des temps en fonction de la distance de propagation. Touati [214] analyse en profondeur la dérivation de cette formule à partir de l'équation (2.30) et en déduit qu'elle surestime la vraie variance et qu'elle n'est valide que si l'hypothèse (2.13), à savoir $l_\parallel \ll X$, est respectée. Cette analyse ne prend cependant pas en compte le terme à l'ordre deux des temps de trajet (troisième terme de l'équation (2.28)). En supposant une perturbation ε de loi gaussienne et par analogie avec l'espérance, j'ai pu calculer analytiquement l'expression de la variance des temps au deuxième ordre dans Iooss et al. [94], ce qui nous donne la propriété suivante :

Propriété 2.3.2 *En milieu aléatoire 3D gaussien, géométriquement anisotrope et transversalement isotrope, la variance des temps de trajet à l'ordre deux vaut*

$$\mathrm{Var}[T_2(\boldsymbol{r})] = \frac{1}{\eta} \frac{\sigma_\varepsilon^4}{8c_0^2} X^4 \frac{l_\parallel^2}{l_\perp^4} \left[\int_0^\infty \frac{C_0'(u)}{u}\, du \right]^2 \,, \tag{2.35}$$

où $\eta = 1$ pour une onde plane et $\eta = 9$ pour une onde sphérique.

Dans le calcul de cette variance, l'hypothèse de normalité de ε est nécessaire pour simplifier le quatrième moment de ε en somme de produits de covariances (Papoulis & Pillai [165]). Grâce à la formule (2.35), on s'aperçoit que le terme à l'ordre deux induit une non linéarité pour l'évolution de la variance des temps en fonction de la distance de propagation (dépendance en X^4).

Négliger l'équation (2.35) devant (2.34) nous a permis de démontrer la condition de validité (2.31) de l'approximation des temps de trajet au premier ordre en milieu anisotrope. Cette condition avait été obtenue en milieu isotrope par Kulkarny & White [129] à l'aide d'une formulation différente de l'optique géométrique en milieu aléatoire. Basée sur les équations des rais en milieu aléatoire, cette formulation amène à la résolution d'un système d'équations différentielles stochastiques linéaires qui permet d'obtenir, entre autres, la probabilité d'apparition des caustiques en fonction de la distance de propagation. Dans Iooss et al. [94], nous avons confirmé le lien entre ces deux approches en reliant la variance des temps de trajet à la probabilité d'apparition des caustiques.

Le comportement non linéaire de la variance des temps à partir d'une certaine distance de propagation a été d'abord observé sur des simulations numériques (basées sur le tracé de rais) par Blanc-Benon et al. [24] et Karweit et al. [113]. Comme prévu, on peut relever la légère surestimation de l'approximation de Chernov par rapport aux simulations pour de courtes distances de propagation puis la croissance non linéaire de la variance des temps simulés. En utilisant une méthode plus robuste que le tracé de rais, à savoir la technique de sommation des faisceaux gaussiens (cf. Fiorina [64] pour son utilisation en milieu aléatoire), Iooss et al. [94] ont obtenu les mêmes résultats entre simulations numériques et prédictions issues de l'équation (2.35).

D'un point de vue expérimental en propagation d'ondes hautes fréquences en milieu turbulent, les études confrontant prédictions théoriques et expériences sont extrêmement rares. Ceci est dû aux difficultés de générer des milieux turbulents stables, de les caractériser proprement et d'extraire des temps de première arrivée des ondes à partir d'enregistrements de signaux souvent bruités. Karweit et al. [113] ont cependant réussi à confirmer expérimentalement la non linéarité de la variance des temps. Récemment, grâce à des moyens performants (pulsations ultrasonores dans une conduite à turbulence bien contrôlée), Andreeva & Durgin [4] ont pu mesurer finement les temps de première arrivée des ondes propagées. Ils ont alors obtenu des résultats expérimentaux proches des prédictions théoriques de l'équation (2.35). La Figure 2.6, extraite d'une présentation de Tatiana Andreeva, l'illustre pour plusieurs degrés de turbulence.

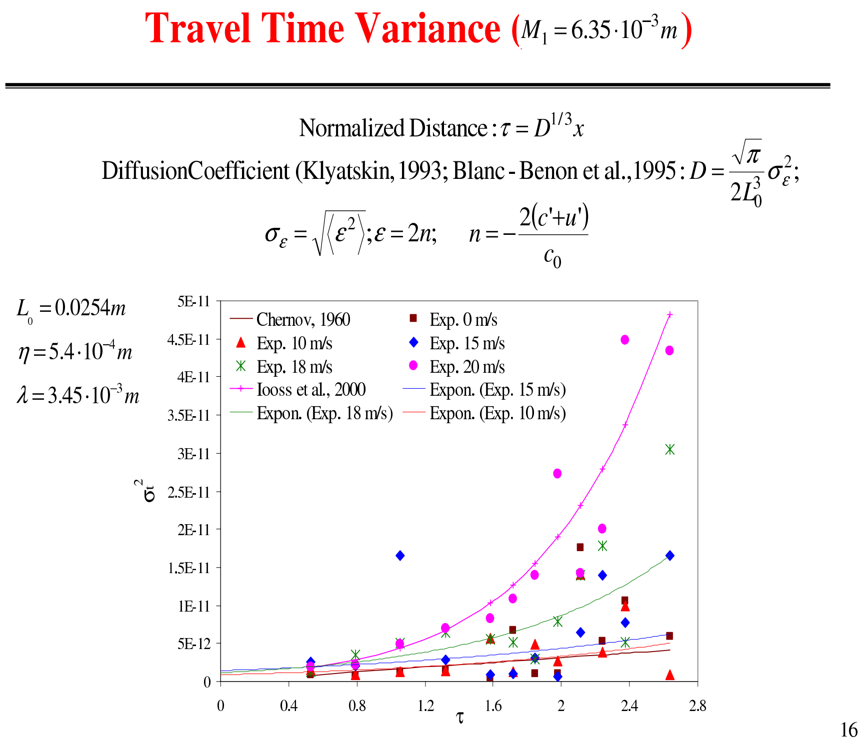


Fig. 2.6 – Comparaisons des variances de temps de trajet expérimentaux et théoriques (figure extraite d'Andreeva & Durgin [4]).

2.4 Tomographie statistique

Le problème inverse en propagation d'ondes consiste à estimer les propriétés physiques d'un milieu à partir d'enregistrements de champs d'ondes qui ont traversé ce milieu (procédure souvent appelée tomographie, cf. Tarantola [208]). Le même problème peut être envisagé en propagation d'ondes en milieu aléatoire : extraire les propriétés statistiques d'un milieu hétérogène à partir des statistiques sur les champs d'ondes qui s'y sont propagées. C'est un axe de recherche important en sismologie et en océanographie depuis les années 1970. Pour ce faire, on peut voir que l'utilisation de l'unique mesure de la variance des temps de trajet d'une onde (Eq. (2.34)) n'est pas suffisante, car elle ne permet pas de distinguer les différents paramètres statistiques de ε. Par contre, cette inversion est possible en travaillant sur la covariance des temps de trajet et des amplitudes (Aki [3], Uscinski [217]).

En me basant sur les travaux de Touati [214] en sismique réflexion et de Müller et al. [154] en sismologie, j'ai proposé une inversion directe en exprimant la covariance du champ de vitesse en fonction de celle des temps. Son application et sa confrontation à d'autres méthodes d'inversion ont été ensuite menées en collaboration avec David Geraets, doctorant m'ayant succédé sur le sujet. Une publication regroupant tous les acteurs du sujet (Iooss et al. [97]) a finalement permis de résumer ces travaux de recherche.

2.4.1 Covariance des temps de trajet

Dans l'approximation de Rytov parabolique et pour une onde sphérique, en utilisant l'approximation au premier ordre (2.22), la covariance entre les temps de trajet en deux points distincts a été obtenue par Ishimaru [105] en milieu isotrope et Rytov et al. [181] en milieu anisotrope. Dans Iooss

[92] et Iooss & Galli [96], celle-ci est exprimée pour un milieu géométriquement anisotrope en fonction de la covariance standardisée (ou plutôt de la densité spectrale standardisée Φ_0). Les propriétés que j'ai démontrées sont les suivantes (résultats donnés en 2D pour simplifier les expressions) :

Propriété 2.4.1 *En milieu 2D géométriquement anisotrope, pour une onde sphérique dont le rayon central est de longueur X, si $\boldsymbol{r}_1$ et $\boldsymbol{r}_2$ sont deux rayons distincts de même longueur séparés d'une distance $\rho = \|\boldsymbol{r}_1 - \boldsymbol{r}_2\| = o(X)$, la covariance des temps de trajet $T(\boldsymbol{r}_1)$ et $T(\boldsymbol{r}_2)$ vaut :*

$$\mathrm{Cov}[T(\boldsymbol{r}_1), T(\boldsymbol{r}_2)] = \frac{\pi\sigma_\epsilon^2}{2c_0^2} X l_\parallel \frac{l_\perp}{\rho} \int_0^{\rho/l_\perp}\int_{-\infty}^{\infty} \exp(i\kappa u)\cos^2\left[\frac{\kappa^2 X}{2k_0}\frac{u}{\rho}\left(\frac{1}{l_\perp} - \frac{u}{\rho}\right)\right]\Phi_0(\kappa)\, d\kappa\, du, \quad (2.36)$$

$$\mathrm{Var}[T(\boldsymbol{r})] = \frac{\pi\sigma_\epsilon^2}{2c_0^2} X l_\parallel \int_0^1 \int_{-\infty}^{\infty} \cos^2\left[\frac{\kappa^2 X}{2k_0}\frac{u}{l_\perp^2}(1-u)\right]\Phi_0(\kappa)\, d\kappa\, du\ . \quad (2.37)$$

Propriété 2.4.2 *Dans le régime asymptotique de l'optique géométrique, en milieu 2D, la covariance des temps de trajet d'une onde sphérique ($\boldsymbol{r}_1 \neq \boldsymbol{r}_2$, $\rho = \|\boldsymbol{r}_1 - \boldsymbol{r}_2\| = o(X)$) vaut :*

$$\mathrm{Cov}[T(\boldsymbol{r}_1), T(\boldsymbol{r}_2)] = C_T(\rho, X) = \frac{\sigma_\varepsilon^2}{2c_0^2} X l_\parallel \frac{l_\perp}{\rho}\int_0^{\rho/l_\perp}\int_0^\infty C_0(\sqrt{u^2+v^2})\, dv\, du\ , \quad (2.38)$$

$$\mathrm{Var}[T(\boldsymbol{r})] = C_T(0, X) = \frac{\sigma_\varepsilon^2}{2c_0^2} X l_\parallel \int_0^\infty C_0(u)\, du\ , \quad (2.39)$$

avec $C_T : (u,v) \in \mathbb{R}_+ \times \mathbb{R}_+ \to C_T(u,v) \in \mathbb{R}$.

Les expressions (2.38) et (2.39) peuvent être déduites des équations (2.36) et (2.37) à partir de la condition (2.26). Dans l'approximation inverse à l'optique géométrique (*i.e.* $\sqrt{\lambda_0 X} \gg l_\perp$), appelée approximation champ lointain ou approximation de Fraunhofer, la covariance et la variance des temps de trajet valent la moitié de celles de l'optique géométrique. De par cette proximité entre les covariances dans l'optique géométrique, l'approximation de Rytov parabolique et l'approximation de Fraunhofer, la mesure des fluctuations de temps de trajet (ou de phase) est connue pour être une mesure plutôt robuste, contrairement à celle sur les amplitudes (Barabanenkov et al. [14]).

La variance des temps de trajet de la propriété 2.4.1 est issue d'une approximation plus générale (l'approximation Rytov parabolique) que celle étudiée au §2.3.4 (l'optique géométrique). Des simulations numériques (par différences finies sur l'équation d'onde acoustique) ont pu reproduire les résultats sur cette variance des temps de trajet (Iooss [92], Iooss & Samuelides [103]). La variance des temps commence par croître linéairement avec la distance de propagation suivant la relation (2.37), en étant encadrée supérieurement par la variance de l'optique géométrique et inférieurement par celle de l'approximation de Fraunhofer. L'évolution de la variance des temps sature ensuite, à partir du moment où l'approximation de Rytov parabolique n'est plus valide. Ces simulations, non limitées au cadre de l'optique géométrique, ne reproduisent pas les effets non linéaires évoqués au §2.3.4 car les phénomènes de diffraction apparaissent plus tôt que ces effets. Pour envisager de reproduire la non linéarité de la variance à l'aide de simulations basées sur l'équation d'onde acoustique, il faudrait être capable de diminuer drastiquement la longueur d'onde de l'onde simulée. Ceci est difficile car le pas de discrétisation spatiale de la méthode des différences finies découle directement de cette valeur de longueur d'onde ; plus le pas de discrétisation est faible, plus la simulation est coûteuse en temps de calcul.

2.4.2 Inversion de la covariance du champ de vitesse

En utilisant des hypothèses restrictives (covariance gaussienne isotrope, ondes planes de même incidence), les sismologues ont cherché dès les années 1970 à extraire les caractéristiques statistiques des milieux traversés à partir d'enregistrements d'ondes télésismiques (Aki [3]). Des ajustements manuels réalisés sur les covariances des phases, les covariances des amplitudes et les covariances croisées (approximation de Rytov parabolique) ont permis ainsi de retrouver des longueurs de corrélation de l'ordre de la dizaine de kilomètres et des écart types des fluctuations de l'ordre de 5% (cf. Sato &

Fehler [193] pour une revue bibliographique). Flatté & Wu [66] ont rendu robuste cette méthode, nommée tomographie stochastique, en relâchant les hypothèses sur le milieu (anisotropie transverse, quasi-stationnarité, introduction de plusieurs couches de covariances différentes) et en permettant de prendre en compte tous les enregistrements d'ondes de même fréquence (possibilité d'incidences différentes entre les ondes). Ce type de champs aléatoires quasi-stationnaires anisotropes a aussi été utilisé récemment par Kravtsov et al. [126] et Kaslilar et al. [114] dans le cadre de la caractérisation du sous-sol avec les temps de trajet d'ondes sismiques réfractées (source et récepteur en surface, trajet courbe sans réflexion).

Plutôt qu'un ajustement indirect des paramètres du milieu, Müller et al. [154] ont proposé une inversion directe en exprimant la covariance des perturbations de vitesse du milieu en fonction de la covariance des temps de trajet d'une onde plane en milieu 2D isotrope. Ils ont nommé cette procédure, développée dans le cadre de l'optique géométrique, tomographie statistique. Dans Iooss [92] et Iooss et al. [97], j'ai étendu cette inversion à une onde sphérique en milieu géométriquement anisotrope. A partir de l'équation (2.38) on peut en effet exprimer directement la covariance des perturbations de vitesse C_0 en fonction de la covariance des temps de trajet d'une onde sphérique C_T. La propriété que j'ai démontrée est la suivante (résultat donné en 2D pour simplifier son expression) :

Propriété 2.4.3 *Pour une onde sphérique en milieu 2D géométriquement anisotrope, dans l'approximation de l'optique géométrique et sous la condition $r = o(X)$ avec $r \in \mathbb{R}_+$, on a*

$$C_0\left(\frac{r}{l_\perp}\right) = \left[\int_r^\infty \frac{\frac{\partial}{\partial u} I(u,X)}{\sqrt{u^2-r^2}}\,du\right] \Bigg/ \left[\int_0^\infty \frac{\frac{\partial}{\partial u} I(u,X)}{u}\,du\right] , \tag{2.40}$$

avec

$$\begin{aligned} I:\ & \mathbb{R}_+ \times \mathbb{R}_+ \to \mathbb{R} \\ & (\rho, X) \mapsto I(\rho,X) = \frac{\partial}{\partial\rho}[\rho C_T(\rho,X)] \, . \end{aligned} \tag{2.41}$$

Ce résultat est obtenu en inversant l'intégrale de

$$I(\rho,X) = \frac{\sigma_\varepsilon^2}{2c_0^2} l_\parallel X \int_0^\infty C_0\left(\sqrt{\frac{\rho^2}{l_\perp^2}+v^2}\right) dv \tag{2.42}$$

par une transformée d'Abel, ce qui donne

$$C_0\left(\frac{r}{l_\perp}\right) = -\frac{4c_0^2}{\pi\sigma_\varepsilon^2}\frac{1}{l_\parallel X}\int_r^\infty \frac{\frac{\partial}{\partial u} I(u,X)}{\sqrt{u^2-r^2}}\,du \, . \tag{2.43}$$

En utilisant la propriété $C_0(0) = 1$, la formule (2.40) est obtenue. Une fois le modèle estimé, l'expression (2.40) donne la portée transverse $l_\perp$ par $C_0(1)$. Des exemples pratiques d'application de ce résultat sont donnés dans Iooss et al. [97].

Les inversions précédentes supposent que les fluctuations $\varepsilon(\boldsymbol{r})$ dont on cherche à retrouver les caractéristiques statistiques se superposent à un champ de vitesse moyen constant. En pratique, il est nécessaire de prendre en compte un macro-modèle de vitesse, qui induit une courbure des rayons sur lesquels on intègre les fluctuations ε, et qui doit être caractérisé de manière déterministe par les méthodes inverses classiques (Tarantola [208]). Quelques auteurs ont proposé des méthodes pour prendre en compte ce macro-modèle, notamment Klimeš [122, 120]. Celui-ci propose également d'utiliser la fonction de covariance estimée pour contraindre correctement le problème inverse déterministe.

2.4.3 Application à la sismique d'exploration

L'adaptation de ces méthodes d'inversion à la problématique de la sismique réflexion (cf. Fig. 2.1) soulève une difficulté particulière : la prise en compte de la réflexion de l'onde sur une interface au cours de la propagation. Touati [214] a proposé une solution en travaillant sur la variance des temps de trajet en fonction de l'offset x (distance source-récepteur) :

$$\mathrm{Var}[T(x)] = \mathrm{Var}[T(\boldsymbol{r}_{\mathrm{down}})] + \mathrm{Var}[T(\boldsymbol{r}_{\mathrm{up}})] + 2\mathrm{Cov}[T(\boldsymbol{r}_{\mathrm{down}}), T(\boldsymbol{r}_{\mathrm{up}})] \ , \qquad (2.44)$$

où $\boldsymbol{r}_{\mathrm{down}}$ et $\boldsymbol{r}_{\mathrm{up}}$ correspondent respectivement au rayon descendant (de la source vers le point de réflexion) et au rayon montant (du point de réflexion vers le récepteur). L'idée est alors de travailler avec une onde sphérique dont la source est située au point de réflexion et d'utiliser dans l'équation (2.44) les expressions des variance et covariance des temps au premier ordre dans l'optique géométrique (Eqs. (2.34) et (2.38)). Dans Touati [214] et Touati et al. [215], une procédure d'inversion indirecte est utilisée pour estimer la longueur de corrélation horizontale des fluctuations du champ de vitesse, dans le cas où le réflecteur est horizontal. Dans Iooss [92, 91], j'ai montré que cette procédure reste applicable dans l'approximation de Rytov parabolique, en dehors du domaine de validité de l'optique géométrique.

Dans le cadre de l'optique géométrique, ces travaux ont été généralisés par Kravtsov et al. [127] à des milieux dont la vitesse moyenne dépend de la profondeur et dont les fluctuations sont anisotropes quasi-stationnaires (dépendance en profondeur des paramètres statistiques).

D'autre part, j'ai montré que l'inversion directe donnée par la formule (2.40) demeure possible dans la géométrie avec réflexion (Iooss [92], Iooss & Galli [96], Iooss et al. [97]). Il suffit alors de remplacer $I(u, X)$ dans (2.40) par $I_R : u \in \mathbb{R}_+ \to I_R(u) \in \mathbb{R}$:

$$I_R(x) = \frac{\partial^2}{\partial x^2}[x\mathrm{Var}[T(x)] \ . \qquad (2.45)$$

Des simulations numériques ont illustré la robustesse de cette inversion directe et son application à des données réelles a démontré sa faisabilité.

Dans Iooss et al. [97], en m'inspirant de Klimeš [122], j'ai également proposé de prendre en compte dans l'inversion un macro-modèle de vitesse en travaillant sur les temps de trajet relatifs, définis comme le rapport du temps de trajet $T(\boldsymbol{r})$ de l'onde et le temps de trajet $T_0(\boldsymbol{r})$ dans le milieu sans fluctuation aléatoire (*i.e.* avec uniquement le macro-modèle de vitesse). En pratique, une première étape de tomographie déterministe est appliquée sur les temps de trajet $T(\boldsymbol{r})$ pour déterminer ce macro-modèle, puis les temps de trajet $T_0(\boldsymbol{r})$ y sont calculés par tracé de rayons. Par ailleurs, l'hypothèse de réflecteur horizontal étant quelque peu limitative, je me suis efforcé d'intégrer dans le modèle des fluctuations possibles du réflecteur, à savoir une composante à variations très lentes, et une composante à faibles fluctuations modélisées par un processus aléatoire (Iooss [90], Iooss & Galli [96]). La procédure d'inversion demeure inchangée car les fluctuations aléatoires du réflecteur n'induisent qu'un terme constant supplémentaire dans l'équation (2.44).

Finalement, j'ai poursuivi ce travail en collaborant avec David Geraets durant sa thèse à l'École des Mines de Paris (Geraets [74], Geraets & Galli [75], Iooss et al. [97] et Geraets et al. [76]). Celui-ci a développé une procédure d'inversion ne nécessitant pas l'utilisation des temps de trajet. En effet, la récupération de ceux-ci est l'une des difficultés majeures de l'application de cette méthodologie en sismique réflexion : la procédure du pointé des temps de trajet sur les sismogrammes est extrêmement délicate du fait du bruit présent dans les données sismiques, d'autant plus qu'on ne sait souvent pas où pointer le temps de trajet sur l'ondelette sismique. David Geraets propose donc une inversion robuste pour retrouver la covariance des perturbations de vitesse des ondes à partir de la covariance des vitesses de stack. Ces dernières sont issues d'un ajustement polynomial par moindres carrés des temps de trajet en fonction de l'offset, et sont donc moins sensibles aux erreurs sur les temps de trajet.

2.5 Conclusion

Les travaux présentés dans ce chapitre ont été réalisés durant mes six premières années d'activités de recherche, entre 1995 et 2001. Les problèmes évoqués rentrent dans le cadre général de la mécanique statistique des milieux continus. Au vu des outils utilisés, ces travaux peuvent être également considérés dans le cadre des applications des mathématiques, et plus spécifiquement de l'application de modèles aléatoires à la physique et aux problèmes d'ingénierie. Cette multiplication des domaines scientifiques concernés a été l'une des difficultés majeures à surmonter durant ces travaux. Il est souvent difficile d'accorder d'un point de vue technique, des sciences aussi différentes telles que la physique statistique, la géostatistique, la géophysique et la mécanique des fluides. Ceci a pu être réalisé grâce à de multiples échanges et collaborations avec des chercheurs dotés d'une grande ouverture d'esprit. Les résultats que j'ai obtenus montrent aussi que cette transversalité permet l'émergence d'idées neuves. Mes travaux s'intègrent également dans l'ensemble des applications modernes des outils géostatistiques (cf. Bilodeau et al. [22] pour une revue récente), dans lesquelles la modélisation des phénomènes physiques n'est plus ignorée. Concernant la propagation d'ondes en milieu aléatoire, mon principal apport résulte de l'introduction de l'anisotropie statistique (en particulier l'anisotropie géométrique) dans les traitements théoriques et les simulations numériques.

Ces recherches qui paraissaient séduisantes théoriquement mais peu exploitables il y a dix ans (car non applicables sur données réelles du fait du bruit des données) ont pris plus de sens ces dernières années, notamment grâce à l'amélioration de certains outils expérimentaux. Au niveau de la communauté internationale, quelques auteurs ont poursuivi l'étude fine des fluctuations de temps de trajet d'ondes acoustiques, aussi bien en sismologie (Baig [11], Saito [183]), qu'en contrôle ultrasonore (Durgin & Andreeva [61]) et en océanographie (Godin [78]). En ce qui concerne l'application de la tomographie statistique (pas forcément dénommée de la même manière par mes homologues) en sismique, on peut noter les travaux de Klimeš [122, 120], Kaslilar et al. [114] et Liu et al. [136].

Ainsi, j'entrevois quelques perspectives et extensions de mes travaux qui pourraient s'avérer intéressantes :

- ▷ intégration de la tomographie statistique dans la gestion des incertitudes en exploration et production pétrolières (Thore et al. [213], Geraets et al. [77]) ;

- ▷ unification de mes travaux en tomographie statistique avec ceux de Kravtsov et al. [126] (formulation en milieu quasi-stationnaire avec fluctuations verticales de vitesse) ;

- ▷ utilisation de la covariance des perturbations de vitesse comme covariance *a priori* dans la tomographie classique (Tarantola [208], Klimeš [120], Vecherin et al. [221]) ;

- ▷ développement de la tomographie statistique dans l'approximation markovienne de l'équation d'onde parabolique en milieu aléatoire (cf. par exemple Rytov et al. [181]), cadre moins restrictif que l'approximation de Rytov de l'équation d'onde parabolique que j'ai utilisée ;

- ▷ dérivation de la variance des temps de trajet au second ordre dans la théorie des rayons complexes ou théorie géométrique de la diffraction (cf. Kravtsov [125] pour une revue récente sur le sujet), extension naturelle de l'optique géométrique ;

- ▷ dérivation rigoureuse de la méthode de sommation des faisceaux gaussiens pour la simulation numérique de la propagation d'ondes acoustiques dans les milieux en mouvement. En effet, dans Iooss et al. [98] je me base sur une méthode heuristique en utilisant, dans le système du tracé de rayons dynamique, la vitesse effective du milieu définie dans la remarque 2.3.1.

La problématique abordée dans ce chapitre se situe finalement dans le contexte de la prise en compte des incertitudes en simulation numérique. Véritable fil conducteur de ce mémoire, celui-ci permet de relier mes deux périodes de recherche. Les travaux présentés dans ce chapitre concernent plus spécifiquement les incertitudes de modèles : négliger l'hétérogénéité des champs de vitesse des ondes revient à simplifier les modèles physiques. Dans le domaine des incertitudes, on cherche souvent à développer des approches génériques (*i.e.* pouvant s'adapter à un grand nombre de problèmes physiques différents). Les outils que j'ai utilisés ici ne sont nullement génériques car, ayant identifié le problème de l'hétérogénéité des champs de vitesse, mon approche a consisté à les modéliser par des champs aléatoires, puis à traiter le problème de manière statistique. Le chapitre suivant expose, quant à lui, des outils génériques pour traiter les problèmes d'incertitude sur les variables d'entrée des modèles numériques.

Chapitre 3

Etudes d'incertitudes de modèles numériques

3.1 Introduction

La simulation numérique désigne le procédé selon lequel on exécute un programme sur un ordinateur en vue de représenter un phénomène physique. La simulation numérique est à présent considérée comme la troisième forme d'étude des phénomènes, après la théorie et l'expérience. Comme une expérience réelle, une simulation peut être extrêmement coûteuse à élaborer (par exemple en préparant le jeu de données) et à réaliser (par exemple plusieurs semaines de temps de calcul). De plus, elle peut ne pas fournir de résultats (par exemple du fait de problèmes de convergence dans la résolution numérique de systèmes d'équations) ou produire des réponses entachées de bruit (par exemple du fait de la discrétisation insuffisante de schémas numériques). Il est d'ailleurs souvent question d'"expérience numérique" pour illustrer l'analogie entre la pratique d'une simulation[1] et la conduite d'une expérience de physique. Au CEA, la R&D dans le domaine de l'énergie et de l'industrie nucléaire est fortement demandeuse de modélisation et simulation numériques. La simulation numérique permet en effet de mieux répondre aux grands enjeux que représentent, entre autres, la compétitivité économique du nucléaire (*e.g.* par une meilleure gestion des ressources), la sûreté des installations nucléaires (*e.g.* par une meilleure estimation des marges de fonctionnement vis-à-vis de conditions accidentelles) et la maîtrise des risques pour l'environnement (*e.g.* par une minimisation du volume des rejets). Pour soutenir les recherches dans ces domaines, des développements logiciels importants ont eu lieu depuis une trentaine d'années dans les différentes physiques concernées. Citons par exemple la neutronique, la thermohydraulique, la thermomécanique, la physique des matériaux et les phénomènes de transfert dans l'environnement.

Pour ces problématiques d'étude et de conception de systèmes complexes à l'aide de modèles numériques prédictifs, on ne peut pas se contenter d'une simulation moyenne sur quelques cas. Il est souvent nécessaire d'estimer précisément les incertitudes sur les prédictions, de connaître les risques d'événements rares, voire d'optimiser les réponses sous contraintes. Ceci rend nécessaire la réalisation d'évaluations probabilisées sur les sorties des modèles numériques. Par ailleurs, la détermination des variables d'entrée qui induisent le plus d'incertitudes sur les réponses permet de définir des voies d'amélioration pour réduire les incertitudes des prédictions. Il peut, par exemple, s'agir d'améliorer la connaissance sur les variables d'entrée influentes ou alors de modifier le système pour éviter la dépendance à ces entrées influentes. La phase de validation d'un code (qui répond à la question de l'adéquation du modèle à la réalité) doit également être particulièrement soignée car les utilisateurs doivent pouvoir se servir du code dans une large gamme de variation de ses entrées. Des méthodes robustes sont alors nécessaires pour définir les domaines de variation des variables d'entrée dans lesquels le modèle représente la réalité. Par conséquent, lors de la phase de développement d'un modèle dont

[1] Dans la littérature, on emploie souvent, et de manière indifférente, les termes "modèle numérique", "simulateur", "programme", "code de calcul", "code" et "logiciel de calcul".

la vocation est de réaliser des évaluations probabilisées, des outils d'analyse vont être nécessaires aux développeurs pour les aider à comprendre leur modèle dans des domaines et des configurations (des entrées) jusque là inexplorés. Pour toutes ces raisons, le développement d'outils génériques d'analyses d'incertitude et d'analyses de sensibilité est d'une importance fondamentale, aussi bien pour les phases de développement, de validation et d'utilisation des codes de calcul.

Le terme "étude d'incertitude" concerne donc l'évaluation de l'influence des sources d'incertitudes (modèle, données d'entrée, ...) sur un modèle représentant un phénomène observé. Phénomène signifie ici un phénomène physique (*e.g.* évolution de la température de gaines de combustible nucléaire ou de la taille de fissures dans une cuve de réacteur) ou chimique, mais peut aussi être un modèle économique ou un modèle de fiabilité ou de disponibilité. Une étude d'incertitude s'inscrit dans un contexte de prise de décision pouvant avoir différents objectifs :

- ⋆ vérifier un critère réglementaire (ce qui impose alors le type de formalisme de modélisation des incertitudes et la nature des indicateurs mathématiques exprimant la variabilité des réponses du modèle) ;
- ⋆ faire un choix parmi plusieurs options ou scénarios pour optimiser la conception, l'exploitation ou le démantèlement d'un système selon des critères (coût, risque, ...) ;
- ⋆ valider, qualifier ou calibrer un modèle ;
- ⋆ améliorer la compréhension du phénomène observé ou sa modélisation (contexte de R&D).

3.1.1 État de l'art

La Figure 3.1 résume les étapes d'une étude d'incertitude et le caractère itératif de la démarche dans de nombreuses applications (de Rocquigny [52, 53], Sudret [206], de Rocquigny et al. [54]). L'ensemble des étapes (certaines optionnelles) d'une telle étude se décompose en :

- ▷ étape A, spécification du problème : définition des objectifs de l'étude, du (des) modèle(s) utilisé(s), des quantités d'intérêt, des variables d'intérêt et des variables d'entrée jugées incertaines ;
- ▷ étape B, quantification des sources d'incertitudes : modélisation des distributions de probabilité des variables d'entrée ou au moins définitions de leurs bornes inférieures et supérieures ;
- ▷ étape C, propagation d'incertitudes : évaluation de la variabilité des sorties du modèle ou variables d'intérêt induite par les incertitudes sur les entrées. Cette variabilité est exprimée sous la forme de quantités d'intérêt. Le résultat de cette phase est fortement conditionné au modèle et à la modélisation des sources d'incertitudes ;
- ▷ étape C', hiérarchisation des sources d'incertitudes : évaluation de l'importance ou de la contribution relative des sources d'incertitudes sur la ou les quantités d'intérêt, phase appelée "analyse de sensibilité" ;
- ▷ à l'issue de ces étapes et selon les résultats obtenus, il est éventuellement nécessaire de redéfinir le problème à résoudre et de revenir sur les différentes étapes de l'étude.

Aux niveaux numérique et probabiliste, les outils permettant de résoudre ces différentes étapes sont relativement anciens et bien connus, comme par exemple les méthodes adjointes (Cacuci [31, 32]), les méthodes de Monte Carlo (Hammersley & Handscomb [81], Rubinstein [178]) et les méthodes fiabilistes (Madsen et al. [139], Ditlevsen & Madsen [59]). La problématique de l'exploration complète de codes de calcul coûteux a été identifiée plus récemment (McKay et al. [150], Sacks et al. [182], Santner et al. [191], Fang et al. [62], Kleijnen [117]), et a éé traitée comme une problématique de planification d'expériences

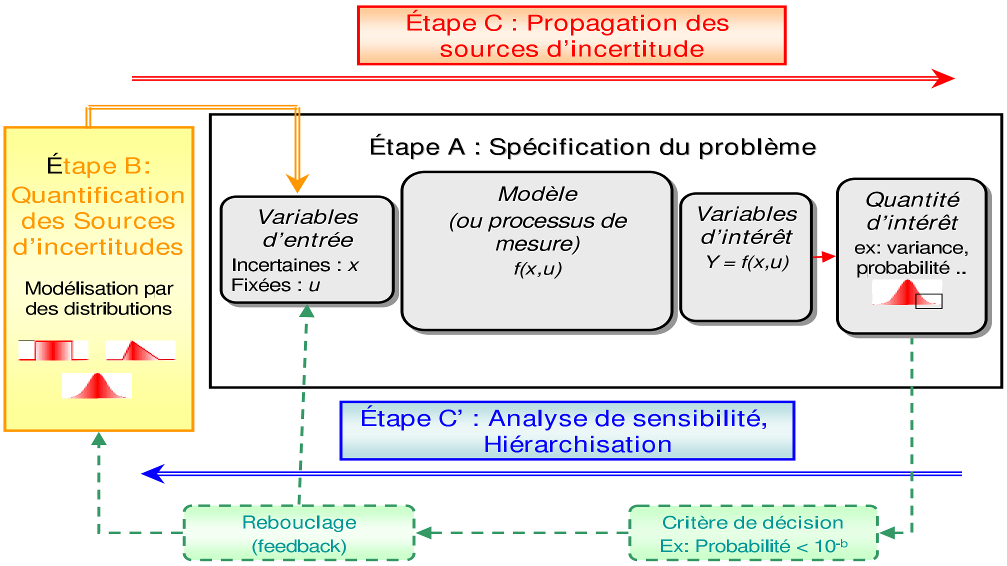


FIG. 3.1 – Cadre général pour les études d'incertitude.

et d'approximation par des surfaces de réponse (appelées aussi métamodèles). Ce thème de recherche, largement développé aux Etats-Unis durant les trois dernières décennies, a énormément progressé en France depuis une dizaine d'années, sous l'impulsion, entre autres, d'instituts de recherche à vocation industrielle (CEA, IFP et EDF R&D notamment). Ces instituts ont financé une grande quantité de thèses se concentrant sur un ou plusieurs aspects spécifiques des études d'incertitude. Bien que non exhaustive, la revue ci-dessous permet d'identifier clairement les différentes thématiques du sujet :

- des problèmes de fiabilité et d'évaluation d'événements rares par des modèles numériques ont été traités dans Devictor [57], Andrieu-Renaud [6], Baroth [15], Berveiller [20] et Cannamela [35] ;

- des problématiques et nouvelles méthodologies d'analyse de sensibilité de modèles ont été étudiées et proposées dans Jacques [106], Da Veiga [48], Petelet [167] et Briand [28] ;

- l'exploration (i.e. l'investigation fine des relations entre sorties et entrées) d'un code de calcul coûteux peut être traitée par la construction d'un métamodèle, fonction mathématique rapide à évaluer et permettant d'approximer la ou les réponses du code. Jourdan [111], Vazquez [219], Scheidt [195] et Marrel [141] se sont intéressés au modèle des processus gaussiens (krigeage) comme métamodèle. Pour les codes de calcul stochastiques (pour lesquels des simulations avec les mêmes entrées produisent des réponses différentes), Zabalza-Mezghani [232] s'est intéressée quant à elle aux modèles linéaires généralisés joints (cf §3.5) ;

- la planification d'expériences simulées (méthodes d'échantillonnage de points dans l'espace des variables d'entrée) a été abordée de manière générale par Jourdan [111] et Feuillard [63], alors que des méthodes de planification adaptative (prenant en compte la sortie de premiers calculs pour planifier de nouveaux calculs) ont été développées par Scheidt [195] (pour la construction d'un modèle de krigeage) et Gazut [72] (en utilisant des méthodes de reéchantillonnage par bootstrap).

De par mon activité transverse aux différents domaines de la physique étudiés en ingénierie nucléaire, j'ai pu être associé aux travaux de plusieurs doctorants, issus de différentes unités du CEA, dans chacune de ces thématiques et dans différents domaines d'application : Claire Cannamela (Département d'Étude des Combustibles) sur la fiabilité des combustibles nucléaires de type particules, Mathieu Petelet (Département de Modélisation des Systèmes et des Structures) sur la simulation numérique du soudage, Amandine Marrel (Département de Technologie Nucléaire) sur les modèles de transferts hydrogéologiques, Julien Jacques, Vincent Feuillard et Benjamin Auder (dans mon laboratoire au sein du Département d'Étude des Réacteurs) sur des problématiques de dosimétrie et de sûreté des réacteurs. Ces travaux de thèse, pour les laboratoires du CEA où ils ont eu lieu, ont constitué et constituent encore d'excellentes bases de travail pour développer des méthodes innovantes au travers de leurs applications.

3.1.2 Contributions

La principale source d'originalité de mes travaux de recherche pour le traitement des incertitudes des codes de calcul provient de certaines applications que j'ai eues à traiter et qui se sont avérées relativement complexes :

- non linéarités, interactions fortes, voire discontinuités dans les modèles. J'ai contribué au développement d'une méthodologie générique pour réaliser une analyse de sensibilité d'un modèle (cf. §3.2 et Fig. 3.2), sans supposer de connaissance *a priori* sur sa régularité (linéarité/monotonie de la réponse par rapport aux entrées, présence de discontinuités, ...) ;

- modèles numériques coûteux en temps de calcul de telle sorte qu'un nombre élevé de simulations (par exemple supérieur à 100) n'est pas possible. Sur ce thème, j'ai développé et étudié les propriétés théoriques de nouveaux algorithmes, basés sur le tirage stratifié et le tirage d'importance, pour estimer les quantiles élevés de sorties de codes de calcul (cf. §3.4.3 et 3.4.5). J'ai également développé un nouvel algorithme pour calculer et simuler des indices de sensibilité à partir du métamodèle processus gaussien (cf. §3.3.4) ;

- code coûteux et nombre important de variables d'entrée incertaines (jusqu'à une cinquantaine). Pour résoudre ce problème, j'ai développé un algorithme pour ajuster le métamodèle processus gaussien lorsqu'il y a peu d'observations et plusieurs dizaines de variables d'entrée (cf. §3.3.3) ;

- processus stochastiques et champs aléatoires en entrée des modèles. Pour traiter ce type de problèmes, j'ai développé une approche basée sur un métamodèle spécifique, le modèle additif généralisé joint (cf. §3.5.1), qui m'a permis de proposer une méthode pour estimer les indices de sensibilité (cf. §3.5.2).

Le problème standard, qui considère des variables d'entrée et une variable de sortie scalaires, se formule de la manière suivante :

$$\begin{aligned} f: \ \mathbb{R}^d &\to \mathbb{R} \\ \mathbf{X} &\mapsto Y = f(\mathbf{X}) \end{aligned} \tag{3.1}$$

où $\mathbf{X} = (X_1, \ldots, X_d)$ est un vecteur aléatoire de d variables d'entrée du code, modélisé par une loi de probabilité, $f(\cdot)$ est la fonction du modèle (le code de calcul), potentiellement inconnue d'un point de vue analytique, et Y est la variable de sortie du modèle qui est par conséquent une variable aléatoire. Le cas d'une sortie vectorielle, voire fonctionnelle, ne sera pas traité dans le cadre de ce mémoire, mais constitue l'un de mes nouveaux sujets d'intérêt. Dans les applications, il est en effet courant d'avoir des champs spatialisés ou des courbes d'évolution temporelle en sortie des modèles (par exemple des cartes de concentrations d'un polluant). On voit dans l'expression (3.1) que, contrairement au cadre statistique classique, nous n'introduisons pas de bruit d'observation car le modèle f est considéré déterministe. Nous verrons au §3.5 comment modéliser une situation différente concernant un code de calcul stochastique (exemple : un code reposant sur un solveur de type Monte Carlo).

La section suivante fait l'objet d'une synthèse des méthodes statistiques d'analyse de sensibilité les plus utilisées.La troisième section décrit mes travaux avec Amandine Marrel qui concernent la mise en œuvre du modèle processus gaussien sur des cas industriels complexes et l'étude des indices de sensibilité obtenus de manière analytique avec ce modèle. La quatrième section décrit les principaux résultats obtenus avec Josselin Garnier[2] et Claire Cannamela concernant le problème de l'estimation des quantiles élevés de codes de calcul. Ce sujet est notamment motivé par des questions de sûreté nucléaire où des critères réglementaires, évalués à l'aide de codes de calcul simulant des phénomènes physiques relativement complexes, doivent être respectés. Enfin, la dernière section s'attarde sur le développement et l'utilisation du modèle additif généralisé joint, réalisé en collaboration avec Mathieu Ribatet[3], pour modéliser les calculs des modèles numériques stochastiques. De tels codes sont par exemple ceux basés sur la méthode de Monte Carlo pour simuler les trajectoires des neutrons dans un cœur de réacteur nucléaire. La conclusion sera l'occasion de mettre en lumière les grands axes de recherche que j'entrevois pour mes futurs travaux et collaborations.

3.2 Analyse de sensibilité de modèles

Lors de la construction et de l'utilisation d'un modèle numérique simulant des phénomènes physiques, les méthodes d'analyse de sensibilité sont des outils précieux. Elles permettent de déterminer quelles sont les variables qui contribuent le plus à la variabilité de la quantité d'intérêt, quelles sont au contraire les variables les moins influentes et quelles variables interagissent avec quelles autres. La quantité d'intérêt peut être la variance d'une variable de sortie du modèle, mais aussi une autre mesure d'information (comme par exemple l'entropie), une probabilité qu'une sortie dépasse un seuil donné, ou toute autre chose. L'analyse de sensibilité est donc une aide à la validation d'un code de calcul, à l'orientation des efforts de R&D, ou encore à la justification en terme de sûreté du dimensionnement d'un système. Saltelli et al. [187] proposent une classification des grands objectifs d'une analyse de sensibilité :

- ▷ hiérarchisation des variables d'entrée ("factors prioritization") : détermination des variables dont la réduction de l'incertitude permettrait d'obtenir la plus forte réduction de l'incertitude sur la quantité d'intérêt ;
- ▷ identification des variables d'entrée non influentes ("factors fixing") : détermination des variables que l'on peut fixer sans altérer le modèle (ce qui permet une simplification du modèle) ;
- ▷ partage de la variance ("variance cutting") : détermination des variables à fixer pour obtenir une réduction donnée de l'incertitude sur la quantité d'intérêt ;
- ▷ cartographie des variables d'entrée ("factors mapping") : détermination des variables les plus influentes dans un domaine de valeurs de la sortie.

L'analyse de sensibilité a longtemps été vue sous un angle local, qui consiste à évaluer les répercussions (sur la valeur de variables de sortie) de petites perturbations des valeurs des entrées autour d'un point nominal. Cette approche déterministe consiste à calculer ou à estimer des indices basés sur les dérivées partielles du modèle en un point précis (Turanyi [216]). La mesure d'importance de chaque variable d'entrée peut alors être calculée en multipliant la dérivée qui lui correspond par son écart type. Découlant des mêmes principes, des méthodes adjointes relativement sophistiquées ont également été développées pour pouvoir traiter de gros systèmes d'équations possédant notamment un très grand nombre de variables d'entrée (Cacuci [31, 32]). Ce type d'approches est par exemple couramment utilisé dans la résolution de gros systèmes environnementaux (climatologie, océanographie, hydrogéologie, cf. Castaings [37]).

A partir de la fin des années 1980 et pour relâcher certaines hypothèses de ces méthodes (hypothèses de linéarité et de normalité, variations locales), de nouvelles méthodes d'analyse de sensibilité ont été développées dans un cadre statistique. Par opposition aux méthodes locales, elles ont été par la suite

[2]Laboratoire de Probabilités et Modèles Aléatoires, Université Paris VII
[3]doctorant, CEMAGREF Lyon

dénommées méthodes globales car elles s'intéressent à l'ensemble du domaine de variation possible des variables d'entrée. Cette distinction local/global semble cependant parfois quelque peu ambiguë. Par exemple, elles sont équivalentes si le comportement du modèle est linéaire de degré un dans l'ensemble du domaine de variation des entrées.

Ces méthodes statistiques, issues de plusieurs horizons, sont composées de techniques bien éprouvées issues de la théorie des plans d'expérience (pour l'exploration des codes de calcul à grand nombre d'entrées), de méthodes de type Monte Carlo rendues possible grâce aux nouvelles capacités informatiques (pour des analyses de sensibilité quantitatives et fines) et de la théorie de l'apprentissage statistique (pour les codes coûteux et complexes). Ce sont toutes ces méthodes d'analyse de sensibilité que le laboratoire dans lequel j'officie a essayé de populariser depuis une dizaine d'années dans divers projets et applications du CEA. La Figure 3.2 présente une synthèse des principales méthodes d'analyse de sensibilité. Cette liste n'est bien entendu pas exhaustive et ne tient pas compte des éventuelles améliorations apportées aux différentes méthodes (c'est le cas par exemple pour la méthode des bifurcations séquentielles).

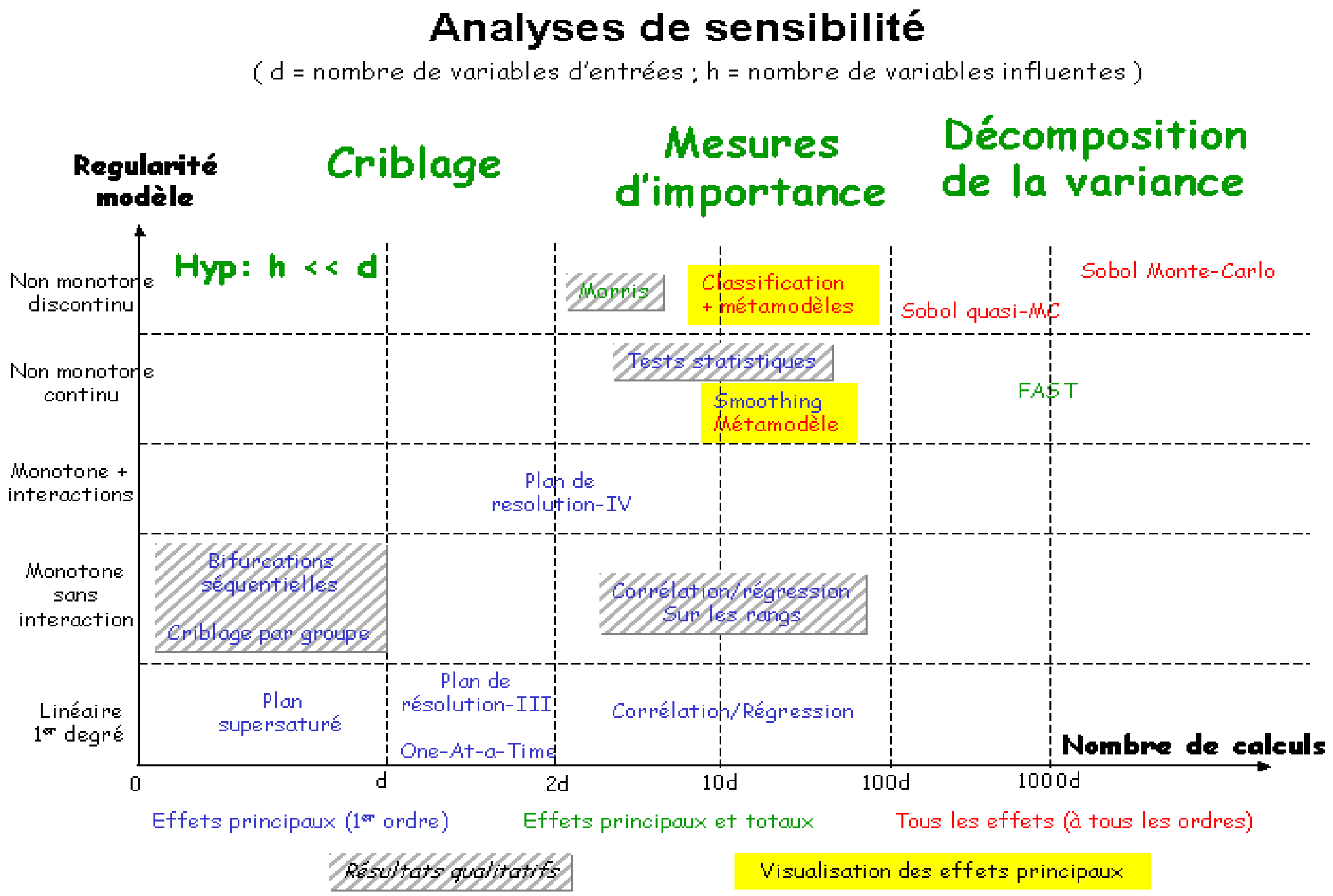


FIG. 3.2 – Synthèse des méthodes d'analyse de sensibilité placées dans un diagramme (coût en nombre d'évaluations du modèle vs. complexité et régularité du modèle). d est le nombre de variables d'entrée du modèle, h est le nombre de variables d'entrée influentes.

Cette figure permet de distinguer cinq classes de méthodes qui correspondent à différents types de problèmes rencontrés en pratique. L'approche méthodologique que je défend consiste à utiliser la méthode la plus simple adaptée au problème posé[4], fonction de l'objectif de l'étude, du nombre d'évaluations du modèle numérique que l'on peut réaliser et de la connaissance que l'on a sur la régularité du modèle étudié. La validation *a posteriori* de la méthode utilisée permet de savoir s'il est

[4]en suivant le principe de parcimonie connue sous le nom de "rasoir d'Occam"

nécessaire d'utiliser une méthode plus performante, en réalisant ou non de nouvelles simulations. Cette approche a été utilisée dans mes publications traitant principalement d'applications environnementales (Iooss et al. [104], Volkova et al. [223]) et dans de nombreuses études réalisées pour différents projets du CEA. Les sections suivantes détaillent les cinq grandes classes de méthodes d'analyse de sensibilité que je distingue.

3.2.1 Criblage à très grande dimension

Les méthodes de criblage ("screening") permettent d'explorer rapidement le comportement des réponses d'un code de calcul coûteux en faisant varier un grand nombre de ses entrées (typiquement plusieurs dizaines voire plusieurs centaines). Certaines techniques issues des plans d'expérience permettent de le faire en réalisant moins de calculs que de variables d'entrée. Celles-ci supposent qu'il n'y a pas d'interaction entre les variables d'entrée, que la variation de la réponse est monotone par rapport à chaque entrée et que le nombre des entrées influentes est très faible devant le nombre total d'entrées (de l'ordre d'une sur dix). Il s'agit en premier lieu des *plans supersaturés* développés dans le contexte de la planification d'expériences réelles (Satterthwaite [194], Lin [134], Dean & Lewis [55]). L'un des plans supersaturés les plus connus résulte de la division en deux parties (à l'aide d'une colonne de branchement) d'une matrice d'Hadamart. En traitant quelques applications, Claeys-Bruno et al. [43] ont montré que ce plan supersaturé est l'un des plus fiables et qu'il faut au moins 5 fois plus de calculs que de variables influentes pour les identifier toutes. Dans Cannamela et al. [36], nous avons appliqué ce type de plan supersaturé en planifiant 30 calculs sur un code possédant 53 variables d'entrée incertaines. Ce plan supersaturé nous a permis d'identifier les 5 entrées les plus influentes, résultat qui a été validé par la suite à l'aide d'un plus grand nombre de simulations et de mesures d'importance basées sur les coefficients de corrélation (cf. §3.2.3).

D'autres approches sont particulièrement bien adaptées aux expériences numériques car elles sont séquentielles et adaptatives, c'est-à-dire qu'elles définissent une nouvelle expérience à réaliser en fonction des résultats des précédentes. On ne sait donc pas *a priori* combien elles vont nécessiter d'expériences. La technique du *criblage par groupe* (Dean & Lewis [55]) consiste à créer un certain nombre de groupes de variables d'entrée et à identifier les plus influents. En répétant l'opération en conservant seulement les groupes influents, on extrait ensuite les variables influentes. La méthode des *bifurcations séquentielles*, mise au point dans un contexte de simulation numérique par Bettonvil & Kleijnen [21], peut être vue comme une méthode de criblage par groupe avec seulement deux groupes. C'est une approche dichotomique où on tente d'éliminer à l'issue de chaque nouveau calcul un groupe de variables. Comme pour le criblage par groupe, son coût dépend donc du nombre de variables influentes, mais aussi de la stratégie de classement, *i.e.* de notre capacité à suspecter quelles sont les variables influentes afin de les rassembler au sein d'un même groupe.

Dans un contexte de criblage pour codes de calcul, Sergent et al. [197] comparent les plans supersaturés, le criblage par groupe et les bifurcations séquentielles et en concluent que la technique des plans supersaturés est nettement plus risquée que les autres mais nécessite le moins d'hypothèses. En effet, il est nécessaire de connaître le sens de variation de la sortie par rapport à chaque entrée pour pouvoir appliquer les bifurcations séquentielles et le criblage par groupe.

3.2.2 Criblage et plans d'expérience

La deuxième classe de méthodes concerne celle issue de la théorie classique des plans d'expérience (Droesbecke et al. [60], Montgomery [152], Azaïs & Bardet [10]). Comme précédemment, les entrées (nommées facteurs) sont discrétisées en plusieurs valeurs (nommées niveaux) de leur domaine de variation. Un *plan factoriel complet* consiste à évaluer le code de calcul pour toutes les combinaisons entre facteurs, ce qui permet l'estimation de tous les effets des facteurs et de leurs interactions. En pratique, le nombre de simulations requis rend ce plan impraticable. En effet, il nécessite par exemple 2^d calculs si on suppose que le modèle est monotone en travaillant avec deux niveaux pour chaque facteur (exemple : min et max).

Si l'on veut estimer de manière non biaisée les effets du premier ordre (aussi appelés effets principaux) de chaque entrée, il faut au minimum simuler $n \geq d+1$ combinaisons des entrées. Par définition, un *plan de résolution trois* (noté RIII) permet cette estimation non biaisée en supposant que les effets des interactions sont nuls, *i.e.* que le modèle est de la forme suivante :

$$Y = \sum_{j=0}^{d} \beta_j X_j + \epsilon \,, \tag{3.2}$$

où $X_0 = 1$, $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_d)^t \in \mathbb{R}^{d+1}$ est le vecteur des effets des entrées et $\epsilon \in \mathbb{R}$ est l'erreur du modèle. L'estimation de ces effets se fait par la méthode des moindres carrés ordinaires.

Le plan d'expériences le plus simple, encore très utilisé par les ingénieurs, est le plan nommé *"One At a Time"* (OAT), qui fait partie de la classe des plans RIII. Le plan OAT consiste à changer le niveau d'une entrée à la fois, en utilisant deux ou trois niveaux par facteur (Kleijnen [117]). Avec deux niveaux, ce plan requiert donc exactement $n = d+1$ calculs (cf. Fig. 3.2), mais ne permet pas de maîtriser la précision que l'on a sur les estimations des effets. La *méthode de Morris*, qui consiste à répéter (entre cinq et dix fois) un plan OAT aléatoirement dans l'espace des variables d'entrée, permet de s'extraire des hypothèses limitatives du plan OAT, mais s'avère bien plus coûteuse en temps de calcul.

Une voie plus raisonnable qu'un simple plan OAT consiste à minimiser la variance des effets estimés, ce qui est l'objectif de la théorie statistique des plans d'expérience. Celle-ci se concentre sur les plans orthogonaux, c'est-à-dire ceux qui satisfont

$$(\boldsymbol{X}_0^n)^t \boldsymbol{X}_0^n = n \boldsymbol{I}_{d+1} \,, \tag{3.3}$$

où $\boldsymbol{X}_0^n = (X_j^{(i)})_{i=1..n, j=0..d}$ est la matrice du plan et $\boldsymbol{I}_{d+1}$ est la matrice identité de dimension $d+1$. Une classe bien connue de plans orthogonaux est celle des *plans factoriels fractionnaires*. Leur construction qui fait appel à la notion d'alias et qui dépasse le cadre de ce mémoire, consiste à confondre des interactions que l'on soupçonne non actives avec des effets principaux.

Il est parfois prudent de supposer que les interactions entre les entrées peuvent avoir des effets importants. Par définition, un *plan de résolution quatre* (noté RIV) permet une estimation non biaisée des effets principaux même si des interactions d'ordre deux sont présentes. Un plan RIV est construit en superposant un plan RIII avec son plan "miroir". Ainsi la taille d'un plan RIV est le double de celle d'un plan RIII. Pour un coût en terme de nombre de calculs de l'ordre de $2 \times d$, un plan RIV permet donc d'identifier les effets principaux des entrées pour des modèles avec interactions. Il existe de nombreux autres types de plan qui assouplissent les hypothèses des plans RIII tout en conservant un nombre de calculs raisonnable.

Remarque 3.2.1 Dans un contexte unique de criblage, ces plans peuvent être intéressants pour les codes de calcul. Malheureusement, dans un contexte de réutilisation ultérieure du plan pour la propagation d'incertitude ou la construction d'un métamodèle complexe (i.e. plus riche qu'une surface de réponse polynomiale), ces plans sont peu recommandés car leurs projections sur les marges sont particulièrement médiocres en terme de recouvrement spatial. Pujol [171] illustre bien ce problème sur les plans de Morris. C'est pourquoi, de nombreux auteurs se sont penchés sur le développement de plans de type "Space Filling Designs" (SFD) qui assurent un bon recouvrement des marges (Fang et al. [62]).

3.2.3 Mesures d'importance basées sur des échantillons

Lorsque l'on dispose d'un échantillon de simulations $(\boldsymbol{X}^n, \boldsymbol{Y}^n)$, où $\boldsymbol{X}^n = (X_j^{(i)})_{i=1..n, j=1..d}$ est la matrice des entrées et $\boldsymbol{Y}^n = (Y_i)_{i=1..n}$ est le vecteur des sorties, il est très facile d'obtenir des indices de la sensibilité de la réponse par rapport aux variables d'entrée en appliquant les techniques de régression linéaire, de régression sur les rangs ou des tests statistiques. On peut parler à présent de mesures d'importance car ces techniques permettent une réelle hiérarchisation de l'influence sur la sortie de toutes les variables d'entrée, contrairement aux techniques de criblage qui ont plutôt pour but de détecter les variables d'entrée non influentes. On rappelle brièvement ci-dessous les principales mesures d'importance que l'on classe dans cette catégorie :

◇ le *coefficient de corrélation linéaire* (nommé communément coefficient de Pearson et noté $\rho(\cdot,\cdot)$) entre X_j et Y :

$$\rho_j = \rho(X_j, Y) = \frac{\text{Cov}(X_j, Y)}{\sqrt{\text{Var}(X_j)\text{Var}(Y)}} \ \ \forall\, j = 1, \ldots, d\,, \tag{3.4}$$

mesure de sensibilité extrêmement simple à calculer à partir d'un échantillon ;

◇ le *coefficient de régression standard* (noté $\text{SRC}(\cdot,\cdot)$) :

$$\text{SRC}_j = \text{SRC}(X_j, Y) = \beta_j \sqrt{\frac{\text{Var}(X_j)}{\text{Var}(Y)}} \ \ \forall\, j = 1, \ldots, d\,, \tag{3.5}$$

où les β_j $(i = 1 \ldots d)$ sont les coefficients de la régression linéaire (cf. Eq. (3.2)) ;

◇ le *coefficient de corrélation partielle* (noté $\text{PCC}(\cdot,\cdot)$),

$$\text{PCC}_j = \text{PCC}(X_j, Y) = \rho(Y - \widehat{Y}, X_j - \widehat{X_j}) \ \ \forall\, j = 1, \ldots, d\,, \tag{3.6}$$

où $\widehat{Y}$ est la prévision du modèle linéaire dans lequel X_j n'est pas présent :

$$Y = \sum_{k=0, k\neq j}^{d} \delta_k X_k + \epsilon_1 \tag{3.7}$$

avec $(\delta_0, \ldots, \delta_{j-1}, \delta_{j+1}, \ldots, \delta_d) \in \mathbb{R}^d$ les coefficients de régression et $\epsilon_1 \in \mathbb{R}$ l'erreur du modèle, et où $\widehat{X_j}$ est la prévision du modèle linéaire qui exprime X_j en fonction des autres entrées :

$$X_j = \sum_{k=0, k\neq j}^{d} \eta_k X_k + \epsilon_2 \tag{3.8}$$

avec $(\eta_0, \ldots, \eta_{j-1}, \eta_{j+1}, \ldots, \eta_d) \in \mathbb{R}^d$ les coefficients de régression et $\epsilon_2 \in \mathbb{R}$ l'erreur du modèle. Contrairement aux coefficients de régression standards, les coefficients de corrélation partielle permettent d'éliminer l'influence des autres variables et sont donc adaptés au cas où les variables d'entrée sont corrélées. Par contre, ils représentent plus une mesure de la linéarité de la sortie Y par rapport à une entrée X_j qu'un indice de sensibilité (Saltelli et al. [185]) ;

◇ le *coefficient de corrélation sur les rangs* des variables (nommé coefficient de Spearman et noté $\rho^S(\cdot,\cdot)$). Si $\boldsymbol{R}_X = (R_{X_1}, \ldots, R_{X_d})$ est le vecteur des rangs des entrées et R_Y est le rang de la sortie, on a :

$$\rho_j^S = \rho^S(X_j, Y) = \rho(R_{X_j}, R_Y) \ \ \forall\, j = 1, \ldots, d\,. \tag{3.9}$$

On calcule ces coefficients après avoir transformé l'échantillon $(\boldsymbol{X}^n, \boldsymbol{Y}^n)$ en un échantillon $(\boldsymbol{R}_X^n, \boldsymbol{R}_Y^n)$ en remplaçant les valeurs par leur rang dans chaque colonne de la matrice (Saporta [192]) ;

◇ le *coefficient de régression standard sur les rangs* (noté $\text{SRRC}(\cdot,\cdot)$), pendant du SRC mais à partir de l'échantillon $(\boldsymbol{R}_X^n, \boldsymbol{R}_Y^n)$:

$$\text{SRRC}_j = \text{SRRC}(X_j, Y) = \text{SRC}(R_{X_j}, R_Y) \ \ \forall\, j = 1, \ldots, d\,. \tag{3.10}$$

◇ le *coefficient de corrélation partielle sur les rangs* (noté $\text{PRCC}(\cdot,\cdot)$), pendant du PCC mais à partir de l'échantillon $(\boldsymbol{R}_X^n, \boldsymbol{R}_Y^n)$:

$$\text{PRCC}_j = \text{PRCC}(X_j, Y) = \text{PCC}(R_{X_j}, R_Y) \ \ \forall\, j = 1, \ldots, d\,. \tag{3.11}$$

◇ les indices calculés à partir des données segmentées. Pour chaque variable d'entrée, un découpage en classes équiprobables permet d'obtenir plusieurs échantillons de données. Des tests statistiques sont alors appliqués pour mesurer l'homogénéité des populations entre les classes : moyennes communes (CMN) basées sur un test de Fisher, médianes communes (CMD) basées sur un test de χ^2, variances communes (CV) basées sur un test de Fisher, localisations communes (CL) basées sur le test de Kruskal-Wallis, ... (Kleijnen & Helton [118], Helton et al. [84]). D'autres mesures peuvent être utilisées pour tester l'homogénéité des classes (comme par exemple l'entropie).

En pratique, on effectue tout d'abord une régression linéaire entre la sortie Y et les entrées $\boldsymbol{X}$ afin de savoir si leur relation est approximativement linéaire. Pour cela, on peut utiliser le coefficient de détermination R^2 :

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2}{\sum_{i=1}^{n}(\overline{Y} - Y_i)^2} \tag{3.12}$$

avec $\widehat{Y}_i = \widehat{Y}(\boldsymbol{X}^{(i)})$ la prévision du modèle linéaire (3.2) et $\overline{Y}$ la moyenne de l'échantillon $(Y_i)_{i=1..n}$. Ce coefficient peut être utilisé dans tout modèle de régression (pas seulement linéaire) comme critère de qualité de la prédiction du modèle. Basé sur l'utilisation des résidus d'observations ayant servies à ajuster le modèle, il est cependant à prendre avec précaution, et à bannir dans les modèles d'interpolation (comme le krigeage, cf §3.3). Il est souvent préférable de travailler avec d'autres critères (cf. par exemple Azaïs & Bardet [10]). En utilisant des résidus de prédiction (issus d'observations n'ayant pas servies à ajuster le modèle), on définit le coefficient de prédictivité Q_2 du modèle :

$$Q_2 = 1 - \frac{\sum_{i=1}^{n_p}[Y_i^p - \widehat{Y}(\boldsymbol{X}^{p(i)})]^2}{\sum_{i=1}^{n_p}(\overline{Y^p} - Y_i^p)^2} \tag{3.13}$$

où $(\boldsymbol{X}^{p(i)}, Y_i^p)_{i=1..n_p}$ est l'échantillon des variables d'entrée et de sortie de la base de prédiction (appelée aussi base de test), de taille n_p, et $\overline{Y^p}$ est la moyenne de $(Y_i^p)_{i=1..n_p}$. Le coefficient Q_2, dont la terminologie est peu employée en statistique, est comparable au coefficient plus connu nommé PRESS ("PRedictive Error Sum of Squares") qui est le numérateur du membre de droite de l'équation (3.13).

Si on juge l'hypothèse de linéarité acceptable (par exemple si $R^2 > 0.8$), alors les indices de sensibilité Pearson, SRC et PCC sont utilisables. Si les variables d'entrée sont indépendantes, la somme des carrés des SRC vaut R^2 et l'ensemble des SRC^2 forment une décomposition de la variance de la réponse : chaque SRC_j^2 exprime la part de variance de la réponse expliquée par le facteur X_j. Dans le cas où la relation entre $\boldsymbol{X}$ et Y n'est pas linéaire mais monotone, les coefficients de corrélation et de régression basés sur les rangs (Spearman, SRRC, PRCC) peuvent être utilisés. L'hypothèse de monotonie doit bien sûr être validée, par exemple à l'aide du coefficient de détermination de la régression sur les rangs (noté R^{2*}) ou du coefficient de prédictivité associé (noté Q_2^*). Enfin, les méthodes basées sur les tests statistiques ne requièrent pas d'hypothèse sur la monotonie de la réponse en fonction des entrées. Par contre, ces méthodes sont définies pour des échantillons de données indépendantes. En théorie, elles ne sont donc applicables que si l'échantillon est purement aléatoire (tirage Monte Carlo simple). On constate cependant qu'elles sont plutôt robustes vis-à-vis de cette hypothèse ; par exemple elles donnent de bons résultats avec des échantillons de type latin hypercube (LHS). Ces méthodes présentent aussi l'inconvénient d'être peu intuitives comparativement aux méthodes de régression.

Tous les indices de sensibilité basés sur la régression sont calculables à partir d'échantillons de taille supérieure à d alors que les indices basés sur les coefficients de corrélation ou les tests statistiques sont calculables à partir d'échantillons de taille quelconque. Sur les applications, on constate cependant qu'il faut souvent des échantillons de taille supérieure à $2d$ pour obtenir des résultats corrects, la confiance que l'on peut avoir sur les indices augmentant avec le nombre de données utilisées pour les calculer.

3.2.4 Décomposition de la variance

Dans le cadre général d'un modèle non linéaire et non monotone, on peut estimer l'importance des variables d'entrée sur la réponse du modèle en utilisant la décomposition de la variance fonctionnelle (appelée aussi représentation ANOVA fonctionnelle). Toute fonction intégrable sur $\Omega = [0,1]^d$ peut être décomposée en somme de fonctions élémentaires :

$$f(X_1,\cdots,X_d) = f_0 + \sum_i^d f_i(X_i) + \sum_{i<j}^d f_{ij}(X_i,X_j) + \ldots + f_{12..d}(X_1,\cdots,X_d)\ , \tag{3.14}$$

où f_0 est une constante et les autres fonctions vérifient les conditions suivantes :

$$\int_0^1 f_{i_1,\ldots,i_s}(x_{i_1},\ldots,x_{i_s})\,dx_{i_k} = 0 \ \ \forall k=1,\ldots,s\ ,\ \ \forall\{i_1,\ldots,i_s\} \subseteq \{1,\ldots,d\}\ . \tag{3.15}$$

Cette décomposition est connue sous le nom de "décomposition de Hoeffding" (présentée dans Hoeffding [86]) et a été introduite par Sobol [200] pour l'analyse de sensibilité (d'où son appellation "décomposition de Sobol" dans ce domaine). Celui-ci a notamment montré que les conditions (3.15) impliquent que la décomposition est unique.

Le terme "ANOVA" est utilisé car l'équation (3.14) fournit la même interprétation qu'une décomposition ANOVA usuelle. En effet, si les X_i sont mutuellement indépendants, l'équation (3.14) permet d'obtenir une décomposition de la variance de la réponse du modèle :

$$\mathrm{Var}\,[Y] = \sum_{i=1}^d V_i(Y) + \sum_{i<j} V_{ij}(Y) + \sum_{i<j<k} V_{ijk}(Y) + \ldots + V_{12..d}(Y)\ , \tag{3.16}$$

où $V_i(Y) = \mathrm{Var}[\mathbb{E}(Y|X_i)]$, $V_{ij}(Y) = \mathrm{Var}[\mathbb{E}(Y|X_iX_j)] - V_i(Y) - V_j(Y)$ et ainsi de suite. La notion d'analyse de variance sur des espaces de fonction a été introduite par Antoniadis [7]. À partir de (3.16), les indices de sensibilité s'obtiennent alors extrêmement naturellement :

$$S_i = \frac{\mathrm{Var}\,[\mathbb{E}\,(Y|X_i)]}{\mathrm{Var}(Y)} = \frac{V_i(Y)}{\mathrm{Var}(Y)},\quad S_{ij} = \frac{V_{ij}(Y)}{\mathrm{Var}(Y)},\quad S_{ijk} = \frac{V_{ijk}(Y)}{\mathrm{Var}(Y)},\quad \ldots \tag{3.17}$$

Ces coefficients, nommés "mesures d'importance basées sur la variance" ou plus simplement *"indices de Sobol"* (appellation non consacrée), peuvent être utilisés pour n'importe quelle fonction $f(\cdot)$. Les indices d'ordre un sont égaux aux carrés des SRC quand le modèle $f(\cdot)$ est purement linéaire. L'indice du second ordre S_{ij} exprime la sensibilité du modèle à l'interaction entre les variables X_i et X_j, et ainsi de suite pour les ordres supérieurs. Compris entre 0 et 1 et leur somme valant 1, les indices de Sobol sont particulièrement faciles à interpréter (en terme de pourcentage de la variance de la réponse expliquée), ce qui explique leur popularité.

Lorsque le nombre de variables d'entrée d augmente, le nombre d'indices de sensibilité croît exponentiellement (il vaut $2^d - 1$) et l'estimation et l'interprétation de tous ces indices deviennent vite impossibles. Homma & Saltelli [89] ont alors introduit la notion d'indice de sensibilité total pour exprimer tous les effets d'une variable d'entrée sur la sortie :

$$S_{T_i} = S_i + \sum_{j\neq i} S_{ij} + \sum_{j\neq i,k\neq i,j<k} S_{ijk} + \ldots = \sum_{l\,\in\,\#i} S_l\ , \tag{3.18}$$

où $\#i$ représente tous les sous-ensembles d'indices contenant l'indice i. Ainsi, $\sum_{l\,\in\,\#i} S_l$ est la somme de tous les indices de sensibilité faisant intervenir i. En pratique, quand d est grand (par exemple $d > 10$), on se contente souvent d'estimer les indices d'ordre un et les indices totaux. Des exemples sont donnés dans les applications que j'ai eu à traiter (Iooss et al. [104], Volkova et al. [223]). Par ailleurs, quand le nombre d'entrées est vraiment trop grand (par exemple $d > 100$), il est également possible de simplifier le problème en traitant les indices de sensibilité par groupe. Sobol [201] a défini la notion d'indices de sensibilité multidimensionnels. Jacques et al. [107] ont utilisé cette définition pour traiter le problème des variables d'entrée corrélées en regroupant tous les ensembles d'entrées corrélées dans des macroparamètres.

Pour estimer les indices de Sobol, des méthodes basées sur des échantillons Monte Carlo ont été développées (Sobol [200], Saltelli [184]). Malheureusement, pour obtenir des estimations précises des indices de sensibilité, ces méthodes sont extrêmement coûteuses en nombre d'évaluations du modèle (taux de convergence en $\sqrt{N}$ où N est la taille de l'échantillon). Il n'est pas rare dans les applications que l'estimation d'un indice de Sobol requiert 10000 évaluations de $f(\cdot)$ pour obtenir une précision de 10%. De plus, les évaluations effectuées pour estimer un indice ne sont pas réutilisées pour les autres indices. L'utilisation d'échantillons déterministes de type quasi Monte Carlo (par exemple les séquences LPτ de Sobol) à la place d'échantillons Monte Carlo permet de réduire d'un facteur 10 le coût de ces estimations (Saltelli et al. [186], Fang et al. [62]). La méthode FAST (Cukier et al. [46]), basée sur une transformée de Fourier multi-dimensionnelle de $f(\cdot)$, est une autre méthode d'estimation des indices, relativement fine et nettement moins coûteuse que la méthode de Monte Carlo. Saltelli et al. [188] l'ont étendu au calcul des indices totaux mais il n'est pas possible avec cette technique de calculer des indices d'ordre multiple.

Les recherches actuelles pour estimer les indices de Sobol se portent sur le développement d'algorithmes qui permettent d'estimer tous les indices du premier ordre avec un coût indépendant de d. Par exemple Tarantola et al. [209] utilisent une technique dite de "Random Balance Design" couplée avec la méthode FAST. Da Veiga [48] et Da Veiga & Gamboa [49] se sont, quant à eux, intéressés au développement d'estimateurs asymptotiquement efficaces des indices de sensibilité, à partir d'estimations non paramétriques d'intégrales de fonctionnelles de densité. Ces nouveaux algorithmes, bien adaptés aux codes de calcul coûteux, doivent encore être testés intensivement et comparés aux méthodes usuelles. L'estimation à moindre coût des indices de Sobol totaux demeure quant à elle un axe de recherche ouvert et de première importance. La section suivante discute d'autres méthodes récentes d'analyse de sensibilité basées sur l'estimation des moments conditionnels par des méthodes de lissage et sur la construction préalable de métamodèles.

3.2.5 Techniques de lissage et métamodèles

Au delà des indices de Sobol qui ne donnent qu'une valeur scalaire pour l'effet d'une variable d'entrée X_i sur la sortie Y, on peut être intéressé par connaître l'influence sur Y de X_i le long de son domaine de variation. Dans la littérature, on parle souvent d'effets principaux, mais pour éviter toute confusion, il est préférable de parler de visualisation (ou graphe) des effets principaux. L'outil graphique des scatterplots (visualisation du nuage de points d'un échantillon quelconque de simulations $(\boldsymbol{X}^n, \boldsymbol{Y}^n)$ à l'aide des d graphes Y vs. X_i, $i = 1, \ldots, d$) remplit cet objectif mais uniquement de manière visuelle, donc quelque peu subjective. Basées sur des méthodes de régression non paramétrique (Hastie & Tibshirani [82]), les techniques de lissage ont pour objectif, quant à elles, d'estimer les moments conditionnels de Y d'ordre un ou plus. En analyse de sensibilité, on se limite souvent à l'espérance conditionnelle et aux ordres un et deux (Santner et al. [191]) pour obtenir :

- les graphes des effets principaux, entre X_i et $\mathbb{E}(Y|X_i) - \mathbb{E}(Y)$ sur tout le domaine de variation de X_i pour $i = 1, \ldots, d$;

- les graphes des effets des interactions, entre (X_i, X_j) et $\mathbb{E}(Y|X_iX_j) - \mathbb{E}(Y|X_i) - \mathbb{E}(Y|X_j) - \mathbb{E}(Y)$ sur tout le domaine de variation de (X_i, X_j) pour $i = 1, \ldots, d$ et $j = i + 1, \ldots, d$.

Storlie & Helton [204] ont effectué une revue relativement complète des méthodes de lissage que l'on peut utiliser pour l'analyse de sensibilité : moyennes mobiles, méthodes à noyaux, polynômes locaux, splines de lissage. Da Veiga et al. [50] discutent des propriétés théoriques des estimateurs par polynômes locaux de l'espérance et de la variance conditionnelles. Ils en déduisent les propriétés théoriques des estimateurs des indices de Sobol par polynômes locaux. Cette approche leur permet de résoudre le problème des entrées corrélées d'une manière nettement moins coûteuse que par les techniques usuelles. Enfin Ratto et al. [174] présentent une méthode de lissage basée sur le filtre de Kalman. Storlie & Helton [204] discutent également des modèles additifs et des arbres de régression pour estimer de manière non paramétrique $\mathbb{E}(Y|X_1, \ldots, X_d)$, ce qui revient à construire ce qu'on appelle une surface de réponse.

La méthode des surfaces de réponse est un outil connu depuis longtemps qui a pour objectif de construire une fonction qui simule le comportement d'un phénomène physique ou chimique dans le domaine de variation des variables influentes, à partir d'un certain nombre d'expériences (Box & Draper [26]). Des généralisations ultérieures ont amené cette méthode à être utilisée pour construire des modèles simplifiés se substituant à l'exécution de codes de calcul nécessitant trop de temps d'exécution ou de ressources (Sacks et al. [182], Fang et al. [62]). Construire un métamodèle a pour objectif d'obtenir un modèle mathématique représentatif du code étudié en terme de qualité d'approximation, ayant de bonnes capacités de prédiction, et dont le temps de calcul pour évaluer une réponse est négligeable. Ce métamodèle est construit et ajusté à partir de quelques simulations du code (correspondant à différents jeux de valeurs des paramètres). Le nombre de simulations nécessaires dépend de la complexité du code et du scénario qu'il modélise, du nombre de variables d'entrée et de la qualité d'approximation souhaitée. Ce métamodèle peut alors être substitué ou associé au code pour différents objectifs :

- ⋆ prédiction rapide de nouvelles réponses ;
- ⋆ analyse de sensibilité et exploration du modèle pour une meilleure compréhension de son comportement ;
- ⋆ résolution de problèmes d'optimisation de la réponse ou de calibration de paramètres (qui nécessitent parfois plusieurs milliers d'évaluations de la réponse du modèle) ;
- ⋆ participation aux phases de validation et de qualification du modèle numérique.

La construction du métamodèle, basée la plupart du temps sur des techniques de moindres carrés, est évidemment réalisée en accord avec son utilisation future qui peut lui imposer des contraintes. La mise à disposition d'un métamodèle est également extrêmement utile si on étudie un système sans bien connaître les incertitudes sur ses variables d'entrée. Si un métamodèle est construit et validé dans un domaine de variation des entrées suffisamment large, différentes études pourront être réalisées en faisant varier les incertitudes des entrées.

Dans la pratique, on s'intéresse à trois principales questions lors de la construction d'un métamodèle :

- ▷ le choix du métamodèle qui peut être issu de tout modèle de régression linéaire, non linéaire, paramétrique ou non paramétrique (Hastie et al. [83]). Parmi les modèles les plus utilisés pour ajuster les réponses de codes de calcul, on peut citer les polynômes, splines, modèles linéaires généralisés, modèles additifs généralisés, le krigeage, la technique MARS, les réseaux de neurones, les SVM, les arbres de régression et le boosting (Simpson et al. [199], Chen et al. [39], Fang et al. [62]). Le choix du métamodèle est un problème en soi, certains étant plus adaptés que d'autres à différents types de situation. Une première stratégie est de privilégier la simplicité, donc de se satisfaire du métamodèle le plus simple possible en adéquation avec les objectifs de l'étude ;
- ▷ la planification des calculs. Les principales qualités requises pour un plan d'expérience sont sa robustesse (capacité d'analyser différents modèles), son efficacité (minimisation d'un critère), la répartition de ses points (remplissage uniforme de l'espace échantillonné) et un coût faible pour sa construction (Santner et al. [191], Fang et al. [62]). Sur ce sujet, qui ne sera pas discuté dans ce mémoire, énormément de travaux sont disponibles. Si on se limite aux plans sans modèle, on peut citer les travaux de Gazut et al. [73] pour les plans adaptatifs basés sur du reéchantillonnage bootstrap pour identifier les zones où de nouveaux calculs sont nécessaires. Concernant les plans non séquentiels sans modèle (donc de type "space filling"), la tendance actuelle est de chercher des plans dont les propriétés se conservent en sous-projections. Dans ses travaux de thèse, Marrel [141] (associée à un stagiaire, Loïc Boussouf) montre que les LHS optimisés à l'aide de certains critères de discrépance sont parmi les meilleurs pour cet objectif ;

▷ la validation du métamodèle. Dans le domaine des plans d'expérience classiques, la validation correcte d'une surface de réponse est un aspect crucial et à soigner particulièrement (Droesbecke et al. [60]). En revanche, dans le domaine des expériences numériques, seul un faible nombre de publications s'attardent sur ce problème (cf. par exemple Kleijnen & Sargent [119], Meckesheimer et al. [151], Reis dos Santos & Porta Nova [175]). La pratique usuelle est d'estimer des critères globaux (erreur quadratique moyenne, erreur en valeur absolue, ...) sur une base de test, par validation croisée, par *leave-one-out*[5] ou par bootstrap (Kleijnen & Sargent [119], Fang et al. [62], Kleijnen [117]). L'un des critères les plus utilisés en pratique est le coefficient de prédictivité, noté Q_2, qui correspond au coefficient de détermination R^2 calculé sur une base de test (cf. Eqs. (3.13) et (3.12)). Dans Iooss et al. [95], j'ai commencé à m'intéresser au problème de la planification de la base de test, pour estimer au mieux les critères d'erreur du métamodèle avec un nombre minimal de calculs supplémentaires. L'utilisation d'un algorithme développé par Feuillard [63] pendant sa thèse permet de placer dans la base de test les points les plus éloignés possibles de ceux de la base d'apprentissage.

Certains métamodèles permettent d'obtenir directement les indices de sensibilité. Par exemple, Sudret [205] et Crestaux et al. [45] ont montré que les indices de Sobol découlent directement d'une décomposition en polynômes de chaos. Les processus gaussiens (krigeage) apparaissent aussi comme un métamodèle particulièrement intéressant. Leur formulation permet d'obtenir les indices de sensibilité sans passer par une estimation de type Monte Carlo (Oakley & O'Hagan [161], Chen et al. [40], Marrel et al. [142]). La section suivante décrit les travaux sur les processus gaussiens réalisés en collaboration avec Amandine Marrel.

3.3 Construction et utilisation du métamodèle processus gaussien

La méthode du krigeage (Krige [128], Matheron [147], Chilès & Delfiner [42], Stein [203]) a été développée en géostatistique pour des problèmes de cartographie dans le but de prendre en compte la structure spatiale de la variable étudiée. La prédiction du krigeage en chaque point du domaine est une combinaison linéaire pondérée des observations de la base d'apprentissage, les poids ne dépendant pas des valeurs des observations mais du plan d'expérience et de la structure de covariance de la variable. L'un des grands atouts du krigeage, par rapport aux autres méthodes de cartographie, est de se placer dans un cadre probabiliste (en modélisant la variable d'intérêt par un processus stochastique), ce qui lui permet d'estimer les incertitudes associées aux prédictions. A la fin des années 1980, Sacks et al. [182] ont introduit les principes du krigeage pour la modélisation et la prédiction de réponses de codes de calcul. En géostatistique linéaire, l'hypothèse gaussienne n'est pas nécessaire pour construire le modèle de krigeage car ses paramètres sont estimés à l'aide d'outils d'analyse de données (les variogrammes). Dans le domaine des expériences numériques, ces outils ne sont plus utilisables du fait de la grande dimension de l'espace des entrées, et les paramètres du modèle sont estimés par des méthodes nécessitant l'hypothèse gaussienne (maximum de vraisemblance). Les arguments en entrée du modèle PG correspondent aux entrées du code de calcul et peuvent être de nature et d'unité très diverses.

Remarque 3.3.1 D'un point de vue terminologique, si l'argument du modèle est le temps, on parle habituellement de processus stochastique gaussien et s'il s'agit de l'espace physique (position 2D ou 3D), on parle de champ aléatoire gaussien. Concernant la simulation numérique, on adopte la terminologie modèle processus gaussien (noté PG).

Le modèle PG est devenu à présent extrêmement populaire dans le domaine des expériences numériques, principalement pour les raisons suivantes :

◇ disposer d'un interpolateur exact (contrairement à un grand nombre d'autres méthodes de régression) est particulièrement attrayant car la grande majorité des applications concernent

[5] procédure appelée aussi "jacknife", cas particulier de la validation croisée où on laisse de côté une seule observation à chaque étape

des codes de calcul déterministes, *i.e.* des codes pour lesquels deux simulations consécutives avec les mêmes entrées fournissent la même réponse. Cette contrainte d'interpolation peut également être relâchée ;

- ◇ ses principes mathématiques sont relativement simples à appréhender ;

- ◇ le prédicteur possède une formulation analytique, ce qui est parfois souhaitable en terme d'interprétabilité, de communication et de transmission du modèle aux physiciens ;

- ◇ le prédicteur est extrêmement rapide à évaluer ;

- ◇ le modèle PG fournit, en plus du prédicteur, l'erreur de celui-ci (le MSE, *i.e.* Mean Square Error). Le MSE peut alors être utilisé efficacement dans la planification adaptative de simulations pour améliorer la prédictivité du métamodèle (Scheidt [195]), pour estimer des quantiles (Oakley [160]) et pour résoudre des problèmes d'optimisation sur la réponse du code (Jones et al. [110]) ;

- ◇ le modèle PG peut être formulé de manière bayésienne (Currin et al. [47], O'Hagan [162]), ce qui permet d'introduire des lois de connaissance *a priori* sur les hyperparamètres ;

- ◇ le cadre gaussien permet de disposer d'outils d'analyse précieux pour valider ou invalider le modèle PG (Bastos & O'Hagan [16]) ;

- ◇ enfin, il existe à présent une bibliographie particulièrement fournie sur l'application de ce modèle aux expériences numériques, avec notamment les deux monographies récentes de Santner et al. [191] et Rasmussen & Williams [173].

Après avoir rappelé la formulation théorique et la construction pratique du modèle PG, on résume dans la prochaine section nos travaux permettant de résoudre les problèmes posés lors de sa construction en grande dimension (Marrel et al. [143]). On explicitera ensuite nos résultats concernant l'utilisation du modèle PG en analyse de sensibilité (Marrel et al. [142]).

3.3.1 Le modèle processus gaussien

On étudie d variables d'entrée $\boldsymbol{x} = (x_1, ..., x_d) \in \mathbb{R}^d$ d'un code de calcul déterministe $f(\cdot)$ et une variable de sortie scalaire $y(\boldsymbol{x}) \in \mathbb{R}$. On dispose de n jeux de simulations des variables d'entrées $\boldsymbol{X}^n = (\boldsymbol{x}^{(1)t}, ..., \boldsymbol{x}^{(n)t})^t$, chaque jeu étant de dimension d : $\boldsymbol{x}^{(i)} = (x_1^{(i)}, ..., x_d^{(i)})$. Les sorties correspondantes du code de calcul sont notées $\boldsymbol{Y}^n = (y_1, ..., y_n)^t$ avec $y_i = y(\boldsymbol{x}^{(i)}),\ i = 1, ..., n$.

La modélisation par les processus gaussiens consiste à considérer la sortie déterministe $y(\boldsymbol{x})$ comme la réalisation d'un processus stochastique gaussien $Y(\boldsymbol{x})$ se décomposant de la manière suivante :

$$Y(\boldsymbol{x}) = m(\boldsymbol{x}) + Z(\boldsymbol{x}) \ , \tag{3.19}$$

où $m : \boldsymbol{x} \in \mathbb{R}^d \to m(\boldsymbol{x}) \in \mathbb{R}$ est la partie déterministe fournissant une approximation de la réponse du simulateur en moyenne et $Z(\boldsymbol{x}) \in \mathbb{R}$ est la partie aléatoire centrée permettant de modéliser les résidus et d'interpoler les réponses.

En ce qui concerne la partie déterministe du modèle PG, on se limite à l'utilisation d'un polynôme de degré 0 ou 1 :

$$m(\boldsymbol{x}) = \beta_0 + \sum_{j=1}^{d} \beta_j x_j = \boldsymbol{F}(\boldsymbol{x})\boldsymbol{\beta} \tag{3.20}$$

où $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_d)^t \in \mathbb{R}^d$ sont les paramètres de régression et $\boldsymbol{F}(\boldsymbol{x}) = (1, x_1, \ldots, x_d)$ est le vecteur de régression au point $\boldsymbol{x}$. Cette formulation se généralise aisément à d'autres bases de fonctions de régression. Cependant, de nombreux auteurs tels que Sacks et al. [182] et Welch et al. [226] conseillent de ne considérer qu'une fonction constante pour la tendance ($m(\boldsymbol{x}) = \beta_0$) arguant que le processus gaussien $Z(\boldsymbol{x})$ est suffisant pour capturer les non linéarités et interactions du modèle. *A contrario*, d'autres auteurs estiment que cette fonction tendance est importante pour limiter le rôle de la partie stochastique du modèle PG à la modélisation des fluctuations rapides du modèle (Jourdan [111]). Martin & Simpson [145] montrent d'ailleurs qu'une fonction de régression suffisamment riche permet d'améliorer la forme de la fonction de vraisemblance lors de l'estimation des hyperparamètres (cf §3.3.2). Enfin, la tendance déterministe offre une belle opportunité d'introduire d'éventuelles informations *a priori*, inhérente à la physique du phénomène simulé (Martin & Simpson [145]).

La partie aléatoire du modèle $Z(\boldsymbol{x})$ est, quant à elle, un processus stochastique gaussien, que l'on choisit stationnaire, donc caractérisé par ses deux premiers moments statistiques :

$$\begin{aligned} \mathbb{E}[Z(\boldsymbol{x})] &= 0 \,, \\ \mathrm{Cov}(Z(\boldsymbol{x}^{(i)}), Z(\boldsymbol{x}^{(j)})) &= R(\boldsymbol{x}^{(i)}, \boldsymbol{x}^{(j)}) = \sigma^2 N(\boldsymbol{x}^{(i)} - \boldsymbol{x}^{(j)}) \,, \end{aligned} \tag{3.21}$$

où $\sigma^2 \in \mathbb{R}$ désigne la variance de Z et $N : \boldsymbol{r} \in \mathbb{R}^d \to N(\boldsymbol{r}) \in \mathbb{R}$ sa fonction de corrélation. Dans le cadre de la modélisation de réponses codes de calcul, l'hypothèse d'isotropie n'est pas du tout appropriée. En effet, les variables d'entrée étant souvent de nature, de dimension et d'unité différentes, il n'est pas raisonnable de considérer les longueurs de corrélation relatives à chacune des entrées comme identiques. En revanche, on accepte souvent l'hypothèse de stationnarité : un comportement non stationnaire du fait d'une dérive de la moyenne pourra être modélisé *via* la partie déterministe du modèle (c'est le cadre du krigeage universel). Vazquez [219] a introduit le modèle plus général du krigeage intrinsèque pour les codes de calcul. Celui-ci ne nécessite plus l'introduction d'une tendance déterministe, il capture la non stationnarité en modélisant Z par une fonction intrinsèque d'ordre k.[6] Des situations non stationnaires peuvent apparaître dans certaines applications (comme lors de la simulation de phénomènes thermo-hydrauliques), où le comportement et la régularité du code peuvent varier fortement dans différents domaines de ses entrées. Pour traiter ces cas, Gramacy [79] propose de relier un arbre de classification (afin de partitionner l'espace des entrées) à différents modèles PG. La construction de modèles PG non stationnaires pour les expériences numériques (donc en grande dimension) demeure néanmoins un problème de R&D ouvert.

Les principales familles de fonctions de covariance paramétriques ont été décrites au §2.2.4. Dans mes travaux, je me suis limité à la fonction exponentielle généralisée (2.9) :

$$N_{\boldsymbol{\theta},\boldsymbol{p}}(\boldsymbol{x}^{(i)} - \boldsymbol{x}^{(j)}) = \prod_{l=1}^{d} \exp(-\theta_l |x_l^{(i)} - x_l^{(j)}|^{p_l}) \tag{3.22}$$

où $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_d)^t \in \mathbb{R}_+^d$ et $\boldsymbol{p} = (p_1, \ldots, p_d)^t \in]0, 2]^d$ sont les paramètres de corrélation. Cette fonction se décompose en un produit de covariances mono-dimensionnelles relatives à chacune des variables d'entrée, ce qui est bien appropriée aux manipulations analytiques pour les analyses d'incertitude et de sensibilité. Chacune de ces covariances mono-dimensionnelles est caractérisée par deux paramètres (un paramètre de forme p_l et un paramètre d'échelle θ_l pour $l = 1, \ldots, d$), ce qui apporte une certaine souplesse au modèle. D'autres auteurs préconisent l'utilisation de la covariance de Matèrn (Eq. (2.10)), dont les propriétés de régularité sont plus intéressantes, mais dont l'expression analytique est plus complexe (cf. §2.2.4).

En dépit du caractère déterministe des codes que l'on considère, les sorties peuvent être assorties de bruits (notamment numériques). Dans ce cas, il peut être intéressant que le modèle PG n'interpole plus les données et on ajoute à sa partie stochastique un bruit blanc $U(\boldsymbol{x})$:

$$Y(\boldsymbol{x}) = m(\boldsymbol{x}) + Z(\boldsymbol{x}) + U(\boldsymbol{x}), \tag{3.23}$$

où $U(\boldsymbol{x}) \in \mathbb{R}$ est un bruit-blanc gaussien : processus aléatoire indépendant gaussien centrée de variance $\sigma^2\tau$, $\tau \in \mathbb{R}_+$. Ce bruit-blanc introduit une discontinuité à l'origine (appelée effet de pépite par Matheron

[6] Il s'agit d'une généralisation de la notion de fonction aléatoire intrinsèque qui correspond au cas $k = 0$ (cf. §2.2.2), pour utiliser des accroissements généralisés (cf. Chilès & Delfiner [42]).

[147]) sur la fonction de covariance :

$$\mathrm{Cov}(Y(\boldsymbol{x}), Y(\boldsymbol{u})) = \sigma^2 \left(N_{\boldsymbol{\theta},\boldsymbol{p}}(\boldsymbol{x}-\boldsymbol{u}) + \tau\delta(\boldsymbol{x}-\boldsymbol{u}) \right) \tag{3.24}$$

où $\delta(\boldsymbol{v}) = \begin{cases} 1 \text{ si } \|\boldsymbol{v}\| = 0, \\ 0 \text{ sinon.} \end{cases}$ Cet effet de pépite peut également être utile pour améliorer le conditionnement de la matrice de covariance à inverser pour construire le prédicteur du modèle PG (cf §3.3.2).

3.3.2 Construction et estimation des paramètres

Sous l'hypothèse du modèle processus gaussien, l'échantillon d'apprentissage $\boldsymbol{Y}^n$ suit une distribution normale multidimensionnelle :

$$\left[\boldsymbol{Y}^n \middle| \boldsymbol{X}^n, \boldsymbol{\beta}, \sigma, \boldsymbol{\theta}, \boldsymbol{p}, \tau \right] \sim \mathcal{N}_n \left(\boldsymbol{F}_n \boldsymbol{\beta}, \boldsymbol{\Sigma}_n \right) , \tag{3.25}$$

où $\boldsymbol{F}_n = [\boldsymbol{F}(\boldsymbol{x}^{(1)})^t, \ldots, \boldsymbol{F}(\boldsymbol{x}^{(n)})^t]^t$ est la matrice de régression et

$$\boldsymbol{\Sigma}_n = \sigma^2 \left(N_{\boldsymbol{\theta},\boldsymbol{p}} \left(\boldsymbol{x}^{(i)} - \boldsymbol{x}^{(j)} \right)_{i,j=1..n} + \tau \boldsymbol{I}_n \right) \tag{3.26}$$

est la matrice de covariance avec $\boldsymbol{I}_n$ la matrice identité de dimension n. Soit $\boldsymbol{x}^* = (x_1^*, ..., x_d^*)$ un nouveau jeu de variables d'entrée, la distribution de probabilité jointe de $(\boldsymbol{Y}^n, Y(\boldsymbol{x}^*))$ s'écrit

$$\left[\begin{pmatrix} \boldsymbol{Y}^n \\ Y(\boldsymbol{x}^*) \end{pmatrix} \middle| \boldsymbol{X}^n, \boldsymbol{\beta}, \sigma, \boldsymbol{\theta}, \boldsymbol{p}, \tau \right] \sim \mathcal{N}_{n+1} \left(\begin{bmatrix} \boldsymbol{F}_n \\ \boldsymbol{F}(\boldsymbol{x}^*) \end{bmatrix} \boldsymbol{\beta}, \begin{bmatrix} \boldsymbol{\Sigma}_n & \boldsymbol{k}(\boldsymbol{x}^*) \\ \boldsymbol{k}(\boldsymbol{x}^*)^t & \sigma^2(1+\tau) \end{bmatrix} \right), \tag{3.27}$$

avec

$$\begin{aligned} \boldsymbol{k}(\boldsymbol{x}^*) &= [\, \mathrm{Cov}(y_1, Y(\boldsymbol{x}^*)), \ldots, \mathrm{Cov}(y_n, Y(\boldsymbol{x}^*)) \,]^t \\ &= \sigma^2 [\, N_{\boldsymbol{\theta},\boldsymbol{p}}(\boldsymbol{x}^{(1)} - \boldsymbol{x}^*) + \tau\delta(\boldsymbol{x}^{(1)} - \boldsymbol{x}^*), \ldots, N_{\boldsymbol{\theta},\boldsymbol{p}}(\boldsymbol{x}^{(n)} - \boldsymbol{x}^*) + \tau\delta(\boldsymbol{x}^{(n)} - \boldsymbol{x}^*) \,]^t. \end{aligned} \tag{3.28}$$

On obtient alors la distribution de $Y(\boldsymbol{x}^*)$ conditionnellement à l'échantillon d'apprentissage $(\boldsymbol{X}^n, \boldsymbol{Y}^n)$, qui est gaussienne de moyenne et variance :

$$\mathbb{E}^*[Y(\boldsymbol{x}^*) | \boldsymbol{Y}^n, \boldsymbol{X}^n, \boldsymbol{\beta}, \sigma, \boldsymbol{\theta}, \boldsymbol{p}, \tau] = \boldsymbol{F}(\boldsymbol{x}^*)\boldsymbol{\beta} + \boldsymbol{k}(\boldsymbol{x}^*)^t {\boldsymbol{\Sigma}_n}^{-1} (\boldsymbol{Y}^n - \boldsymbol{F}_n \boldsymbol{\beta}), \tag{3.29}$$

$$\mathrm{Var}^*[Y(\boldsymbol{x}^*) | \boldsymbol{Y}^n, \boldsymbol{X}^n, \boldsymbol{\beta}, \sigma, \boldsymbol{\theta}, \boldsymbol{p}, \tau] = \sigma^2(1+\tau) - \boldsymbol{k}(\boldsymbol{x}^*)^t {\boldsymbol{\Sigma}_n}^{-1} \boldsymbol{k}(\boldsymbol{x}^*) , \tag{3.30}$$

où $\mathbb{E}^*$ et Var^* sont l'espérance et la variance par rapport à la loi *a posteriori* de $Y(\boldsymbol{x})$.

L'espérance conditionnelle (3.29), notée $\widehat{Y}(\boldsymbol{x}^*)$, est donc utilisée comme prédicteur. La variance conditionnelle (3.30), connue sous le nom de variance du krigeage, peut être utilisée comme un indicateur de la variance locale du modèle. Cette variance correspond d'ailleurs à l'erreur quadratique moyenne du prédicteur que l'on note MSE pour Mean Square Error. Plus généralement, le modèle PG fournit une expression analytique pour la distribution de la variable de sortie en chaque nouveau point de prédiction. Cette distribution peut être utilisée pour la propagation d'incertitude, l'estimation de quantiles, le développement de stratégies d'échantillonnage, ... Son utilisation pourra être analytique (ou du moins en partie) en évitant de passer par des simulations numériques de type Monte Carlo qui s'avèrent souvent coûteuses en temps de calcul.

A partir de la loi jointe énoncée précédemment et des simulations de la base d'apprentissage, on estime les paramètres du modèle PG (3.23), appelés hyperparamètres. Il s'agit des paramètres de régression $\boldsymbol{\beta}$, de corrélation $(\boldsymbol{\theta}, \boldsymbol{p})$ et de variance (σ^2, τ). Pour la covariance exponentielle généralisée et un polynôme de degré un dans la partie régression, il y a donc $3d+3$ paramètres à estimer. Deux méthodes sont utilisées dans la littérature pour estimer ces paramètres : la méthode du maximum de vraisemblance (Sacks et al. [182]) et la technique de validation croisée (Currin et al. [47]). Cette dernière ne nécessite pas l'hypothèse d'une distribution gaussienne de la sortie. Elle consiste à explorer de manière exhaustive les valeurs des hyperparamètres et à estimer les résidus de prédiction (par exemple par *leave-one-out*) pour chaque valeur du vecteur des hyperparamètres (Martin & Simpson [145]). La valeur du vecteur des hyperparamètres retenue est celle qui donne l'erreur quadratique la plus faible, ce qui correspond à une méthode de moindres carrés. Ceci peut donc s'avérer particulièrement coûteux en grande dimension (par exemple pour $d > 10$). De plus, Martin & Simpson [145] ont montré sur des exemples simples que cette technique donne souvent des résultats de moins bonne qualité que celle basée sur le maximum de vraisemblance. Enfin, d'un point de vue théorique, il est connu que l'estimateur des paramètres d'un champ gaussien par moindres carrés n'est pas consistant à partir de $d \geq 2$ (Guyon [80]).

La méthode du maximum de vraisemblance consiste à exprimer la log-vraisemblance de $\boldsymbol{Y}^n$ en supposant que les observations suivent une distribution gaussienne.[7] L'estimateur du maximum de vraisemblance de $\boldsymbol{\beta}$ est l'estimateur des moindres carrés généralisés :

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{F}_n{}^t(\boldsymbol{N}_{\boldsymbol{\theta},\boldsymbol{p}} + \tau \boldsymbol{I}_n)^{-1}\boldsymbol{F}_n)^{-1} \;\; \boldsymbol{F}_n{}^t(\boldsymbol{N}_{\boldsymbol{\theta},\boldsymbol{p}} + \tau \boldsymbol{I}_n)^{-1}\boldsymbol{Y}^n, \tag{3.31}$$

avec la matrice $\boldsymbol{N}_{\boldsymbol{\theta},\boldsymbol{p}} = N_{\boldsymbol{\theta},\boldsymbol{p}}\left(\boldsymbol{x}^{(i)} - \boldsymbol{x}^{(j)}\right)_{i,j=1..n}$. On a alors pour le prédicteur et son MSE :

$$\mathbb{E}^*[Y(\boldsymbol{x}^*)|\boldsymbol{Y}^n, \boldsymbol{X}^n, \sigma, \boldsymbol{\theta}, \boldsymbol{p}, \tau] = \widehat{Y}(\boldsymbol{x}^*) = \boldsymbol{F}(\boldsymbol{x}^*)\widehat{\boldsymbol{\beta}} + \boldsymbol{k}(\boldsymbol{x}^*)^t \boldsymbol{\Sigma}_n{}^{-1}(\boldsymbol{Y}^n - \boldsymbol{F}_n\widehat{\boldsymbol{\beta}}) \; , \tag{3.32}$$

$$\mathrm{MSE}[\widehat{Y}(\boldsymbol{x}^*)|\boldsymbol{Y}^n, \boldsymbol{X}^n, \sigma, \boldsymbol{\theta}, \boldsymbol{p}, \tau] = \sigma^2(1+\tau) - \boldsymbol{k}(\boldsymbol{x}^*)^t \boldsymbol{\Sigma}_n{}^{-1}\boldsymbol{k}(\boldsymbol{x}^*) + \boldsymbol{u}(\boldsymbol{x}^*)(\boldsymbol{F}_n{}^t \boldsymbol{\Sigma}_n{}^{-1}\boldsymbol{F}_n)\boldsymbol{u}(\boldsymbol{x}^*)^t \tag{3.33}$$

avec $\boldsymbol{u}(\boldsymbol{x}^*) = \boldsymbol{F}(\boldsymbol{x}^*) - \boldsymbol{k}(\boldsymbol{x}^*)^t \boldsymbol{\Sigma}_n{}^{-1}\boldsymbol{F}_n$.

Remarque 3.3.2 En pratique, les paramètres de covariance $(\sigma, \boldsymbol{\theta}, \boldsymbol{p}, \tau)$ sont inconnus et estimés à partir de la base d'apprentissage ; la vraie MSE,

$$\mathrm{MSE}[\widehat{Y}(\boldsymbol{x}^*)|\boldsymbol{Y}^n, \boldsymbol{X}^n] = \mathrm{Var}[\widehat{Y}(\boldsymbol{x}^*) - Y(\boldsymbol{x}^*)] + \{\mathbb{E}[\widehat{Y}(\boldsymbol{x}^*) - Y(\boldsymbol{x}^*)]\}^2 \; , \tag{3.34}$$

est donc inconnue. Par décomposition de la variance de $\widehat{Y}(\boldsymbol{x}^*) - Y(\boldsymbol{x}^*)$ par rapport à $[\widehat{Y}(\boldsymbol{x}^*) - Y(\boldsymbol{x}^*)|\sigma, \boldsymbol{\theta}, \boldsymbol{p}, \tau]$, il est facile de voir que la MSE *(3.33) sous-estime la vraie* MSE *(Santner et al. [191], den Hertog et al. [56]). La majorité des auteurs se contentent cependant d'utiliser l'expression (3.33) comme approximation de la* MSE. *Dans un article récent, den Hertog et al. [56] proposent des algorithmes pour estimer la vraie* MSE *à l'aide du bootstrap.*

Le prédicteur (3.32) ne dépend pas de σ^2 mais son MSE (3.33) en dépend. L'estimateur de σ^2 vaut :

$$\widehat{\sigma^2} = \frac{1}{n}(\boldsymbol{Y}^n - \boldsymbol{F}_n\widehat{\boldsymbol{\beta}})^t(\boldsymbol{N}_{\boldsymbol{\theta},\boldsymbol{p}} + \tau\boldsymbol{I}_n)^{-1}(\boldsymbol{Y}^n - \boldsymbol{F}_n\widehat{\boldsymbol{\beta}}). \tag{3.35}$$

Ainsi, $\widehat{\boldsymbol{\beta}}$ et $\widehat{\sigma^2}$ dépendent de $\boldsymbol{\theta}$, $\boldsymbol{p}$ et τ et, en les substituant dans l'expression de la log-vraisemblance, on obtient le choix optimal $(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{p}}, \widehat{\tau})$. Celui-ci minimise

$$\psi(\boldsymbol{\theta}, \boldsymbol{p}, \tau) = \widehat{\sigma^2}|\boldsymbol{N}_{\boldsymbol{\theta},\boldsymbol{p}} + \tau\boldsymbol{I}_n|^{\frac{1}{n}} \; . \tag{3.36}$$

Pour résoudre ce problème d'optimisation, Welch et al. [226] utilisent la méthode du simplexe couplée à une procédure d'introduction progressive des paramètres de corrélation des variables d'entrée pour réduire le plus possible la fonction $\psi(\boldsymbol{\theta}, \boldsymbol{p}, \tau)$. D'autres auteurs se limitent à l'utilisation d'une méthode de gradient conjugué, et parfois couple celle-ci avec une optimisation préalable par simplexe (en la répétant éventuellement avec différents points de départ choisis aléatoirement). Dans le logiciel sous Matlab DACE, Lophaven et al. [137] utilisent l'algorithme d'optimisation stochastique de Hooke & Jeeves (Bazaraa et al. [18]) qui se révèle particulièrement efficace (et que l'on utilise dans notre algorithme d'ajustement d'un modèle PG à grande dimension, cf §3.3.3). Cependant, quelques problèmes conséquents se posent lors de cette phase d'estimation des hyperparamètres :

[7] La méthode du maximum de vraisemblance est connue pour être relativement robuste à l'hypothèse de normalité.

- ▷ l'évaluation de $\psi(\boldsymbol{\theta}, \boldsymbol{p}, \tau)$ qui nécessite l'inversion de la matrice de covariance. Une difficulté apparaît lorsque l'indice de conditionnement de la matrice de covariance est trop grand (système numérique instable pour inverser la matrice). Le conditionnement de la matrice dépend non seulement du plan d'expérience mais aussi du modèle de covariance. La covariance gaussienne implique notamment une matrice de covariance particulièrement mal conditionnée (Ababou et al. [1], Stein [203]). L'effet de pépite $U(\boldsymbol{x})$ que l'on introduit permet de résoudre ce problème de conditionnement ;

- ▷ les nombreux maxima locaux potentiellement présents dans la fonction de vraisemblance. Cette situation est présente lorsqu'il y a peu de données pour exprimer la fonction de vraisemblance. Martin & Simpson [145] ont illustré ce problème sur des exemples simples et ont montré que l'introduction d'une partie déterministe $m(\boldsymbol{x})$ (terme de tendance) suffisamment riche permettait d'améliorer la forme de la log-vraisemblance en rendant plus aisée la détermination de son maximum global ;

- ▷ la forme de la fonction de vraisemblance qui peut présenter une arête relativement large. Le maximum global peut ainsi se trouver dans une zone très plate de la vraisemblance, ce qui induit des difficultés numériques pour les algorithmes de gradient et du simplexe. Li & Sudjianto [133] ont proposé de pénaliser la vraisemblance, par exemple par une technique de type LASSO qui induit une courbure plus forte de la vraisemblance, permettant une estimation précise du maximum global. Dans cette approche particulièrement efficace (nommée krigeage gaussien pénalisé), le terme de pénalisation est mis sur les paramètres de corrélation $\boldsymbol{\theta}$;

- ▷ la grande dimension des entrées qui implique un grand nombre de paramètres à ajuster et consécutivement quelques difficultés dans la procédure d'optimisation par maximum de vraisemblance (coût en temps de calcul, surapprentissage).

La section suivante explicite l'heuristique que j'ai développée pour répondre à ces problèmes.

3.3.3 Méthodologie en grande dimension

L'application du modèle PG à des codes de calcul faisant intervenir un grand nombre de variables d'entrée est une problématique identifiée depuis les travaux de Welch et al. [226]. Le problème semble se poser au delà de 5 variables d'entrée et pour un nombre d'observations du code de l'ordre de quelques centaines (typiquement dix fois plus d'observations que de variables d'entrée). La principale difficulté demeure dans la procédure d'optimisation afin d'estimer les paramètres de corrélation. Une optimisation globale et simultanée de tous les paramètres de corrélation conduit inévitablement à de très mauvaises estimations des hyperparamètres et consécutivement à un modèle PG peu prédictif (Welch et al. [226], Marrel et al. [143]).

Welch et al. [226] utilisent un algorithme *stepwise* pour introduire séquentiellement les variables d'entrée (et donc leurs hyperparamètres) dans la vraisemblance du modèle PG. La première étape consiste à initialiser les hyperparamètres par ceux d'une covariance isotrope (donc avec deux paramètres de corrélation à ajuster) par maximisation de la log-vraisemblance des données. La deuxième étape consiste à réaliser une boucle en réestimant successivement les hyperparamètres de chaque variable d'entrée. Les estimations des hyperparamètres de la variable d'entrée qui ont conduit à la plus forte amélioration de la log-vraisemblance sont conservées et les autres sont reinitialisés à leur valeur initiale. La procédure est répétée jusqu'à un critère d'arrêt sur l'évolution de la log-vraisemblance, ou jusqu'à ce que tous les hyperparamètres aient été estimés. Avec cet algorithme, Welch et al. [226] parviennent à construire un modèle PG à l'aide d'une trentaine de calculs sur un code faisant intervenir 20 variables d'entrée.

Suite à cet algorithme, peu d'auteurs se sont intéressés à ce problème qui ne présente pas d'intérêt théorique mais un intérêt pratique évident. Pour des codes de calcul faisant intervenir quelques dizaines de variables d'entrée, Schonlau & Welch [196] passent en revue un certain nombre de publications où le modèle PG semble avoir donné des résultats satisfaisants. Ils proposent d'utiliser le modèle PG comme une technique de criblage avancée permettant, grâce à la décomposition ANOVA fonctionelle, la visualisation des effets principaux de chaque variable d'entrée (cf. §3.2.5). Le problème de l'estimation des hyperparamètres est ignoré et ramené à un problème de validation du modèle PG (prédicteur et variance) par validation croisée (procédure détaillée par exemple dans Jones et al. [110]). Cette validation croisée par *leave-one-out* semble cependant douteuse car les paramètres de corrélation ne sont pas réestimés à chaque suppression d'observations, ce qui est justifié par le coût d'une réestimation systématique de ceux-ci. Or, lorsque la dimension des entrées n'est pas négligeable devant le nombre de données, chaque observation a une influence potentiellement importante sur l'estimation des hyperparamètres.

Dans Marrel et al. [143], nous proposons d'utiliser une procédure de validation qui évite ce biais en réestimant les hyperparamètres à chaque étape de validation croisée. Le *leave-one-out* n'est alors plus possible car trop coûteux, et une validation croisée en quelques blocs est utilisée. Le choix du nombre de blocs est bien entendu dépendant du nombre d'observations dont on dispose, mais il est usuellement inférieur à 10. Le critère utilisé pour sélectionner le meilleur modèle PG est le coefficient de prédictivité Q_2 (cf. §3.2.3, Eq. (3.13)), le R^2 n'étant d'aucune utilité pour les modèles d'interpolation des observations comme le krigeage.

Notre intérêt pour ce problème d'ajustement de modèles PG à grande dimension vient du fait que l'algorithme de Welch s'est révélé inapproprié et relativement inefficace sur certaines de nos applications. En effet, en restreignant la partie déterministe du modèle PG à une constante (cf §3.3.1), il ne s'intéresse pas à l'estimation séquentielle des paramètres de régression $\boldsymbol{\beta}$ et aux liens entre celle-ci et l'estimation séquentielle des paramètres de covariance. De plus, sa procédure d'initialisation des hyperparamètres est relativement pauvre. Cette initialisation suppose l'isotropie de la covariance, ce qui force les paramètres de corrélation à être égaux entre eux et qui peut conduire à des estimations très éloignées des solutions. Or, celles-ci conditionnent la suite de la procédure d'ajustement car le modèle qui est testé à chaque étape prend en compte toutes les variables d'entrée.

L'algorithme que nous avons développé dans Marrel et al. [143] s'inspire de celui de Welch, en le raffinant afin de pallier aux différents problèmes évoqués précédemment. Les grands principes de l'algorithme proposé sont les suivants :

- ⋆ un tri initial est effectué sur les variables d'entrée afin de les classer par ordre d'influence sur la sortie. Le critère de tri est le coefficient de corrélation linéaire entre chaque entrée et la sortie, calculé à l'aide de l'échantillon d'apprentissage ;

- ⋆ ce tri donne l'ordre d'inclusion progressive des variables d'entrée dans la covariance du modèle PG lors de la procédure d'estimation de ses hyperparamètres. A chaque inclusion d'une nouvelle variable d'entrée, tous les hyperparamètres sont estimés par maximisation de la log-vraisemblance (qui revient à minimiser l'expression (3.36)) ;

- ⋆ la présence d'une partie régression et d'une partie covariance nous oblige à mettre en œuvre une boucle supplémentaire pour sélectionner les termes de la fonction de régression. Le critère d'information d'Akaike (AIC) est bien adapté pour sélectionner un modèle de régression. On utilise ici le critère d'information d'Akaike corrigé (AICC) qui permet de prendre en compte la présence supplémentaire du terme de covariance (Hoeting et al. [87]) :

$$\mathrm{AICC} = -2 l_{\boldsymbol{Y}^n}\left(\widehat{\boldsymbol{\beta}}, \widehat{\sigma}, \widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{p}}, \widehat{\tau}\right) + 2n \frac{m_1 + m_2 + 1}{n - m_1 - m_2 - 2}\,, \tag{3.37}$$

  où m_1 est le nombre de variables qui interviennent dans la partie régression $m(\cdot)$, m_2 est le nombre de variables qui interviennent dans la fonction de covariance et $l_{\boldsymbol{Y}^n}(\cdot)$ est la log-vraisemblance de l'échantillon $\boldsymbol{Y}^n$ pour les estimations des hyperparamètres du modèle PG. Toutes les variables d'entrée ne sont donc pas forcément incluses dans la régression. Le critère AICC étant peu coûteux, cette boucle de sélection des termes de régression est insérée dans la boucle d'estimation des paramètres de covariance ;

- ⋆ pour chaque variable d'entrée incluse dans la covariance, la qualité du modèle PG est estimé à l'aide du coefficient de prédictivité Q_2 (Eq. (3.13)), calculé par validation croisée ;

- ⋆ l'évolution du Q_2 est visualisée en fonction des itérations. Les incréments successifs du Q_2 à chaque ajout de variables (qui trahissent l'influence des variables d'entrée dans le modèle PG) sont utilisés pour offrir un nouveau tri initial des entrées. La procédure séquentielle de construction du modèle PG par inclusion progressive des entrées est alors relancée ;

- ⋆ le modèle sélectionné n'est pas le modèle final mais celui pour lequel le Q_2 est le meilleur. Toutes les variables d'entrée ne sont donc pas forcément incluses dans la covariance.

De manière plus formelle, on note $\mathcal{M}_0 = \left\{e_1^{(0)}, \dots, e_d^{(0)}\right\}$ la liste de toutes les entrées dans leur ordre initial. $\mathcal{M}_1 = \left\{e_1^{(1)}, \dots, e_d^{(1)}\right\}$ (resp. $\mathcal{M}_2 = \left\{e_1^{(2)}, \dots, e_d^{(2)}\right\}$) correspond à la liste des entrées dans leur nouvel ordre après classement avec le critère du coefficient de corrélation (resp. le critère des incréments de Q_2). $\mathcal{M}_{\text{cov}}$ (resp. $\mathcal{M}_{\text{reg}}$) réprésente la liste des entrées apparaissant dans la fonction de covariance (resp. la fonction de régression) à l'étape courante. L'algorithme que j'ai proposé se formalise de la manière suivante :

Etape 1 : classement initial

$$\mathcal{M}_0 = \left\{\mathsf{e}_1^{(0)}, \dots, \mathsf{e}_\mathsf{d}^{(0)}\right\} \Longrightarrow \mathcal{M}_1 = \left\{\mathsf{e}_1^{(1)}, \dots, \mathsf{e}_\mathsf{d}^{(1)}\right\} \Longrightarrow \left\{ \begin{array}{l} \mathcal{M}_{\mathsf{reg}} = \mathcal{M}_1 \\ \mathcal{M}_{\mathsf{cov}} = \mathcal{M}_1 \end{array} \right.$$

Etape 2 : initialisations

Bornes minimales pour chaque composante de $\boldsymbol{\theta}$ et $\boldsymbol{p}$: $\mathsf{lob}_{\boldsymbol{\theta}} = 10^{-8}$, $\mathsf{lob}_{\boldsymbol{p}} = 0$
Bornes maximales pour chaque composante de $\boldsymbol{\theta}$ et $\boldsymbol{p}$: $\mathsf{upb}_{\boldsymbol{\theta}} = 100$, $\mathsf{upb}_{\boldsymbol{p}} = 2$
Valeurs de départ pour chaque composante de $\boldsymbol{\theta}$ et $\boldsymbol{p}$: $\theta^0 = 0.5$, $\mathsf{p}^0 = 1$

Etape 3 : inclusion successive des variables dans la covariance
Pour $\mathsf{i} = 1 \dots \mathsf{d}$
Variables dans la covariance : $\mathcal{M}_{\mathsf{i,cov}} = \mathcal{M}_{\mathsf{cov}}(1, \dots, \mathsf{i})$
Inclusion successive des variables dans la fonction de régression :
Pour $\mathsf{j} = 1 \dots \mathsf{d}$
Fonction de régression : $\mathcal{M}_{\mathsf{j,reg}} = \mathcal{M}_{\mathsf{reg}}(1, \dots, \mathsf{j})$
$\boldsymbol{\theta}^{\mathsf{init}} = (\theta_1{}^{(\mathsf{i}-1),\mathsf{j}}, \dots, \theta_{\mathsf{i}-1}{}^{(\mathsf{i}-1),\mathsf{j}}, \theta^0)^{\mathsf{t}}$
$\boldsymbol{p}^{\mathsf{init}} = (\mathsf{p}_1{}^{(\mathsf{i}-1),\mathsf{j}}, \dots, \mathsf{p}_{\mathsf{i}-1}{}^{(\mathsf{i}-1),\mathsf{j}}, \mathsf{p}^0)^{\mathsf{t}}$
$[\theta^{\mathsf{i,j}}, \mathsf{p}^{\mathsf{i,j}}] = \text{estimation}(\mathcal{M}_{\mathsf{i,cov}}, \mathcal{M}_{\mathsf{j,reg}}, [\boldsymbol{\theta}^{\mathsf{init}}, \boldsymbol{p}^{\mathsf{init}}], [\mathsf{lob}_{\boldsymbol{\theta}}, \mathsf{lob}_{\boldsymbol{p}}], [\mathsf{upb}_{\boldsymbol{\theta}}, \mathsf{upb}_{\boldsymbol{p}}])$
$\mathsf{AICC}(\mathsf{i}, \mathsf{j}) = \mathsf{AICC}(\mathcal{M}_{\mathsf{i,cov}}, \mathcal{M}_{\mathsf{j,reg}})$
Fin de la boucle
Sélection de la fonction de régression optimale : $\mathsf{j}^{\mathsf{optim}}(\mathsf{i}) = \arg\min_{\mathsf{j}} (\mathsf{AICC}(\mathsf{i}, \mathsf{j}))$
Evaluation du Q_2 par validation croisée ou sur une base de validation
$\mathsf{Q}_2(\mathsf{i}) = \mathsf{Q}_2(\mathcal{M}_{\mathsf{i,cov}}, \mathcal{M}_{\mathsf{j}^{\mathsf{optim}}(\mathsf{i}),\mathsf{reg}})$
Fin de la boucle

Etape 4 : détermination du nouveau classement par incrément de Q_2
$\Delta\mathsf{Q}_2(1) = \mathsf{Q}_2(1)$
Pour $\mathsf{k} = 2 \dots \mathsf{d}$
$\Delta\mathsf{Q}_2(\mathsf{k}) = \mathsf{Q}_2(\mathsf{k}) - \mathsf{Q}_2(\mathsf{k} - 1)$
Fin de la boucle
Classement des variables par $\Delta\mathsf{Q}_2$ décroissants : $\mathcal{M}_1 \Longrightarrow \mathcal{M}_2$

Etape 5 : estimation des paramètres à partir du nouveau classement $\begin{cases} \mathcal{M}_{\mathsf{reg}} = \mathcal{M}_1 \\ \mathcal{M}_{\mathsf{cov}} = \mathcal{M}_2 \end{cases}$

Etape 6 : sélection du modèle optimal

$$\mathsf{i}^{\mathsf{optim}} = \arg\max_{\mathsf{i}} \left(\mathsf{Q}_2(\mathsf{i})\right)$$

$$\begin{cases} \mathcal{M}_{\mathsf{cov}}^{\mathsf{optim}} = \mathcal{M}_{\mathsf{cov}}(1, \dots, i^{\mathsf{optim}}) \\ \mathcal{M}_{\mathsf{reg}}^{\mathsf{optim}} = \mathcal{M}_{\mathsf{reg}}(1, \dots, j^{\mathsf{optim}}(i^{\mathsf{optim}})) \end{cases}$$

Etape 7 : validation finale du modèle optimal

$$\mathsf{Q}_2^{\mathsf{final}} = \mathsf{Q}_2(\mathcal{M}_{\mathsf{cov}}^{\mathsf{optim}}, \mathcal{M}_{\mathsf{reg}}^{\mathsf{optim}})$$

Cet algorithme a clairement démontré sa supériorité par rapport aux algorithmes ne faisant pas intervenir de procédure séquentielle sur quelques exemples “jouets”, par exemple sur la fonction g de Sobol définie par

$$g_{\mathrm{Sobol}}(X_1, \dots, X_d) = \prod_{j=1}^{d} g_j(X_j) \text{ où } g_j(X_j) = \frac{|4X_j - 2| + a_j}{1 + a_j} , \tag{3.38}$$

avec $X_j \sim \mathcal{U}[0,1]$ et $a_j = j, \forall j = 1 \dots d$. Avec ces valeurs, la fonction g de Sobol modélise des comportements non linéaires avec des interactions entre les entrées. Sur cette fonction, le tableau 3.1 montre la comparaison de résultats obtenus avec notre algorithme (Marrel et al. [143]) et avec celui proposé par le logiciel GEM-SA (O'Hagan [162]), en faisant varier la dimension des entrées d. La taille de l'échantillon de construction du modèle PG est choisie à $n = 10d$. La procédure de simulation d'échantillons d'apprentissage (par la méthode des hypercubes latins) et de construction des modèles PG est répétée 50 fois pour pouvoir moyenner les résultats. On constate que pour $d \geq 6$, l'algorithme de Marrel donne de bien meilleures performances.

Simulations de g_{Sobol}		Algorithme de GEM-SA		Algorithme de Marrel	
d	n	$\overline{Q_2}$	sd	$\overline{Q_2}$	sd
4	40	0.82	0.08	0.86	0.07
6	60	0.67	0.24	0.85	0.05
8	80	0.66	0.13	0.85	0.04
10	100	0.59	0.25	0.83	0.05
12	120	0.57	0.16	0.84	0.05
14	140	0.60	0.17	0.83	0.03
16	160	0.62	0.11	0.86	0.04
18	180	0.66	0.09	0.84	0.03
20	200	0.64	0.09	0.86	0.02

TAB. 3.1 – Moyenne ($\overline{Q_2}$) et écart type (sd) du coefficient de prédictivité Q_2 pour plusieurs implémentations de la fonction g de Sobol. 50 répétitions sont utilisées pour chaque taille d'échantillon.

Cet algorithme nous a également permis de construire des modèles PG performants sur quelques applications parmi lesquelles un modèle de transport hydrogéologique de polluants à 20 entrées et 300 observations (Marrel et al. [143]) et un modèle d'accident thermohydraulique en sûreté des réacteurs nucléaires à 53 entrées et 200 observations (Cannamela et al. [36]).

Un travail récent (Linkletter et al. [135]) s'est également intéressé au problème de la sélection de variables à l'aide du modèle PG. La technique employée est basée sur l'introduction d'une variable d'entrée inerte (que l'on pourrait aussi appeler inactive ou fictive) dans le modèle. La distribution *a posteriori* du paramètre de covariance θ de cette variable inerte permet de définir un seuil au delà duquel les autres variables d'entrée peuvent être considérées comme influentes. Au final, le but est de sélectionner les variables d'entrée sur lesquelles il est intéressant de réaliser une analyse de sensibilité quantitative (calcul des indices de Sobol par exemple). La section suivante traite de mes travaux, réalisés en collaboration avec Amandine Marrel, Béatrice Laurent et Olivier Roustant, sur l'estimation des indices de Sobol à l'aide du modèle PG.

3.3.4 Calcul des indices de Sobol

A l'aide du prédicteur du modèle PG (Eq. (3.32)), l'estimation d'indices de sensibilité tels que les indices de Sobol (Eqs. (3.17) et (3.18)) peut être réalisée extrêmement rapidement par simulations Monte Carlo (cf. par exemple Santner et al. [191]). C'est d'ailleurs l'une des méthodes usuelles pour estimer les indices de Sobol lorsque l'on a affaire à un code coûteux : construire un métamodèle que l'on utilise de manière intensive pour calculer les indices de Sobol (Marseguerra et al. [144], Iooss et al. [104], Volkova et al. [223]). La part de variance non expliquée par le métamodèle (calculée par $1 - Q_2$) nous permet de connaître ce que l'on perd en utilisant le métamodèle (Sobol [202], Jacques [106]).

Contrairement à bon nombre des métamodèles (réseau de neurones, arbres de régression, ...), la formulation du modèle PG permet également d'obtenir les indices de sensibilité sans passer par une estimation de type Monte Carlo. Deux approches ont été proposées dans la littérature : celle de Chen et al. [40] qui utilisent uniquement l'expression du prédicteur du modèle PG et celle d'Oakley & O'Hagan [161] qui utilisent le modèle PG dans sa globalité, *i.e.* le prédicteur et la structure de covariance du modèle PG conditionnellement à la base d'apprentissage. Les travaux que j'ai réalisés sur le sujet (Marrel et al. [142]) ont consisté tout d'abord à comparer les résultats de ces deux approches, ce qui n'avait pas été fait auparavant. Dans un deuxième temps, nous avons développé une méthode originale pour obtenir des intervalles de prédiction sur les estimations des indices de Sobol.

Dans cette section sur les processus gaussiens, nous avons raisonné jusqu'ici avec des variables d'entrée $\boldsymbol{x}$ déterministes. À présent, nous nous replaçons dans le cadre probabiliste de la section 3.2, où le vecteur des variables d'entrées $\boldsymbol{X}$ est supposé aléatoire, suivant une loi de probabilité G. $Y = f(\boldsymbol{X})$ est donc aussi une variable aléatoire. Le calcul d'un indice de Sobol au premier ordre consiste à estimer l'espérance conditionnelle de Y sachant une variable du vecteur $\boldsymbol{X}$:

$$\mathbb{E}(f(\boldsymbol{X})|X_i) = \int_{\chi_{-i}} f(\boldsymbol{X}) dG_{-i|i}(\boldsymbol{X}_{-i}|X_i) , \tag{3.39}$$

où $G_{-i|i}$ est la distribution conditionnelle du vecteur $\boldsymbol{X}_{-i} = (X_1, \ldots, X_{i-1}, X_{i+1}, X_d)$ sachant X_i et χ_{-i} est l'espace des valeurs possibles pour $\boldsymbol{X}_{-i}$.

On note $Y_{\text{PG}}(\boldsymbol{X}) : \mathbb{R}^d \to \mathbb{R}$ le modèle PG conditionnel construit au §3.3.2. Le modèle PG est défini par sa moyenne $\widehat{Y}(\boldsymbol{X})$ (cf. Eq. (3.29)) et sa covariance :

$$\text{Cov}^* \left[Y_{\text{PG}}(\boldsymbol{x}_1^*), Y_{\text{PG}}(\boldsymbol{x}_2^*)\right] = \sigma^2 \left(N_{\boldsymbol{\theta},\boldsymbol{p}}(\boldsymbol{x}_1^* - \boldsymbol{x}_2^*) + \tau\right) - \boldsymbol{k}(\boldsymbol{x}_1^*)^t \boldsymbol{\Sigma}_n{}^{-1} \boldsymbol{k}(\boldsymbol{x}_2^*) , \tag{3.40}$$

où Cov* est la covariance par rapport à la loi *a posteriori* de $Y(\boldsymbol{x})$. Pour les deux approches, les indices de Sobol sont donnés de la manière suivante :

- ▶ en utilisant seulement le prédicteur $\widehat{Y}(\boldsymbol{X})$, on a

$$S_i = \frac{\text{Var}_{X_i}\{\mathbb{E}_{X_1,\ldots,X_d}[\widehat{Y}(\boldsymbol{X})|X_i]\}}{\text{Var}_{X_1,\ldots,X_d}[\widehat{Y}(\boldsymbol{X})]} \text{ pour } i = 1, \ldots, d \, ; \tag{3.41}$$

- ▶ en utilisant le modèle PG, on a

$$\widetilde{S}_i = \frac{\text{Var}_{X_i}\{\mathbb{E}_{X_1,\ldots,X_d}[Y_{\text{PG}}(\boldsymbol{X})|X_i]\}}{\mathbb{E}^*\{\text{Var}_{X_1,\ldots,X_d}[Y_{\text{PG}}(\boldsymbol{X})]\}} \text{ pour } i = 1, \ldots, d \, . \tag{3.42}$$

$\widetilde{S}_i$ est donc une variable aléatoire dont la moyenne $\mu_{\widetilde{S}_i}$ est considérée comme un indice de sensibilité et la variance $\sigma^2_{\widetilde{S}_i}$ procure la précision que l'on a sur cet indice :

$$\begin{cases} \mu_{\widetilde{S}_i} = \dfrac{\mathbb{E}^*\{\mathrm{Var}_{X_i}\{\mathbb{E}_{X_1,\ldots,X_d}[Y_{\mathrm{PG}}(X)|X_i]\}\}}{\mathbb{E}^*\{\mathrm{Var}_{X_1,\ldots,X_d}[Y_{\mathrm{PG}}(\boldsymbol{X})]\}} \text{ pour } i = 1,\ldots,d\,, \\ \\ \sigma^2_{\widetilde{S}_i} = \dfrac{\mathrm{Var}^*\{\mathrm{Var}_{X_i}\{\mathbb{E}_{X_1,\ldots,X_d}[Y_{\mathrm{PG}}(\boldsymbol{X})|X_i]\}\}}{(\mathbb{E}^*\{\mathrm{Var}_{X_1,\ldots,X_d}[Y_{\mathrm{PG}}(\boldsymbol{X})]\})^2} \text{ pour } i = 1,\ldots,d\,. \end{cases} \quad (3.43)$$

Dans le cas de variables d'entrée indépendantes et pour une covariance qui s'exprime comme un produit de covariances monodimensionnelles (cf. Eq. (3.22)), les expressions analytiques des formules de S_i et $\mu_{\widetilde{S}_i}$ (Eqs. (3.41) et (3.43)) conduisent à des intégrales simples et doubles. Ces intégrations numériques sont donc moins coûteuses à évaluer que les calculs des indices par simulations Monte Carlo. Marrel [141] fournit les expressions analytiques complètes des indices de Sobol (pour S_i, $\mu_{\widetilde{S}_i}$ et $\sigma^2_{\widetilde{S}_i}$).

Sur des fonctions régulières données, nous avons étudié dans Marrel et al. [142] la vitesse de convergence des estimations des indices de Sobol par les deux approches en fonction de la prédictivité du modèle PG (coefficient Q_2). Notre conclusion est que la deuxième approche fournit des résultats plus robustes et moins variables, même quand le modèle PG est peu prédictif ($Q_2 < 80\%$). La structure de covariance du modèle PG conditionnel apporte donc une information utile pour l'estimation des indices de Sobol. Par contre, l'évaluation numérique des intégrales est délicate, ce qui ne permet pas d'envisager d'utiliser cette méthode avec des variables d'entrée corrélées (qui conduisent à des intégrations multidimensionnelles qui posent problème).

Pour la seconde approche, où l'indice de sensibilité $\widetilde{S}_i$ est une variable aléatoire, la distribution de $\widetilde{S}_i$ n'est pas disponible théoriquement mais nous avons proposé un algorithme pour la simuler afin de l'obtenir empiriquement (Marrel et al. [142]). Cette méthode procède en plusieurs étapes :

- ⋆ l'effet principal de X_i,
$$A(X_i) = \mathbb{E}_{X_1,\ldots,X_d}[Y_{\mathrm{PG}}(X)|X_i]\,, \quad (3.44)$$
est un processus gaussien dont la moyenne et la covariance peuvent être écrites explicitement. Ces dernières sont donc calculées numériquement par intégrations simples ;

- ⋆ la variance de $A(X_i)$ relative à X_i,
$$\mathrm{Var}_{X_i}[A(X_i)] = \int_{\chi_i} \left[A(X_i) - \int_{\chi_i} A(X_i) dG_i(X_i) \right]^2 dG_i(X_i) \quad (3.45)$$
(avec χ_i l'espace des valeurs possibles pour X_i et G_i la distribution de X_i), est donc une intégrale aléatoire que l'on choisit de discrétiser, ce qui nous fournit un vecteur gaussien de n_{dis} éléments. Les moyenne et matrice de covariance de ce vecteur de discrétisation peuvent être calculées à l'aide des moyenne et covariance de $A(X_i)$.

- ⋆ finalement, ce vecteur gaussien peut être simulé à l'aide de la méthode basée sur la décomposition de Cholesky de la matrice de covariance. En répétant cette opération k_{sim} fois, on obtient k_{sim} réalisations de l'intégrale aléatoire et donc de l'indice de sensibilité aléatoire $\widetilde{S}_i$. Pour déterminer si le nombre de pas de discrétisation n_{dis} et le nombre de simulations k_{sim} sont suffisants, la convergence de la moyenne et de la variance de $\widetilde{S}_i$ est étudiée empiriquement.

Pour des fonctions tests, nous avons simulé les distributions des indices de sensibilité afin d'obtenir des intervalles de prédiction à 90%. Nous avons ainsi pu comparer l'intervalle théorique aux intervalles observés. Il ressort de nos premières études que les estimations des intervalles de prédiction sont valides pour les indices relativement élevés (supérieurs à 10%) et pour des modèles PG assez prédictifs ($Q_2 > 60\%$). En pratique, un métamodèle dont le coefficient de prédictivité se situe entre 60% et

80% n'est pas considéré comme satisfaisant mais peut être utilisé à l'aide de notre démarche. En revanche, cette démarche est relativement complexe à mettre en œuvre du fait des multiples intégrales à approximer. Une application sur un code industriel a finalement permis d'illustrer l'intérêt pratique de cette nouvelle approche : la mise à disposition d'intervalles de prédiction sur les indices permet d'introduire plus de rigueur et de confiance dans la hiérarchisation des variables d'entrée que l'on peut faire par la suite. Tous ces travaux restent néanmoins à valider par des études plus complètes.

3.4 Estimation de quantiles de codes

Nous considérons à présent le problème de l'estimation des quantiles de la variable de sortie $Y \in \mathbb{R}$ d'un modèle numérique dépendant de variables d'entrée aléatoires $\boldsymbol{X} \in \mathbb{R}^d$ où d est un entier positif. Les quantiles recherchés sont de type élevés (supérieurs à 80%) et le modèle numérique est coûteux en temps de calcul. Ainsi, seul un nombre limité d'appels au code est possible (typiquement moins de $n = 200$), induisant des estimations empiriques relativement imprécises. Les résultats que je vais présenter dans cette section sont issus d'un sujet de recherche que j'ai proposé avec Agnès de Crecy et Pascal Bazin[8] lors de l'école d'été du CEMRACS (Centre d'Été Mathématique de Recherche Avancée en Calcul Scientifique) en 2006 et que j'ai traité en collaboration avec Josselin Garnier et Claire Cannamela (Cannamela et al. [36]).

Mon intérêt pour ce problème est motivé par des questions relatives à la sûreté nucléaire, pour le fonctionnement des centrales nucléaires REP (Réacteur à Eau sous Pression). Lors d'un scénario (hypothétique) d'accident APRP - GB (Accident de Perte de Réfrigérant Primaire - Grosse Brèche), il est impératif que la température de la gaine du combustible reste inférieure à la température de fusion de l'acier de gaine, afin d'éviter tout endommagement du cœur du réacteur. Pour évaluer ce risque, des codes de calcul sont utilisés pour simuler les phénomènes thermohydrauliques intervenant au cours du scénario d'accident, permettant de calculer l'évolution temporelle de la température de la gaine du combustible (Petruzzi et al. [169], Cacuci et al. [33]). L'un des critères de sûreté consiste à montrer que l'estimation du quantile à 95% du premier pic de température de gaine, associé à un niveau de confiance de 95%, est bien inférieur à la limite énoncée précédemment (Nutt & Wallis [159], Zio & Di Maio [233]). Bien entendu, ce problème d'estimation de quantiles de codes est générique et peut être rencontré dans bien d'autres problématiques, comme par exemple la conception aéronautique ou les calculs d'impact environnementaux.

Les première et deuxième sections de ce chapitre présentent les résultats connus sur l'estimation de quantiles par la méthode empirique, les statistiques d'ordre et l'utilisation d'une variable de contrôle. Les trois sections suivantes explicitent les nouveaux résultats que nous avons obtenus sur les estimateurs de quantiles par stratification contrôlée, stratification contrôlée adaptative et tirage d'importance contrôlé. Enfin, la dernière section évoque les nombreuses perspectives de recherche sur ce sujet relativement récent pour moi.

3.4.1 Quantile empirique

Mathématiquement, le problème se pose de la manière suivante. On dispose d'un n-échantillon $(Y_1, \ldots, Y_n)$ de variables aléatoires indépendantes identiquement distribuées (i.i.d.) selon une loi continue, inconnue et à densité $p(y)$. On associe à l'échantillon $(Y_1, \ldots, Y_n)$ les statistiques d'ordre $(Y_{(1)}, \ldots, Y_{(n)})$ tel que $Y_{(1)} \leq \ldots \leq Y_{(n)}$. On cherche un estimateur du α-quantile y_α défini par

$$\mathbb{P}(Y \leq y_\alpha) = \alpha \, . \tag{3.46}$$

L'estimateur classique du α-quantile est le quantile empirique

$$\widehat{Y}_{\mathrm{EE}}(\alpha) = \widehat{Y}_{\alpha,n} = Y_{(\lfloor \alpha n \rfloor + 1)} \, . \tag{3.47}$$

[8]CEA Grenoble, Direction de l'Énergie Nucléaire

où $\lfloor . \rfloor$ est la fonction partie entière. Si la densité $p(y)$ est dérivable en y_α, $\widehat{Y}_{\text{EE}}(\alpha)$ est un estimateur asymptotiquement normal (cf. par exemple David & Nagaraja [51]) :

$$\sqrt{n}(\widehat{Y}_{\text{EE}}(\alpha) - y_\alpha) \overset{n\to\infty}{\longrightarrow} \mathcal{N}(0, \sigma^2_{\text{EE}}), \quad \sigma^2_{\text{EE}} = \frac{\alpha(1-\alpha)}{p^2(y_\alpha)} . \tag{3.48}$$

La variance est donc d'autant plus grande que l'on cherche à évaluer un quantile extrême (la densité au point y_α est alors petite). Dans le contexte de sortie d'un code de calcul, cet estimateur peut être utilisé si une méthode de Monte Carlo non biaisée a été utilisée pour générer les variables d'entrée du code : $(\boldsymbol{X}^{(1)}, \dots, \boldsymbol{X}^{(n)})$ est un n-échantillon de vecteurs aléatoires i.i.d. Cependant, l'estimateur empirique n'est pas satisfaisant dans le cas de notre problème ($y_\alpha = 0.95$ et $n = 200$) où il conduit à des estimations trop imprécises, *i.e.* de variance trop élevée.

En sûreté nucléaire, on veut avoir en plus un certain niveau de confiance $\beta \in]0,1[$ sur le quantile estimé, c'est-à-dire que l'on cherche un estimateur $\widehat{Y}_{\alpha,n}$ tel que

$$\mathbb{P}(\widehat{Y}_{\alpha,n} \geq y_\alpha) \geq \beta . \tag{3.49}$$

Une solution à ce problème est donné par un théorème pour les statistiques d'ordre (David & Nagaraja [51]) qui stipule que le nombre de dépassements d'un seuil y par la suite de variables aléatoires i.i.d $(Y_1, \dots, Y_n)$ suit une loi binomiale de paramètres (n, q), avec $q = \mathbb{P}(Y > y)$. La formule que l'on obtient est connue, dans le domaine de la fiabilité, sous le nom de formule de Wilks (Wilks [228], Nutt & Wallis [159]) et est donnée ci-après.

Théorème 3.4.1 *Si on note r le plus petit entier tel que*

$$\sum_{j=0}^{n(1-\alpha)-r} C_n^j (1-\alpha)^j \alpha^{n-j} \leq 1 - \beta \tag{3.50}$$

alors $\mathbb{P}(Y_{(\lfloor \alpha n \rfloor + r)} > y_\alpha) \geq \beta$, c'est-à-dire que l'estimateur $Y_{(\lfloor \alpha n \rfloor + r)}$ est sûr au niveau β.

La facilité d'utilisation de ce théorème le rend très populaire en pratique. Par exemple, il permet de déterminer le nombre n de calculs (de type Monte Carlo) qu'il faut faire pour obtenir une estimation du quantile d'ordre α avec un niveau de confiance β, grâce à la valeur maximale $Y_{(n)}$ de l'échantillon $(Y_1, \dots, Y_n)$ des réponses du code.

L'estimateur de Wilks souffre, comme l'estimateur empirique, d'une grande dispersion. Dans la suite, on présente les méthodes que j'ai étudiées et qui permettent de réduire la variance de l'estimation du quantile.

3.4.2 Quantile par variable de contrôle

Lors des études d'incertitude des modèles numériques, il est courant de disposer, en plus du code de calcul, d'un code simplifié ou d'un modèle mathématique décrivant sommairement les phénomènes simulés dans le code de calcul. Ce modèle réduit peut aussi être un métamodèle ajusté au préalable sur un certain nombre d'évaluations bien choisies du code (cf. §3.2.5 et §3.3). Par rapport au code de calcul étudié, l'avantage de ce modèle réduit est qu'il est très peu coûteux en temps de calcul ; son inconvénient, par contre, réside dans son degré d'approximation. L'estimation directe (par Monte Carlo) d'un quantile faible ou élevé à partir d'un métamodèle diffère substantiellement du vrai quantile du code de calcul. En effet, le métamodèle est usuellement construit pour imiter le comportement moyen du code de calcul et non pour reproduire son comportement dans des zones de quantiles élevés (Oakley [160], Cannamela et al. [36]). Pour résoudre ce problème, deux stratégies peuvent être envisagées. La première consiste à construire un métamodèle adapté à l'estimation d'un quantile, par exemple en utilisant la régression quantile (Koenker [123]), la construction adaptative d'un métamodèle PG (Oakley [160]), voire les techniques de simulations conditionnelles de différentes réalisations du métamodèle PG (Rutherford [179]). La seconde, celle que nous avons étudiée, consiste à incorporer, dans les stratégies d'estimation de quantiles par Monte Carlo, une information supplémentaire basée sur l'utilisation d'un métamodèle, noté $Z = f_r(\boldsymbol{X})$.

L'estimation par variable de contrôle est une technique classique dans les méthodes de réduction de variance de Monte Carlo (Rubinstein [178]). Elle consiste à soustraire à l'estimateur empirique une fonction faisant intervenir une variable corrélée à la variable étudiée. Ici, il suffit d'utiliser comme variable de contrôle le métamodèle Z et comme fonction de contrôle $g(z) = \mathbf{1}_{z \leq z_\alpha}$, avec z_α le quantile d'ordre α de Z. On obtient alors l'estimation du quantile par variable de contrôle à partir de l'échantillon $(Y_i, Z_i)_{i=1..n}$.

Hesterberg & Nelson [85] se sont intéressés aux propriétés de l'estimateur du quantile par variable de contrôle, noté $\widehat{Y}_{\mathrm{CV}}(\alpha)$. Ils ont montré, en se basant sur les résultats de Nelson [158] concernant les propriétés des estimateurs par variable de contrôle, le théorème asymptotique suivant :

Théorème 3.4.2 *Si $\widehat{Y}_{\mathrm{CV}}(\alpha)$ est l'estimateur du quantile y_α par la méthode de la variable de contrôle, on a*

$$\sqrt{n}(\widehat{Y}_{\mathrm{CV}}(\alpha) - y_\alpha) \overset{n\to\infty}{\longrightarrow} \mathcal{N}(0, \sigma^2_{\mathrm{CV}}) \,, \quad \sigma^2_{\mathrm{CV}} = \frac{\alpha(1-\alpha)}{p^2(y_\alpha)}(1-\rho_I^2) \,, \tag{3.51}$$

où ρ_I est le coefficient de corrélation entre $\mathbf{1}_{Y \leq y_\alpha}$ et $\mathbf{1}_{Z \leq z_\alpha}$:

$$\rho_I = \frac{\mathbb{P}(Y \leq y_\alpha, Z \leq z_\alpha) - \alpha^2}{\alpha(1-\alpha)} \,. \tag{3.52}$$

Ce résultat montre une réduction de variance d'un facteur $(1-\rho_I^2)$ par rapport au quantile empirique : plus les variables aléatoires Y et Z sont corrélées à proximité du quantile recherché, meilleure est la réduction de variance. Ce résultat est intéressant car il est facile d'obtenir un estimateur de ρ_I, en calculant le coefficient de corrélation empirique (à partir de l'échantillon disponible), et donc d'avoir une idée de la réduction de variance.

3.4.3 Une méthode de rejet : la stratification contrôlée

La méthode par variable de contrôle n'utilise cependant pas toute la spécificité du métamodèle car autant de calculs sont réalisés avec le code qu'avec le métamodèle (alors que ce dernier peut être utilisé intensivement). Une autre stratégie consiste à utiliser le modèle réduit non pas pour approcher la réponse du modèle complet $Y = f(\boldsymbol{X})$ dans des configurations exceptionnelles, mais pour sélectionner un échantillon de $\boldsymbol{X}$ dans des zones intéressantes pour l'estimation du quantile. L'idée grossière est simplement de tirer un $\boldsymbol{X}$ selon sa loi originale et de calculer $f_r(\boldsymbol{X})$ par le modèle réduit. Si la réponse du modèle réduit ne nous convient pas (par exemple si elle n'est pas située dans les quantiles proches du quantile d'ordre α du modèle réduit), alors on rejette le $\boldsymbol{X}$ en question (ou plus exactement, on a tendance à le rejeter). Si la réponse du modèle réduit nous convient, alors on calcule $f(\boldsymbol{X})$. Il s'agit donc d'une méthode de rejet.

La méthode que nous avons proposée, la stratification contrôlée, consiste donc à stratifier l'espace des valeurs prises par $Z = f_r(\boldsymbol{X})$ en m intervalles $I_1, \ldots, I_m$, et à forcer le nombre de réalisations de $\boldsymbol{X}$ qui sont telles que $Z = f_r(\boldsymbol{X})$ tombe dans un intervalle I_j. Mathématiquement, on se donne $m+1$ niveaux $0 = \alpha_0 < \alpha_1 < \ldots < \alpha_m = 1$, et les quantiles de Z correspondant $-\infty = z_{\alpha_0} < z_{\alpha_1} < \ldots < z_{\alpha_m} = \infty$. Ces quantiles sont estimables avec précision sans aucun problème car la génération de réalisations Z est peu coûteuse en temps de calcul. On va utiliser les intervalles $]z_{\alpha_{j-1}}, z_{\alpha_j}]$ comme strates. On se donne une suite d'entiers $N_1, \ldots, N_m$ tels que $\sum_{j=1}^m N_j = n$. Pour chaque j, on tire (par une méthode d'acceptation-rejet) N_j réalisations des vecteurs aléatoires d'entrée $(\boldsymbol{X}^{(i)})^{(j)}_{i=1,\ldots,N_j}$ telles que les sorties $Z_i^{(j)}$ correspondantes soient dans $]z_{\alpha_{j-1}}, z_{\alpha_j}]$. Pour chacune de ces N_j réalisations, on calcule $Y_i^{(j)}$. L'estimateur du α-quantile de Y par stratification contrôlée vaut alors (Cannamela et al. [36])

$$\widehat{Y}_{\mathrm{CS}}(\alpha) = \inf\left\{y, \widehat{F}_{\mathrm{CS}}(y) > \alpha\right\} \,, \tag{3.53}$$

où $\widehat{F}_{\mathrm{CS}}(y)$ est l'estimateur par stratification contrôlée de la fonction de répartition de Y :

$$\widehat{F}_{\mathrm{CS}}(y) = \sum_{j=1}^{m} \widehat{P}_j(y)(\alpha_j - \alpha_{j-1}) \ \text{, avec } \ \forall j = 1,\ldots,m \ , \ \widehat{P}_j(y) = \frac{1}{N_j}\sum_{i=1}^{N_j} \mathbf{1}_{Y_i^{(j)} \leq y} \ . \tag{3.54}$$

$\widehat{P}_j(y)$ est l'estimateur de la probabilité conditionnelle $P_j(y) = \mathbb{P}(Y \leq y \,|\, Z \in]z_{\alpha_{j-1}}, z_{\alpha_j}])$.

Nous avons montré dans Cannamela et al. [36] le théorème asymptotique suivant :

Théorème 3.4.3 *Si $\widehat{Y}_{\mathrm{CS}}(\alpha)$ est l'estimateur du quantile y_α par la méthode de stratification contrôlée, on a*

$$\sqrt{n}(\widehat{Y}_{\mathrm{CS}}(\alpha) - y_\alpha) \overset{n\to\infty}{\longrightarrow} \mathcal{N}(0, \sigma^2_{\mathrm{CS}}) \ , \tag{3.55}$$

$$\sigma^2_{\mathrm{CS}} = \frac{\sum_{j=1}^{m} \frac{(\alpha_j - \alpha_{j-1})^2}{N_j/n}\left[P_j(y_\alpha) - P_j^2(y_\alpha)\right]}{p^2(y_\alpha)} \ . \tag{3.56}$$

La réduction de variance par rapport à celle de l'estimateur empirique peut donc être très importante si Y est Z sont fortement corrélés positivement. On a alors intérêt à mettre plus de points dans la queue de distribution de la variable aléatoire de contrôle Z, afin de renforcer le nombre de réalisations potentiellement intéressantes. Plus précisément, on peut montrer que la réduction de variance augmente avec la corrélation entre Y et Z autour du quantile cherché. Sur quelques applications (des fonctions jouées et un cas d'étude industrielle concernant un code de sûreté nucléaire), nous avons pu montrer dans Cannamela et al. [36] que l'efficacité de cette méthode dépend, au moins en partie, de la valeur de ρ_I (cf. Eq. (3.52)). Pour l'objectif d'estimation d'un quantile, il convient donc d'adopter, si possible, une stratégie particulière de construction du métamodèle.

Nous avons testé la situation avec $n = 200$ et $\alpha = 95\%$. Les trois paramètres à choisir pour pouvoir appliquer cette méthode sont le nombre m de strates, les niveaux $(\alpha_j)_{j=0..m}$ de celles-ci et les nombres $(N_j)_{j=1..m}$ de points dans chaque strate. Sur nos tests, la stratification contrôlée en quatre strates, avec $\alpha_1 = 50\%$, $\alpha_2 = 90\%$, $\alpha_3 = 95\%$ et $N_1 = N_2 = N_3 = N_4 = 50$ nous a donné des résultats satisfaisants. Elle permet de réaliser $n/2$ calculs centrés sur le quantile cherché et $n/2$ calculs ailleurs (pour détecter d'éventuelles zones intéressantes non capturées par le métamodèle). D'autres études ont montré qu'une stratégie à trois strates peut également être performante (Bazin [19]).

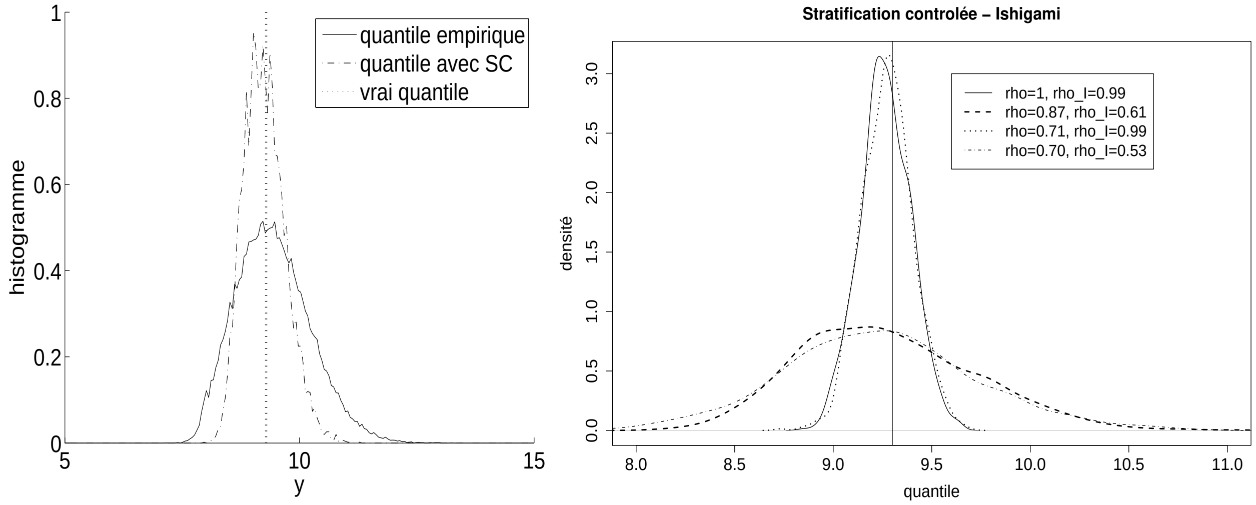
La méthode de stratification contrôlée à 4 strates est illustrée ci-dessous sur la fonction d'Ishigami $f(\cdot)$ et un métamodèle $f_r(\cdot)$ polynomial :

$$\begin{aligned} f(X_1, X_2, X_3) &= \sin(X_1) + 7\sin(X_2)^2 + 0.1X_3^4\sin(X_1) \ , & (3.57)\\ f_r(X_1, X_2, X_3) &= 1.908 + 1.727X_1 + 2.059X_2^2 - 0.276X_1^3 - 0.266X_2^4 + 0.250X_3^2X_1 \ , & (3.58)\end{aligned}$$

avec $X_j \sim \mathcal{U}[-\pi, \pi]$, $j = 1, 2, 3$. Les capacités d'approximation du métamodèle peuvent être mesurées à l'aide du coefficient de prédictivité : $Q_2 = 0.75$. Le coefficient de corrélation linéaire entre $f(\boldsymbol{X})$ et $f_r(\boldsymbol{X})$ vaut quant-à lui $\rho = 0.86$, alors que $\rho_I = 0.63$, ce qui montre une corrélation moyenne à proximité du quantile. Le quantile à 95% du métamodèle $Z = f_r(\boldsymbol{X})$ est $z_\alpha \simeq 8.51$, assez loin du quantile à 95% de $Y = f(\boldsymbol{X})$ qui est $y_\alpha \simeq 9.30$. Le métamodèle peut par contre être utilisé efficacement avec la méthode de stratification contrôlée. La figure 3.3 (a) montre que celle-ci réduit de manière significative la variance de l'estimateur du quantile, par rapport à l'estimateur empirique. Pour juger de l'influence de la qualité du métamodèle sur la variance de l'estimation du quantile obtenue par stratification contrôlée, quatre métamodèles à ρ et ρ_i variables sont utilisés. On constate sur la figure 3.3 (b) que la valeur de ρ_I influe fortement sur la qualité d'estimation par stratification contrôlée : les estimations avec un métamodèle à ρ_I élevé ont des variances nettement plus faibles que celles avec un métamodèle à ρ_I peu élevé.

3.4.4 Quantile par stratification contrôlée adaptative

Il est possible de choisir optimalement les nombres $(N_j^*)_{j=1..m}$ de points dans chaque strate en minimisant la variance (3.56). La répartition des n simulations sur les strates dépend des probabilités



(a) (b)

FIG. 3.3 – Estimations du quantile à 95% de la fonction d'Ishigami à partir d'un échantillon de taille $n = 200$. (a) Comparaison entre les estimateurs empirique et par stratification contrôlée. Les histogrammes des estimateurs sont tracés à partir de 10^4 expériences. (b) Estimations par stratification contrôlée pour quatre métamodèles différents. Les densités correspondent à un lissage des histogrammes obtenus à partir de 10^3 expériences. Le vrai quantile est donné par le trait vertical.

conditionnelles $P_j(y)$ qui sont les quantités que l'on doit estimer. Nous avons alors proposé une procédure adaptative, nommée stratification contrôlée adaptative, pour estimer les nombres de points à allouer par strate (Cannamela et al. [36]). Elle procède en plusieurs étapes :

1. estimation des probabilités conditionnelles $P_j(y)$ $(j = 1, \dots, m)$. On applique la stratification contrôlée avec $\widetilde{n} = n^\gamma$ simulations, $0 < \gamma < 1$, et avec une allocation *a priori* $\beta_j = \dfrac{N_j}{n}$. Une première estimation des probabilités conditionnelles est obtenue :

$$\widetilde{P}_j(y) = \frac{1}{[\beta_j \widetilde{n}]} \sum_{i=1}^{[\beta_j \widetilde{n}]} \mathbf{1}_{Y_i^{(j)} \leq y} \, , j = 1, \dots, m \, , \tag{3.59}$$

qui permet d'obtenir un estimateur du quantile d'ordre α :

$$\widetilde{Y}_\alpha = \inf \left\{ y, \widetilde{F}(y) > \alpha \right\} \ , \ \widetilde{F}(y) = \sum_{j=1}^{m} (\alpha_j - \alpha_{j-1}) \widetilde{P}_j(y) \ ; \tag{3.60}$$

2. estimation de l'allocation de points optimale β_j^* $(j = 1, \dots, m)$:

$$\widetilde{\beta}_j = \frac{(\alpha_j - \alpha_{j-1}) \left[\widetilde{P}_j(\widetilde{Y}_\alpha) - \widetilde{P}_j(\widetilde{Y}_\alpha)^2 \right]^{1/2}}{\sum_{l=1}^m (\alpha_l - \alpha_{l-1}) \left[\widetilde{p}_l(\widetilde{Y}_\alpha) - \widetilde{p}_l(\widetilde{Y}_\alpha)^2 \right]^{1/2}} \ ; \tag{3.61}$$

3. réalisation des $n - \widetilde{n}$ simulations finales en allouant les simulations dans chaque strate pour atteindre les nombres optimaux $[\widetilde{\beta}_j n]$, $j = 1, \dots, m$;

4. estimation du quantile $\widehat{Y}_{\mathrm{ACS}}(\alpha)$:

$$\widehat{Y}_{\mathrm{ACS}}(\alpha) = \inf \left\{ y, \widehat{F}_{\mathrm{ACS}}(y) > \alpha \right\} \tag{3.62}$$

où

$$\widehat{F}_{\mathrm{ACS}}(y)=\sum_{j=1}^{m}\widehat{P}_j(y)(\alpha_j-\alpha_{j-1}) \ \text{, avec } \ \forall j=1,\ldots,m\ ,\ \widehat{P}_j(y)=\frac{1}{[\widetilde{\beta}_j n]}\sum_{i=1}^{[\widetilde{\beta}_j n]}\mathbf{1}_{Y_i^{(j)}\le y}\ . \tag{3.63}$$

Dans Cannamela et al. [36], nous avons obtenu le théorème asymptotique suivant :

Théorème 3.4.4 *Si $\widehat{Y}_{\mathrm{ACS}}(\alpha)$ est l'estimateur du quantile y_α par la méthode de stratification contrôlée adaptative, on a*

$$\sqrt{n}(\widehat{Y}_{\mathrm{ACS}}(\alpha)-y_\alpha)\overset{n\to\infty}{\longrightarrow}\mathcal{N}(0,\sigma^2_{\mathrm{ACS}})\ , \tag{3.64}$$

$$\sigma^2_{\mathrm{ACS}}=\frac{\left\{\sum_{j=1}^{m}(\alpha_j-\alpha_{j-1})\left[P_j(y_\alpha)-P_j^2(y_\alpha)\right]^{\frac{1}{2}}\right\}^2}{p^2(y_\alpha)}\ . \tag{3.65}$$

De premiers tests ont permis de voir qu'il faut que n soit suffisamment élevé pour que cette méthode adaptative soit efficace (en permettant notamment que la première étape soit réellement utile).

3.4.5 Quantile par tirage d'importance contrôlé

L'estimation par tirage d'importance est une autre méthode bien connue pour la réduction de variance de Monte Carlo (Rubinstein [178]). La méthode par tirage d'importance contrôlé, que nous avons proposée dans Cannamela et al. [36], consiste à estimer la densité biaisée pour le tirage d'importance par simulations intensives sur le métamodèle Z, à échantillonner les entrées $\boldsymbol{X}$ selon la densité biaisée, à produire les sorties du code $Y=f(\boldsymbol{X})$ sur cet échantillon, puis à calculer l'estimateur non biaisé du quantile.

La stratégie de tirage d'importance contrôlé pour estimer un quantile consiste à chercher une densité d'importance correcte pour le calcul de l'intégrale suivante :

$$\mathbb{E}\left[\mathbf{1}_{f_r(\boldsymbol{X})\le z_\alpha}\right]=\int \mathbf{1}_{f_r(\boldsymbol{x})\le z_\alpha}q_{\mathrm{ori}}(\boldsymbol{x})d\boldsymbol{x}=\alpha\ , \tag{3.66}$$

où q_{ori} est la densité initiale de $\boldsymbol{X}$. Si les α-quantiles de $f(\cdot)$ et de $f_r(\cdot)$ ne sont pas trop éloignés, trouver la densité q qui minimise la variance de l'estimateur $\displaystyle\frac{1}{n}\sum_{i=1}^{n}\mathbf{1}_{f_r(\boldsymbol{X}^{(i)})\le z_\alpha}\frac{q_{\mathrm{ori}}(\boldsymbol{X}^{(i)})}{q(\boldsymbol{X}^{(i)})}$ permet de s'approcher des régions d'importance de notre code numérique $f(\boldsymbol{X})$. La variance est minimale pour la densité optimale (Rubinstein [178])

$$q^*(\boldsymbol{x})=\frac{\mathbf{1}_{f_r(\boldsymbol{x})\le z_\alpha}q_{\mathrm{ori}}(\boldsymbol{x})}{\int \mathbf{1}_{f_r(\boldsymbol{x}')\le z_\alpha}q_{\mathrm{ori}}(\boldsymbol{x}')d\boldsymbol{x}'}\ . \tag{3.67}$$

On recherche une densité d'importance parmi une famille paramétrique $\mathcal{Q}$ de densités q_γ paramétrées par leurs deux premiers moments $\gamma=(\lambda,C)$. Par simulations intensives sur le métamodèle (échantillon de taille $\widetilde{n}$), on obtient un estimateur $\widehat{\gamma}=(\widehat{\lambda},\widehat{C})$ des paramètres de la densité optimale :

$$\left\{\begin{array}{l}\widehat{\lambda}=\dfrac{\sum_{i=1}^{\widetilde{n}}\boldsymbol{X}^{(i)}\mathbf{1}_{Z_i\le z_\alpha}q_{\mathrm{ori}}(\boldsymbol{X}^{(i)})/q_0(\boldsymbol{X}^{(i)})}{\sum_{i=1}^{\widetilde{n}}\mathbf{1}_{Z_i\le z_\alpha}q_{\mathrm{ori}}(\boldsymbol{X}^{(i)})/q_0(\boldsymbol{X}^{(i)})}\ ,\\[3ex] \widehat{C}=\dfrac{\sum_{i=1}^{\widetilde{n}}\boldsymbol{X}^{(i)}(\boldsymbol{X}^{(i)})^t\mathbf{1}_{Z_i\le z_\alpha}q_{\mathrm{ori}}(\boldsymbol{X}^{(i)})/q_0(\boldsymbol{X}^{(i)})}{\sum_{i=1}^{\widetilde{n}}\mathbf{1}_{Z_i\le z_\alpha}q_{\mathrm{ori}}(\boldsymbol{X}^{(i)})/q_0(\boldsymbol{X}^{(i)})}-\widehat{\lambda}\widehat{\lambda}^t\ ,\end{array}\right. \tag{3.68}$$

où $\boldsymbol{X}^{(i)}\sim q_0$ $(i=1,\ldots,\widetilde{n})$, q_0 étant une densité de probabilité choisie *a priori* ($q_0=q_{\mathrm{ori}}$ en absence d'information *a priori*).

L'estimateur du α-quantile de Y par tirage d'importance contrôlé vaut alors :

$$\widehat{Y}_{\mathrm{CIS}}(\alpha) = \inf\{y, \widehat{F}_{\mathrm{IS}}(y) > \alpha\} \ , \ \widehat{F}_{\mathrm{IS}}(y) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{f(\boldsymbol{X}^{(i)})\leq y} \frac{q_{\mathrm{ori}}(\boldsymbol{X}^{(i)})}{q(\boldsymbol{X}^{(i)})} \ . \tag{3.69}$$

Nous avons montré dans Cannamela et al. [36] le théorème asymptotique suivant :

Théorème 3.4.5 *Si $\widehat{Y}_{\mathrm{CIS}}(\alpha)$ est l'estimateur du quantile y_α par la méthode de tirage d'importance contrôlé, on a*

$$\sqrt{n}(\widehat{Y}_{\mathrm{CIS}}(\alpha) - y_\alpha) \overset{n\to\infty}{\longrightarrow} \mathcal{N}(0, \sigma^2_{\mathrm{CIS}}) \,, \tag{3.70}$$

$$\sigma^2_{\mathrm{CIS}} = \frac{1}{p^2(y_\alpha)}\left(\int \frac{\mathbf{1}_{f(\boldsymbol{x})\leq y_\alpha} q_{\mathrm{ori}}(\boldsymbol{x})^2}{q_{\gamma_r^*}(\boldsymbol{x})} d\boldsymbol{x} - \alpha^2\right) \ , \tag{3.71}$$

où $\gamma_r^* = \lim_{\tilde{n}\to\infty} \widehat{\gamma}$.

Sur certains tests joués, cette méthode a donné d'excellents résultats, parfois meilleurs que ceux des autres méthodes. Par exemple, avec les fonctions $f(\cdot)$ et $f_r(\cdot)$ données par

$$f(X_1, X_2) = 0.95|X_1|X_1\left[1 + \frac{1}{2}\cos(10X_1) + \frac{1}{2}\cos(20X_1)\right] + \frac{7}{10}X_2\left[1 + \frac{3}{5}\cos(X_2) + \frac{3}{10}\cos(14X_2)\right] \ , \tag{3.72}$$

$$f_r(X) = |X_1|X_1 + X_2 \ , \tag{3.73}$$

avec $X_1 \sim \mathcal{N}(0,1)$ et $X_2 \sim \mathcal{N}(0,1)$. La figure 3.4 (a) illustre les densités de $Y = f(\boldsymbol{X})$ et $Z = f_r(\boldsymbol{X})$. Le coefficient de corrélation linéaire entre $f(\boldsymbol{X})$ et $f_r(\boldsymbol{X})$ vaut $\rho = 0.90$, alors que $\rho_I = 0.64$, ce qui montre une corrélation moyenne à proximité du quantile. Le quantile à 95% de $Y = f(\boldsymbol{X})$ est estimé par simulations intensives à $y_\alpha \simeq 2.75$. L'estimateur empirique et l'estimateur par tirage d'importance contrôlé du quantile à 95% de Y, en utilisant $n = 200$ simulations, sont comparés sur la figure 3.4 (b). Pour la densité d'importance, la famille $\mathcal{Q}$ choisie est un ensemble de gaussiennes bidimensionnelles paramétrées par leur moyenne et covariance. La figure 3.4 (b) montre également que les résultats obtenus par tirage d'importance contrôlé sont meilleurs que ceux obtenus par variable de contrôle et par stratification contrôlée.

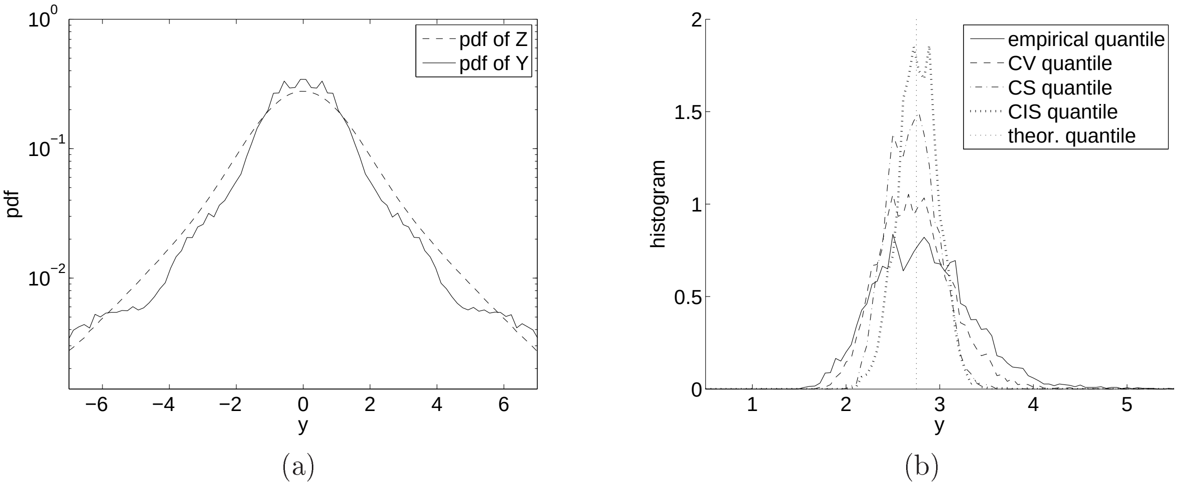


FIG. 3.4 – Étude avec les fonctions (3.72) et (3.73). (a) Densités de Y et Z. (b) Estimations du quantile à 95% de Y à partir d'un échantillon de taille $n = 200$. Comparaisons entre les estimateurs empirique (moyenne 2.83, écart-type 0.52), par variable de contrôle (moyenne 2.74, écart-type 0.38), par stratification contrôlée (moyenne 2.71, écart-type 0.25), et par tirage d'importance contrôlé (moyenne 2.77, écart-type 0.21). Les histogrammes des estimateurs sont tracés à partir de 5000 expériences.

La méthode du tirage d'importance contrôlé souffre cependant de la paramétrisation de la densité d'importance qui limite son applicabilité à l'existence d'une seule région d'importance pour chaque variable d'entrée. Cela signifie que le code de calcul ne doit atteindre les valeurs du quantile recherché que dans un domaine restreint de variation de ses entrées. L'utilisation de mélanges de densité pour la densité d'importance serait une piste intéressante pour remédier à ce problème.

3.4.6 Perspectives

Les méthodes présentées dans cette section supposent la disponibilité d'un métamodèle. Elles ne nécessitent pas que le métamodèle soit une excellente approximation du code de calcul ; sur nos tests, des approximations assez grossières ont donné de bons résultats. Ceci vient du fait que le quantile est estimé à l'aide de simulations sur le code de calcul, le métamodèle guidant juste la planification de ces calculs. L'une des voies de recherche futures serait d'étudier en détail les stratégies d'allocation de calculs entre la construction du métamodèle et l'estimation du quantile. De premières études en ce sens ont été ralisées par Bazin [19]. La stratification contrôlée adaptative pourrait également bénéficier d'une réestimation du métamodèle à l'issue de la première étape. Ces méthodes permettent également d'envisager l'utilisation de codes de calcul simplifiés, par exemple à maillage plus grossier que le code de calcul initial, qui sont souvent disponibles dans les applications industrielles.

Par ailleurs, pour estimer les quantiles de codes, l'utilisation d'un métamodèle tel que le modèle processus gaussien (modèle PG, cf. §3.3) semble assez naturel. L'utilisation de la variance du modèle PG permet d'élaborer des stratégies de planification adaptative des calculs en privilégiant progressivement les simulations du code de calcul dans la région d'intérêt (Oakley [160], Vazquez & Piera-Martinez [220]). La moyenne et la covariance du modèle PG étant connues (Eqs. (3.29) et (3.40)), il est alors aisé de simuler des réalisations du modèle PG et d'estimer un quantile sur chacune de ces réalisations (Oakley [160], Rutherford [179]). Au final, on obtient un intervalle de confiance sur le quantile recherché. Sur des fonctions tests à faible nombre d'entrées, cette approche semble extrèmement efficace. Bien entendu, en plus grande dimension, la validation du modèle PG (prédicteur et covariance) doit être particulièrement soignée, car les quantiles estimés par cette méthode dépendent entièrement du métamodèle, et donc des paramètres estimés de la covariance. L'un de mes sujets de recherche futurs sera de comparer sur des cas concrets, de taille industrielle, les avantages et inconvénients de l'estimation de quantiles entre les méthodes par Monte Carlo contrôlé (variable de contrôle, stratification contrôlée, stratification contrôlée adaptative, tirage d'importance contrôlé) et par le modèle PG.

3.5 Le cas des modèles numériques stochastiques

Dans tous les paragraphes précédents, on s'est intéressé au cadre des codes de calcul déterministes, gouvernés par l'équation (3.1), pour lesquels deux calculs avec un même jeu de variables d'entrée procurent la même réponse. Dans des travaux récents, publiés dans des actes de conférence (Iooss & Ribatet [99, 100]) et prévus pour des revues (Iooss et al. [102], Iooss & Ribatet [101]), je me suis intéressé au cadre des codes de calcul stochastiques, c'est-à-dire aux codes qui possèdent une variabilité non contrôlée. Le modèle pour les codes de calculs stochastiques peut se formuler de la manière suivante :

$$\begin{array}{rll} g: & \mathbb{R}^d & \to \mathbb{R} \\ & \boldsymbol{X} & \mapsto Y = f(\boldsymbol{X}) + \nu \,, \end{array} \tag{3.74}$$

où $\boldsymbol{X}$ est un vecteur aléatoire de d variables d'entrée contrôlables, Y est la réponse du code, $f(\cdot)$ est la partie déterministe du modèle et $\nu \in \mathbb{R}$ est la partie stochastique du modèle (le bruit). Pour signifier que ν dépend d'un (ou de plusieurs) paramètre(s) incontrôlable(s) ε et de $\boldsymbol{X}$ (uniquement à travers ses interactions avec ε), on écrit

$$\nu = \nu(\varepsilon, \mathbf{X} : \varepsilon) \,, \tag{3.75}$$

où ν est supposé centré relativement à ε : $\mathbb{E}_\varepsilon(\nu) = 0$. ε représente l'origine du bruit, par exemple une instabilité numérique (difficile à paramétrer), ou alors un paramètre scalaire réel inhérent au code de calcul et qu'on ne sait pas contrôler (d'où la dénomination "paramètre incontrôlable"), ou alors une variable du modèle que l'on ne sait pas représenter par un nombre raisonnable de paramètres scalaires (comme un champ aléatoire). Dans ce dernier cas, ε pourrait être le germe (qui est un nombre entier) de la simulation du champ aléatoire dont on a besoin à chaque appel du code, et serait donc pleinement contrôlable. Cependant, une petite incrémentation d'un germe conduit à une simulation du champ aléatoire complètement différente, et donc à une réponse du code probablement très différente. La réponse n'est donc pas continue par rapport au germe et on devrait plutôt parler dans ce cas de paramètre à effet chaotique. Pour simplifier, les différentes sources potentielles de bruit sont rangées sous l'appellation générique "paramètres incontrôlables".

Avec le modèle (3.74), différents calculs du code sur un même jeu de variables d'entrée $\boldsymbol{X}$ procurent donc des réponses différentes. Ce cas de figure est rencontré dans des codes qui sont basés sur des équations différentielles stochastiques, qui sont sujets à du bruit numérique non négligeable, ou qui ont des processus stochastiques ou des champs aléatoires comme variables d'entrée. On peut citer par exemple les modèles de file d'attente (Kleijnen [116, 117]), la simulation du transport des neutrons (Hammersley & Handscomb [81]), la résolution d'équations dans des milieux hétérogènes simulés par des techniques géostatistiques (écoulements dans les réservoirs pétroliers, Zabalza-Mezghani [232] et calculs d'impact sur des modèles d'occupation de sols, Tarantola et al. [210]).

Dans mes travaux relatifs aux codes de calcul stochastiques, je me suis intéressé à deux problématiques complémentaires :

- la construction d'un métamodèle adapté au caractère hétéroscédastique de la fonction $g(\cdot)$ (Eq. (3.74)) et pouvant également modéliser sa partie stochastique ;
- l'analyse de sensibilité du modèle $g(\cdot)$, afin d'obtenir non seulement des indices de sensibilité pour les variables d'entrée contrôlables, mais aussi pour les paramètres d'entrée incontrôlables.

Ces deux aspects du problème font l'objet des deux sections suivantes. La dernière section explique comment ces résultats permettent d'aborder le problème des modèles avec des entrées fonctionnelles.

3.5.1 Modélisation jointe

Ajuster un modèle hétéroscédastique sur $Y = g(\boldsymbol{X})$ revient à modéliser les espérance et variance conditionnelles $Y_m(\boldsymbol{X}) = \mathbb{E}(Y|\boldsymbol{X})$ et $Y_d(\boldsymbol{X}) = \mathrm{Var}(Y|\boldsymbol{X})$. Cela consiste à construire deux métamodèles, $Y_m(\boldsymbol{X})$ pour la moyenne et $Y_d(\boldsymbol{X})$ pour la dispersion, que l'on substituera au code de calcul pour réaliser les analyses d'incertitude et de sensibilité par Monte Carlo. La modélisation de l'espérance conditionnelle est un problème classique en statistique, alors que la modélisation de la variance conditionnelle demeure un problème relativement difficile (Antoniadis & Lavergne [8]). Il faut garder à l'esprit la dimension potentiellement importante de $\boldsymbol{X}$ (plusieurs dizaines d'entrées contrôlables dans certaines applications) qui nous restreint sur l'utilisation de certaines méthodes non paramétriques (polynômes locaux par exemple) pourtant bien adaptées à la modélisation des deux composantes (Da Veiga et al. [50]).

Pour construire des métamodèles sur des codes de calcul stochastiques, Zabalza-Mezghani [232] a proposé d'utiliser le formalisme des Modèles Linéaires Généralisés (GLM) joints, développé pour modéliser des données expérimentales par Pregibon [170] et McCullagh & Nelder [149]. Le GLM joint consiste à modéliser la moyenne et la dispersion de données par deux GLM interdépendants. La classe des modèles GLM permet d'étendre la classe des modèles linéaires classiques par l'utilisation d'une distribution appartenant à la famille exponentielle et par l'utilisation d'une fonction lien pour relier les variables explicatives à la variable observée (Nelder & Wedderburn [157]). La première composante du modèle joint porte sur la moyenne ($i = 1, \dots, n$) :

$$\begin{cases} \mathrm{E}(Y_i) & = \mu_i, \quad \eta_i = h(\mu_i) = \sum_{j=1}^{d} \beta_j x_j^{(i)} \ , \\ \mathrm{Var}(Y_i) & = \phi_i v(\mu_i) \ , \end{cases} \tag{3.76}$$

où les observations $(Y_i)_{i=1..n}$ sont supposées indépendantes de moyenne μ_i, $x_j^{(i)}$ sont les observations du paramètre X_j, $(\beta_j)_{j=1..d}$ sont les paramètres de régression à estimer, η_i est le prédicteur linéaire de la moyenne, $h(\cdot)$ est la fonction lien monotone différentiable (*e.g.* identité, racine carrée), ϕ_i est un paramètre de dispersion et $v(\cdot)$ est la fonction variance (*e.g.* constante, identité, carrée). Dans le modèle joint, le paramètre de dispersion ϕ n'est pas supposé constant comme dans un GLM classique, mais au contraire est supposé varier selon le modèle :

$$\begin{cases} \mathrm{E}(d_i) & = \phi_i, \quad \zeta_i = \log(\phi_i) = \displaystyle\sum_{j=1}^{d} \gamma_j x_j^{(i)} \ , \\ \mathrm{Var}(d_i) & = 2\phi_i^2 \ , \end{cases} \tag{3.77}$$

où $(\gamma_j)_{j=1..d}$ sont les paramètres à estimer, ζ_i est le prédicteur linéaire pour la dispersion et d_i est la contribution à la déviance de l'observation Y_i. La déviance $D_{\boldsymbol{Y}^n}$ est définie par la quantité

$$D_{\boldsymbol{Y}^n} = 2\phi[l_{\boldsymbol{Y}^n}(\widehat{\boldsymbol{\beta}}_{\max}) - l_{\boldsymbol{Y}^n}(\widehat{\boldsymbol{\beta}})] \ , \tag{3.78}$$

où $l_{\boldsymbol{Y}^n}$ est la log-vraisemblance des observations $\boldsymbol{Y}^n$ (échantillon $(Y_i)_{i=1..n}$), $\widehat{\boldsymbol{\beta}}_{\max}$ et $\widehat{\boldsymbol{\beta}}$ étant respectivement les estimations par maximum de vraisemblance du vecteur de paramètres sous le modèle saturé[9] et sous le modèle considéré. Il est possible dans ce modèle que les variables explicatives de la dispersion diffèrent de celles de la moyenne.

Le modèle joint est ajusté par maximisation de la quasi-vraisemblance étendue qui se comporte comme une log-vraisemblance pour les paramètres de la moyenne et ceux de la dispersion (Nelder & Pregibon [156]). Pour ajuster un modèle joint à des données, il convient donc d'utiliser une procédure itérative : ajustement d'un GLM sur la moyenne, estimation de $(d_i)_{i=1..n}$, estimation de $(\phi_i)_{i=1..n}$ qui à son tour nous donne les poids pour la prochaine estimation du GLM sur la moyenne. Ce processus doit être itéré suffisamment de fois afin d'ajuster entièrement le modèle joint à l'aide d'un critère d'arrêt. La convergence de cet algorithme n'étant pas garantie, un nombre maximum d'itérations doit être défini. Enfin, les outils de diagnostic disponibles dans l'ajustement d'un GLM (analyse de déviance, tests de Student, analyse des résidus) sont ici disponibles pour chaque composante du GLM (moyenne (3.76) et dispersion (3.77)), ce qui permet de simplifier les expressions des composantes par sélection des termes les plus significatifs.

Remarque 3.5.1 Dans le domaine de la planification robuste, la modélisation de la moyenne et du bruit d'une réponse en fonction de variables de contrôle a été introduite par les travaux de Taguchi (Bursztyn & Steinberg [29]). Elle est aussi connue sous le nom de modélisation duale (Vining & Myers [222]). Celle-ci consiste à construire séparément des modèles polynomiaux (sous l'hypothèse du modèle linéaire gaussien) pour la moyenne et la variance et nécessite de répéter des calculs avec les mêmes jeux de paramètres contrôlables (ce qui n'est pas nécessaire dans la modélisation jointe). Sur quelques exemples, Zabalza-Mezghani [232] et Lee & Nelder [131] ont montré que ce modèle dual est moins intéressant que le modèle joint : il requiert plus de calculs, le modèle de la variance est moins bien estimé (toutes les données sont utilisés pour modéliser la variance dans le modèle joint) et le modèle gaussien est plus restrictif que le cadre des GLM.

Dans le domaine des expériences numériques, la non linéarité des réponses peut être importante et les interactions complexes entre variables d'entrée sont fréquentes. Le cadre paramétrique des GLM induit donc des limitations pour modéliser les réponses de codes de calcul. Dans Iooss et al. [102], j'ai proposé d'utiliser le cadre non paramétrique des Modèles Additifs Généralisés (GAM) pour la modélisation jointe de la moyenne et de la dispersion. Introduits par Hastie & Tibshirani [82], le principe du GAM consiste à remplacer le terme linéaire dans le prédicteur $\eta = \sum_j \beta_j X_j$ de l'équation (3.76) par une somme de fonctions de lissage $\eta = \sum_j s_j(X_j)$. Ces fonctions $s_j(\cdot)$ sont souvent choisies

[9] modèle qui a autant d'inconnues que d'observations (appelé aussi modèle complet) et qui ne cherche donc pas à "synthétiser" l'information. $l_{\boldsymbol{Y}^n}(\widehat{\boldsymbol{\beta}}_{\max})$ fournit donc la plus grande valeur possible de la log-vraisemblance.

dans la famille des splines de lissage (Wahba [225]). Pour des problèmes de régression, il est nécessaire de contrôler les éventuelles problèmes de surapprentissage et on utilise les splines de régression pénalisées (cf. Wood [230] pour une discussion complète sur le sujet). Cette approche permet d'automatiser l'ajustement des splines, réalisée par maximisation de la vraisemblance pénalisée, grâce au critère de validation croisée généralisée. Les GAM fournissent, comme les GLM, des outils de diagnostic pour sélectionner les termes les plus influents. Il est également possible de coupler une partie paramétrique de type GLM à une partie non paramétrique de type splines dans le prédicteur, ainsi que d'introduire des termes de spline bidimensionnelle $s_{ij}(X_i, X_j)$ pour prendre en compte des interactions complexes.

L'extension du GAM au GAM joint se fait naturellement comme l'extension du GLM au GLM joint. Cette idée a été introduite par Rigby & Stasinopoulos [176] qui modélisent la moyenne et la dispersion de données à l'aide de modèles additifs semi-paramétriques (Hastie & Tibshirani [82]). Ce modèle, restreint aux observations gaussiennes, est appelé MADAM (Mean and Dispersion Additive Model). Notre modèle, le GAM joint, peut donc être vu comme une généralisation du MADAM. La procédure d'ajustement des deux composantes du GAM joint est similaire à celle du MADAM et du GLM joint. Pour le GAM joint, on utilise une procédure itérative de maximisation de la quasi-vraisemblance étendue pénalisée, jusqu'à ce que celle-ci devienne stable.

Sur mes applications, le GAM joint s'est révélé satisfaisant pour des dimensions pas trop élevées ($d < 10$). En effet, il devient difficile de trouver les interactions actives en testant toutes les splines d'ordre deux au delà de cette dimension. L'utilisation de réseaux de neurones pour les composantes moyenne et dispersion du modèle joint a été proposée par Juutilainen & Röning [112] pour une application avec 10000 observations et 25 variables d'entrée. Ceux-ci ont donné des résultats plus performants que les modèles GLM joint (trop simple), polynômes locaux (trop coûteux à construire) et MADAM (qui requiert trop d'espace mémoire). Les réseaux de neurones sont bien adaptés pour modéliser des phénomènes fortement non linéaires quand beaucoup d'observations sont disponibles. Sur mes applications, ce modèle joint basé sur les réseaux de neurones ne semble pas intéressant car peu d'observations sont disponibles. De futurs travaux devraient s'atteler à la construction d'un modèle joint basé sur les processus gaussiens (pour les composantes moyenne et dispersion). Le modèle PG est en effet intéressant quand on dispose de peu d'observations et ne nécessite pas une procédure de sélection manuelle des termes d'interaction lorsque la dimension est importante (comme le GAM).

3.5.2 Indices de Sobol pour modèles joints

Une fois le modèle joint obtenu (composantes $Y_m(\boldsymbol{X})$ et $Y_d(\boldsymbol{X})$), il est aisé de réaliser une analyse de sensibilité par les techniques de décomposition de la variance (cf. §3.2.4). De manière générale, on peut décomposer la variance de la variable de sortie Y par

$$\mathrm{Var}(Y) = \mathrm{Var}[\mathrm{E}(Y|\boldsymbol{X})] + \mathrm{E}[\mathrm{Var}(Y|\boldsymbol{X})] = \mathrm{Var}[Y_m(\boldsymbol{X})] + \mathrm{E}[Y_d(\boldsymbol{X})] \ . \tag{3.79}$$

En utilisant la décomposition de la variance fonctionnelle (3.16) de la réponse Y et celle de la réponse Y_m, on montre alors immédiatement (Iooss et al. [102]) que les indices de sensibilité de Y par rapport aux variables d'entrée contrôlables $\boldsymbol{X} = (X_i)_{i=1\ldots d}$ valent

$$S_i = \frac{\mathrm{Var}[\mathbb{E}(Y_m(\boldsymbol{X})|X_i)]}{\mathrm{Var}(Y)}, \quad S_{ij} = \frac{\mathrm{Var}[\mathbb{E}(Y_m(\boldsymbol{X})|X_i X_j)]}{\mathrm{Var}(Y)} - S_i - S_j \ , \quad \ldots \tag{3.80}$$

et peuvent être calculés soit de manière analytique si la forme du métamodèle $Y_m(\boldsymbol{X})$ s'y prête, soit par simulations Monte Carlo sur $Y_m(\boldsymbol{X})$ (cf. §3.2.4).

On a donc estimé tous les termes contenus dans $\mathrm{Var}[Y_m(\boldsymbol{X})]$ de l'équation (3.79) Ainsi, l'indice de sensibilité total de Y par rapport au paramètre incontrôlable ε correspond à ce qu'il reste dans l'équation (3.79), c'est-à-dire à l'espérance de la partie stochastique du modèle joint :

$$S_{T_\varepsilon} = \frac{\mathrm{E}[Y_d(\boldsymbol{X})]}{\mathrm{Var}(Y)} \ . \tag{3.81}$$

Cet indice est positif car $Y_d(\boldsymbol{X})$ est une exponentielle (une fonction lien logarithmique a été prise dans le modèle de dispersion (3.77)). Il faut noter que si on estime le coefficient de prédictivité Q_2 de la composante moyenne $Y_m(\boldsymbol{X})$ par rapport au code stochastique $g(\cdot)$, on a de manière immédiate la relation

$$S_{T_\varepsilon} = 1 - Q_2 \; . \tag{3.82}$$

Cette formule nous donne une autre manière de calculer cet indice de sensibilité total qui ne nécessite pas forcément l'estimation de la composante dispersion $Y_d(\boldsymbol{X})$.

S'il y a plusieurs paramètres incontrôlables, notons que cette méthode regroupe leurs indices totaux dans un seul indice, et ne permet donc pas de faire de distinction entre différents paramètres incontrôlables. De plus, il n'est pas possible de distinguer quantitativement les différentes contributions dans S_{T_ε} (*i.e.* S_ε, $(S_{i\varepsilon})_{i=1..d}$, $(S_{ij\varepsilon})_{i,j=1..d}$, ...). Cependant, l'analyse des termes dans le modèle Y_d et de leur t-value permet d'avoir des contributions qualitatives. Par exemple, si X_i n'intervient pas dans Y_d, alors on sait que son interaction avec le paramètre incontrôlable est nulle ($S_{i\varepsilon} = 0$). Ceci justifie pleinement l'intérêt de modéliser la composante dispersion et donc d'utiliser un modèle joint. Des réflexions sont en cours pour proposer une méthode de quantification plus précise des indices de sensibilité des interactions.

3.5.3 Application aux modèles à entrée fonctionnelle

Les résultats ci-dessus (Eqs. (3.80) et (3.81)) sont relativement simples mais nouveaux. Iooss & Ribatet [101] les illustrent sur la fonction test nommée WN-Ishigami, qui fait intervenir un paramètre d'entrée fonctionnel que l'on traite comme un paramètre d'entrée incontrôlable :

$$Y = f(X_1, X_2, \varepsilon(t)) = \sin(X_1) + 7\sin(X_2)^2 + 0.1[\max_t(\varepsilon(t))]^4 \sin(X_1) \tag{3.83}$$

où $X_i \sim \mathcal{U}[-\pi;\pi]$ pour $i = 1, 2$ et $\varepsilon(t)$ est un bruit-blanc (un processus stochastique i.i.d gaussien de loi $\mathcal{N}(0,1)$), discrétisé sur 100 pas de temps. Il représente le paramètre d'entrée fonctionnel dont on souhaite estimer l'indice de Sobol total. Pour ajuster les modèles joints, un échantillon aléatoire $(X_1, X_2, \varepsilon(t))$ de taille $n = 500$ est généré, et le vecteur de sorties correspondantes est calculé à l'aide de la fonction (3.83). Les résultats obtenus sont les suivants :

- ⋆ un GLM joint (de type polynômial de degré 4) est construit en utilisant les outils d'analyse de déviance et de tests de Student pour conserver uniquement les termes explicatifs dans les composantes moyenne et dispersion. La composante moyenne obtenue s'écrit :

  $$Y_m = 1.77 + 4.75X_1 + 1.99X_2^2 - 0.51X_1^3 - 0.26X_2^4 \; . \tag{3.84}$$

  La déviance expliquée de ce modèle vaut $D_{expl} = 73\%$ et son coefficient de prédictivité (par rapport à la fonction WN-Ishigami) calculé sur un échantillon test vaut $Q_2 = 70\%$. Pour la composante dispersion, aucun terme significatif ne peut être retenu et une simple constante demeure :

  $$\log(Y_d) = 1.97 \; . \tag{3.85}$$

  Le GLM joint se résume donc à un GLM simple et le caractère hétéroscédastique de la fonction WN-Ishigami (dû à l'interaction entre $\varepsilon(t)$ et X_1) n'est pas retrouvé ;

- ⋆ un GAM joint est construit en utilisant les tests de Student (pour la partie paramétrique) et les statistiques de Fisher (pour la partie non paramétrique) pour conserver uniquement les termes explicatifs :

  $$\begin{aligned} Y_m &= 3.76 - 5.54X_1 + s_1(X_1) + s_2(X_2) \; , \\ \log(Y_d) &= 1.05 + s_{d1}(X_1) \; , \end{aligned} \tag{3.86}$$

  où $s_1(\cdot)$, $s_2(\cdot)$ et $s_{d1}(\cdot)$ sont des termes de splines de régression. La déviance expliquée de ce modèle vaut $D_{expl} = 92\%$ et son coefficient de prédictivité (par rapport à la fonction

WN-Ishigami) calculé sur un échantillon test vaut $Q_2 = 77\%$, ce qui montre une nette amélioration par rapport au GLM joint. De plus, la flexibilité du GAM a permis de détecter X_1 comme variable explicative de la composante dispersion, ce qui correspond bien à l'interaction entre X_1 et le paramètre fonctionnel $\varepsilon(t)$ dans la fonction WN-Ishigami.

Les deux modèles joints sont alors utilisés pour estimer les indices de Sobol des variables d'entrée (Eqs (3.80) et (3.81)) par Monte Carlo (cf. §3.2.4). La Figure 3.5 donne les résultats de ces estimations. Différents échantillons d'apprentissage pour le modèle joint sont simulés afin d'obtenir la dispersion des estimations des indices correspondant à la modélisation par modèles joints. On note que pour le GAM joint, l'intervalle interquartile de chaque boxplot contient la valeur de référence, ce qui n'est pas le cas pour le GLM joint. Ceci montre l'intérêt du GAM joint par rapport au GLM joint, mais plus généralement d'un modèle non paramétrique par rapport à un modèle paramétrique pour ajuster des fonctions chahutées. Finalement, les figures du bas montrent que l'estimation de S_{T_ε} à l'aide du coefficient de prédictivité Q_2 de Y_m (figure de droite) est nettement plus précise qu'à l'aide de l'espérance de la composante dispersion (figure de gauche). Ceci illustre la meilleure précision que l'on a sur l'estimation de la composante moyenne que sur l'estimation de la composante dispersion.

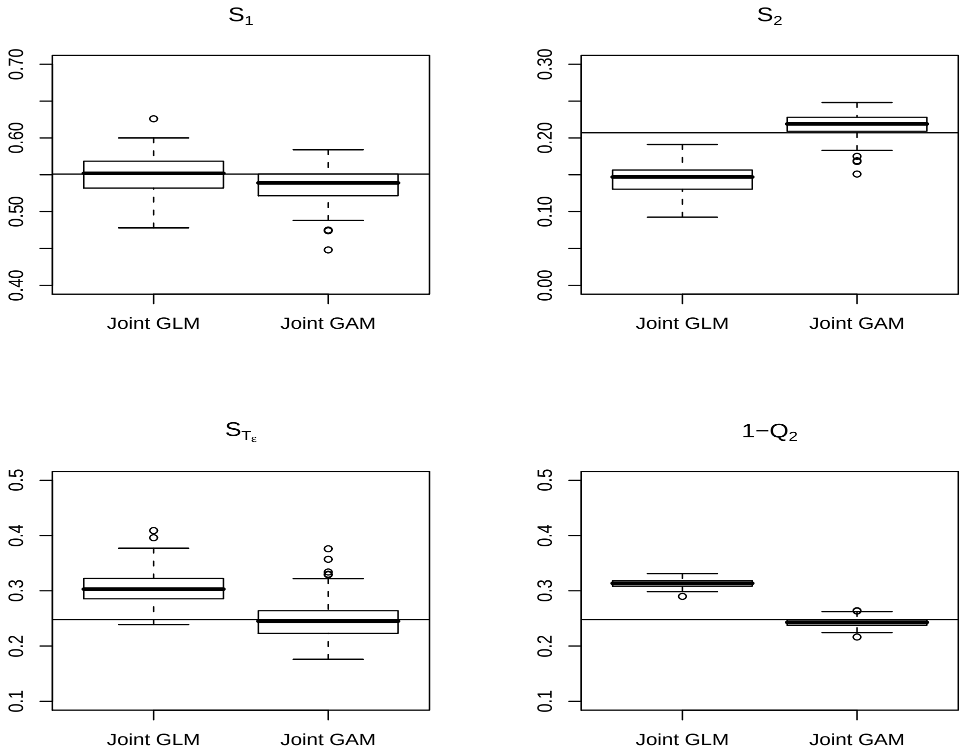


FIG. 3.5 – Boxplots des estimations d'indices de Sobol à l'aide des GLM joint et GAM joint pour la fonction WN-Ishigami (taille de la base d'apprentissage $n = 500$), obtenues à l'aide de 100 bases d'apprentissage différentes. Pour chaque indice, la ligne horizontale est la valeur de référence calculée directement sur la fonction WN-Ishigami par Monte Carlo.

L'approche par modèle joint a été appliquée avec succès sur deux problématiques industrielles où les paramètres incontrôlables correspondaient à des entrées fonctionnelles :

- ▷ un processus stochastique temporel $\varepsilon(t)$ (bruit-blanc) modélisant l'incertitude sur la puissance (évoluant avec le temps) dans un code de simulation de l'irradiation du combustible en réacteur (Iooss & Ribatet [101]) ;
- ▷ un champ aléatoire spatial $\varepsilon(\boldsymbol{r})$ modélisant la perméabilité d'une couche hydrogéologique dans un code simulant le transfert de contaminants dans les sols (Iooss et al. [102]).

D'autres techniques ont été récemment proposées dans la littérature pouvant servir è l'analyse de sensibilité de codes de calcul avec entrées fonctionnelles. Celles-ci peuvent apporter des informations supplémentaires à celles données par la modélisation jointe mais ont des conditions d'utilisation assez sévères :

- ◇ Pour résoudre le problème des variables d'entrée corrélées dans l'estimation des indices de Sobol, Jacques et al. [107] proposent d'utiliser les indices de sensibilité multidimensionnels définis par Sobol [201]. Chaque groupe de variables d'entrée corrélées est examiné comme un seul macroparamètre. Il est également possible de considérer une entrée fonctionnelle comme un macroparamètre. Les indices de Sobol peuvent alors être estimés en utilisant les algorithmes basés sur des tirages Monte Carlo simples. Malheureusement, ce type d'échantillonnage, qui nécessite plusieurs milliers d'observations, rend cette méthode extrêmement coûteuse en nombre d'évaluations du code de calcul et n'a pas pu être envisagée sur mes applications.

- ◇ Tarantola et al. [210] proposent de remplacer l'entrée fonctionnelle par un paramètre "trigger" (qui signifie gâchette) $\xi \sim \mathcal{U}[0;1]$ qui gouverne la simulation ou non du processus stochastique. Lors des simulations de Monte Carlo, si $\xi < 0.5$ alors $\varepsilon = 0$, sinon ε est simulée. L'indice de sensibilité de ξ permet donc de quantifier l'influence du processus stochastique sur la variable de sortie. Par rapport à la méthode précédente, la méthode de Tarantola permet de calculer les indices de Sobol avec tout type d'échantillonnage (*e.g.* Monte Carlo simple, FAST, quasi Monte Carlo, Random Balance Design), ce qui permet de l'envisager pour des codes moyennement coûteux. Cependant, ξ reflète uniquement la présence ou l'absence de ε. $\mathrm{Var}[\mathbb{E}(Y|\xi)]$ ne quantifie donc pas correctement la contribution de la variabilité de l'entrée fonctionnelle sur la variabilité de la réponse Y. Iooss & Ribatet [101] ont illustré ceci sur un exemple joué simple.

- ◇ Récemment, Busby et al. [30] ont proposé de décomposer l'entrée fonctionnelle (un champ aléatoire gaussien dans leur application) dans une base de fonctions (Karhunen-Loève) pour réaliser l'analyse de sensibilité sur les composantes principales de la décomposition. Le nombre restreint de composantes conservées (une trentaine) permet aux auteurs de construire un métamodèle PG à l'aide de 200 simulations du code de calcul. Ainsi, par simulations Monte Carlo sur ce métamodèle, des indices de Sobol sont obtenus pour chaque coefficient des composantes de la décomposition de Karhunen-Loève du champ aléatoire. Cette méthode est donc particulièrement riche en terme de résultats. Elle permet également, contrairement au GAM joint, d'obtenir des indices de sensibilité pour chaque entrée fonctionnelle s'il y en a plusieurs. Elle est subordonnée cependant à la faisabilité de la décomposition de l'entrée fonctionnelle avec un faible nombre de fonctions de base, qui impacte directement la dimension d du problème. Par exemple, elle ne peut pas être appliquée sur le problème de Iooss & Ribatet [101] où l'entrée fonctionnelle est un bruit-blanc (discrétisé sur 3558 valeurs).

Pour l'analyse de sensibilité, les avantages de l'approche par modélisation jointe sont donc sa généralité, sa précision et le faible coût en nombre d'évaluations du modèle qu'elle nécessite. Par contre, elle ne procure pas pour l'instant toutes les informations dont on souhaiterait disposer. Par ailleurs, l'approche par modélisation jointe fournit un métamodèle que l'on peut utiliser pour la propagation d'incertitudes (Zabalza-Mezghani [232]), pour la planification robuste d'expériences (Bates et al. [17]) ou pour la résolution de problèmes d'optimisation. Tous ces thèmes constituent des axes de recherche ouverts.

3.6 Conclusion

Mes six premières années de recherche au CEA s'achèvent donc avec ce chapitre. Durant les trois premières années, je me suis essentiellement consacré à des problèmes d'ingénierie nucléaire, relatifs au traitement des incertitudes. Par la suite, mon approche a consisté à chercher régulièrement, au sein des projets de ma direction, des applications nécessitant des efforts de recherche de ma part, soit du fait de mon inaptitude initiale à les résoudre, soit du fait d'un problème non traité dans la littérature. Par exemple, j'ai commencé à m'intéresser à la thématique des métamodèles (et donc à celle des expériences numériques) en rencontrant des applications nécessitant l'utilisation de codes de calcul coûteux. Au vu de son omniprésence dans la littérature et du fait de ma formation en géostatistique, le cadre du krigeage m'a alors particulièrement intéressé et j'ai œuvré pour le rendre applicable sur des modèles relativement complexes. Par ailleurs, l'analyse et la modélisation statistiques de codes de calcul stochastiques sont des sujets encore peu abordés, sur lesquels j'ai essayé d'apporter de premiers éléments de réponse.

Dans le domaine de la statistique appliquée, l'environnement logiciel R (R Development Core team [172]) est l'un des cadres privilégiés de développement informatique et de collaboration pour la communauté internationale. La plupart des méthodes statistiques que j'ai présentées dans ce chapitre sont ainsi disponibles dans l'environnement R . Je suis d'ailleurs co-auteur avec Mathieu Ribatet du package "JointModeling" qui permet la modélisation de GLM et de GAM joints. Sur les méthodes d'analyse de sensibilité, je suis administrateur du package "sensitivity" développé par Gilles Pujol[10]. Celui-ci contient la plupart des méthodes décrites aux paragraphes 3.2.1, 3.2.3 et 3.2.4. D'autre part, sur la base de mes travaux et de contributions de stagiaires et doctorants, j'ai développé en R le logiciel SSURFER (Iooss [93]) qui supervise diverses techniques de propagation d'incertitudes, de construction de métamodèles et d'analyse de sensibilité. Les techniques de lissage sont, quant à elles, intégrées dans le package "CompModSA" de Curtis Storlie[11]. Une monographie est en cours d'élaboration avec Gilles Pujol, Curtis Storlie et Hervé Monod[12] (qui prépare un package R sur les plans d'expériences) pour réaliser une revue et un guide d'utilisation des méthodes d'analyse de sensibilité dans R .

Parmi les travaux de recherche que j'ai évoqués dans ce chapitre, certains sont loin d'être aboutis. Tout d'abord, il reste beaucoup de travail pour comparer, étudier les propriétés et perfectionner les algorithmes d'estimation de quantiles de codes qui, nous semblent très prometteurs. Par ailleurs, la thématique des codes de calcul stochastiques est sujette à divers axes de recherche ouverts : planification des expériences, comparaison de métamodèles hétéroscédastiques, développement d'outils de diagnostics, ... Concernant les entrées fonctionnelles, j'ai proposé une astuce pour ne pas traiter l'aspect fonctionnel du problème. Des voies de recherche intégrant la structure du processus stochastique en entrée du modèle sont à envisager. Enfin, un axe de recherche complémentaire et peu exploré concerne la planification des calculs intégrant des processus stochastiques ou des champs aléatoires en entrée du code. Sur ce sujet, des premiers travaux ont été réalisés sur des échantillonnages LHS de champs aléatoires gaussiens (Pebesma & Heuvelink [166]), mais de multiples approches restent à explorer.

Dans un futur proche, je prévois de m'intéresser de plus près à l'aspect planification des calculs, phase primordiale pour le succès des analyses ultérieures et sur laquelle je me suis peu penché jusqu'ici. L'aspect séquentiel et adaptatif des plans d'expérience est notamment essentiel pour rendre applicable les techniques d'analyse statistique sur les codes de calcul industriels excessivement coûteux. J'ai également commencé à regarder la planification des calculs de validation d'un métamodèle, qui conditionne nos appréciations sur ses qualités. Il me semble que des efforts plus conséquents devraient être portés sur ce sujet quelque peu oublié. Sur le thème des processus gaussiens, je ne me suis pas encore intéressé à des covariances propres à modéliser des phénomènes non stationnaires. Ce type de comportement "à seuil" des modèles est en effet présent dans certaines de mes applications, par exemple dans les scénarios d'accident de fusion du cœur d'une centrale nucléaire. Cette thématique a été identifiée au sein du GdR MASCOT-NUM[13] comme un axe de recherche important. Enfin, de premiers travaux apparaissent sur la modélisation des variables de sortie fonctionnelles (Campbell et al. [34], Fang et al. [62], Marrel [141]). Les réponses des modèles sont souvent temporelles, voire spatiales, et un trai-

[10] École des Mines de Saint-Etienne

[11] Department of Mathematics and Statistics, University of New Mexico, USA

[12] INRA, Unité de Mathématiques et Informatique Appliquées

[13] Groupement de Recherche du CNRS nommé "Méthodes d'Analyse Stochastique pour les COdes et Traitements NUMériques", cf. chap. 4

tement de l'ensemble des réponses est indispensable pour rendre attrayantes ces techniques d'analyses aux modélisateurs et utilisateurs des codes. Plusieurs thèses ont été récemment initiées sur ce sujet en France, dont celle de Benjamin Auder que je co-encadre avec Gérard Biau[14]. Comme précédemment, les aspects planification, diagnostics et comparaison de modèles devront être étudiés soigneusement.

Vis-à-vis de l'intérêt que porte la Direction de l'Énergie Nucléaire du CEA sur ces méthodes modernes d'analyse d'expériences numériques, de vastes champs de recherches et d'applications peuvent être entrevus :

- ◇ problématiques de conception tenant compte des incertitudes à l'aide de modèles numériques lourds (*e.g.* conception des systèmes nucléaires futurs) ;

- ◇ analyse et validation des gros systèmes numériques, potentiellement spatialisés et/ou temporalisés (*e.g.* qualification des codes de mécanique des fluides, calculs d'impact environnementaux) ;

- ◇ simulations pluridisciplinaires (*e.g.* couplage des codes simulant le cœur d'un réacteur nucléaire : neutronique, mécanique et thermohydraulique) ;

- ◇ simulations multiéchelles (*e.g.* calculs simulant les déformations de matériaux aux différentes échelles nano et micro) ;

- ◇ analyse des systèmes complexes et coûteux (*e.g.* transitoires thermohydrauliques, sûreté des réacteurs de quatrième génération, codes neutroniques de type Monte Carlo) ;

- ◇ traitement des incertitudes pour l'évaluation des marges de sûreté (*e.g.* dans les études d'accidents).

Toutes ces thématiques participent à un enjeu plus global, qui fait partie des grandes préoccupations et prérogatives du Commissariat à l'Énergie Atomique, et qui est celui de la maîtrise et de l'utilisation des outils de la simulation numérique.

[14]Laboratoire de Statistique Théorique et Appliquée, Université Paris VI

Chapitre 4

Bilan et perspectives

Mon activité de recherche s'est donc déclinée autour du traitement des incertitudes en modélisation numérique au cours de deux phases bien distinctes : la première portant sur le thème de la propagation d'ondes en milieu aléatoire, la deuxième ayant trait à la thématique de l'exploration statistique des expériences numériques. Les objectifs de mes travaux de recherche ont été essentiellement portés par les applications, ce qui apparaît clairement dans la présentation que j'en ai faite. L'aspect transverse de mes travaux m'a néanmoins permis d'aborder une importante diversité d'applications, source d'enrichissements professionnels et personnels importants, en permettant notamment de rencontrer des interlocuteurs issus d'horizons multiples (ingénieurs, mathématiciens, physiciens, chimistes, exploitants d'installations, ...). Ceci implique de gros efforts de communication, de compréhension mutuelle et parfois de vulgarisation.

Au niveau de la Direction de l'Énergie Nucléaire du CEA, l'équipe dédiée au traitement des incertitudes dans laquelle j'évolue est en charge de la diffusion d'une culture probabiliste au sein des différents projets où des besoins se font sentir. Elle a notamment été identifiée pour répondre à diverses demandes d'assistance et d'expertise émanant d'autres unités sur des dossiers les plus souvent urgents, comme par exemple des dossiers de sûreté concernant des essais à réaliser ou des installations nucléaires. Cette mission permet de rester en phase avec les besoins réels des ingénieurs et des physiciens "métier" de l'institut. Le travail de R&D proprement dit consiste ainsi souvent à comprendre le besoin d'un demandeur (exploitant, responsable de programmes, ingénieur, ...), à poser le problème de manière rigoureuse, puis à proposer et à développer la solution la plus adéquate dans le temps imparti.

Pour pouvoir perdurer, ces activités de soutien sont bien entendu conditionnées au maintien d'une veille scientifique et d'activités de recherche plus amont dans le domaine de la statistique, afin de continuer à apporter des réponses pertinentes aux questions posées. Il a donc été nécessaire de multiplier les contacts à l'extérieur du CEA, ce qui nous a permis d'établir un grand nombre de collaborations. L'équipe "Incertitudes" à laquelle j'appartiens est à présent investie dans plusieurs cadres collaboratifs :

- ▷ le Groupe de Travail et Réflexion "Incertitudes" de l'Institut de Maîtrise des Risques (IMdR), réseau d'échanges inter-industriels, de valorisation méthodologique et dans lequel participe (entre autres) CEA, EDF, LNE, IRSN, ONERA, Dassault, EADS, CNES, INERIS, SNCF. Dans ce groupe, j'ai notamment participé au montage d'une formation professionnelle sur le traitement des incertitudes, qui fournit à présent un dispositif complet pour former les ingénieurs de nos instituts respectifs ;

- ▷ le Groupement de Recherche (GdR) MASCOT-NUM (Méthodes d'Analyse Stochastique pour les COdes et Traitements NUMériques), structure officielle reconnue par le CNRS, et qui regroupe la quasi totalité des laboratoires de recherche et scientifiques français travaillant sur le sujet. Ce cadre d'échanges a pour vocation de faciliter la réalisation de travaux de recherche entre universitaires et ingénieurs de recherche, notamment grâce à l'organisation régulière de journées scientifiques (dont j'ai eu la charge en 2008) et d'ateliers de travail. Il doit permettre à terme d'atteindre une visibilité scientifique de premier plan au niveau français et international.

Au niveau français, ces collaborations visent clairement à déboucher sur des projets plus intégrés comme ceux lancés par l'Agence Nationale de la Recherche. Cette dernière porte d'ailleurs une attention particulière aux thèmes de la simulation numérique et de la conception sous incertitudes à l'aide de modèles prédictifs.

L'une des prochaines étapes sera de s'inscrire dans un cadre plus international où énormément de recherches ont déjà été effectuées sur le thème des expériences numériques. Je pense notamment aux contacts que je peux d'ores et déjà avoir avec l'unité de statistique du "Joint Research Centre" d'Ispra (Italie) leader dans le domaine de l'analyse de sensibilité. Au niveau américain, les laboratoires nationaux de Sandia (Albuquerque, Nouveau-Mexique) travaillent sur des thématiques extrêmement proches de celles que l'on a au CEA, notamment sur les scénarios de stockage des déchets nucléaires où la gestion des incertitudes est une problématique majeure. Les travaux de Jon Helton sont fondateurs dans ce domaine et de plus amples collaborations sont à prévoir avec les équipes de Sandia.

Pour conclure ce mémoire, je souhaite disserter quelque peu sur trois objectifs importants que je me suis fixé dans mon travail de chercheur : la communication, la collaboration et l'encadrement. Il me semble, qu'outre les nouveaux résultats obtenus et les innovations réalisées, l'un des aboutissements de la recherche est la présentation de ses résultats devant ses pairs. Cela passe bien entendu par la participation à des séminaires et à des conférences mais également par la rédaction de publications dans des revues à comité de lecture bien ciblées. La plupart de mes expériences en ce domaine sont extrêmement positives et les remarques des rapporteurs ont souvent été très constructives. L'une des grandes satisfactions, que j'ai eue jusqu'ici, est celle de voir mon travail de recherche repris, critiqué et poursuivi par d'autres chercheurs. La collaboration correspond aussi au principal plaisir que me procure ce métier, à savoir, travailler en équipe. Ceci permet de côtoyer de multiples personnalités, de mettre en avant ses propres qualités et faiblesses, de se renouveler et finalement d'avancer vers des idées neuves. Les principaux résultats que j'ai obtenus au CEA l'ont été grâce à des collaborations fructueuses avec d'excellents mathématiciens issus de l'université, mais aussi avec des physiciens et des ingénieurs du CEA confrontés à des problématiques passionnantes. Enfin, l'expérience d'encadrement de doctorants ou de stagiaires se révèle être particulièrement enthousiasmante, non seulement au niveau des relations humaines sur lesquelles elle débouche, mais aussi par la remise en question permanente qu'elle provoque. Ce travail de transmission, qui est au cœur du métier de chercheur et qui est lié à nos propres engagements et sens des responsabilités, est à mon avis l'une des tâches les plus difficiles.

English version - Conclusion and perspectives

My research activity deals with uncertainty management in numerical modelling. I have mainly worked on two main subjects. The first one is related to the wave propagation in random media. The second concerns the statistical exploration of numerical experiments. Their objectives were essentially carried by applications. Nevertheless, the transversality of my works allows to tackle a wide variety of applications. That gives me the opportunity to meet researchers coming from very different fields (engineers, mathematicians, physicists, chemists, facilities operator, ...). Of course, it implies important communication and mutual understanding efforts

In Nuclear Energy Division of CEA, my team handles the diffusion of a probabilistic culture within various engineering projects. We give some assistance and expertise on statistical and mathematical questions. These latter are sometimes urgent, as for example in nuclear safety analysis for institutional authorities. This job allows to keep in touch with the real needs of engineers and physicists. Strictly speaking, this job can be decomposed in three steps :

- understanding the need of users (operator, program manager, engineer, ...),
- raising their problem in a rigorous way,
- proposing and developing the most adequate solution within the allocated time.

To bring relevant answers to industrial questions, scientific watch and more theoretical research activities are also required. Thus, it was necessary to establish a large number of collaborations. The team "Uncertainties" to which I belong is invested at the present time in several collaborative frameworks :

▷ the Working Group "Uncertainties" of the Institut de Maîtrise des Risques (IMdR). It is an inter-industrial network, allowing methodological valuation. A lot of french industrial institutes participate in it : CEA, EDF, IRSN, ONERA, Dassault, EADS, CNES, INERIS, SNCF. Our first realization was to organize a professional training on the uncertainty management. At the present time, it supplies a useful tool to train the engineers of our research institutes;

▷ the Research Group (GdR) MASCOT-NUM (Methods of Stochastic Analysis for COdes and NUMerical treatments) of the French National Research Center (CNRS). It includes almost totality of research laboratories and French scientists working on the subject. This official structure facilitates the realization of research works between academics and research engineers. It makes also possible the organization of scientific conferences. One of our objective for this group is to reach a leading scientific visibility in the French and international statistical communities.

These collaborations will also result in integrated projects as those launched by the French National Research Agency. This agency pays moreover a particular attention on our research subjects. For example, many projects concern the conception problems under uncertainties by means of numerical predictive models.

The international community have provided a lot of results on my research themes. One of my next objectives will be to join this international context. I already have exchanges with the statistical unit of the Ispra "Joint Research Centre". For many years, this unit leads the research field of sensitivity analysis. In the United States, the Sandia National Laboratories (Albuquerque, New Mexico) work on similar themes than those of the CEA. For example, Sandia studies for a long time the uncertainty management of the deep radwaste storage scenarios. The resolution of this problem is a major issue for the nuclear industry future. Jon Helton's works are founders in this domain. More ample collaborations would have to be planned between CEA and Sandia teams.

To conclude this report, I wish to talk a little about three important objectives that I settled in my researcher work : communication, collaboration and supervision. Besides the new obtained results, one of the research outcomes is the presentation of these results in front of peers. It naturally involves our participation in seminars and publication writings in peer reviewed journals. Most of my experiences in this domain are extremely positive. The referees remarks were often very constructive. Another satisfaction has been to see my research work resumed, criticized and pursued by other researchers. The collaboration also corresponds to a great pleasure provided by my job. This allows to go alongside to multiple personalities. This also leads to analyze my own qualities and weaknesses, in order to get new ideas. Finally, the experience of PhD students or trainees supervision is particularly compelling. This stands not only on the human relations in which it results. This stands more deeply by the permanent questioning which it provokes. This transmission work is in the core of the researcher's job. It is connected to our own commitments. In my opinion, it is one of our most difficult tasks.

Annexe A

Curriculum vitae en français

Bertrand IOOSS
Né le 6 mars 1973, marié, 2 enfants.
Nationalité : française.

Coordonnées personnelles
69 Hameau de Beaumont
84120 Pertuis, France

Situation professionnelle
Ingénieur-chercheur
Commissariat à l'Energie Atomique (CEA)
Direction de l'Energie Nucléaire (DEN)
Centre de Cadarache (CAD)
Département d'Etude des Réacteurs (DER)
Service d'Etude des Systèmes Innovants (SESI)
Laboratoire de Conduite et de Fiabilité des Réacteurs (LCFR)

Coordonnées professionnelles
CEA Cadarache, DER/SESI/LCFR,
Bât. 212, 13108 Saint-Paul-lez-Durance, France
Téléphone : 04 42 25 72 73
Fax : 04 42 25 24 08
e-mail : bertrand.iooss@cea.fr

Diplômes universitaires

1990-92 : D.E.U.G. Mathématiques-Physique-Mécanique, Université de Nice.
1992-93 : Licence de Mathématiques, Université de Nice.
1993-94 : Maîtrise d'Ingénierie Mathématiques, Université de Nice.

1994-95 : **D.E.A. de Statistiques et Modèles Aléatoires en Finance,** Université Paris VII.
Stage à l'INRIA Sophia-Antipolis, dirigé par D. Talay (chercheur INRIA) et P. Bertrand (professeur Univ. Clermont) sur le thème des *estimateurs de la volatilité dans des modèles de saut pur.*

1995-98 : **Doctorat de l'École des Mines de Paris, spécialité "Géostatistique",** soutenue à l'École des Mines de Paris (Fontainebleau).
Titre : *Tomographie statistique en sismique réflexion : estimation d'un modèle de vitesse stochastique.*
Jury : Alain Galli (directeur de thèse, École des Mines de Paris), Dominique Gibert (rapporteur, Université Rennes I), Jean Virieux (rapporteur, Université Nice), Michel Schmitt (président, École des Mines de Paris), Michel Campillo (examinateur, Univ. Grenoble), Paulo Ruffo (examinateur, AGIP), Pierre Thore (examinateur, Elf).

1999-2000 : Inscriptions sur listes de qualification (fonction maître de conférence) du Conseil National des Universités : section 26 (Mathématiques Appliquées), section 35 (Physique et Chimie de la Terre), section 60 (Mécanique et Génie civil).

Parcours professionnel

1999-2000 : **Post-Doctorat, CEA Cadarache,** Direction des Réacteurs Nucléaires, Département d'Étude des Réacteurs, Service de Systèmes d'Aide à l'Exploitation, Laboratoire de Systèmes de Mesures pour les Réacteurs.

2000-01 : **Ingénieur de recherche, Institut Français du Pétrole** (Rueil Malmaison), Division Géophysique.

Depuis février 2002 : **Ingénieur-chercheur, CEA Cadarache,** Direction de l'Énergie Nucléaire.

ACTIVITES de RECHERCHE

[Réferences bibliographiques en pages 9-13]

Propagation d'ondes en milieu aléatoire

Théorie de la propagation d'ondes hautes fréquences en milieu aléatoire

- Ondes en milieu aléatoire statistiquement anisotrope : étude des domaines de validité des approximations parabolique, de Rytov et de l'optique géométrique.
  {thèse [Im2], rapport [Ir2]}

- Etude par une méthode de perturbation de la moyenne des temps de trajet au second ordre en fonction de la covariance du champ de vitesse des ondes. Validation des résultats théoriques par des simulations numériques (utilisant la méthode des différences finies sur l'équation des ondes). En collaboration avec Y. Samuelides (stagiaire Corps des Mines).
  {thèse [Im2], conf [Ic2]}

- Calcul analytique de la variance des temps de trajet au second ordre. Validation par simulations numériques (technique du tracé de rayons). En collaboration avec P. Blanc-Benon (chercheur CNRS) et C. Lhuillier (ingénieur-chercheur CEA).
  {article [Ia3]}

- Etude de la covariance des temps de trajet, inversion indirecte pour retrouver la covariance du champ de vitesse, validation sur des simulations numériques par différences finies.
  {thèse [Im2], article [Ia1], conf [Ic3]}

Applications

- Tomographie statistique : inversion directe de la covariance, appliquée au cas particulier de la sismique réflexion (prise en compte d'un réflecteur avec fluctuations aléatoires). Couplage avec les méthodes déterministes (tomographie sismique). Validation sur simulations numériques et applications à des données réelles d'exploration pétrolière (contrat AGIP). En collaboration avec M. Touati (doctorant École des Mines de Paris), D. Geraets (doctorant École des Mines de Paris) et A. Galli (chercheur École des Mines de Paris).
  {thèse [Im2], article [Ia5], confs [Ic1, Ic4]}

- Etude statistique des résidus de la procédure déterministe d'inversion des temps de trajet (tomographie sismique).
  {rapport [Ir5]}

- Propagation de l'incertitude due au champ de vitesse sur la localisation des réflecteurs du sous-sol (procédure d'imagerie sismique appelée "migration"). En collaboration avec M. Touati (doctorant École des Mines de Paris) et A. Galli (chercheur École des Mines de Paris).
  {article [Ia2]}

- Simulation de la propagation d'ondes en milieu turbulent en mouvement par les techniques de tracé de rayons et de sommation de faisceaux gaussiens, développement et validation de la méthode. Application à la quantification des incertitudes dues à la turbulence dans la mesure de débit par ultrasons (contrat Électricité De France). En collaboration avec C. Lhuillier (ingénieur-chercheur CEA) et H. Jeanneau (ingénieur de recherche, EDF R&D).
  {article [Ia4], rapports [Ir3, Ir4]}

Incertitudes, planification et analyse d'expériences numériques

Construction et validation de métamodèles

- Etude comparative de modèles de régression et d'apprentissage statistiques (GLM, GAM, processus gaussiens, réseaux de neurones, SVM, boosting d'arbres de régression, forêts aléatoires) pour ajuster les réponses de modèles numériques déterministes. Applications à divers codes de calcul d'ingénierie nucléaire. {rapports [Ir11, Ir19]}

- Etude des méthodes de planification d'expériences numériques (basées sur des critères de dicrépance) en vue de la validation d'un métamodèle. En collaboration avec V. Feuillard (doctorant Univ. Paris VI). {conf [Ic14]}

- Développement d'algorithmes pour modéliser les réponses de codes par processus gaussiens quand la dimension des entrées est élevée (> 10). En collaboration avec A. Marrel (doctorante INSA Toulouse). {article [Ia9], conf [Ic14], rapport [Ir15]}

- Modélisation de réponses fonctionnelles (temporelles, spatiales) de codes de calcul. En collaboration avec A. Marrel (doctorante INSA Toulouse). {conf [Ic21]}

- Modélisation jointe de la moyenne et de la dispersion (par GLM) pour codes de calculs stochastiques. Développement du modèle GAM joint. En collaboration avec M. Ribatet (doctorant CEMAGREF Lyon). {article [Ia12], conf [Ic6], rapport [Ir16]}

Analyse de sensibilité globale

- Développement d'une méthodologie générale pour l'analyse de sensibilité d'un code de calcul. En collaboration avec N. Devictor, M. Marquès et N. Pérot (ingénieur-chercheurs CEA). {articles [Ia6, Ia8], conf [Ic5], rapports [Ir6, Ir12, Ir13, Ir14, Ir19, Ir24]}

- Analyse de sensibilité quand des paramètres d'entrée sont fonctionnels. Utilisation du modèle joint (GLM ou GAM). En collaboration avec M. Ribatet (doctorant CEMAGREF Lyon). {articles [Ia12, Ia13], confs [Ic6, Ic10], rapports [Ir28, Ir33]}

- Analyse de sensibilité pour variables de sortie fonctionnelles. Décomposition en base d'ondelettes et modélisation par processus gaussien. En collaboration avec A. Marrel (doctorante INSA Toulouse). {conf [Ic21]}

- Analyse de sensibilité à partir du modèle des processus gaussiens. En collaboration avec A. Marrel (doctorante INSA Toulouse), B. Laurent (professeur INSA Toulouse) et O. Roustant (chercheur École des Mines de Saint Etienne). {article [Ia14], conf [Ic11], rapport [Ir34]}

- Analyse de sensibilité basée sur l'entropie. En collaboration avec B. Auder (doctorant Univ. Paris VI). {rapport [Ir22], conf [Ic15]

- Synthèse des différentes méthodes d'analyse de sensibilité (criblage, mesures d'importance, méthodes d'exploration fine). En collaboration avec G. Pujol (ingénieur de recherche École des Mines de Saint-Etienne), C. Storlie (professeur assistant Univ. du Nouveau-Mexique) et H. Monod (chercheur INRA). {monographie en préparation [I1]} }

Estimation de quantiles

- Développement de méthodes de Monte Carlo par tirage stratifié et par tirage d'importance pour estimer un quantile élevé (de l'ordre de 95%) d'un code de calcul à l'aide d'une surface de réponse. En collaboration avec J. Garnier (professeur Univ. Paris VII) et C. Cannamela (doctorante Univ. Paris VII). {article [Ia11], conf [Ic20], rapport [Ir21]}

Différentiation automatique

- Application d'un outil de différentiation automatique (TAPENADE, INRIA) sur un modèle de simulation du comportement des combustibles nucléaires sous irradiation. Contribution au développement du mode tangent multi-directionnel de TAPENADE. En collaboration avec L. Hascoët (chercheur INRIA). {rapport [Ir8]}

Ajustement de données nucléaires

- Développement d'une méthode de Monte Carlo pour prendre en compte la matrice de covariance de paramètres non ajustés lors de l'ajustement de sections efficaces netroniques. En collaboration avec C. De Saint Jean et G. Noguère (ingénieur-chercheurs CEA). {articles [Ia10, Ia15], confs [Ic7, Ic8]}

Échantillonnage statistique

- Etude et développement de procédures d'échantillonnage statistique et de cartographie géostatistique dans le cadre du démantèlement des installations nucléaires. En collaboration avec N. Pérot et F. Lamadie (ingénieur-chercheurs CEA) et N. Jeannée (Géovariances). {rapports [Ir26, Ir27, Ir30], confs [Ic18, Ic17]}

PROGRAMMES de COOPÉRATION et COLLABORATION

- Consortium KIM (Kinematic Inverse Methods) de l'Institut Français du Pétrole, consortium de recherche sur l'imagerie et la tomographie en sismique réflexion. Participation en 2000-2001.
  {rapport [Ir5]}

- École d'été 2006 du CEMRACS (Centre d'été Mathématique de Recherche Avancée en Calcul Scientifique) sur le thème "Modélisation de l'aléatoire et propagation d'incertitudes", juillet-Août 2006. Participation au projet CEA "Estimation de quantiles de codes de calcul" proposé par B. Iooss, N. Devictor, A. De Crécy et P. Bazin.
  {publi [Ia11], conf [Ic20], rapport [Ir21]}

- PAMINA (Performance Assessment Methodologies IN Application to guide the development of the safety case), projet européen du 6ème PCRD (Programme Cadre de Recherche et de Développement de la communauté européenne) : apport des modélisations effectuées dans le cadre des évaluations de performance des stockages géologiques pour construire le dossier de sûreté. Participation en 2007-2008.
  {rapports [Ir33, Ir34, Ir35]}

DEVELOPPEMENTS de LOGICIELS

Librairies et logiciels de statistiques en R

- Logiciel SSURFER (Sensibilités et SURFaces de réponse dans l'Environnement R), depuis 2004.
  {rapport [Ir19]}

- Co-auteur et administrateur du package R "JointModeling" avec M. Ribatet (doctorant CEMAGREF Lyon) : modélisation jointe de la moyenne et de la dispersion à l'aide des GLM et des GAM, depuis 2005.

- Administrateur du package R "sensitivity" (auteur : G. Pujol, ingénieur de recherche, École des Mines de Saint-Etienne) : méthodes d'analyse de sensibilité globale, depuis 2006.

Logiciels métiers

- PROUST (PROpagation UltraSonore en milieu Turbulent), logiciel CEA pour le diagnostic ultrasonore dans les cuves et circuits de réacteurs, participation au développement en 1999-2000.
  {rapport [Ir4]}

- LEONAR (Logiciel d'Evaluation de la prObabilité de percemeNt du radier en cas d'Accident gRave), réalisé dans le cadre d'un contrat pour Électricité De France, depuis 2007.
  {conf [Ic19], rapports [Ir23, Ir24, Ir25, Ir31, Ir32, Ir38, Ir39]}

ENCADREMENTS SCIENTIFIQUES

[Réferences bibliographiques en pages 9-13]

Activités d'encadrement de thèse

- A. Marrel (2005-2008) : encadrement avec B. Laurent (directeur de thèse, INSA Toulouse) et M. Jullien (co-encadrant CEA). Sujet : Mise en œuvre et utilisation du métamodèle processus gaussien pour l'analyse de sensibilité de modèles numériques, application à un code de transport hydrogéologique. Soutenance le 3 juillet 2008.
  Publications réalisées : articles [Ia9, Ia14], confs [Ic11, Ic21]

- B. Auder (débutée en janvier 2008) : encadrement avec G. Biau (directeur de thèse, Univ. Paris VI) et M. Marquès (co-encadrant CEA). Sujet : Analyse et modélisation statistiques de sorties fonctionnelles de codes de calcul.

Activités d'encadrement de stage

- S. Campos (INSA Toulouse 5$^{\text{ème}}$ année) : *Différentiation automatique du code combustible MARGARET*, 2002.

- C. Cannamela (DESS Statistique, Informatique et Techniques Numériques, Univ. Lyon 1) : *Etude et comparaison des SVM avec des méthodes classiques de discrimination*, 2003.

- S. Ndao (DESS Statistique, Informatique et Techniques Numériques, Univ. Lyon 1) : *Construction d'un modèle simplifié sur le code combustible MOGADOR*, 2003.

- P-M. Pair (INSA Toulouse 5$^{\text{ème}}$ année) : *Construction de surfaces de réponse non linéaires : étude comparative de nouvelles méthodes de régression*, 2004.

- E. Volkova (en thèse à l'Institut Kurchatov, Moscou) : *Analyse de sensibilité d'un modèle de transport du* ^{90}Sr *en milieu poreux saturé, application à un site de déchets radioactifs*, 2004-05.
  Publication réalisée : article [Ia8]

- A. Marrel (Master 2 Recherche Mathématiques Appliquées, INSA Toulouse) : *Modélisation des codes de calcul dans le cadre des processus gaussiens*, 2005.

- M. Ribatet (en thèse au CEMAGREF Lyon) : *Modélisation de la moyenne et de la dispersion par les GLM*, 2005.
  Publication réalisée : conf [Ic6]

- B. Auder (ENSIMAG 4$^{\text{ème}}$ année) : *Analyse de sensibilité globale basée sur l'entropie*, 2006.
  Publication réalisée : conf [Ic15]

- V. Bakirdjian (Master 2 Pro Génie Scientifique et Informatique, Univ. de Provence) : *Caractérisation géostatistique de cellules contaminées*, 2007.

- D. Barthel (INSA Toulouse 5$^{\text{ème}}$ année) : *Méthodes d'analyse de sensibilité basés sur les tests statistiques*, 2008.

ENSEIGNEMENTS

Cours

- *Cours de Probabilités* (6 heures), D.E.A. “Méthodes quantitatives en Géosciences”, École du Pétrole et des Moteurs, Rueil-Malmaison, octobre 1997.

TD

- *Initiation à R* (4 heures), Licence 3, IUP “Statistique Informatique Décisionnelle”, Université Toulouse III, Toulouse, février 2009.

Interventions dans des formations doctorales

- *Concepts et méthodes de la géostatistique,* Séminaire "Statistique spatiale pour l’industrie et le marketing" organisé par C. Thomas-Agnan, Master 2 Pro “Statistiques et Econométrie”, GREMAQ, Toulouse, février 2006.

- Avec J. Baccou (IRSN Cadarache) : *Analyses d’incertitudes : nouvelles exigences du risque environnemental,* Journée "Incertitudes dans les risques industriels" organisée par E. Pardoux, Écoles Centrales (option mathématiques), Chateau-Gombert, Marseille, février 2007.

Formations professionnelles

- Co-organisation de la formation “Incertitudes” de l’Institut de Maîtrise des Risques (parrainée par la SFdS, la SMAI et Ter@tec), Paris. Animation de la session “Analyse de sensibilité” (cours + TD). Depuis 2007.

RESPONSABILITES

Participations à des jurys de thèse

- Examinateur de la thèse de M. Petelet, Analyse de sensibilité de modèles thermomécaniques de simulation numérique du soudage, 30 octobre 2007, Univ. de Bourgogne.

- Examinateur (en tant que co-encadrant) de la thèse d’A. Marrel, Mise en œuvre et utilisation du métamodèle processus gaussien pour l’analyse de sensibilité de modèles numériques, 3 juillet 2008, INSA Toulouse.

Rapporteurs pour des revues

- *Pure and Applied Geophysics* (1 article en 2001),
- *Journal of Flow Measurement and Instrumentation* (1 article en 2004),
- *Experiments in Fluids* (1 article en 2006),
- *Acta Acustica* (1 article en 2007).

Membre de comités scientifiques de conférences

- ENBIS-EMSE Spring Meeting 2009, “Design and Analysis of Computer Experiments”, Saint-Etienne, France, july 2009.

CONGRES et SEMINAIRES

Exposés oraux ou posters lors de congrès ou conférences

- Journées de Géostatistique, Fontainebleau, France, juin 1998.
- 135th Meeting of the Acoustical Society of America, Seattle, USA, juin 1998.
- 68th Annual Meeting of the Society of Exploration Geophysicists, New Orleans, USA, septembre 1998.
- 4th International Conf. on Sensitivity Analysis of Model Output, Santa Fe, USA, mars 2004.
- Journées Techniques de la CETAMA, Montpellier, France, octobre 2005.
- XXXVIIIèmes Journées de Statistique, Clamart, France, mai-juin 2006.
- 5th International Conf. on Sensitivity Analysis of Model Output, Budapest, Hungary, juin 2007.
- VI Colloqium Chemiometricum Mediterraneum, Saint Maximin, France, septembre 2007.
- ESREL 2008 Annual Conference, Valencia, Espagne, septembre 2008.
- 35rd ESReDA Seminar, Marseille, France, novembre 2008.

Invitations à des conférences

- *Concepts et méthodes de la statistique,* École d'été CEA-EDF-INRIA, Prise en compte des incertitudes en simulation numérique, Saint Lambert des Bois, juin 2005.
- *Concepts et méthodes de la géostatistique,* Journées Techniques de la CETAMA, Echantillonnage et caractérisation "Du prélèvement à l'analyse", Montpellier, octobre 2005.
- *Analyses de sensibilité globales de modèles complexes : quelques applications dans le nucléaire,* Journées organisées par l'INSA Toulouse, Planification d'expériences et analyse d'incertitudes pour les gros codes numériques : approches stochastiques, Toulouse, février 2006.
- *Estimation de quantiles de codes de calcul,* Journées DIF "Incertitudes et simulation", CEA/DAM, Bruyères-le-Châtel, octobre 2007.
- *Uncertainty and reliability study of the creep law to assess the fuel cladding behavior of PWR spent fuel assemblies during interim dry storage,* Session "Uncertainty in industrial practice" organisée par E. de Rocquigny, ESREL 2008 Conference, Valence, Espagne, septembre 2008.

Exposés lors de séminaires

- *Statistical tomography for reflection seismics : stochastic velocity model estimation,* Centre de recherche d'AGIP, Milan, Italie, janvier 1999.
- *Etude des temps de trajet en milieu aléatoire, inversion probabiliste en sismique,* Séminaire du Laboratoire de Mécanique des Fluides et d'Acoustique, École Centrale de Lyon, Lyon, mars 2000.
- *Tomographie statistique en sismique réflexion : estimation d'un modèle de vitesse stochastique,* Séminaire de la Division Géophysique, Institut Français du Pétrole, Rueil-Malmaison, mars 2001.
- *Misfits statistics and data decimation in traveltime inversion,* Consortium KIM annual meeting, Institut Français du Pétrole, Rueil Malmaison, décembre 2001.
- *Traitement des incertitudes,* Institut Français du Pétrole, Rueil Malmaison, novembre 2004.
- *Probabilistic methods for uncertainty studies,* Institut Kurchatov, Moscou, Russie, juin 2005.
- *Analyse de sensibilité globale de modèles numériques à paramètres incontrôlables,* Institut Français du Pétrole, Solaize, avril 2006.
- *Estimation de quantiles de code, applic. thermohydraulique,* Séminaire IMPEC, CEA Saclay, octobre 2006.
- *Traitement des incertitudes, quelques applications en ingénierie nucléaire,* Centre de Géostatistique, École des Mines de Paris, Fontainebleau, octobre 2006.
- *Traitement des incertitudes, quelques applications en ingénierie nucléaire,* Workshop "Management des incertitudes", EADS, Toulouse, octobre 2006.
- *Estimation de quantiles de code, application thermohydraulique,* Département de Mathématiques, École des Mines de Saint Etienne, décembre 2006.
- Avec M. Marquès (CEA Cadarache) : *Sensitivity analysis on the PRHRS performance indicators,* Meeting of the IAEA Project "Status and prospects of development and application of innovative reactor concepts for developing contries", Univ. de Pise, Italie, février 2007.
- *Gaussian processes as metamodels*, Workshop "Sensitivity analysis", Joint Research Centre, Ispra, Italie, février 2008.
- *Two problems in metamodelling and sensitivity analysis of complex numerical models : high dimensional inputs and stochastic computer code*, Seminar for the Operations Research Group of CentER, Tilburg University, Tilburg, Pays-Bas, mai 2008.

ANIMATION SCIENTIFIQUE

Activités d'animations

- Trésorier 1997-98 ABCTEM (Association Bellifontaine des Chercheurs et Thésards de l'École des Mines).

- Depuis 2006, création et co-animation du réseau IMPEC (Incertitudes, Métamodèles et Plans d'Expériences pour les Codes), groupe d'échanges interne du CEA, avec J-M. Martinez et F. Gaudier (ingénieur-chercheurs au CEA Saclay). Co-administrateur (avec F. Gaudier) du site intranet `http://www-impec.cea.fr`

- Depuis 2006, membre du Groupe de Travail et de Réflexion "Incertitudes" de l'Institut de Maîtrise des Risques (responsable : E. de Rocquigny, EDF R&D), URL : `http://www.imdr-sdf.asso.fr/v2/extranet/index.php?page=gtr`

- Depuis 2007, membre du bureau du GdR CNRS MASCOT NUM ("Méthodes d'Analyse Stochastique pour le COdes et Traitements NUMériques"). Responsables : F. Gamboa (Univ. Paul Sabatier, Toulouse) et F. Wahl (IFP Lyon). Administrateur du site internet `http://www.gdr-mascotnum.fr`

- Depuis 2008, membre du ESRA (European Safety and Reliability Association) Technical Committee on Uncertainty and Sensitivity Analysis, URL : `http://www.esrahomepage.org/uncertainty.aspx`

Organisations de séminaires

- Colloque Mines 98, Présentations des doctorants de l'École des Mines, École des Mines de Paris, Fontainebleau, mai 1998.

- Séminaire PAN (Plan d'Action Neuronal), "Plans d'Expériences Numériques", CEA Cadarache, décembre 2005.

- Séminaire IMPEC (Incertitudes Métamodèle et Plans d'Expériences pour les Codes), "Méthodes de krigeage et estimation de quantiles", CEA Saclay, octobre 2006.

- Atelier aux rencontres du GdR CNRS MASCOT NUM (Méthodes d'Analyse Stochastique pour les COdes et traitement NUMériques) avec J-C. Fort (Univ. Paul Sabatier, Toulouse), "Réflexions sur la problématique des données fonctionnelles dans l'analyse et la planification d'expériences numériques", IFP Solaize, mars 2007.
  URL : `http://www.gdr-mascotnum.fr/rencontres/rencontres_mars2007/rencontres_mars2007.html`

- Séminaire IMPEC, "Risque environnemental", CEA Cadarache, juin 2007.

- Séminaire IMPEC, CEA Cadarache, octobre 2008.

- Journée de rencontres entre le GdR MASCOT-NUM, le GT "Incertitudes" de l'IMdR et les étudiants de Master. URL : `http://www.gdr-mascotnum.fr/doku.php?id=thesismeeting`

Organisations de conférences

- Assistant à l'École d'Été d'Analyse Numérique CEA-EDF-INRIA 2005, "Prise en compte des incertitudes en simulation numérique", Saint Lambert des Bois, juin 2005. Coordinateur : J-M. Martinez (CEA).

- Organisation des rencontres du GdR MASCOT NUM, CEA Cadarache, 12 au 14 mars 2008.
  URL : `http://www.gdr-mascotnum.fr/rencontres/rencontres_mars2008/progMASCOTCadmars08.pdf`

Participations à des tables rondes et débats

- *La modélisation des incertitudes*, Journées MAS de la SMAI, Lille, septembre 2006.

English version - Curriculum vitae

Bertrand IOOSS
Date of birth : march 6, 1973, married, 2 children.
Nationality : french.

Personal address
69 Hameau de Beaumont
84120 Pertuis, France

Professional situation
Research engineer
Commissariat à l'Energie Atomique (CEA)
Direction de l'Energie Nucléaire (DEN)
Centre de Cadarache (CAD)
Département d'Etude des Réacteurs (DER)
Service d'Etude des Systèmes Innovants (SESI)
Laboratoire de Conduite et de Fiabilité des Réacteurs (LCFR)

Professional address
CEA Cadarache, DER/SESI/LCFR,
Bât. 212, 13108 Saint-Paul-lez-Durance, France
Phone : 04 42 25 72 73
Fax : 04 42 25 24 08
e-mail : bertrand.iooss@cea.fr

Education

1990-92 : D.E.U.G. Mathematics-Physics-Mecanics, Université de Nice, France.
1992-93 : Licence of Mathematics, Université de Nice, France.
1993-94 : Master 1 of Applied mathematics, Université de Nice, France.

1994-95 : **Master 2 of Statistics and random models in finance,** Université Paris VII, France.
Research traning at INRIA Sophia-Antipolis, supervised by D. Talay (INRIA research director) and P. Bertrand (Univ. Clermont professor) on the subject : *Volatility estimators in pure jump models.*

1995-98 : **PhD thesis at École des Mines de Paris, speciality "Geostatistics",** presented at École des Mines de Paris (Fontainebleau, France).
Title : *Statistical tomography in reflection seismics : stochastic velocity model estimation.*
Jury : Alain Galli (PhD supervisor, École des Mines de Paris), Dominique Gibert (referee, Université Rennes I), Jean Virieux (referee, Université Nice), Michel Schmitt (president, École des Mines de Paris), Michel Campillo (examiner, Univ. Grenoble), Paulo Ruffo (examiner, AGIP), Pierre Thore (examiner, Elf).

1999-2000 : Qualification (for the function assistant professor) by the Conseil National des Universités in section 26 (Applied Mathematics), section 35 (Physics and Chemestry of the Earth), section 60 (Mechanics and Civil Engineering).

Professional experience

1999-2000 : **Post-Doctoral position, CEA Cadarache** (France), Direction des Réacteurs Nucléaires, Département d'Étude des Réacteurs, Service de Systèmes d'Aide à l'Exploitation, Laboratoire de Systèmes de Mesures pour les Réacteurs.

2000-01 : **Research engineer, Institut Français du Pétrole** (Rueil Malmaison, France), Division Géophysique.

Since february 2002 : **Research engineer, CEA Cadarache** (France), Direction de l'Énergie Nucléaire.

RESEARCH ACTIVITIES

[Bibliographical references in pages 9-13]

Wave propagation in random media

Theory of the high frequency wave propagation in random media

- Waves in statistically anisotropic random media : validity domain studies of the parabolic approximation, Rytov approximation and geometrical optics.
  {PhD thesis [Im2], report [Ir2]}

- Study with a perturbation technique of the traveltime mean at second order in function of the wave velocity field covariance. Validation of theoretical results by numerical simulation (via the finite differences method applied on the acoustical wave equation). In collaboration with Y. Samuelides (stagiaire Corps des Mines).
  {PhD thesis [Im2], conf [Ic2]}

- Analytical calculation of the traveltime variance at second order. Validation by numerical simulation (via the ray tracing technique). In collaboration with P. Blanc-Benon (CNRS researcher) and C. Lhuillier (CEA research engineer).
  {article [Ia3]}

- Study of the traveltime covariance, direct inversion to retrieve the wave velocity field covariance, validation by finite differences numerical simulations.
  {PhD thesis [Im2], article [Ia1], conf [Ic3]}

Applications

- Statistical tomography : direct inversion of the covariance, aapplied to the particular case of reflection seismics (by taking into account a random fluctuation reflector). Coupling the statistical tomography with the deterministic seismic tomography. Validation by numerical simulation and application to oil exploration data (AGIP contract). In collaboration with M. Touati (PhD student at École des Mines de Paris), D. Geraets (PhD student at École des Mines de Paris) and A. Galli (École des Mines de Paris researcher).
  {PhD thesis [Im2], article [Ia5], confs [Ic1, Ic4]}

- Statistical study of traveltime seismic tomography residuals.
  {report [Ir5]}

- Uncertainty propagation of the velocity field uncertainty on the seismic reflector localization (seismic imaging process called “migration”). In collaboration with M. Touati (PhD student at École des Mines de Paris) and A. Galli (École des Mines de Paris reseracher).
  {article [Ia2]}

- Simulation of wave propagation in moving turbulent media with the ray tracing and Gaussian beams techniques, development and validation. Application to the uncertainty quantification due to fluid turbulence of the ultrasonic flowmeters (Électricité De France contract). In collaboration with C. Lhuillier (CEA research engineer) and H. Jeanneau (research engineer, EDF R&D).
  {article [Ia4], reports [Ir3, Ir4]}

Uncertainties, design and analysis of numerical experiments

Metamodel construction and validation

- Comparative study of regression and statistical learning models (GLM, GAM, Gaussian process, neural network, SVM, regression trees boosting, random forest) to fit deterministic numerical models outputs. Applications to several nuclear enginnering computer codes.
  {reports [Ir11, Ir19]}

- Study of numerical experiment design methods (based on discrepancy criteria) in order to validate a metamodel. In collaboration with V. Feuillard (PhD student Univ. Paris VI).
  {conf [Ic14]}

- Algorithm development to modelize computer code output by the Gaussian process model, when the inputs dimension is large (> 10). In collaboration with A. Marrel (PhD student INSA Toulouse).
  {article [Ia9], conf [Ic14], report [Ir15]}

- Functional output (temporal, spatial) modelling of computer codes. In collaboration with A. Marrel (PhD student INSA Toulouse).
  {conf [Ic21]}

- Joint modelling of the mean and dispersion of stochastic computer codes. Development of the joint GAM model. In collaboration with M. Ribatet (PhD student CEMAGREF Lyon).
  {article [Ia12], conf [Ic6], report [Ir16]}

Global sensitivity analysis

- Development of a general methodology for the computer code sensitivity analysis. In collaboration with N. Devictor, M. Marquès and N. Pérot (CEA research engineers).
  {articles [Ia6, Ia8], conf [Ic5], reports [Ir6, Ir12, Ir13, Ir14, Ir19, Ir24]}

- Sensitivity analysis for functional input variables via the joint model (GLM or GAM) use. In collaboration with M. Ribatet (PhD student CEMAGREF Lyon).
  {articles [Ia12, Ia13], confs [Ic6, Ic10], reports [Ir28, Ir33]}

- Sensitivity analysis for functional output variables. Wavelet basis decomposition and Gaussian process modelling. In collaboration with A. Marrel (PhD student INSA Toulouse).
  {conf [Ic21]}

- Sensitivity analysis from the Gaussian process model. In collaboration with A. Marrel (PhD student INSA Toulouse), B. Laurent (INSA Toulouse professor) and O. Roustant (École des Mines de Saint Etienne reseracher).
  {article [Ia14], conf [Ic11], report [Ir34]}

- Sensitivity analysis based on entropy. In collaboration with B. Auder (PhD student Univ. Paris VI).
  {report [Ir22], conf [Ic15]

- General synthesis of the sensitivity analysis methods (screening, importance measures, deep exploration methods). In collaboration with G. Pujol (École des Mines de Saint-Etienne research engineer), C. Storlie (New Mexico University (USA) professor assistant) and H. Monod (INRA researcher).
  {monography in preparation [I1]} }

Quantile estimation

- Development of Monte Carlo methods by stratified and importance sampling to estimate a high order quantile (95% order) of a computer code, with the help of a metamodel. In collaboration with J. Garnier (Univ. Paris VII professor) and C. Cannamela (PhD student Univ. Paris VII).
  {article [Ia11], conf [Ic20], report [Ir21]}

Automatic differentiation

- Application of an automatic differentiation tool (TAPENADE, INRIA) on a nuclear fuel irradiation simulation model. Contribution to the development of the multi-directional mode of TAPENADE. In collaboration with L. Hascoët (INRIA researcher).
  {report [Ir8]}

Nuclear data fitting

- Monte Carlo method development to take into account the covariance matrix of non-fitted parameters while the fitting of neutron cross-sections. In collaboration with C. De Saint Jean and G. Noguère (CEA research engineers).
  {articles [Ia10, Ia15], confs [Ic7, Ic8]}

Statistical sampling

- Sudy and development of some statistical sampling methods and geostatistical mapping techniques for nuclear facilities radioactivity characterization. In collaboration with N. Pérot and F. Lamadie (CEA research engineers) and N. Jeannée (Géovariances).
  {reports [Ir26, Ir27, Ir30], confs [Ic18, Ic17]}

COOPERATION and COLLABORATION PROGRAMS

- Consortium KIM (Kinematic Inverse Methods) of Institut Français du Pétrole, research consortium on seismic imaging and seismic tomography for oil exploration. Participation in 2000-2001.
  {report [Ir5]}

- Summer school 2006 of CEMRACS (Centre d'été Mathématique de Recherche Avancée en Calcul Scientifique) on the theme "Random modelling and uncertainty propagation", july-august 2006. Participation to the CEA project "Estimation of computer code quantiles" proposed by B. Iooss, N. Devictor, A. De Crécy and P. Bazin.
  {publi [Ia11], conf [Ic20], report [Ir21]}

- PAMINA (Performance Assessment Methodologies IN Application to guide the development of the safety case), European Commision project of the 6th Framework Program. Participation in 2007-2008.
  {reports [Ir33, Ir34, Ir35]}

SOFTWARE DEVELOPMENTS

Statistical libraries and softwares in R

- Software SSURFER (Sensibilités er SURFaces de réponse dans l'Environnement R), since 2004.
  {report [Ir19]}

- Co-author and maintainer of the R package "JointModeling" with M. Ribatet (PhD student CEMAGREF Lyon) : joint modelling of the mean and dispersion with GLM and GAM, since 2005.

- Maintainer of the R package "sensitivity" (author : G. Pujol, research engineer, École des Mines de Saint-Etienne) : global sensitivity analysis methods, since 2006.

Softwares

- PROUST (PROpagation UltraSonore en milieu Turbulent), CEA software for ultrasonic diagnosis in nuclear reactors, development in 1999-2000.
  {report [Ir4]}

- LEONAR (Logiciel d'Evaluation de la prObabilité de percemeNt du radier en cas d'Accident gRave), Électricité De France contract, since 2007.
  {conf [Ic19], reports [Ir23, Ir24, Ir25, Ir31, Ir32, Ir38, Ir39]}

SCIENTIFIC SUPERVISING

[Bibliographical references in pages 9-13]

Supervision of PhD thesis

- A. Marrel (2005-2008) : supervision with B. Laurent (PhD thesis director, INSA Toulouse) and M. Jullien (CEA). Subject : Implementation and use of Gaussian process metamodel for sensitivity analysis of numerical models : application to an hydrogeological transport computer code. Presentation : 3 july 2008. Publications : articles [Ia9, Ia14], confs [Ic11, Ic21]

- B. Auder (started in january 2008) : supervision with G. Biau (PhD thesis director, Univ. Paris VI) and M. Marquès (CEA). Subject : Statistical analysis and modelling of computer code functional outputs.

Training supervision

- S. Campos (INSA Toulouse 5$^{\text{th}}$ year) : *Automatic differentiation of MARGARET computer code*, 2002.

- C. Cannamela (DESS Statistique, Informatique and Techniques Numériques, Univ. Lyon 1) : *Study and comparison of the SVM technique with classical discrimination methods*, 2003.

- S. Ndao (DESS Statistique, Informatique and Techniques Numériques, Univ. Lyon 1) : *Construction of a simplified model on the computer code MOGADOR*, 2003.

- P-M. Pair (INSA Toulouse 5$^{\text{ème}}$ année) : *Construction of non linear response surfaces : comparative study of new regression techniques*, 2004.

- E. Volkova (in PhD thesis at Kurchatov Institute, Moscow) : *Sensitivity analysis of a* ^{90}Sr *transport model in saturated porous medium, application to a radwaste disposal site*, 2004-05.
  Publication : article [Ia8]

- A. Marrel (Master 2 Recherche Mathématiques Appliquées, INSA Toulouse) : *Computer code modelling with the Gaussian process model*, 2005.

- M. Ribatet (in PhD thesis at CEMAGREF Lyon) : *Mean and dispersion modelling by GLM*, 2005.
  Publication : conf [Ic6]

- B. Auder (ENSIMAG 4$^{\text{ème}}$ année) : *Global sensitivity analysis based on entropy*, 2006.
  Publication : conf [Ic15]

- V. Bakirdjian (Master 2 Pro Génie Scientifique and Informatique, Univ. de Provence) : *Geostatistical characterisation of contaminated premises*, 2007.

- D. Barthel (INSA Toulouse 5$^{\text{ème}}$ année) : *Sensitivity analysis methods based on statistical tests*, 2008.

TEACHING

Cours

- *Probability courses* (6 hours), Master 2 "Quantitative methods for Geosciences", École du Pétrole et des Moteurs, Rueil-Malmaison, october 1997.

TD

- *Initiation to R* (4 hours), License 3, IUP "Statistics computer science for decision", Université Toulouse III, Toulouse, february 2009.

Doctoral courses

- *Concepts and methods of geostatistics,* Seminar "Spatial statistics for industry and marketing" organized by C. Thomas-Agnan, Master 2 Pro "Statistics and Econometry", GREMAQ, Toulouse, february 2006.

- With J. Baccou (IRSN Cadarache) : *Uncertainty analyses : new insights for environmental risk,* Seminar "Uncertainties in industrial risks" organized by E. Pardoux, Écoles Centrales (option mathématiques), Chateau-Gombert, Marseille, february 2007.

Professional courses

- Co-organisation of the formation "Uncertainty management" of the Institut de Maîtrise des Risques, Paris. Animation of the session "sensitivity analysis". Since 2007.

RESPONSABILITIES

Participations to PhD thesis jurys

- Examiner of M. Petelet PhD thesis, Sensitivity analysis of thermomechanical models in welding simulation, 30 october 2007, Univ. de Bourgogne.

- Examiner of A. Marrel PhD thesis, Implementation and use of Gaussian process metamodel for sensitivity analysis of numerical models : application to an hydrogeological transport computer code, 3 july 2008, INSA Toulouse.

Refereeings for international journals

- *Pure and Applied Geophysics* (1 article in 2001),
- *Journal of Flow Measurement and Instrumentation* (1 article in 2004),
- *Experiments in Fluids* (1 article in 2006),
- *Acta Acustica* (1 article in 2007).

Member of conference scientific committees

- ENBIS-EMSE Spring Meeting 2009, "Design and Analysis of Computer Experiments", Saint-Etienne, France, july 2009.

CONGRESSES and SEMINARS

Oral or poster session during congress

- Journées de Géostatistique, Fontainebleau, France, june 1998.
- 135th Meeting of the Acoustical Society of America, Seattle, USA, june 1998.
- 68th Annual Meeting of the Society of Exploration Geophysicists, New Orleans, september 1998.
- 4th International Conf. on Sensitivity Analysis of Model Output, Santa Fe, USA, march 2004.
- Journées Techniques de la CETAMA, Montpellier, France, october 2005.
- XXXVIIIèmes Journées de Statistique, Clamart, France, may-june 2006.
- 5th International Conf. on Sensitivity Analysis of Model Output, Budapest, Hungary, june 2007.
- VI Colloqium Chemiometricum Mediterraneum, Saint Maximin, France, september 2007.
- ESREL 2008 Annual Conference, Valence, Espagne, septembre 2008.
- 35rd ESReDA Seminar, Marseille, France, novembre 2008.

Invitations to conferences

- *Concepts et méthodes de la statistique,* Summer school CEA-EDF-INRIA, Uncertainty management in numerical simulation, Saint Lambert des Bois, june 2005.
- *Concepts et méthodes de la géostatistique,* Journées Techniques de la CETAMA, Echantillonnage et caractérisation "Du prélèvement à l'analyse", Montpellier, october 2005.
- *Analyses de sensibilité globales de modèles complexes : quelques applications dans le nucléaire,* conference organized by INSA Toulouse, Planification d'expériences et analyse d'incertitudes pour les gros codes numériques : approches stochastiques, Toulouse, february 2006.
- *Estimation de quantiles de codes de calcul,* Journées DIF "Incertitudes et simulation", CEA/DAM, Bruyères-le-Châtel, october 2007.
- *Uncertainty and reliability study of the creep law to assess the fuel cladding behavior of PWR spent fuel assemblies during interim dry storage,* Session "Uncertainty in industrial practice" organized by E. de Rocquigny, ESREL 2008 Conference, Valence, Espagne, septembre 2008.

Presentations during seminars

- *Statistical tomography for reflection seismics : stochastic velocity model estimation,* AGIP research center, Milano, Italy, january 1999.
- *Etude des temps de trajet en milieu aléatoire, inversion probabiliste en sismique,* Seminar of the Laboratoire de Mécanique des Fluides and d'Acoustique, École Centrale de Lyon, Lyon, march 2000.
- *Tomographie statistique en sismique réflexion : estimation d'un modèle de vitesse stochastique,* Seminar of the Division Géophysique, Institut Français du Pétrole, Rueil-Malmaison, march 2001.
- *Misfits statistics and data decimation in traveltime inversion,* Consortium KIM annual meeting, Institut Français du Pétrole, Rueil Malmaison, december 2001.
- *Traitement des incertitudes,* Institut Français du Pétrole, Rueil Malmaison, november 2004.
- *Probabilistic methods for uncertainty studies,* Institut Kurchatov, Moscow, Russia, june 2005.
- *Analyse de sensibilité globale de modèles numériques à paramètres incontrôlables,* Institut Français du Pétrole, Solaize, april 2006.
- *Estimation de quantiles de code, applic. thermohydraulique,* IMPEC Seminar, CEA Saclay, october 2006.
- *Traitement des incertitudes, quelques applications en ingénierie nucléaire,* Centre de Géostatistique, École des Mines de Paris, Fontainebleau, october 2006.
- *Traitement des incertitudes, quelques applications en ingénierie nucléaire,* Workshop "Uncertainty management", EADS, Toulouse, october 2006.
- *Estimation de quantiles de code, application thermohydraulique,* Département de Mathématiques, École des Mines de Saint Etienne, december 2006.
- With M. Marquès (CEA Cadarache) : *Sensitivity analysis on the PRHRS performance indicators,* Meeting of the IAEA Project "Status and prospects of development and application of innovative reactor concepts for developing contries", Univ. of Pisa, Italy, february 2007.
- *Gaussian processes as metamodels*, Workshop "Sensitivity analysis", Joint Research Centre, Ispra, Italy, february 2008.
- *Two problems in metamodelling and sensitivity analysis of complex numerical models : high dimensional inputs and stochastic computer code*, Seminar for the Operations Research Group of CentER, Tilburg University, Tilburg, The Netherlands, may 2008.

SCIENTIFIC ANIMATION

Animation activities

- Treasurer 1997-98 of ABCTEM (Association Bellifontaine des Chercheurs and Thésards de l'École des Mines).

- Since 2006, creation and co-animation of IMPEC network (Incertitudes, Métamodèles and Plans d'Expériences pour les Codes), internal research group of CEA, with J-M. Martinez and F. Gaudier (CEA Saclay research engineer). Webmaster (with F. Gaudier) of intranet site `http://www-impec.cea.fr`

- Since 2006, member of the Working Group "uncertainties" of the Institut de Maîtrise des Risques (leader : E. de Rocquigny, EDF R&D),
URL : `http://www.imdr-sdf.asso.fr/v2/extranet/index.php?page=gtr`

- Since 2007, member of the organizing committee of the GdR CNRS MASCOT NUM ("Méthodes d'Analyse Stochastique pour le COdes and Traitements NUMériques"). Leaders : F. Gamboa (Univ. Paul Sabatier, Toulouse) and F. Wahl (IFP Lyon). Webmaster of internet site `http://www.gdr-mascotnum.fr`

- Since 2008, member of the ESRA (European Safety and Reliability Association) Technical Committee on Uncertainty and Sensitivity Analysis, URL : `http://www.esrahomepage.org/uncertainty.aspx`

Seminar organisations

- Colloque Mines 98, Presentations of PhD students of the École des Mines, École des Mines de Paris, Fontainebleau, may 1998.

- PAN (Plan d'Action Neuronal) seminar, "Numerical experimental designs", CEA Cadarache, december 2005.

- IMPEC (Incertitudes Métamodèle and Plans d'Expériences pour les Codes) seminar, "Kriging methods and quantile estimation", CEA Saclay, october 2006.

- Workroom during the GdR CNRS MASCOT NUM meeting (Méthodes d'Analyse Stochastique pour les COdes and traitement NUMériques) with J-C. Fort (Univ. Paul Sabatier, Toulouse), "Functional data problematics in design and analysis of computer code experiments", IFP Solaize, march 2007.
URL : `http://www.gdr-mascotnum.fr/rencontres/rencontres_mars2007/rencontres_mars2007.html`

- IMPEC seminar, "Environmental risk", CEA Cadarache, june 2007.

- IMPEC seminar, CEA Cadarache, october 2008.

- Meeting between GdR MASCOT-NUM, GT "Incertitudes" of IMdR and Master students.
URL : `http://www.gdr-mascotnum.fr/doku.php?id=thesismeeting`

Conference organisations

- Assistant during the Numerical Analysis Summer School CEA-EDF-INRIA 2005, "Uncertainty management in numerical simulation", Saint Lambert des Bois, june 2005. Coordinator : J-M. Martinez (CEA).

- Organisation of GdR MASCOT NUM meeting, CEA Cadarache, 12-14 march 2008.
URL : `http://www.gdr-mascotnum.fr/rencontres/rencontres_mars2008/progMASCOTCadmars08.pdf`

Participations to round tables

- *Uncertainty modelling*, Journées MAS de la SMAI, Lille, september 2006.

Annexe B

Abstracts des publications à revue

B. Iooss. Seismic reflection traveltimes in two-dimensional statistically anisotropic random media. *Geophysical Journal International*, 135 :999-1010, 1998.

Velocity estimation remains one of the main problems when imaging the subsurface with seismic reflection data. Traveltime inversion enables us to obtain large scale structures of the velocity field and the position of seismic reflectors. However, as the media currently under study are becoming more and more complex, we need to know the finer scale structures. The problem is that below a certain range of velocity heterogeneities, deterministic methods become difficult to use, so we turn to a probabilistic approach. With this in view, we characterize the velocity field as a random field defined by its first and second statistical moments. Usually, a seismic random medium is defined as a homogeneous velocity background perturbed by a small random field that is assumed to be stationary. Thus, we make a link between such a random velocity medium (together with a simple reflector) and seismic reflection traveltimes. Assuming that the traveltimes are ergodic, we use 2-D seismic reflection geometry to study the decrease in the statistical traveltime fluctuations, as a function of the offset (the source-receiver distance). Our formulas are based on the Rytov approximation and the parabolic approximation for acoustic waves. The validity and the limits are established for both of these approximations in statistically anisotropic random media. Finally, theoretical inversion procedures are developed for the horizontal correlation structure of the velocity heterogeneities for the simplest case of a horizontal reflector. Synthetic seismograms are then computed (on particular realisations of random media) by simulating scalar wave propagation via finite difference algorithms. There is good agreement between the theoretical and experimental results.

M. Touati, B. Iooss and A. Galli. Quantitative control of migration : a geostatistical attempt. *Mathematical Geology*, 31 :277-295, 1999.

This paper is devoted to a geostatistical attempt at modeling migration errors when localizing a reflector in the ground. Starting with a probabilistic velocity model and choosing the simple geometrical optics background for the wave propagation in such media, we give the expression of the errors. This may be quantified provided the covariance of the velocity field is known. Variance of arrival times at constant offset is related to the covariance of the velocity field at hand. A practical application is given in the same paragraph. After that we give a typical schema for migration and uncertainty modeling : starting with seismic data, we make the weak seismic inversion. We then obtain the covariance of the velocity field that we use for simulating migration errors. The main issues of this methodology are discussed in the last paragraph.

B. Iooss, Ph. Blanc-Benon and C. Lhuillier. Statistical moments of travel times at second order in isotropic and anisotropic random media. *Waves in Random Media*, 10 :381-394, 2000.

We study the high-frequency propagation of acoustic plane and spherical waves in random media. With the geometrical optics and the perturbation approach, we obtain the travel-time mean and travel-time variance at the second order. The main hypotheses are the Gaussian distribution of the acoustic speed perturbation and a factorized form for its correlation function. The second order travel-time variance explains the non linear behavior at large propagation distance observed with numerical experiments based on ray tracing. Usually, homogeneity and isotropy of the refractive index are considered. Using the geometrical anisotropy hypothesis we extend the theory to a general class of statistically anisotropic random media.

B. Iooss, C. Lhuillier and H. Jeanneau. Numerical simulation of transit-time ultrasonic flowmeters : uncertainties due to fluid turbulence. *Ultrasonics*, 40 :1009-1015, 2002.

Flowmeter measurement using the ultrasonic transit-time method is based on the apparent difference of the sound velocity in the flow direction and in the opposite direction. This method gives a flow velocity averaged along a particular acoustical path. To convert this path velocity to a velocity averaged over the entire cross section of the flowing medium, the knowledge of the flow velocity profile is essential. However, the acoustical paths joining the two transducers are supposed to be straight and fluid turbulence phenomena are neglected. In this paper, we describe a numerical procedure to estimate the uncertainties due to these approximations in the case of fully-developed turbulence. The ultrasonic propagation is modelled in 2-D moving inhomogeneous media *via* a ray tracing algorithm. Influence of mean profiles of temperature and velocity is studied on simple examples. Fluid temperature fluctuations and fluid velocity turbulence are considered in the stochastic framework to obtain average uncertainties on the measurements of the liquid flow rate.

B. Iooss, D. Geraets, T. Mukerji, Y. Samuelides, M. Touati and A. Galli. Inferring the statistical distribution of velocity heterogeneities by statistical traveltime tomography. *Geophysics*, 68(5) :1714-1730, 2003.

Understanding the internal heterogeneities of reservoirs is one of the key issues in better recovery and efficient reservoir management. Seismic data is widely used to map subsurface heterogeneities. These heterogeneities can include variations in wave velocity and rock density, which can be used to make interpretations about variations in reservoir properties such as porosity, lithofacies, and fluids. This paper describes a statistical tomography method to infer the spatial statistics of subsurface velocity heterogeneities from seismic data. We consider an acoustic wave propagating in a medium represented as a single macro-model superimposed on statistically stationary random velocity perturbations. While the macro-model is retrieved by classical seismic methods, the picked traveltimes and their fluctuations are used to estimate the covariance function of the spatially varying velocity perturbations. We present a formulation based on ray-theoretical results and describe two algorithms : one using the prestack traveltimes and the other using the stacking velocities. The methods are tested with synthetic seismic reflection data in an idealized medium (with a Gaussian spatial covariance), and with synthetic transmission data in a more geologically realistic medium. Then, the two algorithms are applied on real data. The estimates of the spatial statistics obtained from inverting the traveltime statistics match reasonably well with the true parameters of the heterogeneous media.

B. Iooss, F. Van Dorpe and N. Devictor. Response surfaces and sensitivity analyses for an environmental model of dose calculations. *Reliability Engineering and System Safety*, 91 :1241-1251, 2006.

A parametric sensitivity analysis is carried out on GASCON, a radiological impact software describing the radionuclides transfer to the man following a chronic gas release of a nuclear facility. An effective dose received by age group can thus be calculated according to a specific radionuclide and to the duration of the release. In this study, we are concerned by eighteen output variables, each depending of approximately fifty uncertain input parameters. First, the generation of one thousand Monte-Carlo simulations allows us to calculate correlation coefficients between input parameters and output variables, which give a first overview of important factors. Response surfaces are then constructed in polynomial form, and used to predict system responses at reduced computation time cost ; this response surface will be very useful for global sensitivity analysis where thousands of runs are required. Using the response surfaces, we calculate the total sensitivity indices of Sobol by the Monte-Carlo method. We demonstrate the application of this method to one site of study and to one reference group near the nuclear research Center of Cadarache (France), for two radionuclides : iodine 129 and uranium 238. It is thus shown that the most influential parameters are all related to the food chain of the goat's milk, in decreasing order of importance : dose coefficient "effective ingestion", goat's milk ration of the individuals of the reference group, grass ration of the goat, dry deposition velocity and transfer factor to the goat's milk.

F. Van Dorpe, B. Iooss, V. Semenov, O. Sorokovikova, A. Fokin and Y. Margerit. Atmospheric transfer modeling with 3D Lagrangian dispersion codes compared with SF6 tracer experiments at regional scale. *Science and Technology of Nuclear Installations*, Volume 2007, Article ID 30863, 13 pages, doi :10.1155/2007/30863, 2007.

The results of four gas tracer experiments of atmospheric dispersion on a regional scale are used for the benchmarking of two atmospheric dispersion modeling codes, MINERVE-SPRAY (CEA), and NOSTRADAMUS (IBRAE). The main topic of this comparison is to estimate the Lagrangian code capability to predict the radionuclide atmospheric transfer on a large field, in the case of risk assessment of nuclear power plant for example. For the four experiments, the results of calculations show a rather good agreement between the two codes, and the order of magnitude of the concentrations measured on the soil is predicted. Simulation is best for sampling points located ten kilometers from the source, while we note a divergence for more distant points results (difference in concentrations by a factor 2 to 5). This divergence may be explained by the fact that, for these four experiments, only one weather station (near the point source) was used on a field of 10 000 km2, generating the simulation of a uniform wind field on all the zone of calculation.

E. Volkova, B. Iooss and F. Van Dorpe. Global sensitivity analysis for a numerical model of radionuclide migration from the "RRC Kurchatov Institute" radwaste disposal site. *Stochastic Environmental Research and Risk Assessment*, 22 :17-31, 2008.

Today, in different countries, there exist sites with contaminated groundwater formed as a result of inappropriate handling or disposal of hazardous materials or wastes. Numerical modeling of such sites is an important tool for a correct prediction of contamination plume spreading and an assessment of environmental risks associated with the site. Many uncertainties are associated with a part of the parameters and the initial conditions of such environmental numerical models. Statistical techniques are useful to deal with these uncertainties. This paper describes the methods of uncertainty propagation and global sensitivity analysis that are applied to a numerical model of radionuclide migration in a sandy aquifer in the area of the RRC "Kurchatov Institute" radwaste disposal site in Moscow, Russia. We consider twenty uncertain input parameters of the model and twenty output variables (contaminant concentration in the observation wells predicted by the model for the end of 2010). Monte Carlo simulations allow calculating uncertainty in the output values and analyzing the linearity and the monotony of the relations between input and output variables. For the non monotonic relations, sensitivity analyses are classically done with the Sobol sensitivity indices. The originality of this study is the use of modern surrogate models (called response surfaces), the boosting regression trees, constructed for each output variable, to calculate the Sobol indices by the Monte Carlo method. It is thus shown that the most influential parameters of the model are distribution coefficients and infiltration rate in the zone of strong pipe leaks on the site. Improvement of these parameters would considerably reduce the model prediction uncertainty.

A. Marrel, B. Iooss, F. Van Dorpe and E. Volkova. An efficient methodology for modeling complex computer codes with Gaussian processes. *Computational Statistics and Data Analysis*, 52 :4731-4744, 2008.

Complex computer codes are often too time expensive to be directly used to perform uncertainty propagation studies, global sensitivity analysis or to solve optimization problems. A well known and widely used method to circumvent this inconvenience consists in replacing the complex computer code by a reduced model, called a metamodel, or a response surface that represents the computer code and requires acceptable calculation time. One particular class of metamodels is studied : the Gaussian process model that is characterized by its mean and covariance functions. A specific estimation procedure is developed to adjust a Gaussian process model in complex cases (non linear relations, highly dispersed or discontinuous output, high dimensional input, inadequate sampling designs, ...). The efficiency of this algorithm is compared to the efficiency of other existing algorithms on an analytical test case. The proposed methodology is also illustrated for the case of a complex hydrogeological computer code, simulating radionuclide transport in groundwater.

G. Noguere, D. Bernard, C. De Saint Jean, F. Gunsing, B. Iooss, K. Kobayashi, S. Mughabghab and P. Siegler. Propagation of the ^{237}Np nuclear data uncertainties in integral calculations by Monte Carlo techniques. *Nuclear Science and Engineering*, 160 :108-122, 2008.

A new method to produce covariance or dispersion matrices for the resonance parameters of the neutron cross sections was developed. The technique uses the resonance shape analysis in association with a Monte-Carlo treatment of the uncertainties. The method was implemented in the error propagation tool MCFIT. This program provides a user-friendly textual interface for the shape analysis code REFIT. It was designed to take into account the main sources of uncertainties involved in time-of-flight measurements. Its capability is illustrated with the simultaneous analysis of ^{237}Np capture and transmission data. The covariance matrix obtained in this work was used in the interpretation of oscillation measurements of ^{237}Np samples carried out at in the Minerve reactor located at Cadarache.

C. Cannamela, J. Garnier and B. Iooss. Controlled stratification for quantile estimation. *Annals of Applied Statistics*, 2 :1554-1580, 2008.

In this paper we propose and discuss variance reduction techniques for the estimation of quantiles of the output of a complex model with random input parameters. These techniques are based on the use of a reduced model, such as a metamodel or a response surface. The reduced model can be used as a control variate ; or a rejection method can be implemented to sample the realizations of the input parameters in prescribed relevant strata ; or the reduced model can be used to determine a good biased distribution of the input parameters for the calibration of an importance sampling strategy. The different strategies are analyzed, the asymptotic variances are computed and compared, which show the benefit of an adaptive controlled stratification method. This method is applied to a real example (computation of the peak cladding temperature during a large-break loss of coolant accident in a nuclear reactor).

A. Marrel, B. Iooss, B. Laurent and O. Roustant. Calculations of Sobol indices for the Gaussian process metamodel. *Reliability Engineering and System Safety*, 94 :742-751, 2009.

Global sensitivity analysis of complex numerical models can be performed by calculating variance-based importance measures of the input variables, such as the Sobol indices. However, these techniques, requiring a large number of model evaluations, are often unacceptable for time expensive computer codes. A well known and widely used decision consists in replacing the computer code by a metamodel, predicting the model responses with a negligible computation time and rending straightforward the estimation of Sobol indices. In this paper, we discuss about the Gaussian process model which gives analytical expressions of Sobol indices. Two approaches are studied to compute the Sobol indices : the first based on the predictor of the Gaussian process model and the second based on the global stochastic process model. Comparisons between the two estimates, made on analytical examples, show the superiority of the second approach in terms of convergence and robustness. Moreover, the second approach allows to integrate the modeling error of the Gaussian process model by directly giving some confidence intervals on the Sobol indices. These techniques are finally applied to a real case of hydrogeological modeling.

B. Iooss and M. Ribatet. Global sensitivity analysis of computer models with functional inputs. *Reliability Engineering and System Safety*, in press, 2009.

Global sensitivity analysis is used to quantify the influence of uncertain model inputs on the response variability of a numerical model. The common quantitative methods are appropriate with computer codes having scalar model inputs. This paper aims at illustrating different variance-based sensitivity analysis techniques, based on the so-called Sobol's indices, when some model inputs are functional, such as stochastic processes or random spatial fields. In this work, we focus on large cpu time computer codes which need a preliminary metamodeling step before performing the sensitivity analysis. We propose the use of the joint modeling approach, i.e., modeling simultaneously the mean and the dispersion of the code outputs using two interlinked Generalized Linear Models (GLM) or Generalized Additive Models (GAM). The "mean model" allows to estimate the sensitivity indices of each scalar model inputs, while the "dispersion model" allows to derive the total sensitivity index of the functional model inputs. The proposed approach is compared to some classical sensitivity analysis methodologies on an analytical function. Lastly, the new methodology is applied to an industrial computer code that simulates the nuclear fuel irradiation.

C. De Saint Jean, G. Noguere, B. Habert and B. Iooss. A Monte Carlo approach of nuclear model parameters uncertainties propagation. *Nuclear Science and Engineering*, 161 : 363-370, 2009.

The evaluation of neutron cross sections in the low energy range (electron volt, mega-electronvolt) is based on formal nuclear models having different type of parameters. Some of them may be fitted to reproduce experimental datasets giving rise to an adjusted covariance matrix. In this paper, a Monte-Carlo method is presented to properly consider the influence of the remaining parameters, having a priori uncertainties, on the fitted parameters covariances. This method is based on an exact mathematical description using conditional probabilities. To explain the key points of the methodology, an academic example of average parameters evaluation in the unresolved resonance range is presented using a Hauser-Feshbach model calculations.

B. Iooss, M. Ribatet and A. Marrel. Global sensitivity analysis of stochastic computer models with generalized additive models. *Technometrics*, submitted, 2006.

The global sensitivity analysis, used to quantify the influence of uncertain input parameters on the response variability of a numerical model, is applicable to deterministic computer codes (for which the same set of input parameters gives always the same output value). This paper proposes a global sensitivity analysis method for stochastic computer codes (having a variability induced by some uncontrollable parameters). The mean and dispersion of the code outputs are modeled by two interlinked Generalized Additive Models (GAM). The "mean" model allows to obtain the controllable parameters sensitivity indices, while the "dispersion" model allows to obtain the uncontrollable parameters ones. The relevance of the proposed model is analyzed with two case studies. Results show that the joint modeling approach leads to more accurate sensitivity index estimations, especially for the joint GAM model.

Bibliographie générale

[1] R. Ababou, A.C. Bagtzoglou, and E.F. Wood. On the condition number of covariance matrices in kriging, estimation, and simulation of random fields. *Mathematical Geology*, 26 :99–133, 1994.

[2] P. Abrahamsen. A review of Gaussian random fields and correlation functions. Technical Report 878, Norsk Regnesentral, 1994.

[3] K. Aki. Scattering of P waves under the Montana LASA. *Journal of Geophysical Research*, 78(8) :1334–1346, 1973.

[4] T. Andreeva and W. Durgin. Experimental investigation of the travel-time variance of an acoustic wave propagating through the grid-generated turbulence. *Waves in Random and Complex Media*, 15 :365–374, 2005.

[5] L. C. Andrews and R. L. Phillips. *Laser beam propagation through random media*. SPIE Optical Engineering Press, 1998.

[6] C. Andrieu-Renaud. *Fiabilité mécanique des structures soumises à des phénomènes physiques dépendant du temps*. Thèse de l'Université Blaise Pascal - Clermont II, 2002.

[7] A. Antoniadis. Analysis of variance on function spaces. *Math. Operationsforsch. u. Statist., ser. statist.*, 15 :59–71, 1984.

[8] A. Antoniadis and C. Lavergne. Variance function estimation in regression by wavelet methods. In A. Antoniadis and G. Oppenheim, editors, *Wavelets and statistics*. Springer, 1995.

[9] M. Asch, W. Kohler, G. Papanicolaou, M. Postel, and B. White. Frequency content of randomly scattered signals. *SIAM Review*, 33 :519–625, 1991.

[10] J-M. Azaïs and J-M. Bardet. *Le modèle linéaire par l'exemple*. Dunod, 2005.

[11] A.M. Baig and F.A. Dahlen. Statistics of traveltimes and amplitudes in random media. *Geophysical Journal International*, 158 :187–208, 2004.

[12] F. Bailly, J. F. Clouet, and J. P. Fouque. Parabolic and white noise approximation for waves in random media. *SIAM Journal on Applied Mathematics*, 5 :1445–1470, 1996.

[13] G. Bal. *Lecture notes - Waves in random media*. Department of Applied Physics and Applied Mathematics, Columbia University, New York, NY, USA, 2006.

[14] Yu. N. Barabanenkov, Yu. A. Kravtsov, S.M. Rytov, and V.I. Tatarskii. Status of the theory of propagation of waves in a randomly inhomogeneous medium. *Soviet Physics*, 13 :551–575, 1971.

[15] J. Baroth. *Propagation d'incertitudes dans des modèles mécaniques non linéaires par une méthode d'éléments finis stochastiques*. Thèse de l'Université Blaise Pascal - Clermont II, 2005.

[16] L.S. Bastos and A. O'Hagan. Diagnostics for Gaussian process emulators. *Technometrics*, submitted, 2008.

[17] R.A. Bates, R.S. Kenett, D.M. Steinberg, and H.P. Wynn. Achieving robust design from computer simulations. *Quality Technology and Quantitative Management*, 3 :161–177, 2006.

[18] M.S. Bazaraa, H.D. Sherali, and C.M. Shetty. *Nonlinear programming*. John Wiley & Sons, Inc, 1993.

[19] P. Bazin. *Détermination d'un fractile élevé par l'intermédiaire d'une surface de réponses*. Note Technique CEA/DEN/GRE/DER/SSTH/LDAS DO 114 04/12/07, 2007.

[20] M. Berveiller. *Eléments finis stochastiques : approches intrusive et non intrusive pour des analyses de fiabilité*. Thèse de l'Université Blaise Pascal - Clermont II, 2005.

[21] B. Bettonvil and J.P.C. Kleijnen. Searching for important factors in simulation models with many factors : Sequential bifurcation. *European Journal of Operational Research*, 96 :180–194, 1996.

[22] M. Bilodeau, F. Meyer, and M. Schmitt, editors. *Space, structure, and randomness*. Springer, 2005.

[23] P. Blanc-Benon, D. Juvé, and G. Comte-Bellot. Occurence of caustics for high-frequency waves propagating through turbulent fields. *Theoretical and Computational Fluid Dynamics*, 2 :271–278, 1991.

[24] P. Blanc-Benon, D. Juvé, M. Karweit, and G. Comte-Bellot. Simulation numérique de la propagation des ondes acoustiques à travers une turbulence cinématique. *Journal d'Acoustique*, 3 :1–8, 1990.

[25] P. Blanc-Benon, D. Juvé, V. E. Ostashev, and R. Wandelt. On the appearance of caustics for plane sound-wave propagation in moving random media. *Waves in Random Media*, 5 :183–199, 1995.

[26] G.E. Box and N.R. Draper. *Empirical model building and response surfaces.* Wiley Series in Probability and Mathematical Statistics. Wiley, 1987.

[27] W. Boyse and J. B. Keller. Short acoustic, electromagnetic, and elastic waves in random media. *Journal of the Optical Society of America*, 12 :380–389, 1995.

[28] B. Briand. *Construction d'arbres de discrimination pour expliquer les niveaux de contamination radioactive des végétaux.* Thèse de l'Université Montpellier II, 2008.

[29] D. Bursztyn and D.M. Steinberg. Screening experiments for dispersion effects. In A. Dean and S. Lewis, editors, *Screening - Methods for experimentation in industry, drug discovery and genetics.* Springer, 2006.

[30] D. Busby, T. Romary, M. Feraille, and S. Touzani. An integrated approach for uncertainty and sensitivity analysis in reservoir forecasting. *Computational Geosciences*, submitted, 2008.

[31] D.G. Cacuci. Sensitivity theory for nonlinear systems. I. Nonlinear functional analysis approach. *Journal of Mathematical Physics*, 22 :2794, 1981.

[32] D.G. Cacuci. *Sensitivity and uncertainty analysis - Theory.* Chapman & Hall/CRC, 2003.

[33] D.G. Cacuci, M. Ionescu-Bujor, and I.M. Navon. *Sensitivity and uncertainty analysis - Applications to large-scale systems.* Chapman & Hall/CRC, 2005.

[34] K. Campbell, M.D. McKay, and B.J. Williams. Sensitivity analysis when model ouputs are functions. *Reliability Engineering and System Safety*, 91 :1468–1472, 2006.

[35] C. Cannamela. *Apport des méthodes probabilistes dans la simulation du comportement sous irradiation du combustible à particules.* Thèse de l'Université Denis Diderot - Paris VII, 2007.

[36] C. Cannamela, J. Garnier, and B. Iooss. Controlled stratification for quantile estimation. *Annals of Apllied Statistics*, 2 :1554–1580, 2008.

[37] W. Castaings. *Analyse de sensibilité et estimation de paramètres pour la modélisation hydrologique : potentiel et limitations des méthodes variationnelles.* Thèse de l'Université Joseph Fourier, Grenoble 1, 2007.

[38] S. Chandrasekhar. *Radiative transfer.* Van Nostrand, NJ, 1950.

[39] V.C.P. Chen, K-L. Tsui, R.R. Barton, and M. Meckesheimer. A review on design, modeling and applications of computer experiments. *IIE Transactions*, 38 :273–291, 2006.

[40] W. Chen, R. Jin, and A. Sudjianto. Analytical metamodel-based global sensitivity analysis and uncertainty propagation for robust design. *Journal of Mechanical Design*, 127 :875–886, 2005.

[41] L. A. Chernov. *Wave propagation in a random medium.* Mc Graw-Hill, New York, 1960.

[42] J-P. Chilès and P. Delfiner. *Geostatistics : Modeling spatial uncertainty.* Wiley, New-York, 1999.

[43] M. Claeys-Bruno, M. Dobrijevic, R. Cela, R. Phan-Tan-Luu, and M. Sergent. Supersaturated design : construction, comparison and interpretation. In *VI Colloquium Chemiometricum Mediterraneum*, Saint Maximin La Sainte Baume, France, september 2007.

[44] N.A.C. Cressie. *Statistics for spatial data.* Wiley Series in Probability and Mathematical Statistics. Wiley, 1993.

[45] T. Crestaux, J-M. Martinez, and O. Le Maître. Polynomial chaos expansions for uncertainties quantification and sensitivity analysis. *Reliability Engineering and System Safety*, doi :10.1016/j.ress.2008.10.008, in press, 2009.

[46] H. Cukier, R.I. Levine, and K. Shuler. Nonlinear sensitivity analysis of multiparameter model systems. *Journal of Computational Physics*, 26 :1–42, 1978.

[47] C. Currin, T. Mitchell, M. Morris, and D. Ylvisaker. Bayesian prediction of deterministic functions with applications to the design and analysis of computer experiments. *Journal of the American Statistical Association*, 86(416) :953–963, 1991.

[48] S. Da Veiga. *Analyse d'incertitudes et de sensibilité - Applications aux modèles de cinétique chmique.* Thèse de l'Université Paul Sabatier - Toulouse III, 2007.

[49] S. Da Veiga and F. Gamboa. Efficient estimation of non linear conditional functionals of a density. *Annals of Statistics*, submitted, 2008. Available at URL : http ://hal.archives-ouvertes.fr/hal-00266110/fr/.

[50] S. Da Veiga, F. Wahl, and F. Gamboa. Local polynomial estimation for sensitivity analysis for models with correlated inputs. *Technometrics*, submitted, 2008. Available at URL : http ://fr.arxiv.org/abs/0803.3504.

[51] H.A. David and H.N. Nagaraja. *Order statistics.* Wiley, New-York, third edition, 2003.

[52] E. De Rocquigny. La maîtrise des incertitudes dans un contexte industriel - 1ère partie : une approche méthodologique globale basée sur des exemples. *Journal de la Société Française de Statistique*, 147(3) :33–71, 2006.

[53] E. De Rocquigny. La maîtrise des incertitudes dans un contexte industriel - 2ème partie : revue des méthodes de modélisation statistique physique et numérique. *Journal de la Société Française de Statistique*, 147(3) :73–106, 2006.

[54] E. De Rocquigny, N. Devictor, and S. Tarantola, editors. *Uncertainty in industrial practice.* Wiley, 2008.

[55] A. Dean and S. Lewis, editors. *Screening - Methods for experimentation in industry, drug discovery and genetics.* Springer, 2006.

[56] D. den Hertog, J.P.C. Kleijnen, and A.Y.D. Siem. The correct Kriging variance estimated by bootstrapping. *European Journal of Operational Research*, 57 :400–409, 2006.

[57] N. Devictor. *Fiabilité et mécanique : méthodes FORM/SORM et couplages avec des codes d'éléments finis par des surfaces de réponses adaptatives*. Thèse de l'Université Blaise Pascal - Clermont II, 1996.

[58] P.J. Diggle and P.J. Ribeiro. *Model-based geostatistics*. Springer, 2007.

[59] O. Ditlevsen and H.O. Madsen, editors. *Structural reliability methods*. Wiley & Sons, 1996.

[60] J-J. Droesbecke, J. Fine, and G. Saporta, editors. *Plans d'expériences (Applications à l'entreprise)*. Technip, Paris, 1998.

[61] W. Durgin and T. Andreeva. Experimental investigation of grid-generated turbulence using ultrasonic travel-time technique. *WIT Transactions on Engineering Sciences*, 52 :143–152, 2006.

[62] K-T. Fang, R. Li, and A. Sudjianto. *Design and modeling for computer experiments*. Chapman & Hall/CRC, 2006.

[63] V. Feuillard. *Analyse d'une base de données pour la calibration d'un code de calcul*. Thèse de l'Université Pierre et Marie Curie - Paris VI, 2007.

[64] D. Fiorina. *Application de la méthode de sommation de faisceaux gaussiens à l'étude de la propagation ultrasonore en milieu turbulent*. Thèse de l'École Centrale de Lyon, France, 1998.

[65] S. M. Flatté, R. Dashen, W. H. Munk, K. M. Watson, and F. Zachariasen. *Sound transmission through a fluctuating ocean*. Cambridge University Press, New York, 1979.

[66] S. M. Flatté and R. S. Wu. Small-scale structure in the lithosphere and asthenosphere deduced from arrival time and amplitude fluctuations at NORSAR. *Journal of Geophysical Research*, 93(B6) :6601–6614, 1988.

[67] J-P. Fouque, J. Garnier, G. Papanicolaou, and K. Solna. *Wave propagation and time reversal in randomly layered media*. Springer, 2007.

[68] S. Franchini, A. Sanz-Andrés, and A. Cuerva. Measurement of velocity in rotational flows using ultrasonic anemometry : The flowmeter. *Experiments in Fluids*, 42 :903–911, 2007.

[69] A. Frankel and R. W. Clayton. Finite difference simulations of scattering : implications for the propagation of short-period seismic waves in the crust and models of crustal heterogeneity. *Journal of Geophysical Research*, 91(B6) :6465–6489, 1986.

[70] U. Frisch. Wave propagation in random media. In A. T. Bharucha-Reid, editor, *Probabilistic Methods in Applied Mathematics*, volume 1, pages 75–198. Academic Press, New York, 1968.

[71] U. Frisch. *Turbulence*. Cambridge University Press, 1995.

[72] S. Gazut. *Conception et mise en øeuvre de nouvelles méthodes d'élaboration de plans d'expériences pour l'apprentissage de modèles non linéaires*. Thèse de l'Université Paris Sud XI, 2007.

[73] S. Gazut, J-M. Martinez, G. Dreyfus, and Y. Oussar. Towards the optimal design of numerical experiments. *IEEE Transactions on Neural Networks*, 19 :874–882, 2008.

[74] D. Geraets. *Modélisation stochastique de champs de vitesse géophysique en exploration pétrolière*. Thèse de l'Ecole des Mines de Paris, 2002.

[75] D. Geraets and A. Galli. Statistical traveltime tomography in terms of stacking velocity. *Pure and Applied Geophysics*, 159 :1617–1635, 2002.

[76] D. Geraets, A. Galli, and P. Ruffo. Statistical characterization of a random velocity field using stacking velocity profiles. *Mathematical Geosciences*, 39 :513–527, 2007.

[77] D. Geraets, C. Lajaunie, and P. Ruffo. Conditioned simulations of random velocity fields. *Mathematical Geosciences*, 40 :831–844, 2008.

[78] O.A. Godin, V.U. Zavorotny, A.G. Voronovich, and V.V. Goncharov. Refraction of sound in a horieontally inhomogeneous, time-dependent ocean. *IEEE Journal of Oceanic Engineering*, 31 :384–401, 2006.

[79] R.B. Gramacy. tgp : An R package for Bayesian non stationary semiparametric nonlinear ergression and design by treed Gaussian process models. *Journal of Statistical Software*, 19(9), 2007.

[80] X. Guyon. *Random field on a network - Modeling, statistics, and applications*. Springer-Verlag, 1995.

[81] J.M. Hammersley and D.C. Handscomb. *Monte Carlo methods*. Chapman and Hall, 1964.

[82] T. Hastie and R. Tibshirani. *Generalized additive models*. Chapman and Hall, London, 1990.

[83] T. Hastie, R. Tibshirani, and J. Friedman. *The elements of statistical learning*. Springer, 2002.

[84] J.C. Helton, J.D. Johnson, C.J. Salaberry, and C.B. Storlie. Survey of sampling-based methods for uncertainty and sensitivity analysis. *Reliability Engineering and System Safety*, 91 :1175–1209, 2006.

[85] T.C. Hesterberg and B.L. Nelson. Control variates for probability and quantile estimation. *Management Science*, 44 :1295–1312, 1998.

[86] W. Hoeffding. A class of statistics with asymptotically normal distributions. *Annals of Mathematical Statistics*, 19 :293–325, 1948.

[87] J.A. Hoeting, R.A. Davis, A.A. Merton, and S.E. Thompson. Model selection for geostatistical models. *Ecological Applications*, 16 :87–98, 2006.

[88] W.C. Hoffman. Wave propagation in a general random continuous medium. In *Symposia in Applied Mathematics*, volume 16, pages 117–144. Amer. Math. Soc., 1964.

[89] T. Homma and A. Saltelli. Importance measures in global sensitivity analysis of non linear models. *Reliability Engineering and System Safety*, 52 :1–17, 1996.

[90] B. Iooss. Caractérisation probabiliste de réflecteurs en sismique réflexion. In *Les Cahiers de Géostatistique*, volume 6, pages 61–73. Ecole des Mines de Paris, 1998.

[91] B. Iooss. Seismic reflection traveltimes in two-dimensional statistically anisotropic random media. *Geophysical Journal International*, 135 :999–1010, 1998.

[92] B. Iooss. *Tomographie statistique en sismique réflexion : Estimation d'un modèle de vitesse stochastique*. Thèse de l'École des Mines de Paris, France, 1998. Available at URL : http ://cg.ensmp.fr/Theses/iooss.shtml.

[93] B. Iooss. *Manuel utilisateur du logiciel SSURFER v1.2 : programmes en R d'analyses d'incertitudes, de sensibilités, et de construction de surfaces de réponse.* Note Technique CEA/DEN/CAD/DER/SESI/LCFR/NT DO 6 08/03/06, 2006.

[94] B. Iooss, P. Blanc-Benon, and C. Lhuillier. Statistical moments of travel times at second order in isotropic and anisotropic random media. *Waves in Random Media*, 10 :381–394, 2000.

[95] B. Iooss, L. Boussouf, A. Marrel, and V. Feuillard. Numerical study of algorithms for metamodel construction and validation. In S. Martorell, C. Guedes Soares, and J. Barnett, editors, *Safety, reliability and risk analysis - Proceedings of the ESREL 2008 Conference*, pages 2135–2141, Valencia, Spain, september 2008. CRC Press.

[96] B. Iooss and A. Galli. Statistical tomography for seismic reflection data. In *6th International Geostatistics Congress, Cape Town, South Africa.* [CD-ROM]. s.l. : Geostatistical Association of Southern Africa, 2000.

[97] B. Iooss, D. Geraets, T. Mukerji, Y. Samuelides, M. Touati, and A. Galli. Inferring the statistical distribution of velocity heterogeneities by statistical traveltime tomography. *Geophysics*, 68 :1714–1730, 2003.

[98] B. Iooss, C. Lhuillier, and H. Jeanneau. Numerical simulation of transit-time ultrasonic flowmeters due to flow profile and fluid turbulence. *Ultrasonics*, 40 :1009–1015, 2002.

[99] B. Iooss and M. Ribatet. Analyse de sensibilité globale de modèles numériques à paramètres incontrôlables. In *Proceedings of 38èmes Journées de Statistique*, Clamart, France, may-june 2006.

[100] B. Iooss and M. Ribatet. Global sensitivity analysis of computer models with functional inputs. In *Proceedings of 5th International Conference on Sensitivity Analysis of Model Output*, Budapest, Hungary, june 2007.

[101] B. Iooss and M. Ribatet. Global sensitivity analysis of computer models with functional inputs. *Reliability Engineering and System Safety*, doi :10.1016/j.ress.2008.09.010, in press, 2009.

[102] B. Iooss, M. Ribatet, and A. Marrel. Global sensitivity analysis of stochastic computer models with generalized additive models. *Technometrics*, submitted, 2006. Available at URL : http ://fr.arxiv.org/abs/0802.0443.

[103] B. Iooss and Y. Samuelides. Inversion of velocity statistical parameters from traveltimes. In *Proc. of 135th Meeting of the Acoustical Society of America*, pages 2319–2320, Seattle, 1998.

[104] B. Iooss, F. Van Dorpe, and N. Devictor. Response surfaces and sensitivity analyses for an environmental model of dose calculations. *Reliability Engineering and System Safety*, 91 :1241–1251, 2006.

[105] A. Ishimaru. *Wave propagation and scattering in random media.* IEEE Press, 1997.

[106] J. Jacques. *Contributions à l'analyse de sensibilité et à l'analyse discriminante généralisée.* Thèse de l'Université Joseph Fourier, Grenoble 1, 2005.

[107] J. Jacques, C. Lavergne, and N. Devictor. Sensitivity analysis in presence of model uncertainty and correlated inputs. *Reliability Engineering and System Safety*, 91 :1126–1134, 2006.

[108] L. Jannaud. *Propagation d'onde en milieu aléatoire.* Thèse de l'Université Paris Sud - Orsay, France, 1991.

[109] L. Jannaud, P. M. Adler, and C. G. Jacquin. Wave propagation in random anisotropic media. *Journal of Geophysical Research*, 97 :15277–15289, 1992.

[110] D.R. Jones, M. Schonlau, and W.J. Welch. Efficient global optimization of expensive black-box functions. *Journal of Global Optimization*, 13 :455–492, 1998.

[111] A. Jourdan. *Analyse statistique et échantillonnage d'expériences simulées.* Thèse de l'Université de Pau et des Pays de l'Adour, 2000.

[112] I. Juutilainen and J. Röning. A comparaison of methods for joint modelling of mean and dispersion. In *Proceedings of the 11th Symposium on ASMDA*, Brest, France, may 2005.

[113] M. Karweit, P. Blanc-Benon, D. Juvé, and G. Comte-Bellot. Simulation of the propagation of an acoustic wave through a turbulent velocity field : A study of phase variance. *Journal of the Acoustical Society of America*, 89 :52–62, 1991.

[114] A. Kaslilar, Yu. A. Kravtsov, S. A. Shapiro, S. Buske, R. Giese, and Th. Dickmann. Estimation of the rock statistical parameters from traveltime measurements. *Studia Geophisicae et Geodetica*, 50 :325–336, 2006.

[115] J. B. Keller. Stochastic equations and wave propagation in random media. In *Symposia in Applied Mathematics*, volume 16, pages 145–170. Amer. Math. Soc., 1964.

[116] J.P.C. Kleijnen. Sensitivity analysis and related analyses : a review of some statistical techniques. *Journal of Statistical Computation and Simulation*, 57 :111–142, 1997.

[117] J.P.C. Kleijnen. *Design and analysis of simulation experiments.* Springer, 2008.

[118] J.P.C. Kleijnen and J.C. Helton. Statistical analyses of scatterplots to identify important factors in large-scale simulations, 1 : Review and comparison of techniques. *Reliability Engineering and System Safety*, 65 :147–185, 1999.

[119] J.P.C. Kleijnen and R.G. Sargent. A methodology for fitting and validating metamodels in simulation. *European Journal of Operational Research*, 120 :14–29, 2000.

[120] L. Klimeš. Application of the medium covariance functions to travel-time tomography. *Pure and Applied Geophysics*, 159 :1791–1810, 2002.

[121] L. Klimeš. Correlation functions of random media. *Pure and Applied Geophysics*, 159 :1811–1831, 2002.

[122] L. Klimeš. Estimating the correlation function of a self-affine random medium. *Pure and Applied Geophysics*, 159 :1833–1853, 2002.

[123] R. Koenker. *Quantile regression.* Cambridge University Press, 2005.

[124] A. I. Kon. Qualitative theory of amplitude and phase fluctuations in a medium with anisotropic turbulent irregularities. *Waves in Random Media*, 4 :297–306, 1994.

[125] Y. A. Kravtsov. *Geometrical optics in engineering physics.* Alpha Science, 2005.

[126] Yu. A. Kravtsov, A. Kaslilar, S. A. Shapiro, S. Buske, and T. M. Müller. Estimating statistical parameters of an elastic random medium from traveltime fluctuations of refracted waves. *Waves in Random and Complex Media*, 15 :43–60, 2005.

[127] Yu. A. Kravtsov, T. M. Müller, S. A. Shapiro, and S. Buske. Statistical properties of reflection traveltimes in 3-D randomly inhomogeneous and anisomeric media. *Geophysical journal International*, 154 :841–851, 2003.

[128] S. G. Krige. A statistical approach to some basic valuations problems on the witwatersrand. *J. Chem. Metall. Min. Soc.*, 52(6) :119–139, 1951.

[129] V. A. Kulkarny and B. S. White. Focusing of waves in turbulent inhomogeneous media. *Physics of Fluids*, 25 :1770–1784, 1982.

[130] P.C. Kyriakidis and A. journel. Geostatistical space-time models : A review. *Mathematical Geology*, 31 :651–684, 1999.

[131] Y. Lee and J.A. Nelder. Robust design via generalized linear models. *Journal of Quality Technology*, 35(1) :2–12, 2003.

[132] M. Lesieur. *Turbulence in fluids.* Kluwer Academic Publishers, 1997.

[133] R. Li and A. Sudjianto. Analysis of computer experiments using penalized likelihood in Gaussian kriging models. *Technometrics*, 47 :111–120, 2005.

[134] D.K.J. Lin. A new class of supersaturated design. *Technometrics*, 35 :28–31, 1993.

[135] C. Linkletter, D. Bingham, N. Hengartner, D. Higdon, and K.Q. Ye. Variable selection for Gaussian process models in computer experiments. *Technometrics*, 48 :478–490, 2006.

[136] Y-X. Liu, T. Xu, B. Zhao, and C-C. Liu. Seismic sounding of anisotropic self-similar self-organized medium. *Chinese Journal of Geophysics*, 50 :221–232, 2007.

[137] S.N. Lophaven, H.B. Nielsen, and J. Sondergaard. DACE - A Matlab kriging toolbox, version 2.0. Technical Report IMM-TR-2002-12, Informatics and Mathematical Modelling, Technical University of Denmark, 2002. <http ://www.immm.dtu.dk/∼hbn/dace>.

[138] L. C. Lynnworth. *Ultrasonic measurements for process control.* Academic Press, Inc., 1989.

[139] H.O. Madsen, S. Krenk, and N.C. Lind, editors. *Methods of structural safety.* Prentice Hall, 1986.

[140] L. Margerin. Attenuation, transport and diffusion of scalar waves in textured random media. *Tectonophysics*, 416 :229–244, 2006.

[141] A. Marrel. *Mise en oeuvre et exploitation du métamodèle processus gaussien pour l'analyse de modèles numériques - Application à un code de transport hydrogéologique.* Thèse de l'Université Paul Sabatier - Toulouse III, 2008.

[142] A. Marrel, B. Iooss, B. Laurent, and O. Roustant. Calculations of the Sobol indices for the Gaussian process metamodel. *Reliability Engineering and System Safety*, 94 :742–751, 2009.

[143] A. Marrel, B. Iooss, F. Van Dorpe, and E. Volkova. An efficient methodology for modeling complex computer codes with Gaussian processes. *Computational Statistics and Data Analysis*, 52 :4731–4744, 2008.

[144] M. Marseguerra, R. Masini, E. Zio, and G. Cojazzi. Variance decomposition-based sensitivity analysis via neural networks. *Reliability Engineering and System Safety*, 79 :229–238, 2003.

[145] J.D. Martin and T.W. Simpson. Use of kriging models to approximate deterministic computer models. *AIAA Journal*, 43 :853–863, 2005.

[146] B. Matèrn. *Spatial variation.* Springer-Verlag, 1986.

[147] G. Matheron. *La théorie des variables régionalisées et ses applications*. Les Cahiers du Centre de Morphologie Mathématique de Fontainebleau, Fascicule 5. Ecole des Mines de Paris, 1970.

[148] G. Matheron. Géodésiques aléatoires : application à la prospection sismique. In *Les Cahiers de Géostatistique, Fascicule 1*, pages 1–18. Ecole des Mines de Paris, 1991.

[149] P. McCullagh and J.A. Nelder. *Generalized linear models*. Chapman & Hall, 1989.

[150] M.D. McKay, R.J. Beckman, and W.J. Conover. A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics*, 21 :239–245, 1979.

[151] M. Meckesheimer, A.J. Booker, R.R. Barton, and T.W. Simpson. Computationally inexpensive metamodel assessment strategies. *AIAA Journal*, 40 :2053–2060, 2002.

[152] D.C. Montgomery. *Design and analysis of experiments*. John Wiley & Sons, 2004.

[153] T. Mukerji, G. Mavko, D. Mujica, and N. Lucet. Scale-dependent seismic velocity in heterogeneous media. *Geophysics*, 60 :1222–1233, 1995.

[154] G. Müller, M. Roth, and M. Korn. Seismic-wave traveltimes in random media. *Geophysical Journal International*, 110 :29–41, 1992.

[155] W. H. Munk and F. Zachariasen. Sound propagation through a fluctuating stratified ocean : Theory and observation. *Journal of the Acoustical Society of America*, 59 :818–838, 1976.

[156] J.A. Nelder and D. Pregibon. An extended quasi-likelihood function. *Biometrika*, 74 :221–232, 1987.

[157] J.A. Nelder and R.W.M. Wedderburn. Generalized linear models. *Journal of the Royal Statistical Society A*, 135 :370–384, 1972.

[158] B.L. Nelson. Control variate remedies. *Operation Research*, 38 :974–992, 1990.

[159] W.T. Nutt and G.B. Wallis. Evaluation of nuclear safety from the outputs of computer codes in the presence of uncertainties. *Reliability Engineering and System Safety*, 83 :57–77, 2004.

[160] J.E. Oakley. Estimating percentiles of uncertain computer code outputs. *Applied Statistics*, 53 :83–93, 2004.

[161] J.E. Oakley and A. O'Hagan. Probabilistic sensitivity analysis of complex models : a bayesian approach. *Journal of the Royal Statistical Society, Series B*, 66 :751–769, 2004.

[162] A. O'Hagan. Bayesian analysis of computer code outputs : A tutorial. *Reliability Engineering and System Safety*, 91 :1290–1300, 2006.

[163] V. E. Ostashev. *Acoustics in moving inhomogeneous media*. E & FN Spon, 1997.

[164] G. Papanicolaou. Mathematical problem in geophysical wave propagation. In *Proc. of the International Congress of Mathematicians, Extra Volume*, pages 241–265. Documenta Mathematica, 1998.

[165] A. Papoulis and S.U. Pillai. *Probability, random variables and stochastic processes*. McGraw-Hill, 2002.

[166] E.J. Pebesma and G.B.M. Heuvelink. Latin hypercube sampling of Gaussian random fields. *Technometrics*, 41 :303–312, 1999.

[167] M. Petelet. *Analyse de sensibilité globale de modèles thermoécaniques de simulation numérique du soudage*. Thèse de l'Université de Bourgogne, 2007.

[168] N. V. Petersen. Inverse kinematic problem for a random medium in geometric optics approximation. *Pure and Applied Geophysics*, 132 :417–437, 1990.

[169] A. Petruzzi, F. D'Auria, J-C. Micaelli, A. De Crecy, and J. Royen. The BEMUSE programme (Best-Estimate Methods - Uncertainty and Sensitivity Evaluation). In *Proceedings of the Int. Meet. on Best-Estimate Methods in Nuclear Installation Safety Analysis (BE-2004) IX*, volume 1, pages 225–235, Washington, USA, 2004.

[170] D. Pregibon. Review of "Generalized Linear Models" by McCullagh and Nelder. *Annals of Statistics*, 12 :1589–1596, 1984.

[171] G. Pujol. Simplex-based screening designs for estimating metamodels. *Reliability Engineering and System Safety*, doi :10.1016/j.ress.2008.08.002, in press, 2009.

[172] R Development Core Team. *R : A Language and Environment for Statistical Computing*. ISBN 3-900051-07-0, Vienna, Austria, 2006.

[173] C.E. Rasmussen and C.K.I. Williams. *Gaussian processes for machine learning*. MIT Press, 2006.

[174] M. Ratto, A. Pagano, and P. Young. State dependent parameter metamodelling and sensitivity analysis. *Computer Physics Communication*, 177 :863–876, 2007.

[175] M.I. Reis dos Santos and A.M.O. Porta Nova. Statistical fitting and validation of non-linear simulation metamodels : A case study. *European Journal of Operational Research*, 171 :53–63, 2006.

[176] R.A. Rigby and D.M. Stasinopoulos. A semi-parametric additive model for variance heterogeneity. *Statistics and Computing*, 6 :57–65, 1996.

[177] M. Roth, G. Müller, and R. Snieder. Velocity shift in random media. *Geophysical Journal International*, 115 :552–563, 1993.

[178] R.Y. Rubinstein. *Simulation and the Monte Carlo method*. Wiley, 1981.

[179] B. Rutherford. A response-modeling alternative to surrogate models for support in computational analyses. *Reliability Engineering and System Safety*, 91 :1322–1330, 2006.

[180] S. M. Rytov, Y. A. Kravtsov, and V. I. Tatarskii. *Elements of random fields*, volume 3 of *Principles of statistical radiophysics*. Springer-Verlag, 1987.

[181] S. M. Rytov, Y. A. Kravtsov, and V. I. Tatarskii. *Wave propagation through random media*, volume 4 of *Principles of statistical radiophysics*. Springer-Verlag, 1987.

[182] J. Sacks, W.J. Welch, T.J. Mitchell, and H.P. Wynn. Design and analysis of computer experiments. *Statistical Science*, 4 :409–435, 1989.

[183] T. Saito. Velocity shift in two-dimensional anisotropic random media using the Rytov method. *Geophysical Journal International*, 166 :293–308, 2006.

[184] A. Saltelli. Making best use of model evaluations to compute sensitivity indices. *Computer Physics Communication*, 145 :280–297, 2002.

[185] A. Saltelli, K. Chan, and E.M. Scott, editors. *Sensitivity analysis*. Wiley Series in Probability and Statistics. Wiley, 2000.

[186] A. Saltelli, M. Ratto, T. Andres, F. Campolongo, J. Cariboni, D. Gatelli, M. Salsana, and S. Tarantola. *Global sensitivity analysis - The primer*. Wiley, 2008.

[187] A. Saltelli, S. Tarantola, F. Campolongo, and M. Ratto. *Sensitivity analysis in practice : A guide to assessing scientific models*. Wiley, 2004.

[188] A. Saltelli, S. Tarantola, and K. Chan. A quantitative, model-independent method for global sensitivity analysis of model output. *Technometrics*, 41 :39–56, 1999.

[189] Y. Samuelides. Velocity shift using the Rytov approximation. *Journal of the Acoustical Society of America*, 104 :2596–2603, 1998.

[190] Y. Samuelides and T. Mukerji. Velocity shift in heterogeneous media with anisotropic spatial correlation. *Geophysical Journal International*, 134 :778–786, 1998.

[191] T. Santner, B. Williams, and W. Notz. *The design and analysis of computer experiments*. Springer, 2003.

[192] G. Saporta. *Probabilités, analyse de données et statistique*. éditions Technip, 2ème edition, 2006.

[193] H. Sato and M. C. Fehler. *Seismic wave propagation and scattering in the heterogeneous earth*. Springer, 1998.

[194] F. Satterthwaite. Random balance experimentation. *Technometrics*, 1 :111–137, 1959.

[195] C. Scheidt. *Analyse statistique d'expériences simulées : modélisation adaptative de réponses non régulières par krigeage et plans d'expériences*. Thèse de l'Université Louis Pasteur Strasbourg I, 2006.

[196] M. Schonlau and W.J. Welch. Screening the input variables to a computer model. In A. Dean and S. Lewis, editors, *Screening - Methods for experimentation in industry, drug discovery and genetics*. Springer, 2006.

[197] M. Sergent, B. Corre, and D. Dupuy. Comparison of different screening methods. In *VI Colloquium Chemiometricum Mediterraneum*, Saint Maximin La Sainte Baume, France, september 2007.

[198] S. Shapiro, R. Schwarz, and N. Gold. The effect of random isotropic inhomogeneities on the phase velocity of seismic waves. *Geophysical Journal International*, 127 :783–794, 1996.

[199] T.W. Simpson, J.D. Peplinski, P.N. Kock, and J.K. Allen. Metamodel for computer-based engineering designs : survey and recommandations. *Engineering with Computers*, 17 :129–150, 2001.

[200] I.M. Sobol. Sensitivity estimates for non linear mathematical models. *Mathematical Modelling and Computational Experiments*, 1 :407–414, 1993.

[201] I.M. Sobol. Global sensitivity indices for non linear mathematical models and their Monte Carlo estimates. *Mathematics and Computers in Simulation*, 55 :271–280, 2001.

[202] I.M. Sobol. Theorems and examples on high dimensional model representation. *Reliability Engineering and System Safety*, 79 :187–193, 2003.

[203] M.L. Stein. *Interpolation of spatial data*. Springer, 1999.

[204] C.B. Storlie and J.C. Helton. Multiple predictor smoothing methods for sensitivity analysis : Description of techniques. *Reliability Engineering and System Safety*, 93 :28–54, 2008.

[205] B. Sudret. Global sensitivity analysis using polynomial chaos expansion. *Reliability Engineering and System Safety*, 93 :964–979, 2008.

[206] B. Sudret. *Uncertainty propagation and sensitivity analysis in mechanical models - Contributions to structural reliability and stochastics spectral methods*. Thèse d'Habilitation à Diriger des Recherches de l'Université Blaise Pascal - Clermont II, 2008.

[207] F.D. Tappert. The parabolic approximation method. In J.B. Keller and J.S. Papadakis, editors, *Wave propagation in ocean acoustics*, volume 70 of *Lecture notes in physics*, pages 224–287. Springer Verlag, 1977.

[208] A. Tarantola. *Inverse problem theory and methods for model parameter estimation*. Society for Industrial and Applied Mathematics, SIAM, 2005.

[209] S. Tarantola, D. Gatelli, and T. Mara. Random balance designs for the estimation of first order global sensitivity indices. *Reliability Engineering and System Safety*, 91 :717–727, 2006.

[210] S. Tarantola, N. Giglioli, N. Jesinghaus, and A. Saltelli. Can global sensitivity analysis steer the implementation of models for environmental assesments and decision-making ? *Stochastic Environmental Research and Risk Assesment*, 16 :63–76, 2002.

[211] V. I. Tatarskii. *Wave propagation in a turbulent medium.* Dover Publications, Inc, New York, 1961.

[212] P. D. Thore and C. Juliard. Fresnel-zone effect on seismic velocity resolution. *Geophysics*, 64 :593–603, 1999.

[213] P. D. Thore, A. Shtuka, M. Lecour, T. Ait-Ettajer, and R. Cognot. Structural uncertainties : Determination, management, and applications. *Geophysics*, 67 :840–852, 2002.

[214] M. Touati. *Contribution géostatistique au traitement de données sismiques.* Thèse de l'École des Mines de Paris, France, 1996.

[215] M. Touati, B. Iooss, and A. Galli. Quantitative control of migration : a geostatistical attempt. *Mathematical Geology*, 31 :277–295, 1999.

[216] T. Turanyi. Sensitivity analysis for complex kinetic system, tools and applications. *Journal of Mathematical Chemistry*, 5 :203–248, 1990.

[217] B. J. Uscinski. Acoustic scattering by ocean irregularities : Aspects of the inverse problem. *Journal of the Acoustical Society of America*, 79 :347–355, 1986.

[218] J. Uscinski. *The elements of wave propagation in random media.* Mc Graw-Hill, New York, 1977.

[219] E. Vazquez. *Modélisation comportementale de systèmes non-linéaires multivariables par méthodes à noyaux et applications.* Thèse de l'Université Paris XI Orsay, 2005.

[220] E. Vazquez and M. Piera-Martinez. Estimation du volume des ensembles d'excursion d'un processus gaussien par krigeage intrinsèque. In *Proceedings of 39èmes Journées de Statistique*, Angers, France, june 2007.

[221] S.N. Vecherin, V.E. Ostashev, G.H. Goedecke, D.K. Wilson, and A.G. Voronovich. Time-dependent stochastic inversion in acoustic travel-time tomography of the atmosphere. *Journal of the Acoustical Society of America*, 119 :2579–2588, 2006.

[222] G.G. Vining and R.H. Myers. Combining Taguchi and response-surface philosophies - a dual response approach. *Journal of Quality Technology*, 22 :38–45, 1990.

[223] E. Volkova, B. Iooss, and F. Van Dorpe. Global sensitivity analysis for a numerical model of radionuclide migration from the RRC "Kurchatov Institute" radwaste disposal site. *Stochastic Environmental Research and Risk Assesment*, 22 :17–31, 2008.

[224] H. Wackernagel. *Multivariate geostatistics.* Springer, 1995.

[225] G. Wahba. *Spline models for observational data.* SIAM, 1990.

[226] W.J. Welch, R.J. Buck, J. Sacks, H.P. Wynn, T.J. Mitchell, and M.D. Morris. Screening, predicting, and computer experiments. *Technometrics*, 34(1) :15–25, 1992.

[227] B. S. White. The stochastic caustic. *SIAM Journal of Applied Mathematics*, 44 :127–149, 1984.

[228] S.S. Wilks. Determination of sample sizes for setting tolerance limits. *Annals of Mathematical Statistics*, 12 :91–96, 1941.

[229] P. R. Williamson and M. H. Worthington. Resolution limits in ray tomography due to wave behavior : Numerical experiments. *Geophysics*, 58 :727–735, 1993.

[230] S. Wood. *Generalized Additive Models : An Introduction with R.* CRC Chapman & Hall, 2006.

[231] A.M. Yaglom. *An introduction to the theory of stationary random functions.* Dover edition, 1973.

[232] I. Zabalza-Mezghani. *Analyse statistique et planification d'expérience en ingénierie de réservoir.* Thèse de l'Université de Pau et des Pays de l'Adour, 2000.

[233] E. Zio and F. Di Maio. Bootstrap and order statistics for quantifying thermal-hydraulic code uncertainties in the estimation of safety margins. *Science and Technology of Nuclear Installations*, 2008, Article ID 340164, 9 pages, doi :10.1155/2008/340164, 2008.

Résumé

Contributions au traitement des incertitudes en modélisation numérique : propagation d'ondes en milieu aléatoire et analyse d'expériences simulées

Le présent document constitue mon mémoire d'habilitation à diriger des recherches. Il retrace mon activité scientifique de ces douze dernières années, depuis ma thèse jusqu'aux travaux réalisés en tant qu'ingénieur-chercheur du CEA Cadarache. Les deux chapitres qui structurent ce document correspondent à deux domaines de recherche relativement différents mais se référant tous les deux au traitement des incertitudes dans des problèmes d'ingénierie. Le premier chapitre établit une synthèse de mes travaux sur la propagation d'ondes hautes fréquences en milieu aléatoire. Il concerne plus spécifiquement l'étude des fluctuations statistiques des temps de trajet des ondes acoustiques en milieu aléatoire et/ou turbulent. Les nouveaux résultats obtenus concernent principalement l'introduction de l'anisotropie statistique des champs de vitesse lors de la dérivation des expressions des moments des temps en fonction de ceux du champ de vitesse des ondes. Ces travaux ont été essentiellement portés par des besoins en géophysique (exploration pétrolière et sismologie). Le second chapitre aborde le domaine de l'utilisation des techniques probabilistes pour prendre en compte les incertitudes des variables d'entrée d'un modèle numérique. Les principales applications que j'évoque dans ce chapitre relèvent du domaine de l'ingénierie nucléaire qui offre une grande variété de problématiques d'incertitude à traiter. Tout d'abord, une synthèse assez complète est réalisée sur les méthodes statistiques d'analyse de sensibilité et d'exploration globale de modèles numériques. La construction et l'exploitation d'un métamodèle (fonction mathématique peu coûteuse se substituant à un code de calcul coûteux) sont ensuite illustrées par mes travaux sur le modèle processus gaussien (krigeage). Deux thématiques complémentaires sont finalement abordées : l'estimation de quantiles élevés de réponses de codes de calcul et l'analyse de codes de calcul stochastiques. Une conclusion met en perspective ces travaux dans le contexte plus général de la simulation numérique et de l'utilisation de modèles prédictifs dans l'industrie.

Abstract

Contributions to the uncertainty management in numerical modelisation : wave propagation in random media and analysis of computer experiments

The present document constitutes my habilitation thesis report. It recalls my scientific activity of the twelve last years, since my PhD thesis until the works completed as a research engineer at CEA Cadarache. The two main chapters of this document correspond to two different research fields both referring to the uncertainty treatment in engineering problems. The first chapter establishes a synthesis of my work on high frequency wave propagation in random medium. It more specifically relates to the study of the statistical fluctuations of acoustic wave traveltimes in random and/or turbulent media. The new results mainly concern the introduction of the velocity field statistical anisotropy in the analytical expressions of the traveltime statistical moments according to those of the velocity field. This work was primarily carried by requirements in geophysics (oil exploration and seismology). The second chapter is concerned by the probabilistic techniques to study the effect of input variables uncertainties in numerical models. My main applications in this chapter relate to the nuclear engineering domain which offers a large variety of uncertainty problems to be treated. First of all, a complete synthesis is carried out on the statistical methods of sensitivity analysis and global exploration of numerical models. The construction and the use of a metamodel (inexpensive mathematical function replacing an expensive computer code) are then illustrated by my work on the Gaussian process model (kriging). Two additional topics are finally approached : the high quantile estimation of a computer code output and the analysis of stochastic computer codes. We conclude this memory with some perspectives about the numerical simulation and the use of predictive models in industry. This context is extremely positive for future researches and application developments.