



**HAL**  
open science

# Modélisation statistique de l'exposition humaine aux ondes radiofréquences.

Pierric Kersaudy

► **To cite this version:**

Pierric Kersaudy. Modélisation statistique de l'exposition humaine aux ondes radiofréquences.. Statistiques [stat]. UNIVERSITE PARIS-EST, 2015. Français. NNT : . tel-01271351

**HAL Id: tel-01271351**

**<https://hal.science/tel-01271351>**

Submitted on 11 Feb 2016

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

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

# UNIVERSITÉ PARIS-EST MARNE-LA-VALLÉE

ÉCOLE DOCTORALE MSTIC

## THÈSE

Pour obtenir le grade de :

**Docteur de l'Université Paris-Est**

Spécialité : Électronique, optronique et systèmes

Présentée et soutenue publiquement par :

**Pierric KERSAUDY**

le 12 novembre 2015

dans les locaux d'Orange Labs

## Modélisation statistique de l'exposition humaine aux ondes radiofréquences.

### Jury

Pr. Philippe De Doncker	Université Libre de Bruxelles	Rapporteur
Dr. David Ginsbourger	Idiap Research Institute & University of Bern	Rapporteur
Pr. Bruno Sudret	ETH Zürich	Président du jury
Dr. Shermila Mostarshedi	Université Paris-Est	Examinatrice
Dr. Joe Wiart	Télécom ParisTech, Chaire C2M, Whist Lab	Examineur
Pr. Odile Picon	Université Paris-Est	Directrice de thèse

Directrice de thèse : Odile Picon.

Encadrants de thèse : Joe Wiart et Shermila Mostarshedi.

Laboratoire : ESYCOM - EA 2552



# Remerciements

J'aimerais tout d'abord remercier Joe Wiart, mon encadrant de thèse au sein d'Orange Labs, qui m'a soutenu, m'a accordé toute sa confiance tout au long de ces trois ans et m'a intégré en tant que membre de son équipe à part entière. Ses conseils techniques et professionnels et son exemple me suivront toujours.

Je voudrais également remercier Odile Picon, ma directrice de thèse, qui a toujours été disponible pour m'orienter et émettre des avis critiques qui m'ont permis d'avancer dans ce sujet passionnant qui m'a été proposé.

Mes remerciements vont également à Shermila Mostarshedi, qui m'a accompagné et soutenu pendant ces trois années. Merci pour sa relecture de ce mémoire. Merci pour sa disponibilité sur laquelle j'ai toujours pu compter pour toute sorte de question.

Merci à Bruno Sudret pour sa disponibilité et son aide qui a été très utile pour débloquer des problèmes théoriques et techniques tout au long de mon parcours. Ce travail ne serait pas le même sans ses conseils avisés.

Merci à l'équipe d'Orange Labs, Nadège, Azzedine, Abdelhamid, Thierry, Emmanuelle, Yuan Yuan, Anis, Mouléro et tous ceux que j'oublie pour tous les moments agréables et tous les souvenirs qu'ils me laissent.

Je tiens également à remercier les rapporteurs pour avoir accepté de consacrer du temps à la lecture de ce mémoire et à l'évaluation des travaux qu'il contient. Plus généralement, je remercie l'ensemble du jury pour avoir accepté d'évaluer mon travail.

Ma gratitude va enfin à toute ma famille et à tous mes amis qui ont toujours été présents.



# Table des matières

<b>1</b>	<b>Introduction générale</b>	<b>11</b>
<b>2</b>	<b>L'exposition humaine</b>	<b>15</b>
2.1	L'évaluation de l'exposition . . . . .	16
2.2	La dosimétrie numérique . . . . .	17
2.3	Les modèles de corps humains . . . . .	18
2.4	La modélisation des différentes sources d'exposition . . . . .	19
2.4.1	L'exposition induite par une source proche . . . . .	19
2.4.2	L'exposition induite par une source lointaine . . . . .	20
2.4.3	L'exposition induite par une source <i>intermédiaire</i> . . . . .	21
2.5	Analyse comparative des expositions fœtales induites par des sources intermédiaires et des sources lointaines. . . . .	22
2.5.1	Les modèles anatomiques étudiés . . . . .	23
2.5.2	Comparaison de l'exposition à une onde plane avec celle de la description en ondes sphériques. . . . .	24
2.5.3	Résultats de la femme japonaise enceinte . . . . .	25
2.5.4	Résultats pour Victoria enceinte . . . . .	26
2.6	Conclusion . . . . .	28
<b>3</b>	<b>L'analyse statistique et son application à la problématique de l'exposition.</b>	<b>31</b>
3.1	Etat de l'art sur l'analyse statistique de l'exposition . . . . .	32
3.2	Modélisation statistique de l'espace d'entrée . . . . .	34
3.2.1	Espace probabiliste . . . . .	34
3.2.2	Paramètres d'entrée spatiaux . . . . .	36
3.2.3	Paramètres d'entrée morphologiques . . . . .	37
3.2.4	Transformée isoprobabiliste . . . . .	42
3.3	Etude statistique de la sortie du modèle physique. . . . .	44
3.3.1	Les moments statistiques . . . . .	44
3.3.2	Etude des quantiles de sortie . . . . .	45
3.3.3	Analyse de sensibilité globale de paramètres indépendants . . . . .	45
3.3.4	Analyse de sensibilité globale de paramètres dépendants . . . . .	48
3.4	Conclusion . . . . .	53
<b>4</b>	<b>Analyse statistique par méta-modèle</b>	<b>55</b>
4.1	Le chaos polynomial . . . . .	56
4.1.1	Historique . . . . .	56

4.1.2	La théorie du chaos polynomial . . . . .	58
4.1.3	Les troncatures pleine et hyperbolique . . . . .	59
4.1.4	La troncature LARS . . . . .	61
4.1.5	Sélection et validation de modèle . . . . .	62
4.1.6	Analyse de sensibilité et analyse de signature . . . . .	63
4.1.7	Planification d'expériences . . . . .	64
4.1.8	Exemple d'application : exposition d'un modèle d'enfant à une station de base de femtocell . . . . .	68
4.1.9	Exemple d'application : exposition en zone urbaine : calcul du champ réfléchi par un immeuble à géométrie variable . . . . .	70
4.2	Le Krigeage universel . . . . .	74
4.2.1	Un bref historique . . . . .	74
4.2.2	La formulation du Krigeage universel . . . . .	75
4.2.3	Les fonctions d'autocorrélation . . . . .	76
4.2.4	Le meilleur prédicteur linéaire sans biais (BLUP) . . . . .	78
4.3	Hybridation du chaos polynomial et du Krigeage universel . . . . .	80
4.3.1	Exemple d'application analytique : la fonction d'Ishigami . . . . .	81
4.3.2	Exemple d'application analytique : la fonction de Borehole . . . . .	82
4.3.3	Exemple d'application analytique : la fonction de Sobol . . . . .	83
4.3.4	Application à l'exposition d'un foetus à un système femtocell . . . . .	84
4.3.5	Application à l'étude de l'influence de l'incertitude morphologique . . . . .	87
4.4	Conclusion . . . . .	90
<b>5</b>	<b>Planification d'expérience adaptative pour l'estimation du quantile à 95% de la distribution de sortie</b>	<b>93</b>
5.1	La méthode BOAS ( <i>Bootstrapped Oriented Adaptive Sampling</i> ) . . . . .	96
5.1.1	La théorie du bootstrap et les intervalles de confiance . . . . .	96
5.1.2	Stratégie adaptative de planification d'expériences . . . . .	99
5.1.3	Critères d'arrêt et estimateurs de précision . . . . .	103
5.1.4	Exemple analytique : la fonction d'Ishigami . . . . .	103
5.1.5	Exemple analytique : la fonction de Borehole . . . . .	106
5.1.6	Exemple analytique : la fonction de Sobol . . . . .	108
5.1.7	Exemple d'application : exposition d'un modèle d'enfant à une station de base de femtocell . . . . .	111
5.2	Comparaison de performances avec la méthode GPS ( <i>Gaussian Process Shrunk</i> ) . . . . .	112
5.2.1	La méthode GPS . . . . .	113
5.2.2	Comparaison de performances avec la fonction d'Ishigami . . . . .	114
5.2.3	Comparaison de performances avec la fonction de Borehole . . . . .	115
5.3	Conclusion . . . . .	117
<b>6</b>	<b>Conclusion générale et perspectives</b>	<b>119</b>
	<b>Références</b>	<b>123</b>
	<b>Publications personnelles</b>	<b>131</b>

# Table des figures

1.1	Taux de pénétration du mobile dans la métropole française [ARCEP 2013]	12
2.1	Photographie des calculateurs FDTD utilisés au sein du Whistlab. . . . .	19
2.2	Modèle anatomique : <i>Visible Human</i> . . . . .	19
2.3	Exemple de modèle voxelisé de téléphone portable. . . . .	20
2.4	Vues externe et interne du système femtocell considéré. . . . .	21
2.5	Diagramme de rayonnement du système femtocell considéré en fonction des angle d'azimut $\phi$ et d'élévation $\theta$ . . . . .	22
2.6	Schéma de l'utilisation d'une boîte de Huygens pour l'exposition d'une femme enceinte à une station de base de femtocell. . . . .	23
2.7	Modèles anatomiques de femmes enceintes : (a) femme japonaise enceinte, (b) Victoria enceinte. . . . .	24
2.8	Comparaison des DAS corps entiers du fœtus de la femme japonaise enceinte : en valeur absolue (lignes et axe vertical gauche) et en écart relatif (barres et axe vertical droit). . . . .	25
2.9	Comparaison des DAS moyenné dans le cerveau du fœtus de la femme japonaise enceinte : en valeur absolue (lignes et axe vertical gauche) et en écart relatif (barres et axe vertical droit). . . . .	26
2.10	Comparaison des DAS corps entiers du fœtus de Victoria enceinte : en valeur absolue (lignes et axe vertical gauche) et en écart relatif (barres et axe vertical droit). . . . .	27
2.11	Comparaison des DAS moyenné dans le cerveau du fœtus de Victoria enceinte : en valeur absolue (lignes et axe vertical gauche) et en écart relatif (barres et axe vertical droit). . . . .	27
2.12	Distribution locale du DAS du fœtus de Victoria pour les deux expositions ondes plane et description en ondes sphériques. . . . .	28
3.1	Schéma illustrant la variabilité des sources d'incertitudes influençant l'exposition humaine. . . . .	32
3.2	Chaîne de propagation de l'incertitude . . . . .	35
3.3	Schéma de définition des paramètres morphologique : carrure et hauteur de l'entrejambe . . . . .	38
3.4	Structure spatiale de la copule Gaussienne en deux dimensions pour $\Theta = 0.7$	43
4.1	Schéma explicatif du principe de la construction d'un méta-modèle . . . . .	57
4.2	Evolution du nombre de polynômes $P$ dans la troncature pleine en fonction du degré $p$ pour différents nombres de paramètres d'entrée $M$ . . . . .	60



4.3	Forme d'une troncature hyperbolique en dimension 2 par rapport à une troncature pleine pour $q = 0.5$ .	60
4.4	Exemple de plan LHS en deux dimensions.	66
4.5	Comparaison des méthodes de planification NLHS et séquentiel maximin pour 50 LHS initiaux différents augmenté jusqu'à 160 points pour la fonction d'Ishigami.	67
4.6	Schéma de l'exposition d'Eartha à la station de base de femtocell indiquant les sources d'incertitudes.	69
4.7	Distribution du DAS corps entier d'Eartha.	69
4.8	Illustration de géométrie d'immeuble possible.	70
4.9	Evolution de la force du champ électrique réfléchi par l'immeuble de la Figure 4.8 dans la direction spéculaire en fonction de la distance d'observation.	71
4.10	Intervalles de confiance à 95% en fonction de la distance d'observation obtenus à partir des distributions de sortie du champ réfléchi obtenues par modélisation par chaos polynomial.	72
4.11	Analyse de signature du modèle de chaos polynomial à $d = 100m$ en direction spéculaire.	72
4.12	Intervalles de confiance à 95% en fonction de l'angle d'observation à 100m obtenus à partir des distributions de sortie du champ réfléchi obtenues par modélisation par chaos polynomial.	74
4.13	Analyse de signature du modèle de chaos polynomial à $d = 100$ m pour un angle d'observation de $10^\circ$	75
4.14	Exemple de modèle de Krigeage pour la fonction $y = x \sin(x)$	80
4.15	Comparaison des méthodes de LARS-PC, de LARS-Kriging-PC et de Krigeage ordinaire pour 50 LHS initiaux différents de 40 points augmentés jusqu'à 160 points par la technique NLHS pour la fonction d'Ishigami.	82
4.16	Comparaison des méthodes de LARS-PC, de LARS-Kriging-PC et de Krigeage ordinaire pour 50 LHS initiaux différents de 50 points augmentés jusqu'à 200 points par la technique NLHS pour la fonction de Borehole.	83
4.17	Comparaison des méthodes de LARS-PC, de LARS-Kriging-PC et de Krigeage ordinaire pour 50 LHS initiaux différents de 50 points augmentés jusqu'à 200 points par la technique NLHS pour la fonction de Sobol.	84
4.18	Représentation géométrique des paramètres d'entrée pour le problème de l'exposition du fœtus à une femtocell	85
4.19	Comparaison des méta-modèles LARS PC, LARS-Kriging-PC et Krigeage ordinaire pour le problème de l'exposition du fœtus à la station de base en utilisant l'erreur relative de la validation croisée <i>leave-one-out</i> en fonction du nombre de points.	86
4.20	Schéma du cas test d'application de l'incertitude morphologique.	87
4.21	Comparaison des métamodèles LARS PC, LARS-Kriging-PC et Krigeage ordinaire pour l'incertitude morphologique en utilisant l'erreur relative de la validation croisée <i>leave-one-out</i> en fonction du nombre de points.	89
4.22	Distribution du DAS corps entier obtenu par propagation de l'incertitude morphologique et de l'incertitude de la source à travers le méta-modèle de LARS-Kriging-PC construit à partir du LHS de 300 points.	90
5.1	Schéma de principe de l'utilisation du bootstrap pour un modèle de régression.	98

5.2	Illustration pas à pas de l'application de la méthode BOAS sur la fonction test $f : x \mapsto x \sin(x)$ . En bleu nuit $f$ (tirets) et $q_{95}$ (trait plein), en rouge $\hat{f}$ (trait plein) et les points de $\mathcal{X}^N$ , en bleu clair $U_\alpha$ (tirets) et $\hat{q}_{95U_\alpha}$ (trait plein), en noir $L_\alpha$ (tirets) et $\hat{q}_{95L_\alpha}$ (trait plein) et en vert $\mathbf{x}^{(N+1)}$ (point sur $f$ ) et les points d'échantillonnage Monte Carlo de $\mathcal{D}^{N+1}$ (axe des abscisses) .	102
5.3	Valeurs moyennes de $\epsilon_r$ , $U_r$ et $1 - Q_w^2$ en fonction du nombre de simulations sur les 50 répétitions du LHS initial pour la fonction d'Ishigami. . . . .	104
5.4	Comparaison des performances des méthodes BOAS et NLHS vis-à-vis de $\epsilon_r$ avec 50 plans LHS initiaux différents pour la fonction d'Ishigami. . . . .	105
5.5	Comparaison des performances des méthodes BOAS et NLHS vis-à-vis de $U_r$ avec 50 plans LHS initiaux différents pour la fonction d'Ishigami. . . . .	105
5.6	Valeurs moyennes de $\epsilon_r$ , $U_r$ et $1 - Q_w^2$ en fonction du nombre de simulations sur les 50 répétitions du LHS initial pour la fonction de Borehole. . . . .	106
5.7	Comparaison des performances des méthodes BOAS et NLHS vis-à-vis de $\epsilon_r$ avec 50 plans LHS initiaux différents pour la fonction de Borehole. . . . .	107
5.8	Comparaison des performances des méthodes BOAS et NLHS vis-à-vis de $U_r$ avec 50 plans LHS initiaux différents pour la fonction de Borehole. . . . .	107
5.9	Valeurs moyennes de $\epsilon_r$ , $U_r$ et $1 - Q_w^2$ en fonction du nombre de simulations sur les 50 répétitions du LHS initial pour la fonction de Sobol. . . . .	109
5.10	Comparaison des performances des méthodes BOAS et NLHS vis-à-vis de $\epsilon_r$ avec 50 plans LHS initiaux différents pour la fonction de Sobol. . . . .	110
5.11	Comparaison des performances des méthodes BOAS et NLHS vis-à-vis de $U_r$ avec 50 plans LHS initiaux différents pour la fonction de Sobol. . . . .	110
5.12	Evolution de $U_r$ et $1 - Q_w^2$ en fonction du nombre de points dans le plan d'expériences pour l'exemple d'application de l'exposition d'un modèle d'enfant.	111
5.13	Représentation de la valeur du quantile à 95% de la distribution du DAS corps entier avec ses bornes de confiance. . . . .	112
5.14	Comparaison des performances des méthodes BOAS et GPS vis-à-vis de $\epsilon_r$ avec 50 plan LHS initiaux différents pour la fonction d'Ishigami. . . . .	114
5.15	Comparaison des performances des méthodes BOAS et GPS vis-à-vis de $U_r$ avec 50 plan LHS initiaux différents pour la fonction d'Ishigami. . . . .	115
5.16	Comparaison des performances des méthodes BOAS et GPS vis-à-vis de $\epsilon_r$ avec 50 plans LHS initiaux différents pour la fonction de Borehole. . . . .	116
5.17	Comparaison des performances des méthodes BOAS et GPS vis-à-vis de $U_r$ avec 50 plans LHS initiaux différents pour la fonction de Borehole. . . . .	116

# Liste des tableaux

3.1	Lois marginales retenues pour les paramètres d'entrée ainsi que leur paramètres inférés . . . . .	39
3.2	Matrice de corrélation de Spearman des paramètres morphologiques pour l'homme adulte . . . . .	40
4.1	Domaine de variation des paramètres d'entrée pour le problème de l'exposition du fœtus à une femtocell . . . . .	68
4.2	Indices de Sobol totaux et de premier ordre des paramètres d'entrée pour l'exposition d'Eartha à une station de base de femtocell obtenus à partir de 1 000 000 simulations de Monte Carlo métamodèle de chaos polynomial. . .	70
4.3	Domaine de variation des paramètres d'entrée pour la fonction de Borehole	83
4.4	Domaine de variation des paramètres d'entrée pour le problème de l'exposition du fœtus à une station de bases de femtocell . . . . .	85
4.5	Indices de Sobol totaux et de premier ordre des paramètres d'entrée pour le problème de l'exposition du fœtus à la station de base obtenus à partir de 1 000 000 simulations de Monte Carlo du BLUP du métamodèle LARS-Kriging-PC pour 200 points. . . . .	87
4.6	Lois marginales des paramètres d'entrée. . . . .	88
4.7	Indices de Kucherenko totaux et de premier ordre des paramètres d'entrée pour le problème de l'incertitude morphologique obtenus à partir de 1 000 000 simulations de Monte Carlo du BLUP du méta-modèle LARS-Kriging-PC pour 300 points. . . . .	89

# Chapitre 1

## Introduction générale



urant les deux dernières décennies, les technologies de communication sans fil ont connu un essor de plus en plus important et leurs usages se sont fortement diversifiés et démocratisés. Par exemple, le taux de pénétration de la téléphonie mobile (nombre d'appareils divisé par le nombre de personnes) n'a cessé d'augmenter ces vingt dernières années pour dépasser les 100% récemment comme l'illustre la Figure 1.1 issue du site internet de l'ARCEP <sup>1</sup>. De plus, cet essor a fait émerger via les différentes technologies successives (le GSM<sup>2</sup> en 1993, le Wi-Fi en 1999, l'UMTS<sup>3</sup> en 2003 et maintenant le LTE<sup>4</sup>) des usages et des modes de consommation de plus en plus variés comme le montre aujourd'hui la multiplication des objets connectés sur le marché [TNS-SOFRES 2014]. Ainsi la quasi totalité de la population est équipée en technologies sans fil et les utilise à travers des applications multiples.

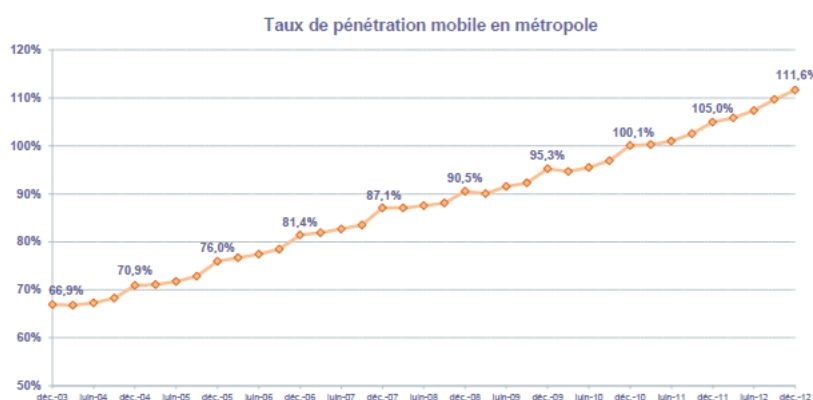


FIGURE 1.1 : Taux de pénétration du mobile dans la métropole française [ARCEP 2013]

En parallèle de ce déploiement massif, des questions relatives aux possibles effets sanitaires des champs électromagnétiques (EMF) émis par les systèmes de télécommunications ont émergé induisant une perception de risque. La gestion et le contrôle de l'exposition des personnes aux EMF est donc une question clef pour les autorités, les sociétés de protection sanitaire et les opérateurs de réseau. Les ondes radios utilisées pour les technologies sans fil sont des rayonnements non ionisants. Afin de prévenir des effets sanitaires connus, l'ICNIRP <sup>5</sup> a défini des limites d'exposition qui sont des *restrictions de base* en  $W/kg$  et des *niveaux de références* dérivés en  $V/m$  [ICNIRP 1998]. En vue d'assurer la conformité des systèmes utilisés à ces limites d'exposition, des scénarii pires cas ont été définis. Mais ces scénarii, s'il permettent de vérifier la conformité aux limites, ne permettent pas d'évaluer précisément l'exposition réelle induite par les systèmes de télécommunication alors que les questions quant à de possibles effets sanitaires en dessous des limites d'exposition restent entières de la part de la population.

Pour répondre aux attentes de la population en matière de connaissance et de compréhension de l'exposition, de gros efforts ont été menés depuis 20 ans dans le domaine

<sup>1</sup>Autorité de Régulation des Communication Electroniques et des Postes

<sup>2</sup>Global system for mobile communication (2G)

<sup>3</sup>Universal mobile telecommunication system (3G)

<sup>4</sup>Long term evolution (4G)

<sup>5</sup>International commission on non-ionizing radiation protection

de la dosimétrie et en particulier dans le développement et l'adaptation des méthodes numériques au bioélectromagnétisme et à la quantification de l'exposition. Aujourd'hui les méthodes numériques telles que les différences finies dans le domaine temporelle (FDTD) permettent de discrétiser et de résoudre les équations de Maxwell [Taflove and Hagness 2005] et d'évaluer l'absorption des ondes électromagnétiques dans le tissu humain caractérisée par le débit d'absorption spécifique (DAS ou SAR en anglais).

Malgré d'énormes progrès réalisés en matière de temps de calcul, le calcul numérique du DAS reste une opération très coûteuse en temps. De plus, les usages de plus en plus diversifiés des technologies sans fil nous conduisent à analyser l'influence de paramètres d'entrée variables sur le calcul du DAS tels que la posture, la fréquence, l'âge etc... Ce management de l'incertitude dans le domaine de la dosimétrie est un vrai challenge et les techniques de propagation d'incertitude comme les techniques de méta-modèles par planification d'expériences numériques sont devenues incontournables dans ce domaine. Par propagation d'incertitude on entend la modélisation de l'influence du caractère aléatoire sur la sortie modèle numérique. En effet, la simulation numérique en dosimétrie étant très coûteuse en temps, le problème de la propagation de l'incertitude ne peut être abordé avec des approches classiques telles que les simulations de Monte Carlo. Le remplacement du modèle numérique par un modèle de substitution beaucoup plus rapide à calculer est apparu dans la dernière décennie comme un puissant outil statistique permettant d'approcher la sortie du code complexe de simulation numérique.

De manière générale, le travail mené dans le cadre de ma thèse vise à étudier la variabilité occasionnée par des paramètres d'entrées incertains sur l'exposition humaine à travers plusieurs techniques de méta-modélisation. L'analyse statistique de l'exposition s'effectuera à travers plusieurs quantités d'intérêt telles que l'analyse de sensibilité qui vise à étudier l'influence relative des paramètres d'entrée sur la sortie du code de calcul ou les quantiles de la distribution de l'exposition. Le budget de calculs étant fortement limité par la simulation numérique, l'intérêt de ce travail réside dans la recherche de parcimonie au niveau de la planification d'expériences pour l'étude de ces quantités d'intérêt.

Ces travaux de thèse s'intègrent dans le cadre de deux projets de recherche : le projet ANR/JST FETUS et le projet ANR CHORUS.

**Le projet FETUS** Le projet FETUS est un projet franco-japonais financé par l'Agence nationale de la Recherche (ANR) en France et par la Japan Science and Technology Agency (JST). Il vise à développer de nouveaux modèles anatomiques de fœtus et de femmes enceintes ainsi que des outils de déformations pour prendre plusieurs postures en compte. Ces nouveaux modèles et outils servent ensuite à étudier l'exposition des fœtus aux émissions électromagnétiques ainsi que l'incertitude associée et l'influence de différents paramètres sur l'exposition.

**Le projet CHORUS** Le projet CHORUS (pour Common Horizon of Open Research in Uncertainty for Simulation) est un projet financé par l'Agence nationale de la Recherche. Il vise à rassembler les nombreuses initiatives autour de la problématique de la gestion et de la propagation de l'incertitude et à donner accès aux mises à jour de la communauté scientifique sur le sujet. Ce projet a pour ambition de solidifier les briques scientifiques et technologiques dans un écosystème adéquat (PME, groupes industriels et universitaires).

Enfin, la démarche générale adoptée dans ce mémoire est de présenter les travaux de la manière dont on aborde la problématique dans le contexte industriel propre au sujet. En

effet, on commence par présenter les outils qui permettent d’aborder la problématique du point de vue physique. Ensuite, on introduit les outils qui nous permettent d’envisager et de construire une analyse statistique autour de la propagation de l’incertitude des paramètres d’entrée à travers le modèle physique présenté préalablement. Enfin, on s’intéresse aux moyens efficaces de propager l’incertitude des paramètres d’entrée qui sont compatibles avec notre exigence de parcimonie. Le corps de ce mémoire se compose de quatre chapitres suivant cette démarche.

Le Chapitre 2 présente tout d’abord le contexte de ces travaux d’un point de vue physique, c’est-à-dire qu’il présente les outils et méthodes numériques permettant de mettre en œuvre des simulations de dosimétrie numérique pour répondre au problème de l’évaluation de l’exposition.

Le Chapitre 3 présente ensuite l’état de l’art de l’analyse statistique dans le domaine de l’exposition humaine aux ondes électromagnétiques et installe le contexte d’une analyse statistique basée sur la propagation de l’incertitude à travers un modèle physique ou numérique. Ainsi les méthodes de modélisation statistique des différents types de paramètres d’entrée que l’on rencontre dans notre contexte et les différentes quantités d’intérêt que l’on utilise pour analyser la distribution de sortie sont détaillées. Cependant aucune considération sur la façon dont on va mettre en œuvre cette propagation pour nos cas d’application n’est abordée en détail dans ce chapitre. Cet aspect est en effet traité dans les chapitres suivants.

Le Chapitre 4, présente la méthodologie de construction de méta-modèles permettant de remplacer le simulateur numérique coûteux en temps de calcul par un modèle mathématique facile à calculer. Ce modèle peut ensuite être utilisé pour la propagation de l’incertitude de manière efficace. Plusieurs approches pour la construction de tels modèles sont présentées : le chaos polynomial, le Krigeage et une méthode d’hybridation des polynômes du chaos polynomial avec le modèle de Krigeage universel qui a fait l’objet d’une publication d’article [Kersaudy et al. 2015]. Les différentes approches sont illustrés par une application sur des cas-test soit analytiques soit issus de problèmes concrets rencontrés en dosimétrie numérique qui ont également fait l’objet de publication [Kersaudy et al. 2014].

Enfin le Chapitre 5 présente une méthodologie de planification d’expériences permettant d’estimer les quantiles à 95% de manière efficace et parcimonieuse. La méthode présentée appelée *BOAS* pour *Bootstraped Oriented Adaptive Sampling*, utilise une méthode de rééchantillonnage des plans d’expériences qui permet de calculer l’incertitude sur l’estimation des quantiles et de restreindre l’espace de recherche pour l’ajout de nouveaux points à la zone d’intérêt pour l’estimation des quantiles. Cette méthode fait également l’objet d’une soumission d’article et est illustrée par des applications sur des exemples analytiques et des cas concrets de dosimétrie comme dans le chapitre précédent.

# Chapitre 2

## L'exposition humaine

### Sommaire

---

<b>2.1</b>	<b>L'évaluation de l'exposition</b> . . . . .	<b>16</b>
<b>2.2</b>	<b>La dosimétrie numérique</b> . . . . .	<b>17</b>
<b>2.3</b>	<b>Les modèles de corps humains</b> . . . . .	<b>18</b>
<b>2.4</b>	<b>La modélisation des différentes sources d'exposition</b> . . . . .	<b>19</b>
2.4.1	L'exposition induite par une source proche . . . . .	19
2.4.2	L'exposition induite par une source lointaine . . . . .	20
2.4.3	L'exposition induite par une source <i>intermédiaire</i> . . . . .	21
<b>2.5</b>	<b>Analyse comparative des expositions fœtales induites par des sources intermédiaires et des sources lointaines.</b> . . . . .	<b>22</b>
2.5.1	Les modèles anatomiques étudiés . . . . .	23
2.5.2	Comparaison de l'exposition à une onde plane avec celle de la description en ondes sphériques. . . . .	24
2.5.3	Résultats de la femme japonaise enceinte . . . . .	25
2.5.4	Résultats pour Victoria enceinte . . . . .	26
<b>2.6</b>	<b>Conclusion</b> . . . . .	<b>28</b>

---





ce chapitre a pour but de présenter le contexte physique des travaux de thèse présentés dans ce mémoire. Premièrement les métriques d'évaluation de l'exposition humaine aux ondes électromagnétiques seront présentées. Ensuite la dosimétrie numérique et ses outils seront introduits suivis d'une réflexion sur la modélisation des différents types de source d'ondes électromagnétiques. Enfin une étude comparative des expositions induites dans des modèles anatomiques de fœtus par des sources dites *lointaines* et des sources dites *intermédiaires* sera présentée.

## 2.1 L'évaluation de l'exposition

L'exposition humaine aux ondes électromagnétiques est quantifiée par le débit d'absorption spécifique (DAS). Il est exprimé en Watt par kilogramme et est défini localement par l'expression suivante à un endroit donné dans un tissu donné :

$$DAS = \frac{\sigma E^2}{2\rho} \quad (2.1)$$

où  $\sigma$  et  $\rho$  représentent respectivement la conductivité et la masse volumique du tissu dans lequel le DAS est calculé et  $E$  représente l'amplitude du champ électrique dans le tissu considéré. Le DAS quantifie la puissance électromagnétique absorbée par le tissu par unité de masse. On peut distinguer deux grandes catégories dans l'évaluation de l'exposition. D'une part, on peut s'intéresser à l'évaluation d'une exposition globale sur un corps humain. On quantifie cette exposition par ce que l'on appelle le "DAS corps entier" qui est défini comme la puissance électromagnétique totale absorbée par le corps divisé par la masse de ce corps :

$$DAS_{CE} = \frac{P_{abs}(corps)}{\mathcal{M}(corps)} \quad (2.2)$$

Cette définition de l'exposition globale peut se décliner à une échelle plus locale si on est intéressé par le DAS global dans un organe par exemple. Dans ce cas on effectue la division entre la puissance totale absorbée par cet organe et sa masse.

D'autre part, une définition locale de l'exposition peut être donnée avec ce que l'on appelle le "DAS10g" où encore le "DAS1g". Ces deux indicateurs d'exposition sont calculés de la même manière que les DAS globaux mais sur des masses de 10 grammes et 1 gramme entourant le point de calcul. Dans les limites d'exposition définies par l'ICNIRP, le maximum du DAS10g dans le corps humain ne doit pas excéder des seuils qui prennent des valeurs différentes selon la partie du corps étudiée ou bien selon que l'on s'intéresse à l'exposition du grand public ou des travailleurs. Par exemple, le DAS10g ne doit pas excéder 4 W/kg sur tout le corps pour le grand public et le seuil est porté à 20 W/kg pour les travailleurs.

En pratique, l'étude de l'exposition passe soit directement par l'évaluation du champ électrique dans les tissus, soit indirectement par l'évaluation du champ incident émis par les antennes et la comparaison aux niveaux de référence défini par l'ICNIRP. Dans le cas d'une source proche tel qu'un téléphone portable, l'évaluation du DAS est permise par la mesure [CENELEC 2001] ou bien par la simulation numérique (FDTD) [Taflove and Hagness 2005]. Dans le cas d'une source lointaine comme une station de base, l'étude de l'exposition passe par l'assimilation du rayonnement de la station de base à une onde plane

et/ou par l'étude du champ induit dans différents milieux de propagation (ex : milieu urbain). Le but de ce chapitre est de présenter la dosimétrie et l'évaluation du DAS induit dans le corps humain par différents types de sources électromagnétiques.

## 2.2 La dosimétrie numérique

La dosimétrie repose sur l'estimation du DAS dans les tissus via la valeur du champ électrique induit dans les tissus biologiques. Dans le cas de la dosimétrie numérique, ces champs induits sont calculés à l'aide de méthodes numériques. Ces 20 dernières années, de gros efforts ont été fournis dans la communauté scientifique pour développer des méthodes numériques permettant de résoudre temporellement et spatialement les équations de Maxwell. On peut citer par exemple la méthode des éléments finis dans le domaine fréquentiel (FEFD), la méthode de Galerkin discontinue dans le domaine temporel (DGTD) ou bien encore la méthode des différences finies dans le domaine temporel (FDTD). Ce travail de thèse est réalisé dans le cadre de l'utilisation de la FDTD comme outil d'évaluation du champ électrique induit. Cette méthode est basée sur une résolution explicite d'équations aux dérivées partielles dépendant du temps. Cette méthode numérique a été développée par Yee en 1966 [Yee 1966] pour résoudre les équations de Maxwell et a depuis prouvé toute son efficacité dans la résolution de problèmes électromagnétiques [Taflove and Hagness 2005]. Elle est majoritairement utilisée dans les études dosimétriques. Dans l'approche de la FDTD, l'espace de calcul  $\mathcal{D} \subset \mathbb{R}^3$  borné est maillé à l'aide de parallélépipèdes orthogonaux  $\mathbf{v} \subset \mathcal{D}$  de dimensions  $\Delta x(\mathbf{v})$ ,  $\Delta y(\mathbf{v})$  et  $\Delta z(\mathbf{v})$  qui servent à représenter des objets en trois dimensions à géométrie complexe et constitués de matériaux hétérogènes ; chaque voxel est ainsi associé à des propriétés diélectriques particulières. Dans ce chapitre, la notation  $\mathbf{v}$  désignera invariablement un voxel parallélépipédique du domaine de calcul  $\mathcal{D}$ . Le calcul du champ électrique s'effectue itérativement dans le domaine temporel sur un intervalle de temps  $[0, T]$  avec un pas de temps  $\Delta T$  constant qui doit respecter les conditions de stabilité [Taflove and Hagness 2005] afin d'assurer la convergence du schéma explicite de la FDTD. La condition de stabilité s'exprime telle que suit :

$$\Delta T \leq \frac{1}{\nu \sqrt{\frac{1}{\Delta x(\mathbf{v})^2} + \frac{1}{\Delta y(\mathbf{v})^2} + \frac{1}{\Delta z(\mathbf{v})^2}}} \quad \forall \mathbf{v} \subset \mathcal{D} \quad (2.3)$$

$\nu$  étant la vitesse de propagation de l'onde dans le milieu. La résolution du maillage doit quant à elle respecter à la fois le critère de stabilité ainsi que le critère de dispersion afin de minimiser la dispersion numérique due à la discrétisation des équations de Maxwell [Taflove and Hagness 2005]. Le critère de dispersion s'exprime de la façon suivante :

$$\frac{\lambda_{min}}{10} \geq \max_{\mathbf{v} \subset \mathcal{D}}(\Delta x(\mathbf{v}), \Delta y(\mathbf{v}), \Delta z(\mathbf{v})) \quad (2.4)$$

$\lambda_{min}$  étant la longueur d'onde minimale des ondes se propageant dans le domaine de calcul. La finitude du domaine de calcul impose de mettre des conditions absorbantes afin d'éviter les réflexions à la frontière  $\partial \mathcal{D}$  de  $\mathcal{D}$ . Dans le travail présenté dans ce rapport, les conditions absorbantes utilisées sont les *perfectly matched layer* (PML) qui ont une grande capacité d'absorption [Berenger 1994]. On considère maintenant un sous-domaine  $\mathcal{D}_t \subset \mathcal{D}$  sur lequel on veut calculer un DAS global tel que défini dans la Section 2.1.  $\mathcal{D}_t$  peut ainsi être un modèle anatomique de corps humain, de fœtus ou encore d'un organe

particulier. Le DAS global sur ce domaine s'exprime alors de la manière suivante dans le cas d'un maillage uniforme :

$$DAS_{\mathcal{D}_t} = \frac{P_{abs}(\mathcal{D}_t)}{\mathcal{M}(\mathcal{D}_t)} = \frac{\sum_{\mathbf{v} \in \mathcal{D}_t} \sigma(\mathbf{v}) E(\mathbf{v})^2}{2 \sum_{\mathbf{v} \in \mathcal{D}_t} \rho(\mathbf{v})} \quad (2.5)$$

en reprenant les mêmes notations que dans les Eq. (2.1) et (2.2).

En ce qui concerne le calcul du DAS local (on prend l'exemple du DAS 10g), on considère un volume de voxels  $\mathcal{D}_{v_c}$  entourant un voxel central  $v_c$  et représentant 10 grammes de tissus, i.e. :

$$\sum_{\mathbf{v} \in \mathcal{D}_{v_c}} \rho(\mathbf{v}) \Delta x(\mathbf{v}) \Delta y(\mathbf{v}) \Delta z(\mathbf{v}) = 10^{-2} \text{ kg} \quad (2.6)$$

On a alors l'expression suivante pour exprimer l'exposition locale moyennée sur 10 grammes de tissus suivant l'expression générale de l'Eq. (2.5) :

$$DAS_{10g}(v_c) = \frac{\sum_{\mathbf{v} \in \mathcal{D}_{v_c}} \sigma(\mathbf{v}) E(\mathbf{v})^2}{2 \sum_{\mathbf{v} \in \mathcal{D}_{v_c}} \rho(\mathbf{v})} \quad (2.7)$$

En général, la valeur d'intérêt que l'on utilise est le maximum de ce DAS 10g sur le sous-domaine d'intérêt  $\mathcal{D}_t$ . Cette définition peut se décliner sur le même modèle pour le DAS 1g. Dans le cadre de ces travaux de thèse, les calculs FDTD sont effectués à l'aide d'un code C développé au sein de l'équipe d'Orange Labs.

Lorsque l'on maille des corps humains entiers pour évaluer leur exposition, les domaines de calcul représentent plusieurs milliards de mailles. Il est donc nécessaire de recourir à des calculateurs puissants et à des méthodes de calcul parallèle pour effectuer les simulations FDTD dans un temps raisonnable. L'équipe du Whistlab au sein de laquelle j'ai réalisé mes travaux dispose ainsi de plusieurs calculateurs multiprocesseurs dédiés au calcul FDTD dont une photo est présentée en Figure 2.1. Sur ces machines, le calcul parallèle s'effectue à l'aide de plusieurs cartes graphiques GPU qui permettent de réduire significativement le temps de calcul d'une simulation FDTD. Malgré ces moyens importants, la durée d'une simulation reste de l'ordre d'une ou plusieurs heures.

## 2.3 Les modèles de corps humains

Dans le cadre de l'exposition humaine aux ondes électromagnétiques, des modèles anatomiques de corps humains sont voxelisés et intégrés dans le domaine de calcul FDTD pour évaluer les champs électriques induits dans les tissus humains. Depuis les années 1960, plusieurs modèles anatomiques ont été développés [Xu and Eckerman 2010]. Le développement de l'imagerie médicale et en particulier de l'imagerie par résonance magnétique (IRM) dans les années 1980 a permis la réalisation de modèles anatomiques de plus en plus réalistes et ayant des résolutions de plus en plus fines. Aujourd'hui un certain nombre de modèles sont disponibles et représentent aussi bien des adultes [Ackerman 1995, Dimbylow 2005] que des enfants [Christ et al. 2010] ou encore des femmes enceintes [Nagaoka et al. 2007, Bibin et al. 2010]. En guise d'exemple, une image du célèbre *Visible Human* [Ackerman 1995] est donnée en Figure 2.2.



FIGURE 2.1 : Photographie des calculateurs FDTD utilisés au sein du Whistlab.



FIGURE 2.2 : Modèle anatomique : *Visible Human*.

## 2.4 La modélisation des différentes sources d'exposition

Pour l'évaluation numérique de l'exposition, la diversité des sources de rayonnements électromagnétiques, leurs usages et leurs localisations diverses supposent des techniques de modélisation adaptées à chaque type de source. Les sous-sections suivantes traitent de cette problématique et de la contribution de mes travaux à ce domaine.

### 2.4.1 L'exposition induite par une source proche

Une source proche est définie ici comme une source de rayonnement utilisée à de très faibles distances du corps humain (i.e. de l'ordre d'une dizaine de centimètres). Cette catégorie de source englobe les téléphones portables, les tablettes tactiles, les ordinateurs portables ainsi que tous les objets connectés qui peuvent être situés à proximité ou dans le corps

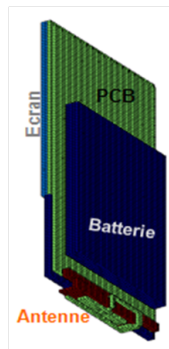


FIGURE 2.3 : Exemple de modèle voxelisé de téléphone portable.

humain. On modélise généralement ce type de source électromagnétique directement dans le domaine FDTD. Différents modèles voxelisés de téléphones portables, de tablettes ou d'ordinateurs ont été réalisés [Ghanmi 2012], un exemple de modèle de téléphone portable est présenté en Figure 2.3. Les antennes de ces modèles sont excitées électriquement par des courants lors de la simulation numérique, ce qui permet d'évaluer le champ induit (et par conséquent le DAS induit) par ces sources dans un modèle anatomique. L'exposition humaine à ces sources, en particulier l'exposition induite par les téléphones portables, a fait l'objet de nombreuses études telles que [Christ and Kuster 2005, Beard et al. 2006].

### 2.4.2 L'exposition induite par une source lointaine

Quand on parle de sources lointaines, on considère que la source du rayonnement électromagnétique est située à de grandes distances du corps humain. Typiquement une telle source lointaine peut être une station de base de macrocell qui sera localisée à des distances variant de la centaine de mètres à quelques kilomètres du corps humain. La taille caractéristique du corps humain étant faible comparée à ces distances, le rayonnement à l'endroit où se situe la personne peut être localement assimilé à une onde plane, c'est-à-dire une onde dont les fronts d'onde sont des plans infinis et perpendiculaires à une même direction de propagation. Pour simuler numériquement ce rayonnement dans la FDTD, on utilise la technique dite de la boîte de Huygens au travers du principe d'équivalence [Merewether et al. 1980, Holland and Williams 1983]. Le principe est tout d'abord d'insérer le modèle anatomique dans une boîte virtuelle ayant approximativement les mêmes dimensions que le domaine FDTD et appelée boîte de Huygens. Ensuite, à l'aide de la description du rayonnement par une onde plane, on calcule les courants équivalents correspondant à cette onde plane tout autour de la boîte de Huygens avec une résolution égale à celle du maillage utilisé. Enfin en utilisant le principe d'équivalence, l'excitation des courants équivalents permet d'évaluer le champ induit par un tel rayonnement dans tout le domaine inclus dans la boîte de Huygens. L'exposition humaine aux ondes planes a également fait l'objet de nombreuses études telles que [Hirata et al. 2009, Conil et al. 2011].



FIGURE 2.4 : Vues externe et interne du système femtocell considéré.

### 2.4.3 L'exposition induite par une source *intermédiaire*

Avec le développement des nouveaux systèmes de télécommunications sans fil, certaines sources de rayonnement se trouvent aujourd'hui trop loin pour ne pas permettre un domaine de calcul incluant la source et la personne (c'est le cas des sources lointaines). Parmi ces sources, on appelle ici sources *intermédiaires* les cas où la distance à la personne reste trop faible pour considérer que le champ incident a une structure d'onde plane. Ces sources peuvent généralement être situées à des distances de l'ordre du mètre du corps humain. Typiquement, les points d'accès wifi et les stations de base de femtocell entrent dans cette catégorie. Dans de nombreuses études de dosimétrie, une approximation en onde plane est utilisée sans être certain que l'hypothèse d'onde plane soit valide. Une des contributions de mon travail présenté dans ce rapport a donc été de proposer une méthode de modélisation de l'exposition induite par les sources *intermédiaires*. Ce travail s'intéresse tout particulièrement aux stations de base de femtocell. Une station de base de femtocell est un point d'accès sans fil connecté au réseau WLAN (*wireless local area network*). Elle offre une couverture radio réduite dans les bandes de fréquences 3G et 4G. Nous nous intéresserons ici à un modèle de station de base NEC dont des vues interne et externe sont présentées en Figure 2.4.

Pour modéliser l'exposition induite par de tels systèmes, la première étape est de modéliser leur rayonnement. Pour se faire, la théorie de décomposition en modes sphériques est utilisée [Hansen 1988]. Le principe est de décomposer le rayonnement d'un système électromagnétique sur une base de modes orthogonaux constituant une base de l'espace électromagnétique. A l'aide de cette décomposition, on peut décrire précisément le rayonnement en espace libre du système considéré par les expressions des Eq. (2.8) et (2.9).

$$\mathbf{E}(r, \theta, \phi) = \frac{k}{\sqrt{\eta}} \sum_{s=1}^2 \sum_{n=1}^{\infty} \sum_{m=-n}^n Q_{smn} \mathbf{F}_{smn}(r, \theta, \phi) \quad (2.8)$$

$$\mathbf{H}(r, \theta, \phi) = -ik\sqrt{\eta} \sum_{s=1}^2 \sum_{n=1}^{\infty} \sum_{m=-n}^n Q_{smn} \mathbf{F}_{3-s,m,n}(r, \theta, \phi) \quad (2.9)$$

Dans les équations ci-dessus,  $\mathbf{E}(r, \theta, \phi)$  et  $\mathbf{H}(r, \theta, \phi)$  sont respectivement les vecteurs des champs électrique et magnétique émis par le système exprimés en fonction des coordonnées sphériques ( $r$  étant le rayon entre le système et le point de calcul,  $\theta$  l'angle d'élévation et  $\phi$  l'angle d'azimut).  $k$  et  $\eta$  sont respectivement la constante de propagation et l'admittance spécifique de l'air.  $Q_{smn}$  et  $\mathbf{F}_{smn}$  sont le coefficient et le mode sphérique d'indice  $s$ , d'ordre  $m$  et de degré  $n$ . En théorie, les émissions électromagnétiques d'un



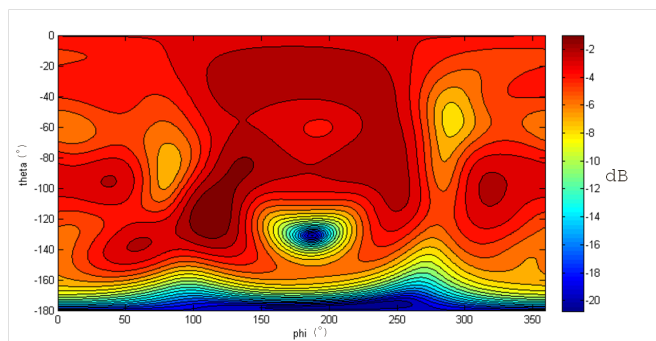


FIGURE 2.5 : Diagramme de rayonnement du système femtocell considéré en fonction des angle d'azimut  $\phi$  et d'élévation  $\theta$ .

système sont décrites en utilisant un nombre infini de modes mais en pratique seul un nombre limité de modes sont nécessaires à une description précise des émissions du système. Ainsi la troncature de l'expression des champs des Eq. (2.8) et (2.9) se fait selon des règles empiriques décrites dans [Hansen 1988] et [Jensen et al. 2004] telles que le nombre de modes considéré soit suffisamment grand pour décrire précisément les émissions.

Hansen a développé en 1988, une méthodologie s'appuyant sur la théorie des modes sphériques permettant de déduire les champs émis par un système émetteur à partir de mesures en champ proche [Hansen 1988]. Cette méthodologie a ici été mise en œuvre pour déduire l'expression des émissions de la station de base de femtocell considérée en espace libre. Le diagramme de rayonnement de ce système a donc pu être calculé et est présenté en Figure 2.5 en fonction des angles d'élévation  $\theta$  et d'azimut  $\phi$ . Ce diagramme est assez omnidirectionnel à part dans deux zones correspondant à  $\theta = -180^\circ$  et  $\theta = -130^\circ$  pour  $\phi = 190^\circ$ . Ces zones correspondent au dessous du système femtocell.

Pour estimer l'exposition induite dans des modèles anatomiques par ce type de système, on utilise ensuite la technique de la boîte de Huygens décrite dans la Section 2.4.2. En effet, la connaissance du rayonnement du système en espace libre nous permet de calculer les courants équivalents de la boîte de Huygens et d'estimer l'exposition induite par utilisation du principe d'équivalence [Merewether et al. 1980, Holland and Williams 1983]. Un schéma de l'utilisation de la boîte de Huygens dans le cas de l'exposition d'une femme enceinte à une station de base de femtocell est donné en Figure 2.6. Ma contribution au développement de la modélisation des sources intermédiaires a été de modifier et d'adapter les codes de description des modes sphériques existants au sein d'Orange Labs au problème de la dosimétrie numérique.

## 2.5 Analyse comparative des expositions fœtales induites par des sources intermédiaires et des sources lointaines.

Cette section analyse l'erreur que l'on peut générer sur la simulation de l'exposition de modèles de fœtus en modélisant une source intermédiaire par une onde plane comme cela peut se faire classiquement au lieu de la modéliser avec une description en ondes sphériques

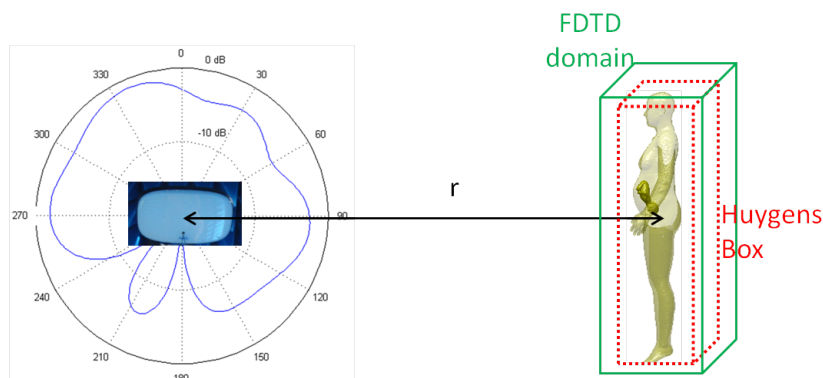


FIGURE 2.6 : Schéma de l'utilisation d'une boîte de Huygens pour l'exposition d'une femme enceinte à une station de base de femtocell.

comme expliquée dans la section 2.4.3. Cette analyse a été conduite dans le cadre du projet franco-japonais FETUS. Ici, on s'intéresse à l'exposition de modèles anatomiques de femmes enceintes à la station de base de femtocell montrée en Figure 2.4. Cette station de base émet à la fréquence de 2100 MHz correspondant à la bande descendante 3G en France, on a donc le diagramme de rayonnement de la Figure 2.5.

### 2.5.1 Les modèles anatomiques étudiés

Les modèles anatomiques utilisés dans ces travaux sont deux modèles de femmes enceintes à 26 semaines d'aménorrhée présentés en Figure 2.7. Le premier modèle de femme enceinte à été réalisé grâce à la déformation d'un modèle hétérogène de femme japonaise construit à partir de données IRM [Nagaoka et al. 2007]. Après cette déformation, un modèle de fœtus hétérogène a été inséré dans l'abdomen du modèle féminin. Dans la suite de cette analyse, ce modèle sera invariablement appelé *femme japonaise enceinte*. Le second modèle utilisé dans cette analyse est un modèle développé dans à la fois dans un projet antérieur appelé FEMONUM et dans le projet FETUS [Bibin et al. 2010, Dahdouh et al. 2014]. La réalisation du modèle de fœtus est basé sur la segmentation d'images ante-natales. Le modèle de fœtus ainsi réalisé est ensuite inséré dans un modèle féminin synthétique et déformable qui est basé sur un modèle féminin homogène appelé Victoria. Ce modèle homogène est seulement muni de tissus basiques tels que la peau, le muscle, la graisse sous-cutanée et les os ajoutés dans le tissu homogène de manière réaliste. Ce second modèle de femme enceinte sera par la suite invariablement appelé *Victoria enceinte*. Concernant les propriétés diélectriques des fœtus, les valeurs des adultes ont été utilisées. Pour le placenta, le liquide amniotique et le cordon ombilical, nous avons utilisé les données de [Peyman et al. 2011] pour définir leurs propriétés diélectriques. Il est intéressant de noter que le fœtus de la femme japonaise enceinte est localisé plus profondément dans l'abdomen que le fœtus de Victoria enceinte. Ces deux modèles fournissent donc deux représentations différentes de l'exposition fœtale.



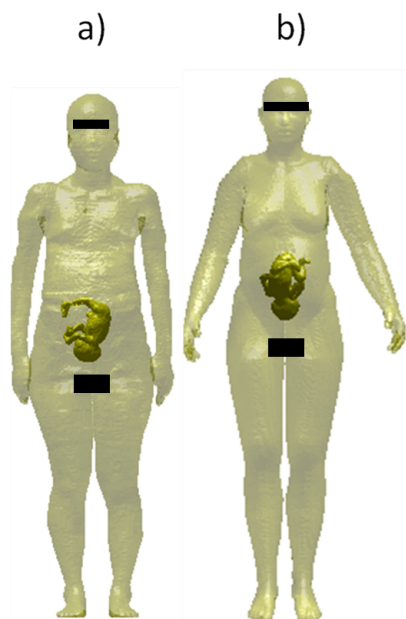


FIGURE 2.7 : Modèles anatomiques de femmes enceintes : (a) femme japonaise enceinte, (b) Victoria enceinte.

### 2.5.2 Comparaison de l'exposition à une onde plane avec celle de la description en ondes sphériques.

Dans le cadre de cette étude, la station de base de femtocell est placée en exposition frontale et à mi-hauteur du modèle anatomique. Cette configuration est visible sur la Figure 2.6. On place ainsi le fœtus en exposition directe et la tête des fœtus se situe à la même hauteur que la station de base. La station de base est orientée de telle manière que la direction définie par les angles  $\theta = 90^\circ$  et  $\phi = 0^\circ$  du diagramme de rayonnement de la Figure 2.5 soit normale à la face avant de la boîte de Huygens. Cette orientation correspond à une configuration d'exposition réaliste où la station de base est posée sur son pied. La puissance rayonnée par la station de base est ici normalisée de telle sorte que la puissance totale rayonnée soit égale à 1 W.

Ainsi pour différentes distances séparant la station de base et les modèles anatomiques, on utilise la description en ondes sphériques des émissions de la station de base et l'exposition directe du fœtus avec la simulation FDTD est simulée via la boîte de Huygens. Les sorties étudiées du code de calcul sont le DAS corps entier du fœtus et le DAS moyen dans son cerveau. On veut ensuite comparer cette exposition de la station de base à l'exposition induite par une onde plane d'amplitude équivalente qui pourrait aussi servir de modèle pour la station de base pour une grande distance. Ici, le terme amplitude équivalente signifie que la puissance de l'onde plane reçue par le fœtus doit être du même ordre de grandeur que la puissance venant de la station de base venant du rayonnement de la station de base décomposé en ondes sphériques. Sans l'assignation de cette amplitude équivalente à l'onde plane, la comparaison entre les deux expositions n'aurait pas de sens. Étant donné que nous nous intéressons à l'exposition du fœtus, que l'exposition est frontale et que la station de base est positionnée à la hauteur du fœtus, on moyenne le module au carré

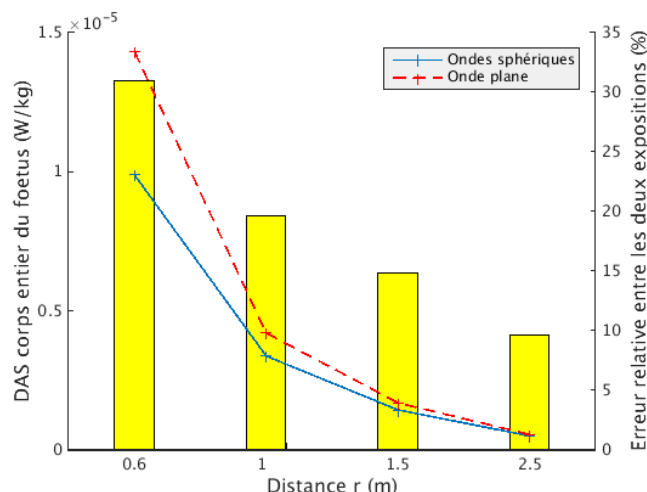


FIGURE 2.8 : Comparaison des DAS corps entiers du fœtus de la femme japonaise enceinte : en valeur absolue (lignes et axe vertical gauche) et en écart relatif (barres et axe vertical droit).

du champ électrique sur une surface rectangulaire encadrant la projection du fœtus sur la face avant de la boîte de Huygens. Ce champ moyen est ainsi assigné à l'amplitude de l'onde plane. En faisant cette moyenne, seule la puissance reçue par le fœtus est prise en compte dans le calcul de l'amplitude équivalente de l'onde plane. Enfin, on compare les deux expositions pour plusieurs distances séparant le modèle anatomique et la station de base :  $r = 0.6$ ,  $1$ ,  $1.5$  et  $2.5$  mètres.

### 2.5.3 Résultats de la femme japonaise enceinte

On présente dans ce paragraphe les résultats de la comparaison entre l'exposition induite par la description en modes sphériques de la station de base et l'exposition induite par l'onde plane équivalente pour le modèle de femme japonaise enceinte. La Figure 2.8 montre la comparaison pour le DAS corps entier du fœtus aux différentes distances.

La représentation en lignes pleines (axe gauche des ordonnées) présente les valeurs absolues de DAS pour la description en ondes sphériques et pour l'onde plane équivalente, la représentation en barres (axe droit des ordonnées) présente l'écart relatif entre les deux valeurs absolues à la distance correspondante. On peut observer que l'écart relatif entre les deux modélisations diminue avec l'augmentation de la distance. Le modèle en onde plane semble donc se rapprocher de la description en ondes sphériques au fur et à mesure que l'on éloigne la station de base du modèle anatomique. Ceci était prévisible car plus la distance entre le modèle anatomique et la station de base est grande, plus l'étalement des valeurs de champs sur la fenêtre servant au calcul du champ moyen se rétrécit. L'amplitude équivalente de l'onde plane qui est la moyenne du champ sur la surface rectangulaire encadrant la projection du fœtus sur la face avant de la boîte de Huygens devient donc de plus en plus représentative de la distribution du champ sur cette surface. On passe d'un écart relatif de 30% sur l'évaluation de l'exposition du corps entier du fœtus à 0.6 m à un écart de 10% à 2.5 mètres. La Figure 2.9 présente les données pour le DAS moyenné dans le cerveau du fœtus sur le même schéma que pour la Figure 2.8. Les valeurs absolues

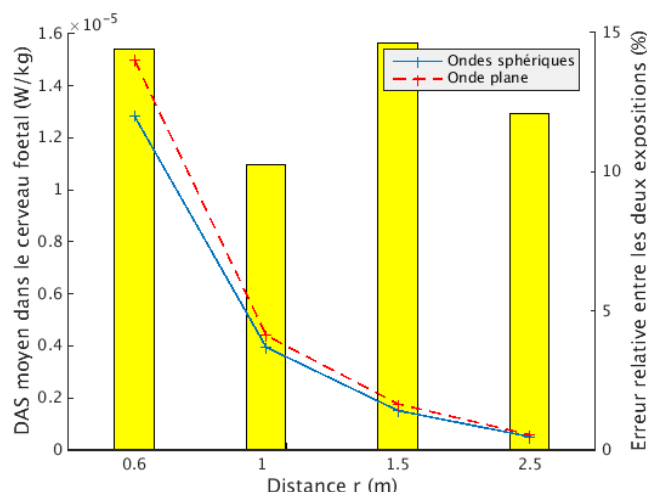


FIGURE 2.9 : Comparaison des DAS moyenné dans le cerveau du fœtus de la femme japonaise enceinte : en valeur absolue (lignes et axe vertical gauche) et en écart relatif (barres et axe vertical droit).

de DAS sont dans ce cas du même ordre de grandeur que celles du DAS corps entier mais les écarts relatifs sont beaucoup plus faibles que dans le cas précédent et n'excèdent pas 15%. Ces valeurs basses ne peuvent bien entendu pas être généralisées car elles sont fortement dépendantes de la configuration du fœtus. En effet, dans le cas présent, la tête de fœtus est à la même hauteur que la station de base. Une autre configuration (la tête en haut par exemple), aurait probablement donné des résultats très différents. Enfin, il est intéressant de noter que dans ce cas l'exposition induite par l'onde plane équivalente surestime l'exposition induite par la description en ondes sphériques.

#### 2.5.4 Résultats pour Victoria enceinte

L'analyse de la précédente section a également été conduite pour le modèle anatomique Victoria enceinte. Les Figures 2.10 et 2.11 présentent les résultats sur le même schéma que les Figures 2.8 et 2.9. La Figure 2.10 présente des valeurs absolues de DAS corps entier significativement supérieures à celles de la Figure 2.8. Ceci peut être dû au fait que pour Victoria enceinte, une grande partie du fœtus (le bas du corps) est située moins profondément dans l'abdomen de Victoria que pour le cas de la femme japonaise enceinte. Ainsi l'épaisseur plus importante de tissu séparant l'air du fœtus de Victoria absorbe une quantité plus importante et la puissance arrivant au fœtus est moindre. Les têtes des deux fœtus sont quant à elles approximativement à la même profondeur, ce qui explique que pour les Figures 2.11 et 2.9, les valeurs absolues soient du même ordre de grandeur.

Concernant le DAS corps entier, la Figure 2.10 présente des écarts relatifs qui diminuent avec la distance comme dans le cas de la femme japonaise enceinte (de 40% à 0.6 m à 8% à 2.5 m). Pour le DAS moyenné dans le cerveau du fœtus, on observe à nouveau des valeurs d'écart relatif très faibles comparées à celle du corps entier (en dessous de 6%). La même explication que dans le cas précédent peut être avancée. Enfin, de même que précédemment, l'exposition induite par l'approximation équivalente en onde plane surestime l'exposition induite par la description en ondes sphériques.

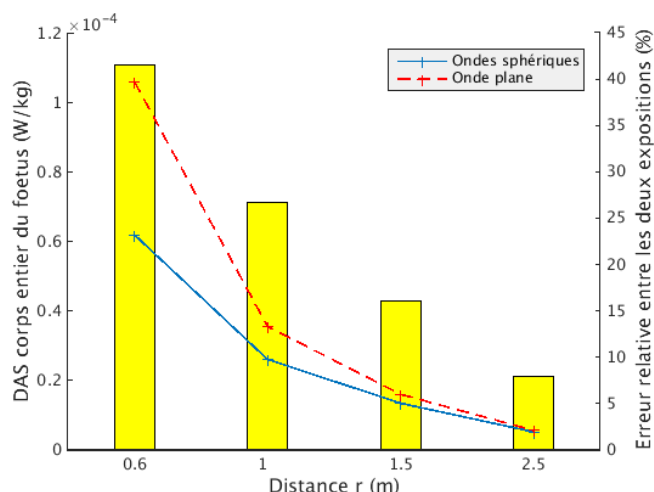


FIGURE 2.10 : Comparaison des DAS corps entiers du fœtus de Victoria enceinte : en valeur absolue (lignes et axe vertical gauche) et en écart relatif (barres et axe vertical droit).

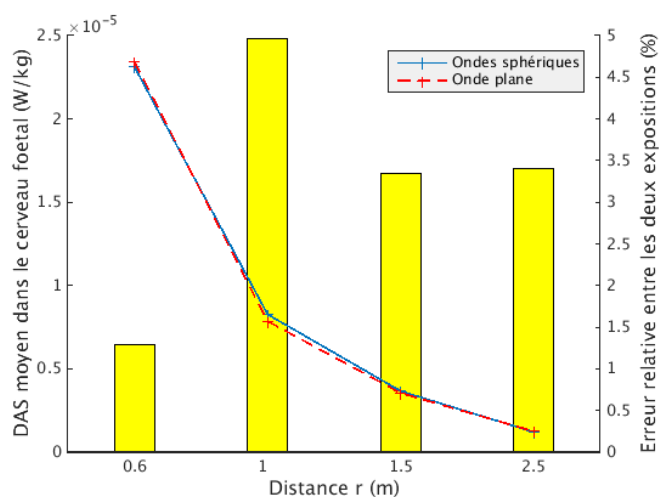


FIGURE 2.11 : Comparaison des DAS moyenné dans le cerveau du fœtus de Victoria enceinte : en valeur absolue (lignes et axe vertical gauche) et en écart relatif (barres et axe vertical droit).

On s'intéresse maintenant à l'exposition locale du fœtus de Victoria pour voir si les expositions convergent avec l'augmentation de la distance. La Figure 2.12 présente la distribution locale de DAS dans le fœtus de Victoria. Cette distribution est représentée pour chaque exposition à toutes les distances étudiées. Les valeurs de DAS sont exprimés en  $10^{-3}$  W/kg et l'échelle de DAS est la même pour les deux expositions à chaque distance. La Figure 2.12 nous montre que les zones d'absorption principales sont les mêmes pour la description en ondes sphériques et pour l'onde plane. Elles sont localisées sur la main droite du fœtus, sur la jambe droite et sur le cordon ombilical. Néanmoins, les niveaux de DAS peuvent être assez disparates et les valeurs maximales de DAS ne sont pas localisées au même endroit en fonction de la distance. En effet, pour l'exposition induite par l'onde

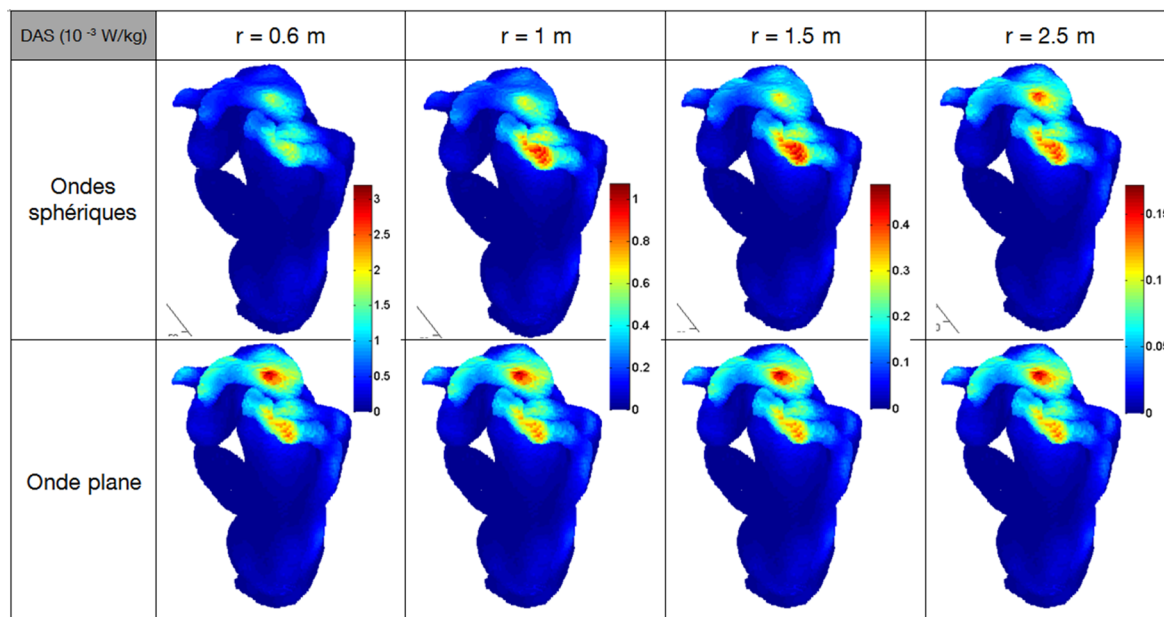


FIGURE 2.12 : Distribution locale du DAS du fœtus de Victoria pour les deux expositions ondes plane et description en ondes sphériques.

plane, la jambe est la zone la plus exposée alors que dans le cas de la description en ondes sphériques, à 1 m et à 1.5 m, la zone la plus exposée est le cordon ombilical. Cet ordre d'intensité de DAS semble s'inverser à 2.5 m et se rapprocher de l'exposition de l'onde plane équivalente. Des simulations additionnelles montrent que pour des distances supérieures, les deux expositions deviennent identiques. A partir de 2.5 m, l'onde plane semble être une bonne approximation de l'exposition induite par une station de base de femtocell si l'on se fie à sa description en ondes sphériques.

## 2.6 Conclusion

Pour conclure, ce chapitre a développé la problématique de l'évaluation de l'exposition ainsi que certaines des réponses apportées par la communauté scientifique à cette problématique. Le principe de la dosimétrie numérique ainsi que les différents types de sources et leur modélisation ont été exposés. Ce chapitre met en particulier l'accent sur la modélisation des sources dites *intermédiaires*. On utilise pour cela une description de rayonnement de ces sources à l'aide d'ondes sphériques et via une boîte de Huygens et d'après le principe d'équivalence, on excite les sources équivalente de la boîte de Huygens dans le domaine FDTD pour modéliser l'exposition induites par ces sources dans les modèles anatomiques. D'après l'étude comparative entre la modélisation avec une onde plane et la modélisation avec la description en ondes sphériques conduite par la suite, il ressort qu'à des distances en deçà de 2.50 mètres, les erreurs potentielles occasionnées par la modélisation en onde plane peuvent être conséquentes. Ce qui nous pousse à conclure sur la pertinence d'une description en ondes sphériques pour modéliser le rayonnement de sources intermédiaires. Néanmoins, on voit sur cet exemple que seules quelques positions relatives de la station de base et des modèles anatomiques sont analysées et qu'il n'est pas possible en l'état

de caractériser avec précision l'exposition induite par les sources étudiées. Cette situation vient du fait que dans la réalité, la position et l'orientation des sources sont aléatoires et qu'il faut prendre en compte cette incertitude pour obtenir la distribution de l'exposition. Ce caractère aléatoire provient ici du fait que dans la réalité la station de base n'est pas toujours située au même endroit en fonction des habitations munies de tels systèmes et que le déplacement des personnes par rapport à ces systèmes modifie constamment les distances et les orientations relatives. Un moyen de prendre en compte cette incertitude est donc de considérer la position et l'orientation relatives comme des paramètres aléatoires qui influencent l'exposition à travers le modèle physique de la dosimétrie numérique. La propagation de l'incertitude des ces paramètres aléatoires que l'on qualifera de *paramètres d'entrée* à travers le modèle physique occasionne donc la représentation de l'exposition elle aussi comme une variable aléatoire que l'on qualifera de *variable de sortie* qu'il convient d'analyser afin de la caractériser. A cette condition, nous pourrions fournir des éléments précis sur la distribution de l'exposition induite dans les modèles anatomiques et sur les influences relatives des paramètres d'entrée. Pour mener à bien une telle entreprise, l'enjeu est double. D'une part, il convient de définir des outils et des quantités d'intérêt qui permettent d'étudier cette distribution. D'autre part, comme évoqué dans la Section 2.2, une simulation permettant l'évaluation du DAS nécessite un temps non négligeable de calcul (de plusieurs dizaines de minutes à plusieurs heures). En conséquence, l'approche de Monte Carlo qui consiste en la génération d'un grand nombre de configurations des paramètres d'entrées indépendantes et identiquement distribuées et en la propagation de ces configurations à travers le modèle physique afin d'obtenir une estimation de la distribution de la variable de sortie, devient inadaptée pour de telles durée de simulation. Par exemple, pour un cas d'application intégrant une station de base de femtocell et un modèle anatomique, le calcul de 100 000 simulations de DAS effectué à partir de 100 000 positions et orientations différentes de la station de base nécessite un temps de calcul total qui n'est pas envisageable.

L'analyse statistique de l'exposition sera donc traitée dans la suite de ce chapitre selon ces deux aspects. Le Chapitre 3 s'intéressera à l'état de l'art de l'analyse statistique de l'exposition humaine aux ondes électromagnétiques ainsi qu'aux outils et quantités d'intérêt nécessaires à l'analyse de la distribution de la variable de sortie sans considérer le problème de la propagation de l'incertitude en lui-même. Ensuite, les Chapitres 4 et 5, s'intéresseront au problème de la propagation d'incertitude et au moyen de répondre à la contrainte de temps de calcul propre à nos cas d'application.



# Chapitre 3

## L'analyse statistique et son application à la problématique de l'exposition.

### Sommaire

---

<b>3.1</b>	<b>Etat de l'art sur l'analyse statistique de l'exposition</b>	<b>32</b>
<b>3.2</b>	<b>Modélisation statistique de l'espace d'entrée</b>	<b>34</b>
3.2.1	Espace probabiliste	34
3.2.2	Paramètres d'entrée spatiaux	36
3.2.3	Paramètres d'entrée morphologiques	37
3.2.4	Transformée isoprobabiliste	42
<b>3.3</b>	<b>Etude statistique de la sortie du modèle physique.</b>	<b>44</b>
3.3.1	Les moments statistiques	44
3.3.2	Etude des quantiles de sortie	45
3.3.3	Analyse de sensibilité globale de paramètres indépendants	45
3.3.4	Analyse de sensibilité globale de paramètres dépendants	48
<b>3.4</b>	<b>Conclusion</b>	<b>53</b>

---



Comme expliqué dans la conclusion précédente, avant toute considération du problème de la propagation de l'incertitude en lui-même dans notre contexte de dosimétrie numérique, ce chapitre s'intéresse à la fois à la présentation générale de la problématique de l'analyse statistique de l'exposition humaine aux ondes électromagnétiques, à l'état de l'art de cette analyse statistique, et aux outils et quantités d'intérêt nécessaires pour l'étude et la caractérisation de l'espace des paramètres d'entrée d'une part et à l'analyse de la distribution de sortie d'autre part.

### 3.1 Etat de l'art sur l'analyse statistique de l'exposition

Si les progrès en terme de capacité de calcul ont été très significatifs dans le domaine de la dosimétrie numérique ces dernières années, le calcul numérique du DAS induit dans des modèles anatomiques reste une procédure couteuse en temps de calcul comme expliqué dans la Section 2.2. Ceci ajouté aux facteurs de variabilité de plus en plus nombreux dus à la diversification des usages et des technologies et à la variabilité humaine (morphologie, posture etc.) comme illustré en Figure 3.1, le management de l'incertitude en dosimétrie devient un vrai challenge.

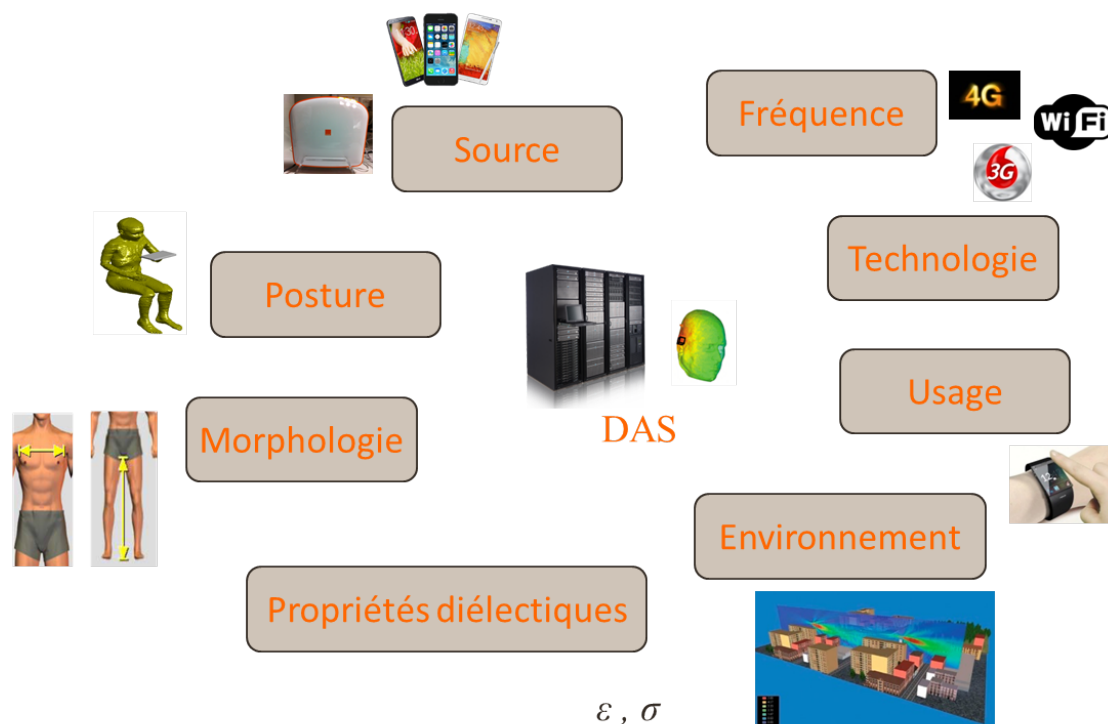


FIGURE 3.1 : Schéma illustrant la variabilité des sources d'incertitudes influençant l'exposition humaine.

En effet, les temps de simulation pouvant être de l'ordre de plusieurs heures, l'incertitude du DAS due aux paramètres d'entrée ne peut être étudiée via une approche classique telle que les simulations de Monte Carlo. Des méthodes avancées de propagation d'incertitude ont donc été proposées pour répondre à ce défi de la variabilité. Dans nos problèmes

où la durée de simulation est longue et où les paramètres spatiaux sont très divers, les approches qui se sont montrées efficaces dans le traitement de l'incertitude sont les approches dites non-intrusives. Par opposition aux méthodes intrusives qui modifient les équations du code de calcul du simulateur numérique (exemple : la FDTD) pour prendre en compte l'incertitude des paramètres d'entrée, l'objectif de telles méthodes est de considérer le modèle FDTD comme une boîte noire et de construire un méta-modèle approchant le comportement de cette boîte noire à partir d'un nombre limité d'observations de la sortie.

En 2008, Silly-Carrete présente dans ces travaux de thèse [Silly-Carette 2008] l'application d'une méthode de projection spectrale pour l'étude de la propagation de paramètres d'entrée incertains tels que l'orientation d'une onde plane ou la variabilité des propriétés diélectriques à travers le modèle physique de la FDTD. La méthode utilisée pour la propagation d'incertitude se base sur la décomposition des champs électriques et magnétiques sur une base de polynômes orthogonaux multivariés fonctions des paramètres d'entrée. Cette méthode est appelée *développement par chaos polynomial*. Dans ces travaux, le calcul des coefficients des polynômes dans cette base se fait par projection sur les sous-espaces engendrés par les polynômes. Dans la résolution d'un problème de projection se basant sur le calcul intégral, des quadratures Gaussiennes sont utilisées pour le calcul des intégrales de projection. Cette approche, si elle permet d'évaluer avec précision la valeur des coefficients de la base polynomiale, présente l'inconvénient de voir le nombre de points de calcul nécessaire à l'estimation de l'intégrale exploser avec le nombre et le degré des polynômes de base (même dans le cas d'utilisation de quadratures creuses). Bien qu'élégante, la mise en œuvre de cette approche est très difficile dans notre cas à cause du budget de calcul conséquent qui nécessiterait des temps de calcul irréalistes.

Si la méthode de résolution ne convenait pas à nos cas d'application, l'approche du développement par chaos polynomial a été reprise en 2012 par Ghanmi [Ghanmi 2012] avec une méthode de résolution différente. En effet, plutôt que de s'orienter vers une approche de projection, ces travaux considèrent la décomposition sur la base polynomiale comme un problème de régression. Les coefficients sont alors estimés par une méthode des moindres carrés. Ces travaux abordent également la question d'ensembles creux de polynômes conservant les polynômes de bases les plus influents sur la sortie dans le but de réduire le coût de calcul. Cette approche, bien que nécessitant une planification d'expériences itérative gouvernée par la qualité du méta-modèle généré, présente l'avantage, via une troncature creuse, de requérir un budget de simulation moins important que la projection. Cette méthodologie est appliquée à l'exposition d'une tête humaine et des parties génitales d'un modèle d'enfant à un téléphone mobile ayant plusieurs degrés de liberté spatiaux.

Avec une approche différente, les travaux de thèse de Habachi [EL Habachi 2011] se consacrent à l'étude de l'influence de la variabilité de la morphologie humaine sur l'exposition corps-entier induite par une onde plane dans des modèles anatomiques déformés. En particulier, la thèse s'attache au problème de l'estimation du quantile à 95% de la distribution de sortie du DAS fonction des paramètres morphologiques d'entrée. Cette estimation de quantile repose ici sur une stratégie de planification d'expériences séquentielle basée sur l'inférence bayésienne d'un modèle paramétrique permettant de raffiner la région du quantile.

Dans cette même optique d'estimation de quantile, Jala développe dans ses travaux [Jala 2013] des méthodologies de planification séquentielle d'expériences orientées vers l'estimation de quantiles des distributions de sortie basées sur la modélisation par processus

Gaussiens (également connue sous le nom de Krigeage dans un cadre plus général). Ces méthodologies sont appliquées à l'évaluation de l'exposition de modèles de fœtus.

Enfin, plus récemment, Jawad [Jawad 2014] utilise le Krigeage pour caractériser l'exposition électromagnétique du corps humain et associe le modèle réduit obtenu à un modèle de canal de propagation stochastique.

Dans mes travaux développés dans ce rapport de thèse, j'ai également utilisé des approches non-intrusives pour traiter le problème de la propagation de l'incertitude des paramètres d'entrée à travers le modèle physique de la FDTD. Les approches utilisées reprennent celles développées dans [Ghanmi 2012] et [Jala 2013].

Mais tout d'abord, avant d'étudier plus en détail la façon dont on propage l'incertitude à l'aide de ces méthodes, il convient de présenter la modélisation de l'espace incertain des paramètres d'entrée ainsi que les différents traitements statistiques envisagés pour la variable de sortie. La suite de ce chapitre est donc consacrée à la modélisation statistique de l'espace d'entrée et à l'analyse statistique de la sortie.

## 3.2 Modélisation statistique de l'espace d'entrée

### 3.2.1 Espace probabiliste

On se place dans le cadre de la propagation de l'incertitude de paramètres d'entrée à travers un modèle physique déterministe dans l'objectif d'analyser le comportement statistique de la sortie de ce modèle. Considérons  $M$  paramètres d'entrée  $\mathbf{x} = (x_1, \dots, x_M)^T \in \mathbb{R}^M$  liés à une variable de sortie  $y$  par un modèle physique déterministe  $f$  tel que  $y = f(\mathbf{x})$ . Soit  $\mathbb{X} \subset \mathbb{R}^M$  l'espace des paramètres d'entrée. Dans le cas de l'exposition humaine, le modèle déterministe est typiquement le simulateur numérique de la FDTD qui calcule le DAS en fonction de la configuration d'exposition. Afin de modéliser l'incertitude des paramètres d'entrée, on se situe dans l'espace probabiliste  $(\Omega, \mathcal{F}, \mathbb{P})$ . Dans cet espace,  $\Omega$  est l'univers,  $\mathcal{F}$  est le corps de Borel sur  $\Omega$ , i.e. l'ensemble des événements et  $\mathbb{P} : \mathcal{F} \rightarrow [0, 1]$  est la mesure de probabilité. Ainsi, l'incertitude affectant  $\mathbf{x}$  nous conduit à sa représentation sous la forme d'un vecteur aléatoire  $\mathbf{X}$  associant à chaque événement élémentaire  $\omega$  un vecteur à valeurs réelles appartenant à l'espace des paramètres d'entrée :

$$\begin{aligned} \mathbf{X} &: \Omega \rightarrow \mathbb{X} \\ \omega &\mapsto \mathbf{X}(\omega) \end{aligned} \quad (3.1)$$

Le vecteur aléatoire  $\mathbf{X}$  est composé des variables aléatoires  $X_1, \dots, X_M$  tel que  $\mathbf{X} = (X_1, \dots, X_M)^T$ . Chaque variable aléatoire est associée à une distribution statistique qui dans le cas de variables aléatoires à valeurs réelles est entièrement décrite par sa fonction de répartition :

$$\begin{aligned} F_{X_i} &: \mathbb{X}_i \subset \mathbb{R} \rightarrow [0, 1] \\ x &\mapsto F_{X_i}(x) = \mathbb{P}(X_i \leq x) \end{aligned} \quad (3.2)$$

$\mathbb{X}_i$  étant le support de la variable aléatoire  $X_i$ . La fonction de répartition au point  $x$  représente donc la probabilité de tirer une observation inférieure à  $x$  en simulant la variable aléatoire  $X_i$ . On définit également la densité de probabilité de la variable aléatoire  $X_i$  comme la fonction  $p_{X_i} : \mathbb{X}_i \rightarrow \mathbb{R}^+$  telle que, pour tout  $A \subset \mathbb{X}_i$  mesurable, l'intégrale suivante ait un sens :

$$\mathbb{P}(X_i \in A) = \int_A p_{X_i}(x) dx \quad (3.3)$$

Les distributions associées à chaque composante du vecteur aléatoire  $\mathbf{X}$  sont appelées *distributions marginales*. Dans l'ensemble de ce mémoire, deux lois de probabilité usuelles sont régulièrement utilisées, la loi uniforme  $\mathcal{U}_{[a,b]}$  et la loi normale  $\mathcal{N}_{\{\mu,\sigma\}}$  aussi appelée loi Gaussienne. La densité de probabilité de la loi uniforme sur  $[a, b]$  avec  $a < b$  se définit comme :

$$p_X(x, a, b) = \begin{cases} \frac{1}{b-a} & \text{si } x \in [a, b] \\ 0 & \text{sinon} \end{cases} \quad (3.4)$$

La densité de la loi normale de moyenne  $\mu$  et d'écart-type  $\sigma$  se définit quant à elle comme :

$$p_X(x, \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x - \mu)^2}{2\sigma^2}\right) \quad (3.5)$$

Dans les travaux présentés dans ce mémoire, on a régulièrement dû générer des échantillons de nos variables aléatoires marginales. Si l'on connaît l'expression analytique de nos fonctions de répartition marginales et que l'on sait inverser ces fonction de répartitions, un moyen simple de générer un échantillon de la variable  $X_i$  est de le faire à partir d'une variable uniforme entre 0 et 1  $\mathcal{U}_{[0,1]}$  que l'on peut aisément générer à partir de tout logiciel d'informatique scientifique (e.g. fonction `rand` sous Matlab). On a en effet les expressions directes et inverses (en se limitant aux cas où  $F_{X_i}$  est inversible et dérivable) :

$$F_{X_i}(X_i) = \mathcal{U}_{[0,1]} \quad \text{et} \quad F_{X_i}^{-1}(\mathcal{U}_{[0,1]}) = X_i \quad (3.6)$$

Si les densités de probabilité marginales suffisent à décrire entièrement les distributions marginales, elles ne sont pas suffisantes pour décrire la distribution statistique du vecteur aléatoire  $\mathbf{X}$  appelée *distribution jointe*. Cette distribution jointe est entièrement décrite par sa fonction de répartition jointe :

$$F_{\mathbf{X}} : \begin{array}{l} \mathbb{X} \rightarrow [0, 1] \\ \mathbf{x} = (x_1, \dots, x_M)^T \mapsto F_{\mathbf{X}}(\mathbf{x}) = \mathbb{P}(X_1 \leq x_1, \dots, X_M \leq x_M) \end{array} \quad (3.7)$$

La densité de probabilité jointe est alors obtenue par dérivation partielle :

$$p_{\mathbf{X}} : \begin{array}{l} \mathbb{X} \rightarrow \mathbb{R}^+ \\ \mathbf{x} \mapsto p_{\mathbf{X}}(\mathbf{x}) = \frac{\partial^M F_{\mathbf{X}}(\mathbf{x})}{\partial x_1 \dots \partial x_M} \end{array} \quad (3.8)$$

Le modèle physique  $f$  étant déterministe, le caractère aléatoire du vecteur  $\mathbf{X}$  se propage à travers  $f$  et affecte la variable de sortie qui est par conséquent elle aussi considérée comme aléatoire. La variable aléatoire de sortie sera donc notée  $Y = f(\mathbf{X})$ . Un schéma de la chaîne d'étude est présenté en Figure 3.2.

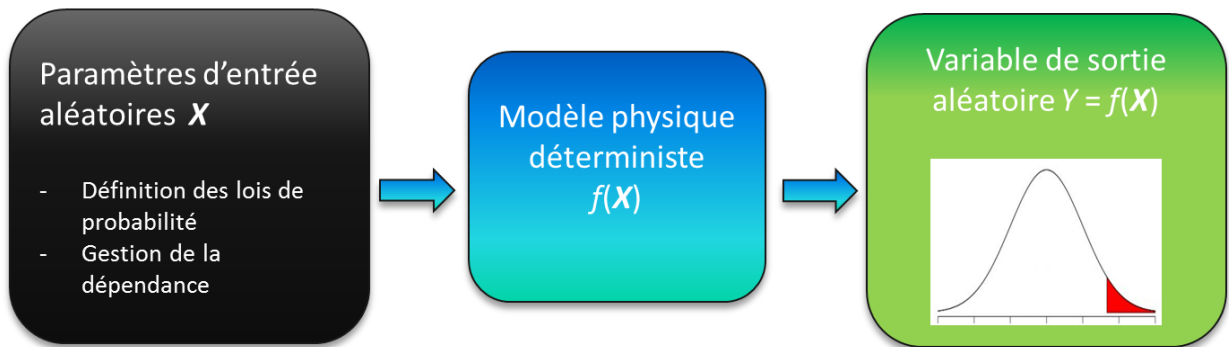


FIGURE 3.2 : Chaîne de propagation de l'incertitude

Si on suppose que l'on possède un générateur d'observations du vecteur  $\mathbf{X}$ , on peut théoriquement propager cette incertitude par Monte Carlo. Le principe est que la distribution statistique de la variable de sortie peut être obtenue par image des observations de  $\mathbf{X}$  par le modèle physique. Cette méthode a l'avantage de fournir des estimateurs des quantités d'intérêt à partir d'un échantillonnage des paramètres d'entrée mais requiert des échantillons de très grande taille pour que obtenir une précision élevée. Dans le cas présent d'étude de l'exposition, le modèle physique de la FDTD impose un temps de calcul très important. Calculer l'image d'un échantillon de plusieurs milliers (voire de plusieurs millions) d'observations par la FDTD n'est pas réalisable. D'autres méthodes doivent donc être employées pour la propagation de l'incertitude qui seront développées dans le chapitre suivant.

### 3.2.2 Paramètres d'entrée spatiaux

On peut distinguer une première grande famille de paramètres d'entrée qui sont traités dans ces travaux de thèse : les paramètres spatiaux. Ces paramètres regroupent toutes les incertitudes liées à la position ou à l'orientation des sources électromagnétiques. Par exemple pour l'exposition induite par une onde plane, il est possible de considérer ses angles d'azimut et d'élévation définissant son orientation par rapport au modèle anatomique comme sources d'incertitude. Pour une source intermédiaire de type station de base de femtocell, on considère sa position par rapport au modèle anatomique et son orientation comme sources d'incertitude. Pour définir les variables aléatoires associées à ces incertitudes, il convient de définir leur distribution statistique comme expliqué dans le paragraphe précédent. Une démarche classique pour définir ces distributions statistiques est d'abord de collecter un échantillon représentatif d'observations des paramètres aléatoires d'entrée. Ces observations sont généralement collectées par la mesure dans un cas réel d'utilisation des systèmes électromagnétiques (si on se situe dans le cadre de ces travaux). On rencontre ce cas de figure avec l'exemple des paramètres liés aux propriétés diélectriques des tissus humain. Des mesures sont effectuées sur des tissus animaux dont la composition est jugée proche de ceux des animaux et la variabilité des propriétés diélectriques est extraite des mesures faites sur ces tissus animaux. Les lois statistiques régissant ces paramètres d'entrée sont ensuite déduites à partir de ces observations par ajustement de familles de lois usuelles. Dans le cas présent des paramètres spatiaux, il est très compliqué de collecter ce genre d'observations. Pour l'onde plane qui modélise une exposition lointaine à une antenne, il faudrait être capable de mesurer en temps réel l'orientation de l'onde incidente sur les personnes alors que dans le cas d'une exposition réelle, l'exposition est multi-composantes du fait des multiples réflexions et des multiples sources. De même dans le cas de l'exposition induite par une station de base de femtocell, il faudrait pouvoir mesurer au cours du temps la position relative de la station de base par rapport à plusieurs personnes et pour plusieurs individus. Par conséquent, dans le cas de ces paramètres spatiaux, on doit définir les distributions statistiques des paramètres d'entrée en se basant sur des hypothèses. On choisit dans le cas présent de définir les paramètres d'entrée de telle sorte que l'on puisse vraisemblablement postuler l'indépendance entre les paramètres. La densité de probabilité jointe peut donc s'exprimer aisément comme le produit des densités marginales :

$$p_{\mathbf{X}} = \prod_{i=1}^M p_{X_i} \quad (3.9)$$

Ensuite, en ce qui concerne la définition des lois marginales, en l'absence d'échantillons d'observations des paramètres d'entrée, la stratégie adoptée a été de définir les bornes de variation des paramètres selon un jugement d'expert et de considérer des lois uniformes au sein de ces intervalles. Par exemple dans le cas d'une femtocell cette hypothèse s'explique facilement. On considère l'exposition directe de la station de base quand le modèle anatomique et la station de base sont supposés être dans la même pièce. Une zone est définie autour du modèle anatomique ayant des dimensions du même ordre de grandeur que la taille d'une pièce d'habitation. Après la définition de paramètres géométriques indépendants au sein de cette zone, on leur affecte des lois uniformes munies des bornes adéquates. L'hypothèse de lois uniformes peut paraître forte mais l'absence de données à disposition nous amène à faire de telles hypothèses. Il est néanmoins important de souligner qu'avec de telles hypothèses des situations paradoxales peuvent arriver où une loi uniforme sur un domaine ne correspond pas toujours à des lois uniformes sur les marginales. On peut prendre l'exemple d'une loi uniforme sur un disque qui donne une loi uniforme sur l'angle mais pas sur le rayon. Mais la question se pose de la pertinence du temps passé à définir la bonne loi marginale du rayon qui respecte l'uniformité sur le disque pour parvenir au respect stricte d'hypothèses qui pourront possiblement être inexactes en regardant ce qui arrive dans la réalité. Le choix a donc été fait de rester avec ces hypothèses fortes dans ces travaux. La même approche de définition de paramètres uniformes a été appliquée dans les travaux de Ghanmi [Ghanmi 2012] pour la prise en compte de l'incertitude de la position d'un téléphone portable lors d'un appel voix et de son influence sur le DAS local dans des modèles anatomiques de têtes ou encore dans la prise en compte de l'incertitude de la position du téléphone et de son influence sur la puissance électromagnétique absorbée par des parties génitales de modèles anatomiques d'enfants.

### 3.2.3 Paramètres d'entrée morphologiques

Les paramètres morphologiques des modèles anatomiques constituent une autre grande famille de paramètres d'entrée traitée dans ces travaux. En effet, la variabilité morphologique des êtres humains apparaît comme une source d'incertitude pouvant influencer la puissance électromagnétique absorbée par le corps humain. Des premiers travaux sur l'incertitude morphologique et l'estimation des quantiles des distributions de sortie ont été menés par El Habachi [El Habachi et al. 2010, EL Habachi 2011]. Ses travaux ont montré une influence des morphologies externes et internes sur la valeur du DAS corps entier des modèles anatomiques. Pour générer différentes configurations de la morphologie externe, des techniques de morphing qui modifient la géométrie des modèles anatomiques [Hirata and Fujiwara 2007, Conil et al. 2008, El Habachi et al. 2010] peuvent être utilisées. Le principe des techniques de morphing est de déformer par partie un modèle anatomique existant pour l'amener vers une nouvelle configuration morphologique. Chaque partie du corps est déformée par homothétie. Si ces méthodes permettent une déformation externe de modèle anatomique, les proportions relatives des tissus internes sont conservées avec la transformation. De façon générale, il est très complexe de prendre en compte une morphologie interne, encore plus de la relier à la morphologie externe. Dans les travaux présentés dans ce mémoire, seule l'influence de la morphologie externe sera étudiée. La suite de cette section s'attache à la détermination des paramètres morphologiques externes, à l'inférence des lois marginales et à l'inférence de la loi jointe de ces paramètres morphologiques.



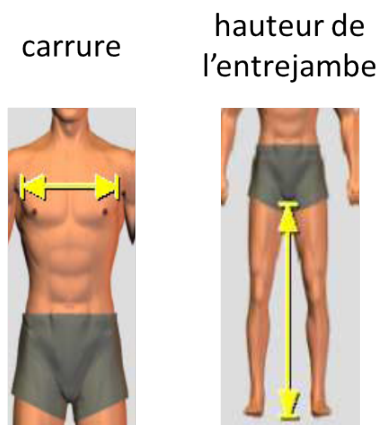


FIGURE 3.3 : Schéma de définition des paramètres morphologique : carrure et hauteur de l'entrejambe

### Inférence des lois marginales

Une campagne antérieure de mesure menée en 2009 a permis de constituer une base de données anthropométriques regroupant des mesures de différents paramètres morphologiques externes de 3887 individus. L'échantillon de ces individus est constitué de 1640 hommes, 2234 femmes et de 644 enfants. Parmi les paramètres morphologiques mesurés, nous en avons retenu 5 pouvant possiblement avoir une influence sur le DAS corps entier des modèles anatomiques pour les travaux présentés dans ce mémoire :

- La stature : la taille de l'individu.
- Le tour de poitrine.
- Le tour de ceinture.
- La carrure de l'individu : voir Figure 4.20
- La hauteur de l'entrejambe : voir Figure 4.20

Comme exemple, le cas de la variabilité morphologique externe des hommes adultes dont l'échantillon est constitué de 1640 individus sera développé ici. A partir de cet échantillon on veut inférer les lois statistiques marginales des paramètres d'entrée. Pour cela des familles de lois statistiques usuelles (loi normale, loi lognormale etc.) sont utilisées et les paramètres de ces lois sont ajustés pour qu'elles soit le plus représentatif possible de l'échantillon à disposition. Afin d'inférer les paramètres des lois marginales, on utilise le principe du maximum de vraisemblance. Considérons un échantillon d'observations indépendantes et identiquement distribuées (i.i.d.)  $\mathbf{x}_i = \{x_i^{(1)} \dots x_i^{(N)}\}$  de la variable marginale  $X_i$ , ainsi qu'une famille de lois de probabilité ayant pour densité de probabilité  $p_{X_i}(x, \boldsymbol{\theta})$  dont on veut inférer le vecteur des paramètres  $\boldsymbol{\theta}$  de taille  $n_i$  (e.g. on fait l'hypothèse que la variable suit une loi normale et on veut inférer sa moyenne et son écart type). La méthode consiste alors en la maximisation de la probabilité d'avoir observé les différentes valeurs de l'échantillon avec la loi en question. Pour cela, la fonction de vraisemblance est définie

de la manière suivante :

$$L(\mathbf{x}_i|\boldsymbol{\theta}) = \prod_{k=1}^N p_{X_i}(x_i^{(k)}, \boldsymbol{\theta}) \quad (3.10)$$

L'inférence du vecteur des paramètres revient alors à résoudre le problème d'optimisation suivant :

$$\hat{\boldsymbol{\theta}} = \arg \min_{\boldsymbol{\theta} \in \mathbb{R}^{n_i}} - \sum_{k=1}^N \ln \left( p_{X_i}(x_i^{(k)}, \boldsymbol{\theta}) \right) \quad (3.11)$$

En pratique, avec le même échantillon, la vraisemblance est ajustée pour plusieurs familles de lois de probabilité usuelles. Si les différentes familles ont le même nombre de paramètres à optimiser (ce qui est le cas ici), on sélectionne la famille ayant la valeur de vraisemblance la plus élevée après ajustement. Dans le cas où les différentes familles n'ont pas le même nombre de paramètres, des critères d'ajustement de vraisemblance faisant intervenir le nombre de paramètres pour sélectionner la famille adéquate peuvent être utilisés. On peut citer par exemple le critère BIC pour *Bayesian Information Criterion* qui tient compte à la fois du nombre de paramètres à inférer et de la taille de l'échantillon qui a servi à inférer la loi [Schwarz 1978]. Enfin, pour juger de la bonne ou mauvaise adéquation entre la loi paramétrique ajustée et l'échantillon, il est recommandé d'utiliser des tests statistiques tels que le test de *Kolmogorov Smirnov* ou encore le test d'*Anderson-Darling*. Il est aussi possible d'utiliser des outils graphiques tels que le *QQ-plot* pour juger de l'adéquation. Le jugement d'expert de l'adéquation à partir de la représentation visuelle des lois est également valable.

Dans le cas de nos paramètres morphologiques, cette méthode a été appliquée et les différentes lois marginales ont été inférées. Il en ressort que les lois utilisées pour modéliser les distributions des paramètres sont des lois de Gumbel et des lois lognormales. Ces deux lois impliquent deux paramètres  $\boldsymbol{\theta} = [\theta_1, \theta_2]$ . La loi lognormale est définie avec la densité de probabilité suivante :

$$p_X(x, \boldsymbol{\theta}) = \frac{1}{x\theta_2\sqrt{2\pi}} \exp \left( -\frac{(\ln(x) - \theta_1)^2}{2\theta_2^2} \right) \quad (3.12)$$

Et la densité de probabilité de la loi de Gumbel se définit comme :

$$p_X(x, \boldsymbol{\theta}) = \exp \left( -\exp \left( \frac{\theta_1 - x}{\theta_2} \right) \right) \quad (3.13)$$

Les résultats des ajustements des différentes lois statistiques sont présentés dans la Table 3.1.

Paramètre d'entrée	Loi	$\theta_1$	$\theta_2$
Stature	Lognormale	5.175	0.044
Tour de poitrine	Gumbel	96.51	7.919
Tour de ceinture	Gumbel	83.68	8.098
Carrure	Lognormale	3.736	0.094
Hauteur de l'entrejambe	Lognormale	4.370	0.066

TABLE 3.1 : Lois marginales retenues pour les paramètres d'entrée ainsi que leur paramètres inférés



### Mesure de la dépendance entre paramètres

Si les distributions marginales sont suffisantes pour définir la loi jointe dans le cas de paramètres indépendants, elle ne le sont absolument pas dans le cas dépendant. Dans le cas des paramètres morphologiques externes, les paramètres peuvent être plus ou moins dépendants les uns des autres. Une première approche pour estimer la dépendance pourrait être de calculer la corrélation linéaire de *Pearson* entre les paramètres. Le coefficient de corrélation de Pearson entre deux paramètres se définit comme la covariance de ces deux paramètres divisée par la racine du produit des deux variances :

$$\rho(X_1, X_2) = \frac{\text{Cov}[X_1, X_2]}{\sqrt{\text{Var}[X_1]\text{Var}[X_2]}} \quad (3.14)$$

Les notions de variance et de covariance seront définies en détail ultérieurement dans la Section 3.3.1 et l'absence temporaire de leur définition ne gêne pas la compréhension générale de cette section. Pour en revenir à la mesure de dépendance, la corrélation de Pearson ne peut pas être utilisée comme une véritable mesure de dépendance car d'une part elle ne décrit que le caractère linéaire de la dépendance et d'autre part sa valeur est dépendante de la modélisation probabiliste des lois marginales. Une alternative réside dans le coefficient de corrélation des rangs de *Spearman*. A l'inverse du coefficient de Pearson qui prend en compte les valeurs des observations de l'échantillon, le coefficient de Spearman prend en compte leur position au sein de l'échantillon. Ce coefficient se définit comme :

$$\rho_s(X_1, X_2) = \rho(F_1(X_1), F_2(X_2)) \quad (3.15)$$

Où  $F_1$  et  $F_2$  sont les deux fonctions de répartition marginales de respectivement  $X_1$  et  $X_2$  et  $\rho$  est la fonction du coefficient de Pearson. On peut également citer le  $\tau$  de *Kendall* [Nelsen 1999] comme mesure de dépendance. Ce coefficient de Spearman est donc calculé pour nos 5 paramètres morphologiques et la matrice de corrélation de Spearman est présentée en Table 3.2

	S	TP	TC	C	LE
S	1	0.18	0.24	0.19	0.89
TP		1	0.80	0.65	0.06
TC			1	0.57	0.07
C				1	0.08
LE					1

TABLE 3.2 : Matrice de corrélation de Spearman des paramètres morphologiques pour l'homme adulte

Il est remarquable que certains paramètres sont extrêmement dépendants et que l'on atteint des corrélations de Spearman supérieures à 0.8. Il apparaît donc important de prendre en compte la structure de dépendance des paramètres pour construire la loi jointe.

### La théorie des copules

La théorie des copules fait référence à une méthodologie permettant de modéliser la dépendance entre variables aléatoires et ainsi d'en déduire la loi de probabilité jointe. Cette

théorie est développée en détail dans le livre [Nelsen 1999]. On en fera ici une description de principe pour illustrer la construction d'une loi de probabilité jointe. Le mot copule apparait pour la première fois dans le théorème de Sklar [Sklar 1959] qui introduit une fonction qui permet de coupler des distributions unidimensionnelles pour construire une distribution multi-dimensionnelle. La version du théorème présentée ici est une traduction de la version énoncée dans [Sklar 1973] mise en forme avec le formalisme de ce mémoire.

**Théorème 1** (Sklar). *Pour  $M \geq 2$ , soit  $F$  une fonction de répartition de dimension  $M$  et les fonctions de répartition marginales unidimensionnelles associées  $F_1, \dots, F_M$ . Alors il existe une fonction  $C$  en dimension  $M$  appelée copule telle que :*

$$\forall \mathbf{x} \in \mathbb{X}, \quad F_{\mathbf{X}}(x_1, \dots, x_M) = C(F_1(x_1), \dots, F_M(x_M)) \quad (3.16)$$

*Inversement, soit  $C$  une copule de dimension  $M$  et  $F_1, \dots, F_M$  un  $M$ -uplet de fonctions de répartition unidimensionnelles. Soit une fonction  $F$  de dimension  $M$  à valeurs réelles définie via l'Eq. (3.16). Alors  $F$  est une fonction de répartition de dimension  $M$  ayant pour fonctions marginales  $F_1, \dots, F_M$ .*

Dans le cas où les fonctions de répartition marginales sont continues sur leur domaine de définition, la copule  $C$  est unique [Sklar 1973]. Les copules sont des fonctions définies sur  $[0, 1]^2$  vérifiant les propriétés suivantes :

1.  $\forall (x, y) \in [0, 1]^2, C(x, 0) = C(0, y) = 0$
2.  $\forall (x, y) \in [0, 1]^2, C(x, 1) = x$  et  $C(1, y) = y$
3. Soit un pavé ayant pour sommets  $(a_1, b_1)$ ,  $(a_1, b_2)$ ,  $(a_2, b_2)$  et  $(a_2, b_1)$  vérifiant  $a_1 \leq a_2$  et  $b_1 \leq b_2$  alors :

$$C(a_1, b_1) - C(a_1, b_2) - C(a_2, b_1) + C(a_2, b_2) \geq 0 \quad (3.17)$$

La première propriété indique que  $C$  vaut zéro sur la borne inférieure de son domaine de définition. La seconde montre que  $C$  est bornée. La troisième fait référence à l'uniformité des marginales et la quatrième indique que les copules sont des fonctions 2-croissantes (généralisable en  $M$ -croissantes). Une autre propriété intéressante est qu'il est possible de lier la copule d'une distribution multivariée à sa mesure de dépendance, i.e. son coefficient de Spearman :

$$\rho_S(X_1, X_2) = 12 \int \int_{[0,1]^2} C(u, v) dudv - 3 \quad (3.18)$$

En pratique, il est très difficile connaître la copule exacte d'une distribution jointe à partir d'un échantillon. Ainsi, de la même manière que pour l'inférence des loi marginales, des familles usuelles de copules sont sélectionnées et ajustées à l'échantillon à disposition. On définit tout d'abord deux copules importantes que l'on nomme les bornes de Fréchet-Hoeffding :

- $W(x_1, x_2) = \max(x_1 + x_2 - 1, 0)$
- $M(x_1, x_2) = \min(x_1, x_2)$

Ces copules essentielles définissent les copules extrémales de toute copule, ainsi pour toute copule  $C$  on a :

$$\forall (x_1, x_2) \in [0, 1]^2, \quad W(x_1, x_2) \leq C(x_1, x_2) \leq M(x_1, x_2) \quad (3.19)$$

Il existe plusieurs familles de copules. On ne présente ici que celle qui nous a permis de modéliser la structure de dépendance de nos paramètres morphologiques : la copule *normale* ou *Gaussienne*. Elle est définie en 2D par l'expression :

$$C(u_1, u_2; \Theta) = \Phi_2(\Phi^{-1}(u_1), \Phi^{-1}(u_2), \Theta) \quad (3.20)$$

Avec  $\Phi$  qui représente la fonction de répartition normale standard univariée (loi normale de moyenne nulle et d'écart-type égal à 1).  $\Phi_2$  représente ici la fonction de répartition de la loi Gaussienne bivariée dont la densité de probabilité s'écrit :

$$\phi_2(x, y, \Theta) = \frac{1}{2\pi\sqrt{|\Theta|}} \exp\left(-\frac{(x \ y)\Theta^{-1}(x \ y)^T}{2}\right) \quad (3.21)$$

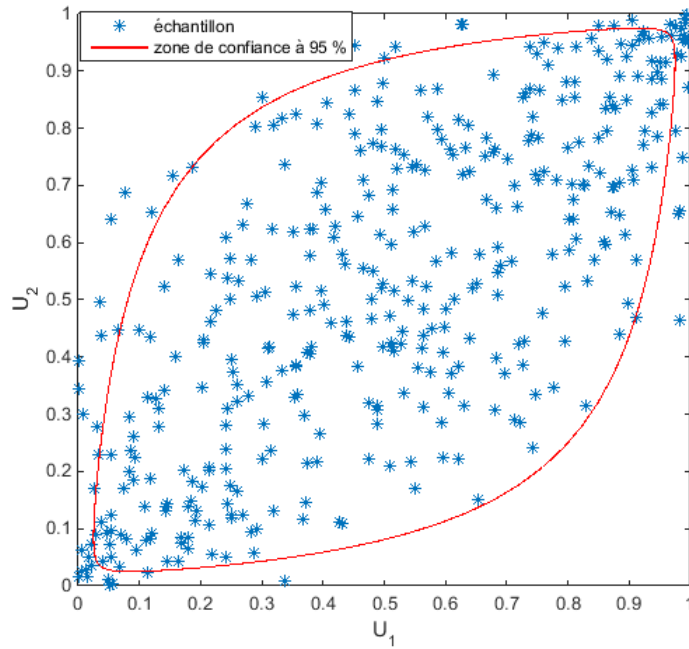
Dans cette équation,  $\Theta$  est la matrice-paramètre de dépendance (prenant ici la place de la matrice de corrélation de la fonction de répartition bivariée) entre les deux variables et s'exprime en fonction du coefficient de Spearman par résolution de l'Eq. (3.18) :

$$\Theta = \begin{bmatrix} 1 & 2 \sin(\frac{\pi}{6}\rho_S) \\ 2 \sin(\frac{\pi}{6}\rho_S) & 1 \end{bmatrix} \quad (3.22)$$

La Figure 3.4 présente la structure spatiale de la copule Gaussienne en montrant un échantillon entouré de la courbe d'isodensité de la copule délimitant un région de probabilité à 95 %. Ainsi, lorsqu'on a un échantillon des paramètres d'entrée comme dans notre cas pour les paramètres morphologiques, il est possible d'approcher sa copule en postulant son appartenance à une certaine famille sur la base de l'observation de la forme des différentes projections du nuage de points. Pour le cas morphologique, on postule une copule Gaussienne. En prenant comme point de départ l'estimation empirique de  $\hat{\Theta}$  faite à partir de la matrice de corrélation de Spearman, la vraisemblance de l'échantillon est ensuite maximisée et la matrice-paramètre  $\Theta$  est inférée de la même manière que les paramètres des lois marginales. Enfin, un test statistique de Cramer-Von Mises est effectué pour juger de la bonne adéquation de la copule postulée. On obtient donc finalement une modélisation de la distribution jointe de nos 5 paramètres aléatoires morphologiques.

### 3.2.4 Transformée isoprobabiliste

Bien que nous ayons défini la manière d'obtenir les lois jointes et marginales des paramètres d'entrée, dans la suite de ce mémoire, on voudra régulièrement pouvoir se ramener à un espace d'entrée indépendant. C'est en effet un cadre pratique pour la construction de ce qu'on appelle les *plans d'expériences*. Il faut donc introduire des fonctions permettant de passer de l'espace d'entrée réel à un espace équivalent indépendant. C'est ce que l'on appelle les *transformées isoprobabilistes*. La transformée de Nataf [Nataf 1962], qui permet de se ramener à un espace standardisé à partir de l'espace physique des paramètres d'entrée dont la copule est supposée Gaussienne, est ici utilisée.


 FIGURE 3.4 : Structure spatiale de la copule Gaussienne en deux dimensions pour  $\Theta = 0.7$ 

**Definition 3.2.1** (Transformée de Nataf). Soit  $\mathbf{X} = (X_1 \dots X_M)^T \in \mathbb{X}$  un vecteur aléatoire muni d'une fonction de répartition jointe  $F_{\mathbf{X}}$  définie par ses fonctions marginales  $F_1, \dots, F_M$  et sa copule Gaussienne  $C_{\Theta}$ . La transformée de Nataf est définie par :

$$\begin{aligned} T_N &: \mathbb{X} \rightarrow \mathbb{R}^M \\ \mathbf{X} &\mapsto \mathbf{U} = T_N(\mathbf{X}) = T_2 \circ T_1(\mathbf{X}) \end{aligned} \quad (3.23)$$

Avec  $T_1$  et  $T_2$  définies respectivement par :

$$\begin{aligned} T_1 &: \mathbb{X} \rightarrow \mathbb{R}^M \\ \mathbf{X} &\mapsto \mathbf{Y} = \begin{pmatrix} \Phi^{-1} \circ F_1(X_1) \\ \vdots \\ \Phi^{-1} \circ F_M(X_M) \end{pmatrix} \end{aligned} \quad (3.24)$$

$$\begin{aligned} T_2 &: \mathbb{R}^M \rightarrow \mathbb{R}^M \\ \mathbf{X} &\mapsto \mathbf{U} = \mathbf{L}^{-1}\mathbf{Y} \end{aligned} \quad (3.25)$$

Dans les équations ci-dessus,  $\Phi$  représente la fonction de répartition normale standard univariée,  $\circ$  est l'opérateur composition de fonctions et  $\mathbf{L}$  représente la décomposition de Cholesky de la matrice de corrélation  $\Theta$ .

L'espace de  $\mathbf{U}$  est donc caractérisé comme l'espace *standardisé*, composé de variables indépendantes avec des distributions Gaussiennes standards. L'espace d'origine  $\mathbb{X}$  est alors appelé l'espace *physique*. La transformée de Nataf implique l'utilisation d'une copule normale. Dans le cadre des applications traitées dans ce mémoire, cette limitation à une copule

normale est suffisante pour modéliser les fonctions de répartition multivariées. Il est néanmoins intéressant d'ajouter que la définition cette transformée de Nataf peut être généralisée pour toute copule elliptique [Lebrun and Dutfoy 2009]. Comme expliqué précédemment, la matrice  $\Theta$  est ici estimée par maximum de vraisemblance avec pour initialisation son estimation à partir de la matrice de Spearman estimée. Pour toute loi jointe construite avec une copule elliptique, on peut donc se ramener à un espace équivalent indépendant et standardisé qui sera plus simple d'utilisation pour la suite. Il est évidemment possible de faire le chemin dans l'autre sens en prenant la transformée inverse.

### 3.3 Etude statistique de la sortie du modèle physique.

Avant de s'intéresser à la façon dont on propage l'incertitude de ces variables d'entrées à travers le modèle physique de la FDTD, on présente les objectifs poursuivis quant à la variable aléatoire de sortie du DAS, i.e. les traitements statistiques que l'on veut effectuer sur la distribution de sortie.

#### 3.3.1 Les moments statistiques

Commençons par présenter les moments statistiques principaux : l'espérance et la variance de la variable de sortie. Ces moments donnent des premières indications sur la situation et l'étalement de la distribution de sortie. L'espérance (ou premier moment) de la distribution est définie par :

$$\mathbb{E}[Y] = \mu_Y = \int_{\mathcal{Y}} y \mathbb{P}_Y(dy) \quad (3.26)$$

Où  $\mathcal{Y}$  est le support de la variable de sortie. Dans le cas où la variable aléatoire admet une densité de probabilité on peut écrire l'espérance sous la forme :

$$\mathbb{E}[Y] = \int_{\mathcal{Y}} y p_Y(y) dy \quad (3.27)$$

Cette espérance est également appelée valeur moyenne de la distribution. En supposant la variable aléatoire de carré intégrable, on définit ensuite la variance (ou second moment) de la distribution de sortie par :

$$\text{Var}[Y] = \sigma_Y^2 = \mathbb{E}[(Y - \mathbb{E}[Y])^2] = \mathbb{E}[Y^2] - \mathbb{E}[Y]^2. \quad (3.28)$$

En développant l'expression à l'aide de l'Eq. (3.27) toujours dans le cas où la variable aléatoire  $Y$  admet une densité de probabilité, il vient :

$$\text{Var}[Y] = \int_{\mathcal{Y}} (y - \mu_Y)^2 p_Y(y) dy \quad (3.29)$$

On parle aussi couramment d'*écart type* pour les variables aléatoires. Cet écart type noté  $\sigma_Y$  n'est rien d'autre que la racine carrée de la variance :  $\sigma_Y = \sqrt{\text{Var}[Y]}$ . Il est alors possible de définir ce que l'on appelle le *coefficient de variation* de la distribution de sortie :

$$CV_Y = \frac{\sigma_Y}{\mu_Y} \quad (3.30)$$

Ce coefficient de variation caractérise l'étalement de la distribution de sortie relativement à la valeur de sa moyenne. Si on considère un couple de variables aléatoires  $\{Y_1, Y_2\}$ , il est important d'introduire la notion de covariance de ces deux variables qui se définit comme :

$$\text{Cov}[Y_1, Y_2] = \mathbb{E}[(Y_1 - \mathbb{E}[Y_1])(Y_2 - \mathbb{E}[Y_2])] \quad (3.31)$$

Cette covariance sert à définir la variance du couple de variables aléatoires qui s'exprime sous la forme d'une matrice appelée la matrice de covariance du couple de variable (qui est proportionnelle à la matrice de corrélation de Pearson dans le cas de matrices de diagonales constantes – cf Eq. (3.14)). Enfin, il peut également être intéressant de calculer des moments d'ordre supérieur. Ainsi le moment d'ordre  $n$  de la distribution de sortie noté  $\mu_Y^n$  est défini par :

$$\mu_Y^n = \mathbb{E}[(Y - \mathbb{E}[Y])^n] \quad (3.32)$$

On peut par exemple citer le moment d'ordre 3 communément appelé le *coefficient d'asymétrie* (ou *skewness*) qui comme son nom l'indique, permet de quantifier l'asymétrie de la distribution, ou encore le moment d'ordre 4 communément appelé *coefficient d'aplatissement* (ou *kurtosis*) qui permet de mesurer l'aplatissement de la distribution. En supposant l'existence de la densité de probabilité de la variable de sortie, ces deux moments nous renseignent donc sur sa forme.

### 3.3.2 Etude des quantiles de sortie

Le quantile  $y_\xi$  est défini tel que la probabilité d'observer une valeur supérieure à  $y_\xi$  soit égale à  $1 - \xi$  :

$$\mathbb{P}(Y \leq y_\xi) = F_Y(y_\xi) = \xi \quad (3.33)$$

Dans le contexte de ces travaux, on veut également étudier certains quantiles de la distribution de sortie, en particulier les quantiles élevés (et plus précisément encore le quantile à 95 %). En effet dans le cas de notre problématique d'exposition, il est intéressant de connaître la valeur de DAS dont on est certain que 95% des configurations rencontrées donneront lieu à une exposition inférieure à cette valeur. L'approche classique pour estimer un quantile est de le faire à partir d'une approximation de la fonction de répartition  $F_Y$ . Si un échantillon i.i.d. de  $n$  observations de la variable  $Y$   $\{y_1, \dots, y_n\}$  est à disposition, on peut construire un estimateur empirique de la fonction de répartition à partir de ces observations :

$$\hat{F}_Y(y) = \frac{1}{n} \sum_{i=1}^n \mathbf{1}_{y_i \leq y} \quad (3.34)$$

L'estimateur standard du quantile  $y_\xi$  est alors :

$$\hat{y}_\xi = \inf \left\{ y, \hat{F}_Y(y) \geq \xi \right\} = y_{(\lceil \xi n \rceil)} \quad (3.35)$$

Où  $\lceil \xi n \rceil$  est la valeur entière supérieure de  $\xi n$  et  $y_{(i)}$  est la  $i$ -ème statistique d'ordre.

### 3.3.3 Analyse de sensibilité globale de paramètres indépendants

En considérant la chaîne  $Y = f(\mathbf{X})$ , il est également pertinent d'étudier les influences relatives des paramètres d'entrée sur le comportement de la variable de sortie. Pour cela,

une analyse de sensibilité est menée pour hiérarchiser les influences des paramètres d'entrée sur la variance de sortie  $\text{Var}[Y]$ . Pour estimer ces influences relatives, les indices de Sobol' [Sobol' 2001] sont utilisés. Dans la théorie originale de Sobol', on considère que le modèle physique  $f$  supposé continu est défini sur le cube unité  $I^M = [0, 1]^M$  et que les paramètres d'entrée sont statistiquement indépendants. Il est intéressant de noter que si les paramètres sont originellement indépendants, l'étude de paramètres d'entrée associés à des distributions quelconques est équivalente à l'étude de paramètres uniformes sur le cube unité par transformée isoprobabiliste. En cas de dépendance, on ne peut utiliser la transformée de Nataf pour calculer les indices de Sobol' car les indices calculés dans l'espace équivalent ne représentent plus les mêmes paramètres à cause de la dépendance qui modifie la structure spatiale de l'espace d'entrée. On se limitera donc dans cette section au cas de paramètres indépendants et uniformes sur le cube unité comme décrit dans [Sobol' 2001]. Le calcul des indices de Sobol' repose sur ce que l'on appelle la *décomposition de Sobol'* de  $f(\mathbf{x})$  :

$$f(x_1, \dots, x_M) = f_0 + \sum_{i=1}^M f_i(x_i) + \sum_{1 \leq i < j \leq M} f_{ij}(x_i, x_j) + \dots + f_{1,2,\dots,M}(x_1, \dots, x_M) \quad (3.36)$$

Dans l'équation ci-dessus,  $f_0$  est une constante qui correspond à la valeur moyenne de  $f$  sur  $I^M$  et l'intégrale de chaque composante  $f_{i_1, \dots, i_s}(x_{i_1}, \dots, x_{i_s})$  sur n'importe lequel de ses paramètres d'entrée est égale à zéro :

$$\int_0^1 f_{i_1, \dots, i_s}(x_{i_1}, \dots, x_{i_s}) dx_{i_k} = 0 \quad k \in \llbracket 1, s \rrbracket \quad (3.37)$$

Cette décomposition de Sobol' suit plusieurs propriétés :

1. Le nombre de composantes de la décomposition est égal à

$$\sum_{i=1}^M \binom{M}{i} = 2^M \quad (3.38)$$

2. Avec l'Eq. (3.37), les composantes sont orthogonales les unes aux autres au sens du produit scalaire  $\langle \phi, \psi \rangle = \mathbb{E}[\phi(\mathbf{X})\psi(\mathbf{X})]$  :

$$\int_{I^M} f_{i_1, \dots, i_s}(x_{i_1}, \dots, x_{i_s}) f_{j_1, \dots, j_s}(x_{j_1}, \dots, x_{j_s}) d\mathbf{x} = 0 \quad \text{for } \{i_1, \dots, i_s\} \neq \{j_1, \dots, j_s\} \quad (3.39)$$

Dans l'intégrale précédente,  $p_{\mathbf{X}}(\mathbf{x})$  n'apparaît pas car les paramètres étant uniformes et indépendants sur le cube unité, ce terme est égal à 1 sur  $I^M$ .

Avec les notations de la décomposition de l'Eq. (3.36), la variance totale  $V_T = \text{Var}[Y]$  de  $f(\mathbf{X})$  s'exprime comme suit :

$$V_T = \int_{\mathbf{X}} f^2(\mathbf{x}) d\mathbf{x} - f_0^2 \quad (3.40)$$

Ainsi, en intégrant l'Eq. (3.36) mise au carré, on peut décomposer la variance totale suivant l'expression :

$$V_T = \sum_{i=1}^M V_i + \sum_{1 \leq i < j \leq M} V_{ij} + \dots + V_{1,2,\dots,M} \quad (3.41)$$

Avec chaque composante de *variance partielle* qui s'écrit :

$$V_{i_1, \dots, i_s} = \int_{I^M} f_{i_1, \dots, i_s}^2(x_{i_1}, \dots, x_{i_s}) d\mathbf{x} \quad \text{for } \{i_1, \dots, i_s\} \neq \{0\} \quad (3.42)$$

Les indices de Sobol' sont alors définis comme :

$$S_{i_1, \dots, i_s} = \frac{V_{i_1, \dots, i_s}}{V_T} \quad (3.43)$$

On utilise couramment ce que l'on appelle les *indices du premier ordre* et les *indices totaux* pour estimer l'influence de chaque paramètre d'entrée sur la variance totale de sortie. Les indices du premier ordre  $S_i$  évaluent l'influence de chaque paramètre pris individuellement et les indices totaux évaluent l'effet total du paramètre d'entrée sur la variance totale. L'indice total associé à la  $i$ -ème variable s'exprime comme :

$$S_{T_i} = \frac{\sum_{\mathcal{I}_i} V_{i_1, \dots, i_s}}{V_T} = 1 - S_{\sim i} \quad (3.44)$$

où l'on a :

$$\mathcal{I}_i = \{(i_1, \dots, i_s) : \exists k, k \in \llbracket 1, s \rrbracket, i_k = i\} \quad (3.45)$$

et où  $S_{\sim i}$  représente la somme de tous les  $S_{i_1, \dots, i_s}$  n'incluant pas le paramètre  $x_i$ . Ainsi une différence importante entre les indices de premier ordre et les indices totaux est interprétée comme une forte influence des termes de variances partielles croisés incluant le paramètre  $x_i$  et donc comme autant d'interactions entre le paramètre  $x_i$  et les autres paramètres.

Il est également possible de considérer les indices de sensibilité d'ensemble de variables prises conjointement. Soit le sous-ensemble de variables d'entrée  $\mathbf{u} = \{x_{i_1}, \dots, x_{i_s}\}$ , l'ensemble des indices correspondant  $K = \{i_1, \dots, i_s\}$  et le sous ensemble complémentaire  $\mathbf{v} = \{x_{i_{M-s}}, \dots, x_{i_M}\}$ . La variance correspondant au sous ensemble de variables  $\mathbf{u}$  est définie comme suit :

$$V_{\mathbf{u}} = \sum_{J \subset K} V_J \quad (3.46)$$

Où si  $J = \{i_1, \dots, i_j\}$ ,  $V_J = V_{i_1, \dots, i_j}$ . L'indice de sensibilité associé s'écrit logiquement :

$$S_{\mathbf{u}} = \frac{V_{\mathbf{u}}}{V_T} \quad (3.47)$$

L'indice total correspondant s'exprime quant à lui comme :

$$S_{T_{\mathbf{u}}} = 1 - \frac{V_{\mathbf{v}}}{V_T} \quad (3.48)$$

Afin de faciliter la compréhension, on prend l'exemple de sous-ensemble de variables  $\{x_1, x_2\}$  avec un nombre total de 3 variables, les indices ci-dessus se définissent alors de la manière suivante :

$$S_{\{x_1, x_2\}} = S_1 + S_2 + S_{12} \quad (3.49)$$

$$S_{T_{\{x_1, x_2\}}} = S_1 + S_2 + S_{12} + S_{13} + S_{23} + S_{123} = 1 - S_3 \quad (3.50)$$

Des formules pour le calcul des indices par simulations de Monte Carlo sont disponibles dans [Sobol' 2001]. Ces formules ont l'avantage de ne pas nécessiter la connaissance de



la décomposition de l'Eq. (3.36). Pour déduire ces formules, il est montré dans [Sobol' 2001] que la variance partielle  $V_{\mathbf{u}}$  s'exprime selon une formulation intégrale avec deux réalisations indépendantes  $\mathbf{x} = (\mathbf{u}, \mathbf{v})$  et  $\mathbf{x}' = (\mathbf{u}', \mathbf{v}')$  du vecteur aléatoire  $\mathbf{X}$  :

$$V_{\mathbf{u}} = \int_{I^{2M-s}} f(\mathbf{x})f(\mathbf{u}, \mathbf{v}')d\mathbf{x}d\mathbf{v}' - f_0^2 \quad (3.51)$$

Considérons maintenant deux échantillons distincts de  $N$  simulations de Monte Carlo  $\mathcal{X} = \{\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(N)}\}$  et  $\mathcal{X}' = \{\mathbf{x}'^{(1)}, \dots, \mathbf{x}'^{(N)}\}$  des paramètres d'entrée qui peuvent évidemment se décomposer comme  $\mathcal{X} = \{\mathcal{U}, \mathcal{V}\}$  et  $\mathcal{X}' = \{\mathcal{U}', \mathcal{V}'\}$ . Les formules de Monte Carlo des indices de Sobol' (de premier ordre et totaux) s'expriment donc comme suit [Sobol' 2001] :

$$\hat{S}_{\mathbf{u}} = \frac{\frac{1}{N} \sum_{i=1}^N (f(\mathbf{u}^{(i)}, \mathbf{v}^{(i)})f(\mathbf{u}^{(i)}, \mathbf{v}'^{(i)})) - \left(\frac{1}{N} \sum_{i=1}^N f(\mathbf{u}^{(i)}, \mathbf{v}^{(i)})\right)^2}{\hat{V}_T} \quad (3.52)$$

$$\hat{S}_{T_{\mathbf{u}}} = \frac{\frac{1}{N} \sum_{i=1}^N (f(\mathbf{u}^{(i)}, \mathbf{v}^{(i)}) - f(\mathbf{u}'^{(i)}, \mathbf{v}^{(i)}))^2}{2\hat{V}_T} \quad (3.53)$$

où  $\hat{V}_T$  est l'estimateur empirique de Monte Carlo de la variance totale de la sortie :

$$\hat{V}_T = \frac{1}{N} \sum_{i=1}^N f(\mathbf{u}^{(i)}, \mathbf{v}^{(i)})^2 - \left(\frac{1}{N} \sum_{i=1}^N f(\mathbf{u}^{(i)}, \mathbf{v}^{(i)})\right)^2 \quad (3.54)$$

Par la suite on verra que l'on aura d'autres moyens que la méthode de Monte Carlo pour estimer ces indices. Néanmoins, chaque fois que l'on parlera d'estimation d'indice par Monte Carlo, on fera référence à ces formules.

### 3.3.4 Analyse de sensibilité globale de paramètres dépendants

Comme évoqué dans le paragraphe précédent l'analyse de sensibilité avec des paramètres dépendants représente une difficulté. En effet, pour la propagation d'incertitude et le calcul de moments statistiques, on peut se ramener à un espace standardisé équivalent et les quantités d'intérêt peuvent alors être interprétées de la même manière. Si on se place dans le cadre d'une analyse de sensibilité, la transformation de l'espace d'entrée par la transformée de Nataf va modifier spatialement l'espace d'entrée et donc modifier la formule de décomposition de l'Eq. (3.36) d'une manière que l'on ne peut anticiper. Les indices calculés avec l'espace standardisé ne pourront donc pas être associés à une quelconque influence d'un des paramètres d'entrée. Il faut donc trouver d'autres moyens pour mener une analyse de sensibilité avec des paramètres dépendants en gardant l'espace physique d'entrée. Parmi les tentatives menées pour mettre en place une telle analyse de sensibilité, on peut citer celle de [Borgonovo et al. 2007] qui a développé une mesure d'importance caractérisant la modification de la distribution de sortie lorsqu'un paramètre d'entrée est fixé. Cette méthode donne une définition très générale d'indices de sensibilités et est donc compatible avec des paramètres dépendants mais elle ne représente pas une influence sur la variance comme les indices de Sobol', elle représenterait plutôt une distance à la distribution de sortie. Parmi les approches faisant intervenir des indices qui représentent une influence sur la variance de sortie, on peut citer la méthode ANCOVA [Chastaing et al. 2010] et les indices de Kucherenko [Kucherenko et al. 2012]. Comme on le verra dans

la suite, si la méthode ANCOVA est trop restrictive pour nos cas d'application, il est tout de même intéressant de développer la philosophie de la méthode car l'approche est assez élégante. Cette section développera donc succinctement la méthode ANCOVA et plus en détail les indices de Kucherenko que l'on utilisera dans la suite de ce mémoire.

### La méthode ANCOVA

Cette méthode a d'abord été introduite par [Li and Rabitz 2010] puis ensuite réécrite par [Chastaing et al. 2010]. L'objectif est ici de généraliser la définition des indices de Sobol' vus dans la partie précédente. On part de la décomposition de l'Eq. (3.36) qui reste valable pour des paramètres dépendants et que l'on simplifie sous la forme :

$$f(\mathbf{x}) = f_0 + \sum_{\mathbf{u} \subset \{1 \dots M\}} f_{\mathbf{u}}(\mathbf{x}_{\mathbf{u}}) \quad (3.55)$$

Le principe est de réécrire la décomposition de la variance de l'Eq. (3.41) sous la forme de la covariance entre la variable  $Y$  et elle-même, d'où le nom d'ANCOVA :

$$V_T = \text{Var}[Y] = \text{Cov} \left[ Y, \sum_{\mathbf{u} \subset \{1 \dots M\}} f_{\mathbf{u}}(\mathbf{X}_{\mathbf{u}}) \right] \quad (3.56)$$

On peut ensuite décomposer l'expression de la covariance sous la forme :

$$V_T = \sum_{\mathbf{u} \subset \{1 \dots M\}} \left[ \text{Var}[f_{\mathbf{u}}(\mathbf{X}_{\mathbf{u}})] + \text{Cov} \left[ f_{\mathbf{u}}(\mathbf{X}_{\mathbf{u}}), \sum_{\mathbf{v} \subset \{1 \dots M\}, \mathbf{v} \cap \mathbf{u} = \emptyset} f_{\mathbf{v}}(\mathbf{X}_{\mathbf{v}}) \right] \right] \quad (3.57)$$

Les indices suivants sont alors définis à partir des différentes composantes des Eq. (3.56) et (3.57) :

$$S_{\mathbf{u}} = \frac{\text{Cov}[Y, f_{\mathbf{u}}(\mathbf{X}_{\mathbf{u}})]}{V_T} \quad (3.58)$$

$$S_{\mathbf{u}}^U = \frac{\text{Var}[f_{\mathbf{u}}(\mathbf{X}_{\mathbf{u}})]}{V_T} \quad (3.59)$$

$$S_{\mathbf{u}}^C = \frac{\text{Cov} \left[ f_{\mathbf{u}}(\mathbf{X}_{\mathbf{u}}), \sum_{\mathbf{v} \subset \{1 \dots M\}, \mathbf{v} \cap \mathbf{u} = \emptyset} f_{\mathbf{v}}(\mathbf{X}_{\mathbf{v}}) \right]}{V_T} \quad (3.60)$$

Le premier indice représente la part totale de variance expliquée par  $\mathbf{X}_{\mathbf{u}}$ , le deuxième représente la partie non corrélée de cette part de variance et la troisième représente la partie corrélée. On remarque que l'on a l'égalité suivante :

$$S_{\mathbf{u}} = S_{\mathbf{u}}^U + S_{\mathbf{u}}^C \quad (3.61)$$

On a donc des indices de sensibilité composés de deux parties, l'une venant de l'effet pur et l'autre venant de l'effet de corrélation. Si cette méthode ANCOVA généralise la décomposition de la variance pour des paramètres dépendants, elle a cependant le désavantage de nécessiter la connaissance de la décomposition de l'Eq. (3.36). En effet si l'indice  $S_{\mathbf{u}}^U$  peut être estimé par des méthodes de Monte Carlo similaires au cas indépendant, le calcul de la covariance de  $S_{\mathbf{u}}^C$  ne peut être fait qu'avec la connaissance précise de la décomposition.

La question de l'unicité de la décomposition se pose alors. Dans [Canoui 2012], la solution envisagée est d'utiliser un méta-modèle de régression de développement en séries polynomiales qui va directement être une telle décomposition. Le calcul des indices est alors assez direct par Monte Carlo. De plus, la construction de ce méta-modèle peut être assez compliquée puisque ce dernier doit être l'image directe de tout l'espace d'entrée, alors que la dépendance entre les paramètres permet de ne s'intéresser qu'à une partie de l'espace physique d'entrée puisque certaines zones ont une probabilité minimale d'occurrence. On peut aussi se poser la question de la précision réelle du méta-modèle pour l'estimation d'indices globaux. En effet, étant donné que le méta-modèle construit doit être l'image directe de tout l'espace d'entrée, il n'intègre pas intrinsèquement la structure de dépendance. En conséquence, lors de la construction du méta-modèle à l'aide de son erreur de généralisation – que l'on exposera plus tard et qui se définit comme l'espérance de l'écart du modèle à la réalité au carré – la probabilité jointe des paramètres d'entrée n'est pas prise en compte dans le calcul cette erreur de généralisation. L'écart à la réalité donné par la construction du méta-modèle n'est donc pas son écart réel.

Au delà de ces interrogations, dans le cas des travaux présentés dans ce mémoire, la limitation aux méta-modèles de régression n'est pas envisageable. Il nous faut donc utiliser une autre méthode pour l'estimation d'indices de sensibilités dans le cas dépendant.

### Les indices de Kucherenko

Les *indices de Kucherenko* ont été introduits dans [Kucherenko et al. 2012]. De la même façon que la méthode ANCOVA, ils se présentent eux aussi comme une généralisation des indices de Sobol' dans le cas dépendant. Dans son article, Kucherenko propose une décomposition de la variance et des indices sur le modèle des indices de Sobol' qui ne nécessitent pas l'utilisation automatique de méta-modèles pour leur estimation par simulations de Monte Carlo.

Dans ce paragraphe, les notations seront volontairement modifiées pour des raisons de simplification d'écriture. On rappelle tout d'abord notre cadre d'étude : considérons le vecteur des  $M$  paramètres d'entrées  $\mathbf{x} = (x_1, \dots, x_M)^T$  supposés ici définis sur  $\mathcal{R}^M$  pour des raisons de simplification. Ces paramètres sont associés à une densité de probabilité jointe  $p_{\mathbf{X}}(x_1, \dots, x_M)$  que l'on note  $p(x_1, \dots, x_M)$  pour simplifier l'écriture. Les densités conditionnelles sont également notées  $p(u, \bar{v}|u)$  au lieu de  $p_{v|u}(u, v)$  toujours pour des raisons de simplification d'écriture. Ainsi, quand on parle d'un échantillon de la variable aléatoire  $\bar{V}$ , on parle bien de la variable conditionnelle  $V|U$ .

Les paramètres d'entrée sont toujours liés à la variance de sortie supposée finie par un modèle physique  $f(x_1, \dots, x_M)$ . Pour cette analyse de sensibilité globale, on veut connaître l'influence du sous ensemble de variables d'entrée  $\mathbf{u} = \{x_{i_1}, \dots, x_{i_s}\}$  sur la variance totale avec  $s \in \llbracket 1, M \rrbracket$ . Considérons dans le même temps le sous ensemble complémentaire  $\mathbf{v} = \{x_{i_{s+1}}, \dots, x_{i_M}\}$ . On a donc  $\mathbf{x} = (\mathbf{u}, \mathbf{v})$  en remettant les composantes dans l'ordre initial.

Dans ce paragraphe, les notations  $\mathbf{u}$  et  $\bar{\mathbf{u}}$  sont distinguées.  $\mathbf{u}$  représente la réalisation d'un vecteur aléatoire généré à partir de la densité jointe  $p(\mathbf{u}, \mathbf{v})$  et  $\bar{\mathbf{u}}$  représente la réalisation d'un vecteur aléatoire généré depuis la densité conditionnelle  $p(\bar{\mathbf{u}}|\mathbf{v}, \mathbf{v})$  : dans ce cas le vecteur  $\bar{\mathbf{u}}$  est donc généré en connaissant la valeur de  $\mathbf{v}$  générée depuis sa loi jointe  $p(\mathbf{v})$ . La variance totale de sortie  $V_T = \text{Var}[Y]$  peut être décomposée selon l'expression suivante :

$$V_T = V_{\mathbf{u}} [E_{\mathbf{v}}[f(\mathbf{u}, \bar{\mathbf{v}})]] + E_{\mathbf{u}} [V_{\mathbf{v}}[f(\mathbf{u}, \bar{\mathbf{v}})]] \quad (3.62)$$

Où chaque terme s'exprime comme suit :

$$E_v[f(\mathbf{u}, \bar{\mathbf{v}})] = \int_{\mathbb{R}^{M-s}} f(\mathbf{u}, \bar{\mathbf{v}})p(\mathbf{u}, \bar{\mathbf{v}}|\mathbf{u})d\bar{\mathbf{v}} \quad (3.63)$$

$$V_u[E_v[f(\mathbf{u}, \bar{\mathbf{v}})]] = \int_{\mathbb{R}^s} [E_v[f(\mathbf{u}, \bar{\mathbf{v}})]]^2 p(\mathbf{u})d\mathbf{u} - f_0^2 \quad (3.64)$$

$$V_v[f(\mathbf{u}, \bar{\mathbf{v}})] = \int_{\mathbb{R}^{M-s}} f(\mathbf{u}, \bar{\mathbf{v}})^2 p(\mathbf{u}, \bar{\mathbf{v}}|\mathbf{u})d\bar{\mathbf{v}} - f_0^2 \quad (3.65)$$

$$E_u[V_v[f(\mathbf{u}, \bar{\mathbf{v}})]] = \int_{\mathbb{R}^s} V_v[f(\mathbf{u}, \bar{\mathbf{v}})]p(\mathbf{u})d\mathbf{u} \quad (3.66)$$

Dans ces équations  $f_0$  représente la moyenne de la sortie comme dans les cas précédents. Kucherenko définit alors l'indice de sensibilité de premier ordre de l'ensemble de paramètres  $\mathbf{u}$  comme :

$$S_u = \frac{V_u[E_v[f(\mathbf{u}, \bar{\mathbf{v}})]]}{V_T} \quad (3.67)$$

Par opposition, il définit l'indice de sensibilité total de l'ensemble  $\mathbf{v}$  qui fait intervenir l'autre terme de la décomposition de l'Eq. (3.62) :

$$S_{T_v} = \frac{E_u[V_v[f(\mathbf{u}, \bar{\mathbf{v}})]]}{V_T} \quad (3.68)$$

On en déduit donc l'expression de l'indice de sensibilité total de l'ensemble  $\mathbf{u}$  :

$$S_{T_u} = \frac{E_v[V_u[f(\bar{\mathbf{u}}, \mathbf{v})]]}{V_T} \quad (3.69)$$

Il est intéressant de remarquer que si les paramètres sont indépendants, on retrouve les expressions des indices de Sobol' du premier ordre et totaux des Eq. (3.47) et (3.48). Les indices dans la théorie de Sobol' n'auront pas la même signification que dans celle de Kucherenko. En effet, la dépendance modifie la manière d'interpréter des indices de sensibilité car la corrélation entre les variables prend une part dans l'influence de la variance. L'avantage de la théorie de Kucherenko est qu'il est possible d'obtenir des estimateurs de Monte Carlo de ces indices. En développant la formule de l'indice de Kucherenko du premier ordre, il vient :

$$S_u = \frac{1}{V_T} \left( \int_{\mathbb{R}^s} p(\mathbf{u})d\mathbf{u} \left( \int_{\mathbb{R}^{M-s}} f(\mathbf{u}, \bar{\mathbf{v}})p(\mathbf{u}, \bar{\mathbf{v}}|\mathbf{u})d\bar{\mathbf{v}} \right)^2 - f_0^2 \right) \quad (3.70)$$

Si l'on veut estimer les intégrales de l'Eq. (3.70) par Monte-Carlo, on voit qu'il y a une double boucle qui va rendre les calculs lourds même pour une fonction  $f$  facile à évaluer. En revanche, si on développe l'expression :

$$S_u = \frac{1}{V_T} \left( \int_{\mathbb{R}^s} p(\mathbf{u})d\mathbf{u} \left( \int_{\mathbb{R}^{M-s}} f(\mathbf{u}, \bar{\mathbf{v}})p(\mathbf{u}, \bar{\mathbf{v}}|\mathbf{u})d\bar{\mathbf{v}} \times \int_{\mathbb{R}^{M-s}} f(\mathbf{u}, \bar{\mathbf{v}})p(\mathbf{u}, \bar{\mathbf{v}}|\mathbf{u})d\bar{\mathbf{v}} \right) - f_0^2 \right) \quad (3.71)$$

en utilisant la formule de Bayes, il est possible de simplifier l'expression qui regroupe les deux premières intégrales en une seule. La formule de l'indice s'écrit alors de la manière suivante :

$$S_u = \frac{1}{V_T} \left( \int_{\mathbb{R}^M} f(\mathbf{u}, \mathbf{v})p(\mathbf{u}, \mathbf{v})d\mathbf{x} \times \int_{\mathbb{R}^{M-s}} f(\mathbf{u}, \bar{\mathbf{v}})p(\mathbf{u}, \bar{\mathbf{v}}|\mathbf{u})d\bar{\mathbf{v}} - f_0^2 \right) \quad (3.72)$$

Dans l'Eq. (3.72), l'expression est bien moins lourde à estimer par simulations de Monte Carlo. Cette estimation requiert alors de générer un premier échantillon de réalisations de  $(\mathbf{U}, \mathbf{V})$  à partir de la loi jointe  $p(\mathbf{u}, \mathbf{v})$ . Ceci peut se faire facilement à partir de la transformée inverse de Nataf (cf. section 3.2.4) en partant d'un échantillon indépendant. En revanche, générer un échantillon de la loi conditionnelle  $\bar{\mathbf{V}}$  est un peu moins évident. On détaille ici uniquement la forme de la fonction de répartition Gaussienne conditionnelle multivariée. Il est possible d'utiliser cette forme pour la génération d'échantillon dans le cas d'une copule Gaussienne. On rappelle la forme de la fonction de répartition Gaussienne multivariée en dimension  $M$  qui généralise l'Eq. (3.21) :

$$\Phi_M(\mathbf{x}) = \frac{1}{(2\pi)^{M/2} \sqrt{|\boldsymbol{\Sigma}|}} \exp\left(-\frac{(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})}{2}\right) \quad (3.73)$$

où le vecteur moyenne  $\boldsymbol{\mu}$  et la matrice de corrélation  $\boldsymbol{\Sigma}$  sont exprimés en faisant apparaître les composantes  $(\mathbf{U}, \mathbf{V})$  dont on veut étudier la sensibilité :

$$\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_U \\ \boldsymbol{\mu}_V \end{pmatrix} \quad \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_U & \boldsymbol{\Sigma}_{UV} \\ \boldsymbol{\Sigma}_{VU} & \boldsymbol{\Sigma}_V \end{pmatrix} \quad (3.74)$$

On a alors la fonction de répartition jointe conditionnelle de  $\bar{\mathbf{V}}$  qui est aussi Gaussienne :

$$\Phi_{M-s}(\mathbf{u}, \bar{\mathbf{v}}|\mathbf{u}) = \frac{1}{(2\pi)^{(M-s)/2} \sqrt{|\boldsymbol{\Sigma}_{\bar{\mathbf{V}}}|}} \exp\left(-\frac{(\bar{\mathbf{v}} - \boldsymbol{\mu}_{\bar{\mathbf{V}}})^T \boldsymbol{\Sigma}_{\bar{\mathbf{V}}}^{-1} (\bar{\mathbf{v}} - \boldsymbol{\mu}_{\bar{\mathbf{V}}})}{2}\right) \quad (3.75)$$

avec

$$\boldsymbol{\mu}_{\bar{\mathbf{V}}} = \boldsymbol{\mu}_V + \boldsymbol{\Sigma}_{UV} \boldsymbol{\Sigma}_U^{-1} (\mathbf{u} - \boldsymbol{\mu}_U) \quad (3.76)$$

et

$$\boldsymbol{\Sigma}_{\bar{\mathbf{V}}} = \boldsymbol{\Sigma}_V + \boldsymbol{\Sigma}_{UV} \boldsymbol{\Sigma}_U^{-1} \boldsymbol{\Sigma}_{VU} \quad (3.77)$$

De plus, il est intéressant de préciser que la fonction de répartition marginale du vecteur  $\mathbf{U}$  a la forme :

$$\Phi_s(\mathbf{u}) = \frac{1}{(2\pi)^{s/2} \sqrt{|\boldsymbol{\Sigma}_U|}} \exp\left(-\frac{(\mathbf{u} - \boldsymbol{\mu}_U)^T \boldsymbol{\Sigma}_U^{-1} (\mathbf{u} - \boldsymbol{\mu}_U)}{2}\right) \quad (3.78)$$

Ainsi, pour générer un échantillon de la variable conditionnelle  $\bar{\mathbf{V}}$ , on génère d'abord un échantillon de  $\mathbf{U}$  avec la transformée de Nataf inverse en utilisant  $\Phi_s(\mathbf{u})$ . On génère ensuite un échantillon de  $\bar{\mathbf{V}}$  à partir de cet échantillon de  $\mathbf{U}$  et de la fonction de répartition conditionnelle  $\Phi_{M-s}(\mathbf{u}, \bar{\mathbf{v}}|\mathbf{u})$  toujours en utilisant la transformée de Nataf inverse. Considérons donc maintenant des échantillons de Monte Carlo de taille  $N$  :  $\{\mathbf{u}^{(1)} \dots \mathbf{u}^{(N)}\}$ ,  $\{\mathbf{v}^{(1)} \dots \mathbf{v}^{(N)}\}$  et  $\{\bar{\mathbf{v}}^{(1)} \dots \bar{\mathbf{v}}^{(N)}\}$  obtenus comme expliqué précédemment. La formule de Monte Carlo de l'indice  $S_u$  s'écrit alors de la manière suivante :

$$\hat{S}_u = \frac{\frac{1}{N} \sum_{i=1}^N (f(\mathbf{u}^{(i)}, \mathbf{v}^{(i)}) f(\mathbf{u}^{(i)}, \bar{\mathbf{v}}^{(i)})) - \left(\frac{1}{N} \sum_{i=1}^N f(\mathbf{u}^{(i)}, \mathbf{v}^{(i)})\right)^2}{\hat{V}_T} \quad (3.79)$$

Par une procédure semblable, il vient la formule d'estimation de l'indice total par Monte Carlo :

$$\hat{S}_{T_u} = \frac{\frac{1}{N} \sum_{i=1}^N (f(\mathbf{u}^{(i)}, \mathbf{v}^{(i)}) - f(\bar{\mathbf{u}}^{(i)}, \mathbf{v}^{(i)}))^2}{2\hat{V}_T} \quad (3.80)$$

Les indices de Kucherenko fournissent ainsi une généralisation des indices de Sobol' dans le cas de paramètres dépendants. Une particularité de ces indices pointée dans [Kucherenko et al. 2012] est qu'en présence d'une forte corrélation entre deux paramètres, leurs indices totaux tendent vers 0. On peut trouver une explication en prenant le cas d'une corrélation totale entre deux paramètres. Dans ce cas, les deux paramètres ne représentent en fait qu'un seul et même paramètre équivalent. Si on essaie d'isoler l'effet total de chaque paramètre, on voit que la corrélation totale fait que les paramètres n'ont pas d'effets totaux séparés (ces effets totaux n'auraient pas de sens s'ils n'étaient pas nuls). Une solution pour isoler l'effet total de ces deux paramètres est donc de considérer l'indice des deux paramètres regroupés. Dans la suite de ce mémoire, on utilise ces indices de Kucherenko pour analyser la sensibilité des paramètres morphologiques.

### 3.4 Conclusion

Pour conclure, ce chapitre a présenté d'une part les méthodes de modélisation de l'espace d'entrée à travers l'estimation de la loi de probabilité jointe des paramètres d'entrée en distinguant les différents types de paramètres que l'on peut rencontrer dans le contexte de la dosimétrie numérique. D'autre part, les outils d'analyse de la distribution de sortie tels que les moments statistiques, les quantiles de la distribution et les indices de sensibilité de Sobol' pour le cas indépendant et de Kucherenko pour le cas dépendant ont été introduits et seront utilisés dans la suite de ce chapitre pour l'analyse des variables sortie (le DAS principalement). Ce cadre étant posé, nous pouvons maintenant nous intéresser en détail à la propagation de l'incertitude à travers le modèle physique dans le contexte particulier de la dosimétrie numérique et plus généralement dans tout contexte où une simulation nécessite un coût de calcul important. Ainsi, le Chapitre 4 présentera, à travers l'utilisation de plusieurs méthodes, la construction de méta-modèles approchant le modèle physique et nécessitant un coût de calcul faible qui peuvent être utilisés pour la propagation Monte Carlo. Le Chapitre 5, présentera quant à lui une méthode adaptative de planification d'expériences dans le but d'estimer les quantiles à 95% des distributions de sortie de manière efficace et parcimonieuse.



# Chapitre 4

## Analyse statistique par méta-modèle

### Sommaire

---

<b>4.1 Le chaos polynomial</b> . . . . .	<b>56</b>
4.1.1 Historique . . . . .	56
4.1.2 La théorie du chaos polynomial . . . . .	58
4.1.3 Les troncatures pleine et hyperbolique . . . . .	59
4.1.4 La troncature LARS . . . . .	61
4.1.5 Sélection et validation de modèle . . . . .	62
4.1.6 Analyse de sensibilité et analyse de signature . . . . .	63
4.1.7 Planification d'expériences . . . . .	64
4.1.8 Exemple d'application : exposition d'un modèle d'enfant à une station de base de femtocell . . . . .	68
4.1.9 Exemple d'application : exposition en zone urbaine : calcul du champ réfléchi par un immeuble à géométrie variable . . . . .	70
<b>4.2 Le Krigeage universel</b> . . . . .	<b>74</b>
4.2.1 Un bref historique . . . . .	74
4.2.2 La formulation du Krigeage universel . . . . .	75
4.2.3 Les fonctions d'autocorrélation . . . . .	76
4.2.4 Le meilleur prédicteur linéaire sans biais (BLUP) . . . . .	78
<b>4.3 Hybridation du chaos polynomial et du Krigeage universel</b> . . . . .	<b>80</b>
4.3.1 Exemple d'application analytique : la fonction d'Ishigami . . . . .	81
4.3.2 Exemple d'application analytique : la fonction de Borehole . . . . .	82
4.3.3 Exemple d'application analytique : la fonction de Sobol . . . . .	83
4.3.4 Application à l'exposition d'un fœtus à un système femtocell . . . . .	84
4.3.5 Application à l'étude de l'influence de l'incertitude morphologique . . . . .	87
<b>4.4 Conclusion</b> . . . . .	<b>90</b>

---





omme indiqué dans les chapitres précédents, notre problématique intègre un modèle physique qui comprend un simulateur numérique très coûteux en temps de calcul (FDTD). Avec des temps de simulation allant de quelques dizaines de minutes à quelques heures, les traitements statistiques réalisés sur la sortie du code de calcul ne peuvent être réalisés par propagation directe de simulations Monte Carlo à travers le modèle physique.

Une alternative intéressante réside dans les modèles de substitution ou méta-modèles. Le principe du méta-modèle est de remplacer le simulateur numérique coûteux par un modèle beaucoup plus facile à calculer. Ainsi à partir de quelques observations réelles de la sortie, un modèle est ajusté aux valeurs de la sortie à disposition et la précision du modèle est estimée. Lorsqu'on obtient un méta-modèle suffisamment précis, il est alors possible de propager l'incertitude des paramètres d'entrée à travers le méta-modèle et de réaliser les traitements statistiques présentés dans le chapitre précédent sur la nouvelle sortie estimée.

Les méthodes de modèles de substitution ou méta-modèles dites *non intrusives* sont apparues dans les dernières décennies comme de puissants outils statistiques permettant de modéliser la sortie d'un code de calcul complexe sous une forme beaucoup moins coûteuse en temps de calcul. Parmi les méthodes de modélisation, on peut citer le développement par chaos polynomial et le Krigage (également appelé modélisation par processus Gaussiens). Ce chapitre s'attachera d'abord à présenter respectivement les techniques de modélisation du chaos polynomial et du Krigage universel. Enfin, ce chapitre présentera la contribution de mes travaux en proposant une combinaison des deux techniques présentées afin d'optimiser la construction du méta-modèle. Comme notifié précédemment, le simulateur numérique FDTD étant très coûteux, l'objectif est d'ajuster ce méta-modèle en me basant sur un nombre minimal d'observations de la sortie réelle, l'enjeu est donc la parcimonie. Un schéma explicatif du principe des méta-modèles intégrant le caractère spécifique de parcimonie propre à notre problématique est présenté en Figure 4.1. On se situe donc toujours dans le même environnement mathématique où l'on a  $M$  paramètres d'entrée  $\mathbf{x} = (x_1, \dots, x_M)^T \in \mathbb{R}^M$  liés à une variable de sortie  $y$  par un modèle physique déterministe  $f$  coûteux à évaluer tel que  $y = f(\mathbf{x})$ . Soit  $\mathbb{X} \subset \mathbb{R}^M$  l'espace des paramètres d'entrée.

## 4.1 Le chaos polynomial

### 4.1.1 Historique

A l'origine, la théorie du chaos polynomial a été introduite par Wiener comme le *chaos polynomial en dimension finie* [Wiener 1938] dans le cas de variables d'entrée aléatoires Gaussiennes. Dans ce cadre, la sortie est explicitement exprimée dans un espace adéquat à l'aide d'une base constituée de polynômes d'Hermite multivariés orthogonaux au sens de la densité de probabilité jointe des variables aléatoires d'entrée. Ce développement en séries polynomiales a été plus tard étendu à d'autres types de variables aléatoires à l'aide de différentes bases de polynômes [Xiu and Karniadakis 2002, Lucor and Karniadakis 2004] amenant le développement du *chaos polynomial généralisé*. Concernant l'estimation des coefficients déterministes du développement en séries polynomiales, un travail d'ampleur a été mené dans le cadre des méthodes dites *intrusives* et plus particulièrement de la méthode des *éléments finis stochastiques spectraux* [Ghanem and Spanos 2003, Soize and

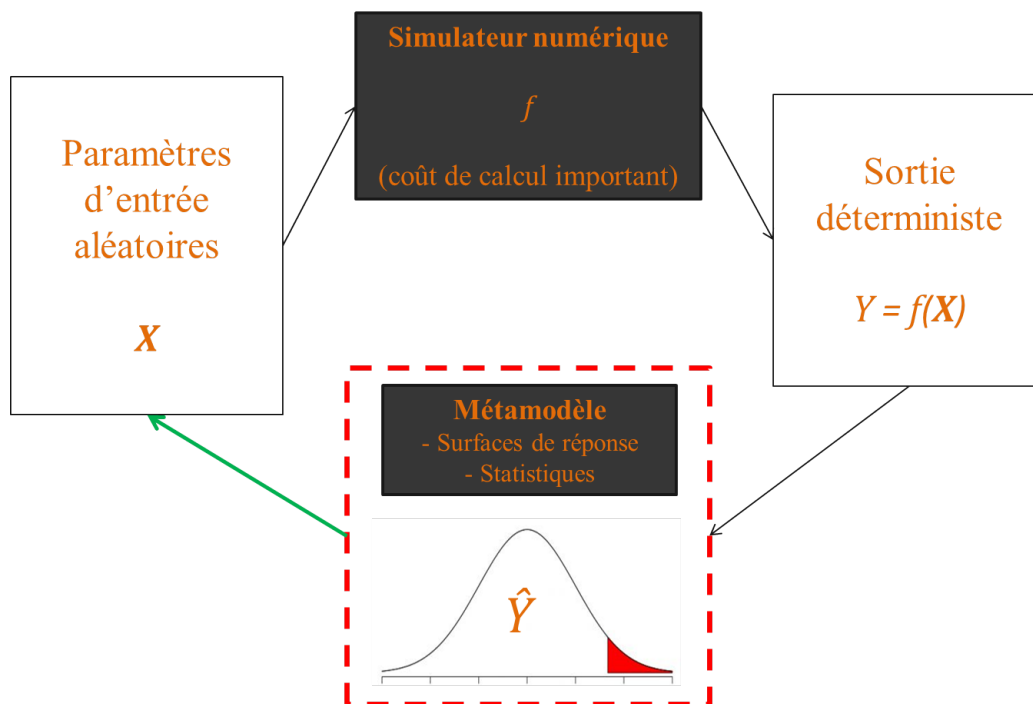


FIGURE 4.1 : Schéma explicatif du principe de la construction d'un méta-modèle

Ghanem 2004] qui utilise une combinaison du développement de Karhunen-Loeve [Karhunen 1947] avec la méthode des éléments finis pour des systèmes physiques modélisés par des problèmes linéaires elliptiques aux conditions limites. Plus tard, les méthodes dites *non intrusives* sont apparues comme une alternative dans ce domaine. Dans le cadre des méthodes non intrusives, deux approches sont classiquement distinguées pour l'estimation des coefficients déterministe du méta-modèle :

- La méthode de *projection* [Le Maitre et al. 2002, Silly-Carette 2008, Gilli et al. 2013] qui est basée sur l'orthogonalité des polynômes du chaos. Le principe est de projeter tout le développement polynomial sur le sous-espace associé au polynôme dont on veut calculer le coefficient. Cette projection nécessite pour chaque coefficient le calcul d'une intégrale. En pratique cette approche est assez coûteuse car le calcul intégral nécessite des schémas de quadrature basés sur des méthodes de quasi-Monte Carlo. Même si des techniques de quadratures creuses ont pu être développées Matthies and Keese [2005], l'approche par projection reste une approche coûteuse en temps de calcul pour nos cas d'application. L'avantage de cette méthode reste néanmoins le fait que la précision de l'estimation du coefficient est uniquement liée à la précision du calcul d'intégral. On a donc, en théorie, accès à la valeur exacte des coefficients.
- La méthode de *régression* introduite par [Isukapalli 1999] dont le principe est de tronquer le développement et de réduire l'estimation des coefficients à un problème de régression [Blatman and Sudret 2010, Berveiller et al. 2006]. C'est cette approche plus appropriée à nos cas d'application qui sera privilégiée et détaillée dans mes travaux.

### 4.1.2 La théorie du chaos polynomial

Dans la théorie du chaos polynomial, comme dans la Section 3.2.1, l'incertitude affectant le vecteur des paramètres d'entrée  $\mathbf{x}$  conduit à sa représentation sous la forme d'un vecteur aléatoire  $\mathbf{X}$  dans un espace probabiliste  $(\Omega, \mathcal{F}, \mathbb{P})$ . On note toujours  $p_{\mathbf{X}}(\mathbf{x})$  la densité de probabilité (PDF) associée au vecteur  $\mathbf{X}$ . La variable de sortie affectée elle aussi par l'incertitude  $Y = f(\mathbf{X})$  est considérée de variance finie et appartenant à l'espace  $L^2(\mathbb{R}^M, \mathcal{B}_M, \mathbb{P}_{\mathbf{X}})$  des fonctions de  $\mathbf{X}$  de carré intégrables au sens de  $p_{\mathbf{X}}(\mathbf{x})$ . Cet espace est un espace de Hilbert muni du produit scalaire :

$$\langle \psi, \phi \rangle = \mathbb{E}[\psi(\mathbf{X})\phi(\mathbf{X})] = \int_{\mathbb{X}} \psi(\mathbf{x})\phi(\mathbf{x})p_{\mathbf{X}}(\mathbf{x})d\mathbf{x} \quad (4.1)$$

Dans un soucis de simplification, les paramètres d'entrée sont considérés indépendants, un espace d'entrée dépendant pouvant être ramené à un espace standardisé indépendant équivalent par la transformée de Nataf en faisant l'hypothèse de l'utilisation d'une copule Gaussienne comme nous l'avons vu dans la Section 3.2.4.

$Y$  étant de variance finie, on peut montrer qu'elle peut s'exprimer sur une base de fonctions orthogonales [Soize and Ghanem 2004]. Une méthode pour obtenir une telle base est d'utiliser des polynômes multivariés, ce qui donne le développement en séries polynomiales suivant :

$$Y = \sum_{\alpha \in \mathbb{N}^M} \beta_{\alpha} \psi_{\alpha}(\mathbf{X}) \quad (4.2)$$

Où  $\alpha = (\alpha_1 \dots \alpha_M)$  est le vecteur des multi-indices, les  $\beta_{\alpha}$  sont des coefficients déterministes à estimer, et les  $\psi_{\alpha}$  sont des polynômes multivariés orthogonaux appelés polynômes de base du chaos polynomial. Il est donc théoriquement possible de remplacer n'importe quel simulateur numérique coûteux en temps de calcul par un développement polynomial. Certes ce développement est infini, mais si on arrive à l'approcher avec un nombre fini de polynômes comme on le verra par la suite, on crée un méta-modèle polynomial qui sera évidemment beaucoup plus rapide à estimer.

L'indépendance des paramètres d'entrée permet de construire les polynômes de base comme une tensorisation de polynômes univariés et orthogonaux au sens des densités marginales. Considérons une famille  $\{\pi_j^{(i)}, j \in \mathbb{N}\}$  de polynômes univariés orthogonaux au sens de la marginale  $p_{X_i}(x_i)$ , l'indice  $j$  se référant au degré du polynôme et l'indice  $i$  se référant à la variable aléatoire considérée. On obtient par tensorisation la famille de polynômes multivariés  $\psi_{\alpha}$ ,  $\alpha \in \mathbb{N}^M$  définie par :

$$\psi_{\alpha}(\mathbf{x}) = \pi_{\alpha_1}^{(1)}(x_1) \otimes \dots \otimes \pi_{\alpha_M}^{(M)}(x_M) \quad (4.3)$$

Si les marginales suivent des lois uniformes sur  $[-1, 1]$ , une telle famille de polynômes orthogonaux univariés sera la famille des polynômes de Legendre définis entre -1 et 1. Si les marginales suivent des lois normales standards, une telle famille sera celle des polynômes d'Hermite. L'approche de régression utilisée dans ces travaux consiste en la troncature du développement infini de l'Eq. (4.2) afin de se ramener à une série finie de polynômes dont les multi-indices appartiennent à un certain ensemble fini  $\mathcal{A} \subset \mathbb{N}^M$  :

$$Y = \sum_{\alpha \in \mathcal{A}} \beta_{\alpha} \psi_{\alpha}(\mathbf{X}) + \epsilon \quad (4.4)$$

Dans cette formulation, on estime les coefficients  $\beta_\alpha$  en minimisant l'erreur quadratique entre la troncature et la sortie  $Y$ .

$$\hat{\beta} = \{\hat{\beta}_\alpha, \alpha \in \mathcal{A}\} = \arg \min_{\beta \in \mathbb{R}^{\text{card}(\mathcal{A})}} \mathbb{E} \left[ \left( \sum_{\alpha \in \mathcal{A}} \beta_\alpha \psi_\alpha(\mathbf{X}) - Y \right)^2 \right] \quad (4.5)$$

Soit un plan d'expériences de  $N$  points  $\mathcal{X}^N = \{\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(N)}\}$ , les coefficients des polynômes de la troncature sont estimés avec l'estimateur des moindres carrés ordinaires :

$$\hat{\beta}_{OLS} = (\Psi^T \Psi)^{-1} \Psi^T \mathbf{y} \quad (4.6)$$

En notant les multi-indices composant l'ensemble  $\mathcal{A}$  sous forme vectorielle  $(\alpha^{(0)}, \dots, \alpha^{(P-1)})^T$  avec  $P = \text{card}(\mathcal{A})$ ,  $\hat{\beta}_{OLS} = (\hat{\beta}_{\alpha^{(0)}}, \dots, \hat{\beta}_{\alpha^{(P-1)}})^T$  est le vecteur des coefficients estimés par moindres carrés,  $\Psi = (\psi_{\alpha^{(j)}}(\mathbf{x}^{(i)}))_{1 \leq i \leq N, 0 \leq j \leq P-1}$  est la matrice des polynômes calculée aux points du plan d'expériences, et  $\mathbf{y} = (f(\mathbf{x}^{(1)}), \dots, f(\mathbf{x}^{(N)}))^T$  est le vecteur de la sortie aux points correspondants. Cette approche nécessite de définir une troncature et un plan d'expériences adaptatif dirigé par la qualité du modèle obtenu. En effet, pour un plan d'expériences donné de  $N$  observations, la résolution des moindres carrés suppose que le nombre  $P$  de polynômes dans la troncature soit inférieur à  $N$  et choisi de telle sorte que la matrice d'information soit bien conditionnée [Blatman and Sudret 2010]. Par conséquent, si la qualité (au sens de la capacité de généralisation) du modèle obtenu n'est pas suffisante, le nombre de polynômes dans la troncature doit être augmenté ainsi que le nombre de points dans le plan d'expériences.

### 4.1.3 Les troncatures pleine et hyperbolique

Parmi les troncatures possibles, celle classiquement utilisée est la troncature en *base pleine* [Blatman and Sudret 2008]. En considérant un degré  $p$  donné, cette troncature conserve les polynômes de l'ensemble suivant :  $\{\psi_\alpha, |\alpha| \leq p\}$  où  $|\cdot|$  est la norme L1. Dans ce cas, le nombre  $P$  de polynômes dans la troncature est égal à :

$$\text{card}(\mathcal{A}^{M,p}) = P = \binom{M+p}{p} \quad (4.7)$$

Dans ce type de troncature,  $P$  augmente rapidement avec le degré  $p$  et le nombre de paramètres d'entrée  $M$  comme le montre la Figure 4.2. Par conséquent, le nombre  $N$  d'observations nécessaires dans le plan d'expériences peut rapidement devenir ingérable d'un point de vue coût de calcul pour un degré  $p$  important. Une première solution proposée dans [Blatman and Sudret 2011] est de considérer une troncature hyperbolique de facteur  $q$ . Une telle troncature intègre les polynômes dont le multi-indice fait partie de l'ensemble suivant :

$$\mathcal{A}^{M,p,q} = \{\alpha \in \mathbb{N}^M, \|\alpha\|_q \leq p\} \quad (4.8)$$

où

$$\|\alpha\|_q = \left( \sum_{i=1}^M (\alpha_i)^q \right)^{1/q} \quad (4.9)$$

L'objectif de cette troncature hyperbolique est de "creuser" avec plus ou moins d'intensité la troncature pleine équivalente en fonction du facteur  $q$ . La troncature pleine équivalente

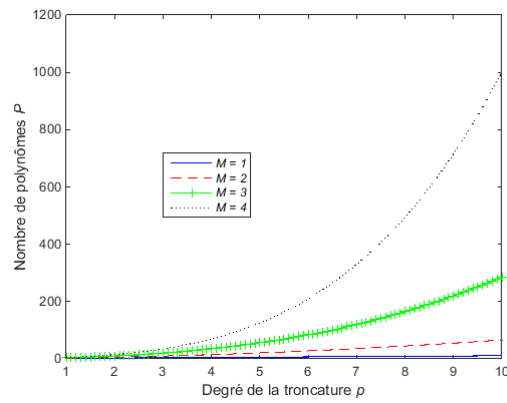


FIGURE 4.2 : Evolution du nombre de polynômes  $P$  dans la troncature pleine en fonction du degré  $p$  pour différents nombres de paramètres d'entrée  $M$

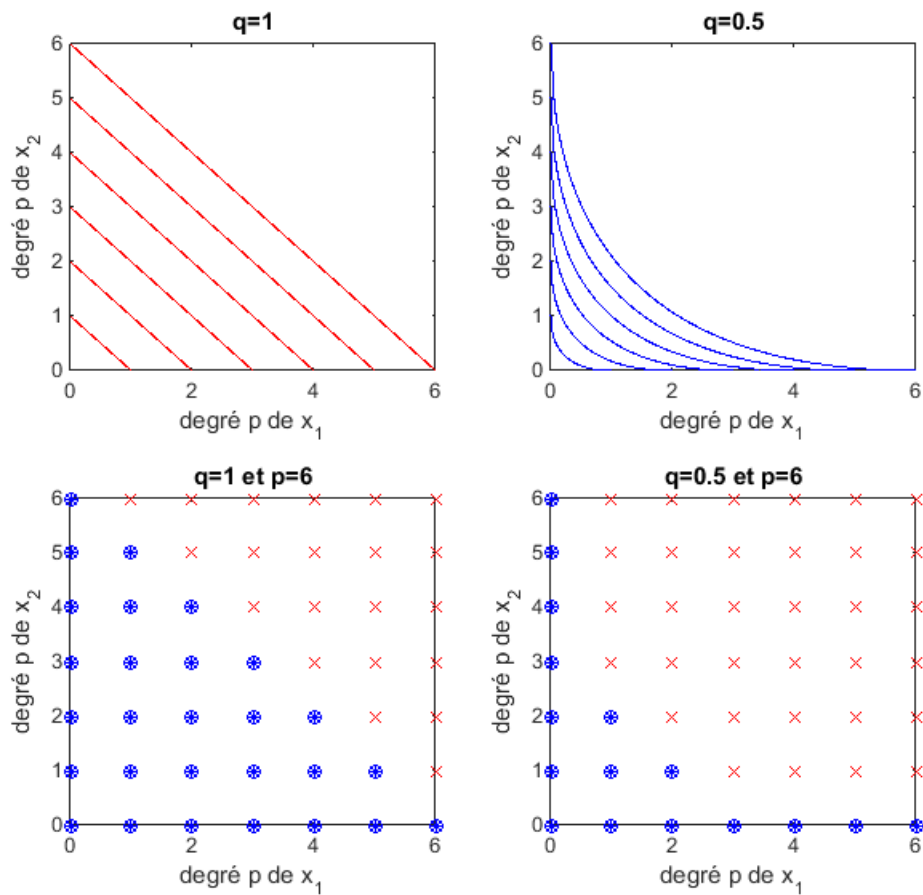


FIGURE 4.3 : Forme d'une troncature hyperbolique en dimension 2 par rapport à une troncature pleine pour  $q = 0.5$ .

correspond bien entendu à  $q = 1$ . Cela permet d'atteindre des degrés élevés avec un

nombre réduit de polynômes comme on peut le voir sur la Figure 4.3 qui montre la forme d'une troncature hyperbolique avec  $q = 0.5$  et pour deux paramètres d'entrée  $x_1$  et  $x_2$ . Sur cette figure, les deux premiers graphes montrent la forme continue de la troncature tandis que les deux graphes suivants montrent les polynômes sélectionnés caractérisés par les ronds bleus pour chaque valeur de  $q$ . L'inconvénient est qu'en faisant cela on enlève des polynômes d'interaction d'ordre relativement élevé qui peuvent possiblement avoir une influence significative sur la sortie. Il faut donc être certain de l'absence d'interaction entre paramètres d'entrée si l'on veut utiliser cette troncature. Dans le cas de l'analyse statistique de l'exposition aux ondes, on verra que des phénomènes peuvent faire intervenir beaucoup d'interactions, ce type de troncature ne sera par conséquent pas privilégié.

#### 4.1.4 La troncature LARS

Une autre approche pour constituer une troncature est de faire l'hypothèse que la variable de sortie est expliquée par un nombre réduit, mais inconnu, de polynômes influents. Le challenge est alors d'identifier ces polynômes. C'est le principe de la troncature LARS. Cette méthode permet de construire une troncature creuse en utilisant la méthode de sélection *Least Angle Regression* (LARS) [Blatman and Sudret 2011, Efron et al. 2004]. Cette méthode initialement introduite dans [Efron et al. 2004], permet de sélectionner dans la troncature les polynômes ayant le plus d'influence sur la sortie du modèle. Parmi un très grand ensemble de polynômes formant une troncature pleine, l'objectif est de sélectionner de façon itérative ceux ayant le plus d'impact en considérant leur corrélation avec le résidu courant. Ainsi l'algorithme choisit un par un les polynômes dans l'ordre d'influence décroissant supposé. On a donc autant de troncatures possibles que de polynômes sélectionnés par l'algorithme, ces troncatures ayant une taille croissante. Dans la théorie du LARS, la sortie est centrée et les prédicteurs (ici les polynômes) sont centrés et normalisés sur l'espace d'entrée. Par conséquent le polynôme de degré total  $|\alpha| = 0$  (le terme constant) ne peut être sélectionné par l'algorithme. En pratique, ce polynôme est préalablement retiré de l'ensemble de sélection et ajouté à la fin de la procédure de sélection. L'algorithme de sélection des polynômes est le suivant :

1. Choisir un degré  $p$  d'une troncature pleine fournissant un ensemble de  $P - 1$  polynômes comme vu en Eq. (4.7).
2. Mettre à 0 les coefficients  $\beta_{\alpha^{(1)}}, \dots, \beta_{\alpha^{(P-1)}}$  et le vecteur résidu initial à  $\mathbf{r}^{(0)} = \mathbf{y}$ .
3. Sélectionner le polynôme  $\psi_{\alpha^{(j_1)}}$  le plus corrélé avec  $\mathbf{r}^{(0)} = \mathbf{y}$  au sens de la corrélation de Pearson.
4. Déplacer la valeur  $\beta_{\alpha^{(j_1)}}$  dans la direction du polynôme  $\psi_{\alpha^{(j_1)}}$  jusqu'à ce que le résidu courant  $\mathbf{r}^{(1)} = \mathbf{y} - \hat{\beta}_{\alpha^{(j_1)}} \psi_{\alpha^{(j_1)}}(\mathcal{X})$  soit autant corrélé avec  $\psi_{\alpha^{(j_1)}}$  qu'avec un autre polynôme  $\psi_{\alpha^{(j_2)}}$ .
5. Sélectionner ce second polynôme  $\psi_{\alpha^{(j_2)}}$  et définir  $\mathbf{u}^{(2)}$  comme étant le vecteur unitaire directeur de la bissectrice de  $\psi_{\alpha^{(j_1)}}$  et de  $\psi_{\alpha^{(j_2)}}$ .
6. Pour  $k$  allant de 3 à  $\min(P, N - 2)$  :

- (a) Déplacer la valeur de  $\boldsymbol{\beta} = (\beta_{\alpha^{(j_1)}}, \dots, \beta_{\alpha^{(j_{k-1})}})$  dans la direction de  $\mathbf{u}^{(k-1)}$  jusqu'à ce que le résidu  $\mathbf{r}^{(k-1)}$  soit autant corrélé avec  $\mathbf{u}^{(k-1)}$  qu'avec un autre polynôme  $\psi_{\alpha^{(j_k)}}$ .
  - (b) Sélectionner ce polynôme  $\psi_{\alpha^{(j_k)}}$  et définir  $\mathbf{u}^{(k)}$  comme étant le vecteur unitaire directeur de la bissectrice de  $\mathbf{u}^{(k-1)}$  et de  $\psi_{\alpha^{(j_k)}}$ .
7. Ajouter le polynôme constant égal à 1 (de degré total  $|\alpha| = 0$ ) aux polynômes précédemment sélectionnés.

On obtient par conséquent  $L = \min(P, N - 2) + 1$  bases possibles  $\{\psi_{\alpha^{(0)}}, \psi_{\alpha^{(j_1)}}, \dots, \psi_{\alpha^{(j_k)}}\}$  où  $k \in \llbracket 1, L \rrbracket$ . L'intérêt de la méthode est de pouvoir choisir  $P \geq N - 1$  pour intégrer des polynômes de degré élevé tout en construisant une troncature creuse. Pour chaque base possible, les coefficients du développement polynomial sont obtenus par l'estimateur des moindres carrés ordinaires calculé à partir des polynômes sélectionnés par le LARS. Cette approche constitue une variante du LARS d'origine appelée méthode *LARS OLS hybrid* et exposée dans [Efron et al. 2004]. Cette variante se montre en effet plus performante pour obtenir une erreur quadratique empirique plus faible. On utilise donc le LARS uniquement comme méthode de sélection et non pas pour estimer les coefficients du développement. Cette troncature fournit par conséquent une représentation parcimonieuse du développement de l'Eq. (4.2) et a prouvé son efficacité en terme de gain de points dans le plan d'expériences [Blatman and Sudret 2011]. Elle sera celle utilisée dans la suite de ces travaux.

#### 4.1.5 Sélection et validation de modèle

Les troncatures creuses étant construites par régression à partir d'un plan d'expériences, il est fondamental d'évaluer la qualité des méta-modèles générés et d'estimer leur précision. Pour extraire une qualité globale de ces méta-modèles, on considère l'erreur de généralisation suivante :

$$Err = \mathbb{E} \left[ \left( f(\mathbf{X}) - \hat{f}(\mathbf{X}) \right)^2 \right] \quad (4.10)$$

où  $f$  est le modèle physique,  $\hat{f}$  est le méta-modèle construit et  $\mathbb{E}$  est l'opérateur d'espérance. Un estimateur intuitif de cette erreur de généralisation est l'erreur moyenne quadratique aux points de simulation. Pour un plan d'expériences  $\mathcal{X} = \{\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(N)}\}$  de taille  $N$ , on a l'estimateur suivant :

$$MSE = \frac{1}{N} \sum_{i=1}^N \left( f(\mathbf{x}^{(i)}) - \hat{f}(\mathbf{x}^{(i)}) \right)^2 \quad (4.11)$$

Dans le cas d'un plan d'expériences avec un faible nombre de points, cette estimateur MSE peut largement sous-estimer l'erreur réelle de généralisation. Pour régler ce problème de sous-estimation, on utilise dans ces travaux la validation croisée *leave-one-out* (LOOCV) [Molinaro et al. 2005]. La LOOCV est une méthode fréquemment utilisée en *machine learning* pour estimer la qualité d'un méta-modèle et/ou pour sélectionner le meilleur méta-modèle parmi des possibles. Le principe de la LOOCV est de retirer un point  $\mathbf{x}^{(i)}$  du plan d'expériences et de calculer le méta-modèle  $\hat{f}^{(-i)}$  à partir du plan d'expériences



réduit  $\mathcal{X} \setminus \{\mathbf{x}^{(i)}\}$ . Il est alors possible de calculer l'écart du méta-modèle à la réalité au point manquant :

$$\Delta^{(i)} = f(\mathbf{x}^{(i)}) - \hat{f}^{(-i)}(\mathbf{x}^{(i)}) \quad (4.12)$$

Cet écart est alors calculé pour tous les points du plan d'expériences et l'erreur de généralisation est estimée en prenant la moyenne de tous les écarts quadratiques. Cette estimation de l'erreur de généralisation est appelé l'erreur *leave-one-out* :

$$Err_{LOO} = \frac{1}{N} \sum_{i=1}^N \Delta^{(i)2} \quad (4.13)$$

Même si cet estimateur peut être encore optimiste par rapport à l'erreur de généralisation réelle, il l'est beaucoup moins qu'une erreur quadratique moyenne classique [Molinaro et al. 2005]. Il est néanmoins dépendant du plan d'expérience et de la répartition des points dans l'espace d'entrée. De plus, cette erreur *leave-one-out* peut être directement calculée analytiquement à partir du modèle de base [Blatman and Sudret 2010], ce qui gomme l'aspect calculatoire couteux de cette méthode. Pour faciliter l'interprétation de cette erreur, on peut calculer un coefficient de détermination  $Q^2$  à partir de cette erreur :

$$Q^2 = 1 - \frac{Err_{LOO}}{\hat{\sigma}_y^2} \quad (4.14)$$

Où  $\hat{\sigma}_y^2$  est la variance estimée de la variable de sortie. Ainsi, plus  $Q^2$  est proche de 1, plus la capacité de généralisation du méta-modèle sera élevée.

Cette estimation de la qualité du méta-modèle obtenu est utile non seulement pour estimer l'erreur à la réalité mais aussi pour la sélection de troncature dans le cas de l'approche par régression du chaos polynomial. En effet, dans le cas de la troncature LARS utilisée dans mes travaux, la LOOCV permet de sélectionner le nombre de polynômes LARS aboutissant au méta-modèle le plus précis. Plus précisément, à chaque itération de l'algorithme LARS, une LOOCV est réalisée afin d'évaluer la qualité du modèle courant. A la dernière itération de l'algorithme LARS, on vient sélectionner le nombre de polynômes LARS qui donne la valeur de  $Q^2$  la plus élevée, i.e. le modèle avec la meilleure capacité de généralisation.

#### 4.1.6 Analyse de sensibilité et analyse de signature

Le méta-modèle obtenu par régression du développement tronqué du chaos polynomial peut ensuite être utilisée de façon très directe pour estimer des quantités d'intérêt ou pour réaliser une analyse de sensibilité sur des paramètres indépendants en utilisant les indices de Sobol' vus dans la Section 3.3.3. Dans cette section, on suppose que les polynômes ont été préalablement centrés et réduits exceptés bien entendu le polynôme constant égal à 1. Ainsi, à partir de l'expression de l'Eq. (4.4), on estime la moyenne et la variance de la variable de sortie par les expressions suivantes :

$$\hat{\mu}_{Y,\mathcal{A}} = \beta_{\mathbf{0}} \quad (4.15)$$

$$\hat{\sigma}_{Y,\mathcal{A}}^2 = \sum_{\alpha \in \mathcal{A} \setminus \{\mathbf{0}\}} \beta_{\alpha}^2 \quad (4.16)$$



De plus, on peut effectuer une analyse de sensibilité globale en estimant les indices de Sobol' de premier ordre et totaux [Sobol' 2001]. La troncature du développement du chaos polynomial permet d'estimer ces indices à l'aide des expressions suivantes [Sudret 2008] des indices de Sobol' d'ordre  $s$   $S_{i_1, \dots, i_s, \mathcal{A}}$  et total  $S_{T_i, \mathcal{A}}$  du paramètre d'entrée  $x_i$ . L'indice d'ordre  $s$  s'exprime :

$$S_{i_1, \dots, i_s, \mathcal{A}} = \frac{1}{\hat{\sigma}_{Y, \mathcal{A}}^2} \sum_{\alpha \in \mathcal{I}_{i_1, \dots, i_s}} \beta_{\alpha}^2 \quad (4.17)$$

où l'on a :

$$\mathcal{I}_{i_1, \dots, i_s} = \left\{ \alpha \in \mathcal{A}, \begin{array}{l} \alpha_k \neq 0 \quad k \in \{i_1, \dots, i_s\} \\ \alpha_j = 0 \quad j \notin \{i_1, \dots, i_s\} \end{array} \right\} \quad (4.18)$$

L'ensemble  $\mathcal{I}_{i_1, \dots, i_s}$  regroupe ici les multi-indices de la troncature tels que toutes leur composantes relative aux variables étudiées ( $i_1, \dots, i_s$ ) soient non nulles et tels que toutes leurs autres composantes soient nulles. On ne s'intéresse qu'aux polynômes faisant intervenir l'ensemble des variables d'intérêt et uniquement celles-ci. Quant à l'indice total, il s'exprime :

$$S_{T_i, \mathcal{A}} = \frac{1}{\hat{\sigma}_{Y, \mathcal{A}}^2} \sum_{k \in \mathcal{I}_i} \beta_{\alpha}^2 \quad (4.19)$$

Où  $\mathcal{I}_i$  représente l'ensemble des indices de la troncature dont la composante du multi-indice associée à la variable d'entrée considérée est non nulle :

$$\mathcal{I}_i = \{ \alpha \in \mathcal{A}, \alpha_i \neq 0 \} \quad (4.20)$$

L'approche par régression permet donc une estimation beaucoup plus rapide des indices de sensibilité que l'approche classique par simulations de Monte Carlo à partir du méta-modèle [Sobol' 2001].

Pour aller plus loin dans l'étude de l'influence des paramètres d'entrée sur la variable de sortie, une des contributions de mes travaux a été de proposer une analyse de signature du développement en séries polynomiales tronqué. Le but de l'analyse, inspirée par l'analyse de signature modale utilisée en électronique, est d'évaluer la variance de sortie expliquée par chaque polynôme de la troncature. Pour cela, on calcule la magnitude relative de chaque coefficient. La magnitude relative du coefficient  $\beta_{\alpha^{(k)}}$  est définie comme suit :

$$m_{\alpha, \mathcal{A}} = \frac{\beta_{\alpha}^2}{\hat{\sigma}_{Y, \mathcal{A}}^2} \quad (4.21)$$

Ainsi, la magnitude relative est directement image de l'importance du polynôme associé dans la variance de la variable de sortie. Grâce à l'analyse de signature on peut avoir accès de façon détaillée à l'ordre des polynômes les plus influents, à la nature des interactions entre les différents paramètres d'entrée et à la complexité du méta-modèle généré, ce que l'analyse de sensibilité ne permet pas. On peut donc ensuite effectuer une analyse physique en se basant sur la distribution modale du méta-modèle.

#### 4.1.7 Planification d'expériences

La définition du plan d'expériences est un élément déterminant dans la mise en oeuvre du chaos polynomial par régression et plus généralement de beaucoup de méthodes de modélisation non intrusives. Un plan d'expériences mal défini peut mener à une détérioration de

la précision de l'approximation. Pour obtenir un méta-modèle globalement précis sur tout le domaine d'entrée, on utilise généralement les techniques de planification d'expériences dites de *space-filling*. Ces techniques visent à remplir l'espace d'entrée  $\mathbb{X}$  avec un nombre fini  $N$  de points dans le but de capter le plus d'information possible pour approcher la sortie. Dans cette section, le domaine d'entrée  $\mathbb{X}$  est supposé être l'ensemble fermé  $[0, 1]^M$  et les paramètres d'entrée sont supposés uniformes et indépendants sur cet espace. L'étude de configurations différentes ne pose pas de problème en pratique car n'importe quel espace d'entrée muni de lois de probabilité quelconques peut en théorie être ramené à cet espace par transformation isoprobabiliste.

### Echantillonnage par Monte Carlo

Une première méthode de planification classique et facile d'implémentation est l'échantillonnage uniforme par Monte Carlo qui est défini par :

$$\mathcal{X}_{MCS} = \left\{ x_i^{(j)}, i = 1 \dots M, j = 1 \dots N \right\} \quad (4.22)$$

où  $x_i^{(j)}$  est la réalisation d'une variable aléatoire uniforme sur  $[0, 1]$  et  $M$  étant le nombre de variables. Il est important de préciser que les variables aléatoires dont les points sont des réalisations sont indépendantes. Si elle est facile à implémenter, cette méthode nécessite un grand nombre de points pour remplir l'espace d'entrée de manière uniforme et les méta-modèles construits à partir de tels plans ont en pratique une erreur de généralisation très aléatoire pour un nombre de points donné faible.

### Plans Hypercubes Latins

Une alternative souvent employée sont les *plans hypercubes latins* ou *latin hypercube sampling* (LHS) [McKay et al. 1979]. Ces plans assurent que les points du plan d'expériences soit répartis de manière uniforme sur les axes de chaque paramètre d'entrée. Sa facilité d'implémentation ajoutée à cette propriété intéressante en a fait une technique très populaire et très utilisée dans le domaine de la planification d'expériences. Un plan LHS de taille  $N$  défini sur  $[0, 1]^M$  est défini de la façon suivante :

$$\mathcal{X}_{LHS} = \left\{ x_i^{(j)} = \frac{\pi_i^{(j)} - 1 + u_i^{(j)}}{N}, i = 1 \dots M, j = 1 \dots N \right\} \quad (4.23)$$

où  $\pi_i^{(j)}$  est le  $j$ -ème élément de la  $i$ -ème permutation aléatoire de la séquence  $\{1 \dots M\}$  et  $u_i^{(j)}$  est la réalisation d'une variable aléatoire uniforme sur  $[0, 1]$ . Un exemple de plan LHS en dimension 2 est présenté en Figure 4.4.

Si les plans Latin Hypercube ont la propriété d'être distribués de manière uniforme sur les axes des paramètres d'entrée, ils n'assurent pas cette propriété pour des projections de dimensions supérieures ce qui peut conduire à des plans sous-optimaux pour des sorties incluant des interactions fortes entre les paramètres d'entrée. Des plans Latin Hypercube dits optimaux permettent de limiter ce problème. De nombreuses références sont disponibles sur le sujet des plan hypercubes latin optimaux dans la littérature, que les plans visent à être orthogonaux pour les effets principaux [Owen 1994, Tang 1998] ou bien que les plans soit optimaux vis-à-vis d'un critère assurant la bonne répartition des points dans

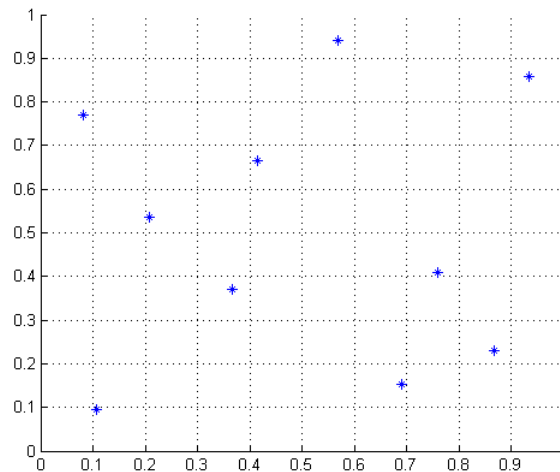


FIGURE 4.4 : Exemple de plan LHS en deux dimensions.

l'espace d'entrée [Johnson et al. 1990, Park 1994]. On peut citer plus particulièrement les plans LHS *maxmin* [Johnson et al. 1990] qui visent à maximiser la distance euclidienne minimum entre chaque couple de points du plan. Ces plans *maxmin* seront ceux utilisés dans la suite de ce rapport à chaque fois que l'on parlera de plans LHS.

Les plans Latin Hypercube offrent également une possibilité d'enrichissement séquentiel grâce à la technique du *nested Latin Hypercube sampling* (NLHS) [Wang 2003]. Le principe de cette technique est de compléter un plan LHS existant par de nouveaux points en conservant, au moins approximativement, la configuration LHS du plan. Cette technique nous sera par la suite utile pour enrichir des plans existants et construire de nouveaux méta-modèles plus précis.

### La planification séquentielle suivant un critère *maxmin*

Afin d'optimiser l'enrichissement de plans d'expériences existants, une des pistes explorées par ces travaux a été de considérer un enrichissement séquentiel d'un plan existant suivant un critère *maxmin*. A partir d'un plan d'expériences existant, le but est d'ajouter un point choisi dans l'espace d'entrée de sorte que l'on maximise la distance euclidienne minimum entre ce point et tous ceux du plan d'expériences initial. Si on considère un plan d'expériences  $\mathcal{X}^N$  de taille N, le plan augmenté correspondant sera  $\mathcal{X}^{N+1}$  :

$$\mathcal{X}^{N+1} = \mathcal{X}^N \cup \{\mathbf{x}^{(N+1)}\} \quad (4.24)$$

tel que :

$$\mathbf{x}^{(N+1)} = \max_{\mathbf{x} \in \mathbb{X}} \min_{\mathbf{x}^{(i)} \in \mathcal{X}^N} \|\mathbf{x} - \mathbf{x}^{(i)}\| \quad (4.25)$$

Cette méthode séquentielle permet de remplir l'espace d'entrée de façon séquentielle et uniforme. Pour ajouter des points selon ce critère, on génère un échantillon Monte Carlo de grande taille et on sélectionne les points parmi cet échantillon qui maximisent le critère. On ne sélectionne pas le point correspondant à l'optimum mais le meilleur point parmi un très grand nombre de possibles. On compare maintenant les performances de cette

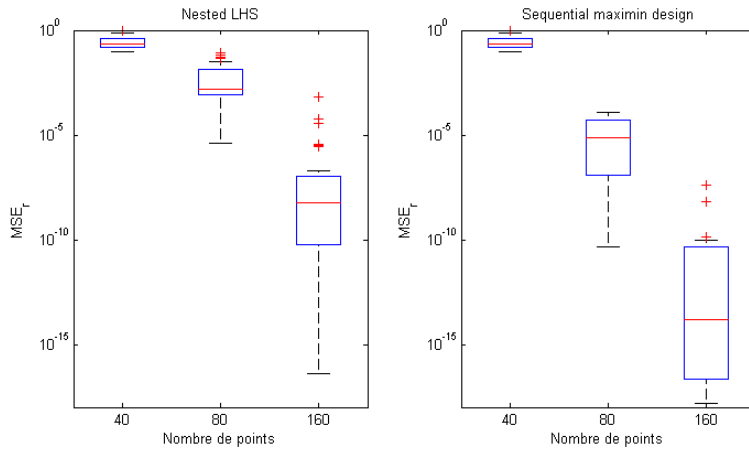


FIGURE 4.5 : Comparaison des méthodes de planification NLHS et séquentiel maximin pour 50 LHS initiaux différents augmenté jusqu'à 160 points pour la fonction d'Ishigami.

méthode de remplissage séquentielle avec les performances de la technique NLHS à l'aide d'une fonction test facile à calculer. La fonction test utilisée est la fonction d'Ishigami. La fonction d'Ishigami est une fonction analytique de test couramment utilisée en analyse de sensibilité [Saltelli et al. 2000]. C'est une fonction de trois paramètres d'entrée  $\{x_1, x_2, x_3\}$  indépendants que l'on équipe de distributions uniformes sur  $[-\pi, \pi]$  et s'exprimant comme suit :

$$f(\mathbf{x}) = \sin x_1 + 7 \sin^2 x_2 + 0.1x_3^4 \sin x_1 \quad (4.26)$$

Cette fonction étant très rapide à calculer, on n'utilisera pas l'erreur *leave-one-out* comme outil d'évaluation de la prédiction mais l'erreur moyenne quadratique relative de validation de l'Eq. (4.27) sur  $N_m = 100,000$  points simulés par Monte Carlo.

$$MSE_r = \frac{\sum_{i=1}^{N_m} \left( f(\mathbf{x}^{(i)}) - \hat{f}(\mathbf{x}^{(i)}) \right)^2}{\sum_{i=1}^{N_m} \left( f(\mathbf{x}^{(i)}) - \hat{\mu}_y \right)^2} \quad (4.27)$$

où  $\hat{\mu}_y$  est la moyenne estimée de la variable de sortie. Cette erreur sera calculée pour les deux techniques de planification d'expériences. On partira d'un plan LHS initial de 40 points. D'un côté ce plan sera augmenté par la technique NLHS jusqu'à 80 et 160 points ; de l'autre côté, le plan initial sera augmenté par la méthode séquentielle de remplissage maximin et l'erreur présentée dans l'Eq. (4.27) sera calculée et comparée au NLHS pour les mêmes nombres de points. Cette analyse est répétée 50 fois dans le but d'évaluer la variabilité des résultats de ces méthodes de planification et des boxplots illustrant cette variabilité sont présentées en Figure 4.5. Ces boxplots présentent la médiane, les quartiles et les valeurs extrêmes de l'erreur de l'Eq. (4.27) pour les 50 générations indépendantes des LHS initiaux. La Figure 4.5 montre que la technique de planification séquentielle maximin donne des résultats plus précis en terme de valeur médiane ainsi que pour les quartiles et les valeurs extrêmes. Il semble donc que l'échantillonnage séquentiel maximin ait une convergence plus rapide que le NLHS pour la prédiction à l'aide de modèles de chaos polynomial. La limitation de cette méthode est qu'elle a tendance à mettre des points sur les bords du domaine qui représentent souvent les distances les plus élevées aux

autres points. Par conséquent, en grande dimension, cette méthode peut passer du temps à remplir les bords du domaines avant de venir en ajouter dans le centre qui dans certains cas va être plus utile à la précision du modèle.

#### 4.1.8 Exemple d'application : exposition d'un modèle d'enfant à une station de base de femtocell

Pour illustrer les cas d'utilisation du chaos polynomial, on présente un exemple cas-test issu de la problématique de la dosimétrie. L'objectif est d'analyser statistiquement l'exposition d'un modèle anatomique d'enfant à une station de base de femtocell. Le modèle physique sera celui que l'on a présenté dans la Section 2.4.3. Le modèle anatomique d'enfant étudié ici est celui d'une fille de 8 ans nommée Eartha qui est issu de la Virtual Family [Christ et al. 2010]. En termes d'incertitudes, on s'intéresse ici à l'incertitude de la position relative de la station de base de femtocell par rapport à la position du modèle anatomique. Par position relative, on entend que l'on veut capter par cette modélisation l'influence de l'incertitude de la position de la station de base dans une pièce par rapport à la position d'une personne en position debout. On définit alors quatre paramètres géométriques représentant un domaine spatial d'une taille de l'ordre de grandeur d'une pièce standard :

- trois paramètres spatiaux correspondant aux coordonnées cylindriques  $(r, \phi, z)$  de la station de base par rapport au modèle anatomique.
- un paramètre angulaire  $\alpha$  correspondant à la rotation de la station de base sur elle-même autour de son axe vertical.

En l'absence de données réelles sur ces incertitudes, on fait l'hypothèse que les paramètres suivent tous des lois uniformes dont les bornes d'intervalle sont données dans la Table 4.1 :

Paramètre d'entrée	Intervalle
$r$	$[1; 3]$ mètres
$\phi$	$[0, 360]$ degrés
$z$	$[-0.75, 1.75]$ mètres
$\alpha$	$[0, 360]$ degrés

TABLE 4.1 : Domaine de variation des paramètres d'entrée pour le problème de l'exposition du fœtus à une femtocell

Pour une meilleure compréhension, les sources d'incertitude sont représentées graphiquement dans la Figure 4.6. La variable de sortie d'intérêt est le DAS corps entier du modèle anatomique. La station de base rayonne à une puissance normalisée de 1 W. On crée un plan d'expériences LHS initial constitué de 100 points à partir desquels on construit un méta-modèle de chaos polynomial avec une troncature LARS. Pour améliorer la précision du modèle et être sûr du résultat, le plan d'expérience est complété par un enrichissement maxmin jusqu'à 360 points où l'on arrive à une valeur de  $Q^2 = 0.98$  satisfaisante. A partir du modèle final, il est alors possible d'effectuer des simulations de Monte Carlo et d'estimer la distribution de la variable de sortie. La Figure 4.7 présente l'estimation de cette distribution. On remarque que cette distribution présente une modalité importante qui

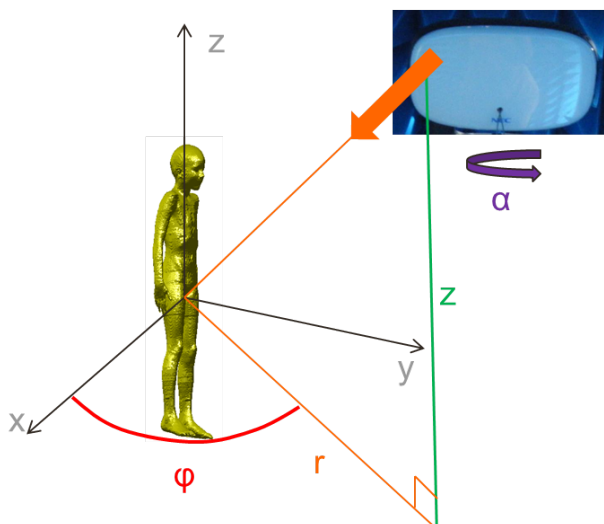


FIGURE 4.6 : Schéma de l'exposition d'Eartha à la station de base de femtocell indiquant les sources d'incertitudes.

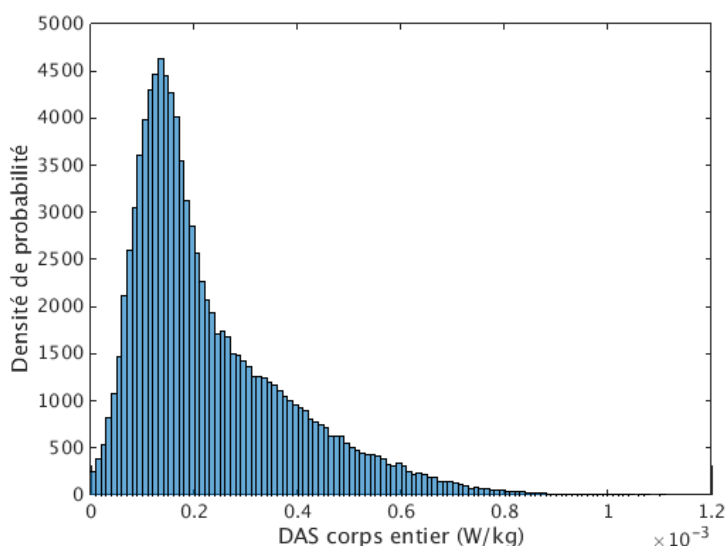


FIGURE 4.7 : Distribution du DAS corps entier d'Eartha.

s'explique par le fait que la décroissance en  $1/r^2$  de la puissance reçue en fonction de la distance génère plus de valeurs de DAS dans la partie inférieure du domaine de variation. Le coefficient de variation de la distribution est  $CV = 0.68$  (cf Eq. (3.30)). A partir de ce modèle de chaos polynomial, on calcule les indices de Sobol' totaux et de premier ordre avec les formules présentées dans la Section 4.1.6. Les résultats sont présentés dans la Table 4.2. Les résultats montrent que le paramètre ayant la plus grande influence sur la variance de sortie est la distance radiale  $r$  entre la station de base et le modèle anatomique. On voit que les autres paramètres contribuent peu séparément mais néanmoins, si on prend la contribution totale de tous ces paramètres en même temps, on obtient une valeur non

Paramètre d'entrée	Indice de premier ordre	Indice total
$r$	0.61	0.77
$\phi$	0.07	0.12
$z$	0.03	0.11
$\alpha$	0.09	0.18

TABLE 4.2 : Indices de Sobol totaux et de premier ordre des paramètres d'entrée pour l'exposition d'Eartha à une station de base de femtocell obtenus à partir de 1 000 000 simulations de Monte Carlo métamodèle de chaos polynomial.

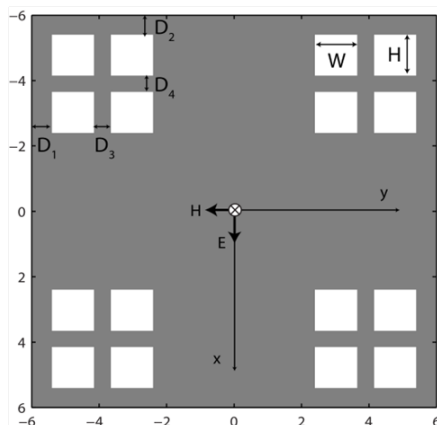


FIGURE 4.8 : Illustration de géométrie d'immeuble possible.

négligeable. Il est également remarquable que le phénomène étudié est assez peu interagissant puisque la différence entre les indices totaux et de premier ordre est assez faible, excepté peut-être pour le paramètre de hauteur  $z$ .

Ce cas-test illustre ainsi les possibilités de l'utilisation du chaos polynomial pour les problèmes de dosimétrie.

#### 4.1.9 Exemple d'application : exposition en zone urbaine : calcul du champ réfléchi par un immeuble à géométrie variable

L'approche par régression du chaos polynomial est ici appliquée à un exemple d'étude issu du domaine de l'électromagnétisme. Le problème est non pas d'étudier l'exposition en terme de puissance absorbée mais d'étudier l'exposition de la voie descendante à travers l'évaluation de la valeur du champ électrique en milieu urbain. En effet, la forte densification des réseaux de téléphonie mobile en milieu urbain a conduit à l'évaluation de l'exposition induite par la voie descendante de ces réseaux. Cette exposition se quantifie généralement par la force du champ électrique mesuré ou simulé en un point de la zone urbaine étudiée. Des modèles empiriques, statistiques ou encore théoriques basés sur des techniques de propagation et de lancer de rayons existent pour simuler cette exposition [Iskander and Yun 2002]. Cependant la plupart de ces simulateurs n'intègrent pas la variabilité des détails architecturaux dans leur prédiction du champ diffracté par les bâtiments. Pour étudier l'influence de l'architecture, on utilise dans cette section les travaux qui ont

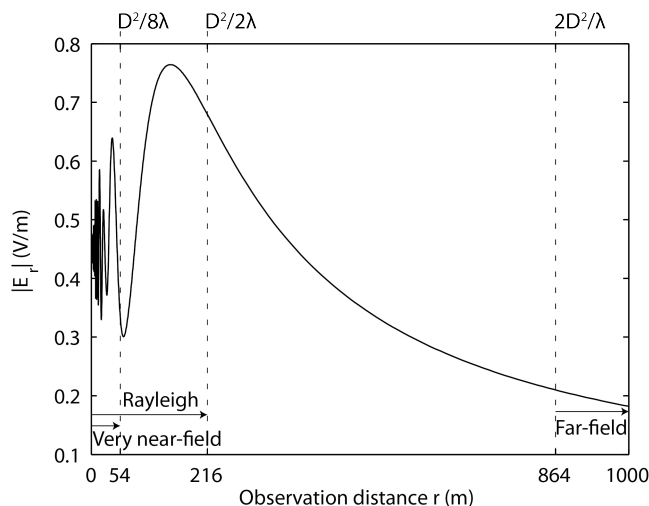


FIGURE 4.9 : Evolution de la force du champ électrique réfléchi par l'immeuble de la Figure 4.8 dans la direction spéculaire en fonction de la distance d'observation.

abouti à la construction d'un simulateur de champ diffracté par un immeuble à géométrie variable éclairé par une onde plane [Mostarshedi et al. 2010]. Ce simulateur repose sur le calcul du champ réfléchi par un immeuble dont on peut modifier les paramètres architecturaux. Une illustration d'une géométrie possible de l'immeuble est donné en Figure 4.8. L'immeuble est uniquement constitué de béton et de verre. Les paramètres modifiables de l'architecture de l'immeuble sont les permittivités du béton et du verre, la largeur et la hauteur des fenêtres, les écarts verticaux et horizontaux entre les fenêtres et les distances verticales et horizontales séparant les fenêtres des bords de l'immeuble. L'immeuble généré est éclairé par une onde plane dont on spécifie l'angle d'incidence. Le calcul du champ diffracté repose sur l'utilisation de fonctions de Green [Mostarshedi et al. 2010]. La force de ce champ peut alors être calculée à différentes distances de l'immeuble et pour différents angles d'observation. La Figure 4.9 présente le résultat du calcul de la force du champ diffracté par l'immeuble de la Figure 4.8 éclairé par une onde plane en incidence normale en fonction de la distance d'observation. On voit qu'en fonction de la zone de rayonnement, l'évaluation de la force du champ diffracté en fonction de la distance n'est pas la même. En effet, on constate une évolution très perturbée du champ électrique en zone de champ très proche, suivi par une augmentation de la force du champ dans la zone de Rayleigh et terminée par une décroissance dans les zones de Fresnel et de Fraunhofer [Balanis 1988].

On se propose ici d'étudier l'influence de la variabilité des paramètres architecturaux sur la valeur du champ dispersé par l'immeuble à travers une décomposition du champ par chaos polynomial. L'approche par régression présentée plus haut dans la Section 4.1 est utilisée ainsi qu'une troncature creuse obtenue à l'aide de l'algorithme LARS. Le choix est fait ici de se placer en direction spéculaire par rapport à l'immeuble et d'évaluer l'impact des paramètres architecturaux sur le champ dispersé pour plusieurs distances. Avec un pas de 50 m, on construit un modèle de chaos LARS pour chaque distance entre  $d = 50m$  et  $d = 800m$  à partir d'un plan d'expériences LHS de 300 points qui nous donne une précision par validation croisée que l'on juge suffisante pour chaque modèle construit ( $Q^2 \geq 0.95$ ). Le choix est volontairement fait ici de ne pas considérer la zone perturbée



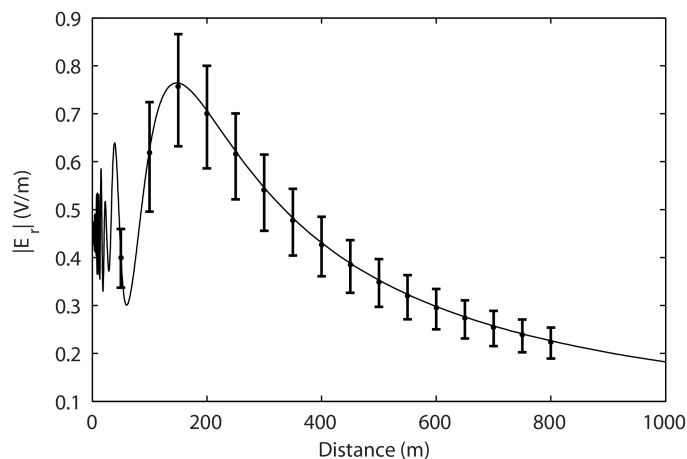


FIGURE 4.10 : Intervalles de confiance à 95% en fonction de la distance d’observation obtenus à partir des distributions de sortie du champ réfléchi obtenues par modélisation par chaos polynomial.

en dessous de 50 m. En effet, la trop forte instabilité de cette zone rend la modélisation par chaos polynomial irréalisable pour un budget de calcul raisonnable. La distribution de sortie du champ réfléchi est ensuite obtenue pour chaque distance par simulations de Monte Carlo sur le méta-modèle obtenu par chaos polynomial. On trace ensuite la courbe moyenne de l’amplitude du champ réfléchi avec les intervalles de confiance à 95% issus des distributions de sortie estimées à partir des méta-modèles. Les résultats sont présentés en Figure 4.10. Cette figure montre une diminution de la largeur de l’intervalle de confiance avec l’augmentation de la distance. Ce résultat est assez intuitif car il est assez vraisemblable que les effets des paramètres architecturaux soient plus atténués à grande distance de l’immeuble.

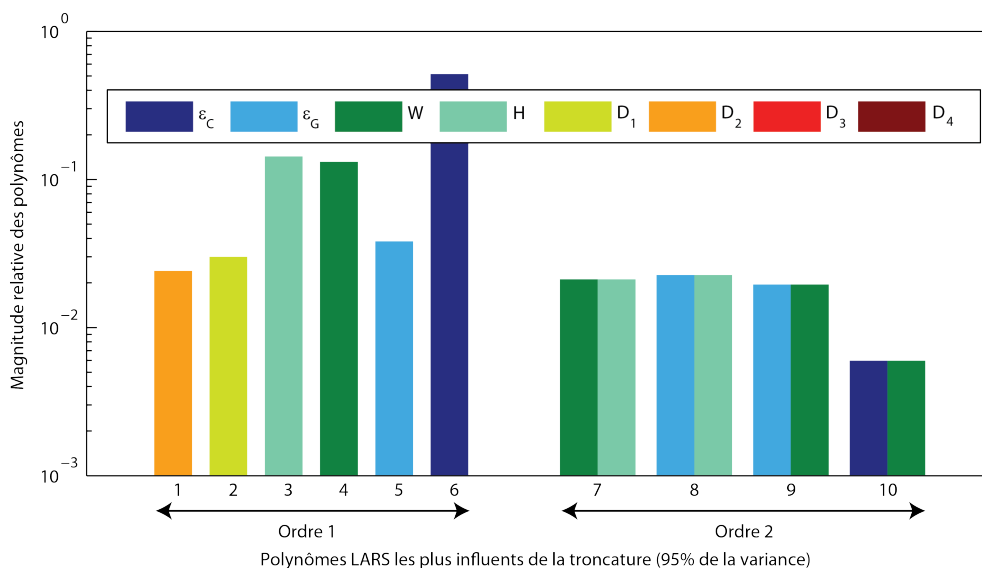


FIGURE 4.11 : Analyse de signature du modèle de chaos polynomial à  $d = 100m$  en direction spéculaire.

On se place maintenant à  $d = 100m$  de l'immeuble en direction spéculaire et on veut évaluer l'influence relative de chaque paramètre sur le champ réfléchi. Pour cela, on effectue une analyse de signature sur le modèle de chaos polynomial généré comme présenté à la Section 4.1.6. Les résultats de cette analyse de signature sont présentés en Figure 4.11. Sur cette figure, l'axe des abscisses représente les polynômes du méta-modèle de chaos LARS ordonnés dans l'ordre de sélection de l'algorithme. Pour plus de lisibilité, ne sont présentés sur cette figure que les polynômes ayant le poids le plus élevé et exprimant à eux seuls 95% de la variance de sortie du méta-modèle. L'axe des ordonnées de gauche représente la magnitude relative de chaque polynôme qui apparaît sous la forme d'une barre dont la hauteur représente la magnitude. Dans le même temps, les barres sont divisées horizontalement pour représenter la nature de chaque polynôme par un code couleur. Plus la barre est épaisse, plus l'ordre total  $|\alpha|$  du polynôme est élevé ( $\alpha$  étant le multi-indice du polynôme considéré). La largeur de chaque division de barre représente les poids relatifs  $\alpha_i$  de chaque variable dans le polynôme ( $\alpha_i$  étant la  $i$ -ème composante de  $\alpha$ ). Chaque couleur correspond à un paramètre d'entrée spécifique. Cette figure montre que les polynômes les plus influents sur le champ réfléchi à 100 m sont des polynômes d'ordre 1 univariés fonction des permittivités du ciment et du verre, de la largeur et la hauteur de fenêtre et de l'espacement entre les fenêtres. Il apparaît également que les paramètres représentant la distance des fenêtres aux bords de l'immeuble ( $D_3$  et  $D_4$ ) sont absents des polynômes les plus influents). On remarquera aussi que ne figurent pas de polynôme univarié d'ordre total  $|\alpha| = 2$  et qu'il n'y a que des polynômes d'interactions pour cet ordre total. Le phénomène physique possède donc des interactions importantes qui semblent être constituées en majorité de l'interaction entre une permittivité et un paramètre géométrique (3 polynômes sur 4 dans ce cas parmi les plus influents).

Intéressons-nous maintenant à l'influence du décalage angulaire de l'observation du champ avec une onde qui arrive toujours en incidence normale sur l'immeuble. La distance d'observation est fixée à 100 mètres de l'immeuble et l'angle d'observation est décalé de 0 à 10°. On répète le processus que l'on a mis en place pour la Figure 4.10 et on construit des méta-modèles de chaos polynomial pour plusieurs angles. La Figure 4.12 présente la courbe moyenne du champ réfléchi en fonction de l'angle d'observation à 100 m et les intervalles de confiance à 95% correspondants. De même que précédemment, pour l'évolution de l'intervalle de confiance en fonction de la distance, l'intervalle diminue avec l'augmentation de l'écart angulaire d'observation. Les effets des paramètres architecturaux sont là encore atténués avec l'angle d'observation. Ensuite, la distance reste toujours fixée à 100m et l'écart angulaire est lui fixé à une valeur de 10°. On veut faire l'analyse de signature précédente pour ce point d'observation pour étudier l'évolution de l'influence des différents paramètres par rapport à un angle nul. L'analyse de signature est présentée en Figure 4.13. Cette figure affiche cette fois-ci uniquement les polynômes représentant 90% de la variance pour des raisons de clarté d'affichage. La lecture de la figure se fait de la même façon que pour la Figure 4.11. Il apparaît clairement que le phénomène à modéliser est plus complexe que dans le cas de la direction spéculaire puisque beaucoup plus de polynômes ont une influence significative sur la variance. Cette complexité est aussi révélée par le fait que des polynômes d'ordre total 9 et 10 sont impliqués significativement dans la variance. Si le phénomène est plus complexe que précédemment, on retrouve néanmoins la prédominance des polynômes linéaires univariés dans la variance totale. De même, la présence significative des polynômes d'interaction nous conforte dans notre choix de base pleine comme ensemble de sélection pour le LARS. Enfin, on remarque que les variables correspondant

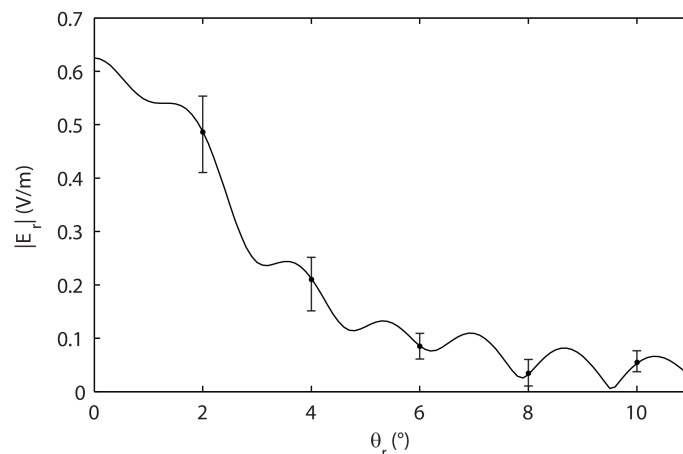


FIGURE 4.12 : Intervalles de confiance à 95% en fonction de l'angle d'observation à 100m obtenus à partir des distributions de sortie du champ réfléchi obtenues par modélisation par chaos polynomial.

à un paramètre géométrique représentant une distance horizontale sont beaucoup plus présents dans les polynômes d'ordre élevé que les paramètres verticaux. Il y a donc une très forte complexité de l'effet de ces paramètres horizontaux. Cela s'explique par le fait qu'en augmentant l'angle d'observation (décalage horizontal d'un côté de l'immeuble), les diffractions selon l'axe horizontal deviennent de plus en plus complexes alors que les diffractions venant de l'axe vertical restent à peu près les mêmes car on se situe à la même hauteur de l'immeuble (on ne se décale pas verticalement).

Les travaux effectués dans le cadre de cette application et la présentation de l'analyse de signature ont fait l'objet d'une publication d'article dans la revue *IEEE Transaction on Antenna Propagation* [Kersaudy et al. 2014].

Dans le cadre de la construction de méta-modèles basés sur le chaos polynomial pour la résolution de problèmes concrets, j'ai également contribué à la publication d'un article traitant de la gestion de l'incertitude dans le domaine des méta-matériaux où la réalisation pratique de tels matériaux conduit généralement à des différences de performances par rapport à celles prévues par la simulation. Le chaos polynomial permet ainsi de prendre en compte l'influence de cette incertitude de réalisation [Arya et al. 2015].

## 4.2 Le Krigeage universel

### 4.2.1 Un bref historique

Le Krigeage universel (également appelé dans un cadre particulier modélisation par processus Gaussiens) a été initialement introduit par Matheron en 1969 [Matheron 1969] dans le domaine des géostatistiques comme un outil d'interpolation de données ponctuelles considérées comme les points d'une trajectoire d'un champ aléatoire. Cette approche a été largement utilisée par la suite dans le domaine des expériences numériques comme outil de méta-modélisation [Sacks et al. 1989b;a, Rasmussen and Williams 2006, Santner et al. 2003], dans le domaine de la planification séquentielle d'expériences [Bect et al. 2012, Picheny et al. 2010, Dubourg et al. 2011] ou encore dans le domaine de l'optimisation globale

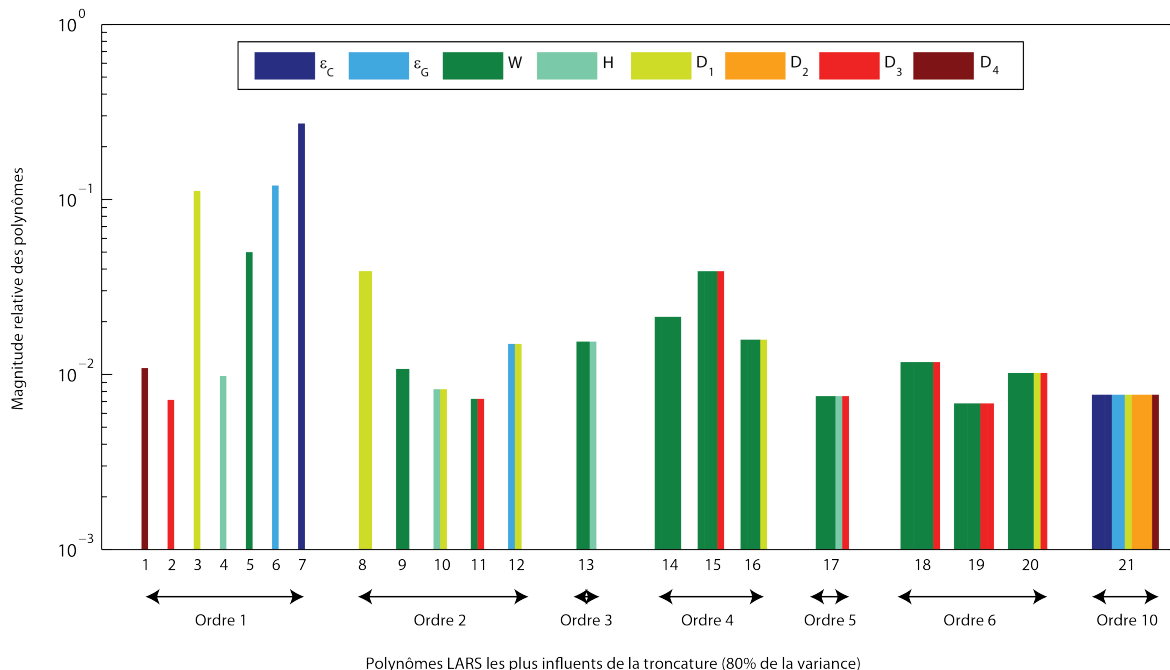


FIGURE 4.13 : Analyse de signature du modèle de chaos polynomial à  $d = 100$  m pour un angle d’observation de  $10^\circ$

[Jones et al. 1998].

### 4.2.2 La formulation du Krigeage universel

Dans un cadre entrées-sortie identique à celui présenté au début de la Section 4, le Krigeage universel traite la sortie déterministe du modèle physique  $y = f(\mathbf{x})$  comme la réalisation d’un processus stochastique  $Y(\mathbf{x})$  décomposé comme la somme d’une partie linéaire déterministe et d’un processus stochastique  $Z(\mathbf{x})$  de moyenne nulle indexé par  $\mathbf{x} \in \mathbb{X}$ . Pour mémoire, un processus stochastique est une variable aléatoire fonction d’une ou plusieurs variables scalaires (correspondant ici aux paramètres  $\mathbf{x}$ ). Toute réalisation d’un tel processus est également appelée *trajectoire*. Dans ces travaux, ce processus sera Gaussien, i.e. la loi statistique régissant la variable aléatoire  $Y(\mathbf{x})$  sera une loi normale et plus généralement pour n’importe quel  $N$ -uplet de points  $\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(N)}$  dans le domaine d’entrée, le vecteur  $(Z(\mathbf{x}^{(1)}), \dots, Z(\mathbf{x}^{(N)}))$  possède une loi normale multivariée. Notons  $Y(\mathbf{x})$  le processus et  $Y(\mathbf{x}, \omega)$  une réalisation de ce processus où  $\omega$  appartient à l’espace des évènements  $\Omega$  sous-jacent. On obtient alors la formulation du Krigeage universel suivante :

$$f(\mathbf{x}) = Y(\mathbf{x}, \omega) = \sum_{k=0}^{P-1} \beta_k \psi_k(\mathbf{x}) + Z(\mathbf{x}, \omega) \quad \mathbf{x} \in \mathbb{X}, \omega \in \Omega \quad (4.28)$$

Dans cette équation,  $\psi = \{\psi_k, k = 0 \dots P - 1\}$  est un ensemble de fonctions de régression,  $\beta = \{\beta_k, k = 0 \dots P - 1\}$  sont les coefficients de régression correspondants et  $Z(\mathbf{x})$  est le processus Gaussien [Sacks et al. 1989a, Rasmussen and Williams 2006] indexé par  $\mathbf{x} \in \mathbb{X}$ . Ce processus est ici stationnaire à l’ordre 2 (le processus étant Gaussien il est également stationnaire au sens strict), i.e. sa moyenne et sa variance ne dépendent pas de  $\mathbf{x}$  et sa

covariance ne dépend que de l'incrément  $\mathbf{x} - \mathbf{x}'$  de l'espace d'entrée et non des valeurs de  $\mathbf{x}$  et  $\mathbf{x}'$  en elles-mêmes. La loi de ce processus est alors entièrement caractérisée par sa moyenne et sa fonction (ou noyau) de covariance  $k$  :

$$\mathbb{E}[Z(\mathbf{x})] = 0 \quad \forall \mathbf{x} \in \mathbb{X} \quad (4.29)$$

$$\text{Cov}[Z(\mathbf{x}), Z(\mathbf{x}')] = k(\mathbf{x}, \mathbf{x}') = \sigma^2 R(\mathbf{x} - \mathbf{x}') \quad \forall \{\mathbf{x}, \mathbf{x}'\} \in \mathbb{X}^2 \quad (4.30)$$

Dans les équations ci-dessus,  $\sigma^2 = \text{Var}[Z(\mathbf{x})]$  est la variance constante du processus et  $R$  est ce que l'on appelle la *fonction d'autocorrélation*.

### 4.2.3 Les fonctions d'autocorrélation

Dans la théorie du Krigeage, il y a classiquement deux approches pour modéliser la structure de dépendance entre les valeurs du processus aux différents points de l'espace d'entrée. On distingue premièrement l'approche dite du *variogramme*. Cette approche largement utilisée dans le domaine géostatistique [Matheron 1969] repose sur l'utilisation d'une fonction  $\gamma$  que l'on appelle *semi-variogramme* et définie comme suit :

$$2\gamma(h) = \text{Var}[Z(\mathbf{x} + h) - Z(\mathbf{x})] \quad (4.31)$$

L'hypothèse de stationnarité du processus n'étant pas forcément valide pour les applications rencontrées en géostatistique, une hypothèse plus faible de *stationnarité intrinsèque* est utilisée. On entend par stationnarité intrinsèque le fait que le processus  $\Delta Z(h) = \{Z(\mathbf{x} + h) - Z(\mathbf{x}), \mathbf{x} \in \mathbb{X}\}$  est stationnaire à l'ordre 2. Il est possible de lier ce semi-variogramme  $\gamma$  à covariance du processus en développant l'expression de l'Eq. (4.31) :

$$2\gamma(h) = \text{Var}[Z(\mathbf{x} + h)] + \text{Var}[Z(\mathbf{x})] - 2\text{Cov}[Z(\mathbf{x}), Z(\mathbf{x}')] \quad (4.32)$$

La construction du modèle de Krigeage est alors sous-tendue à la connaissance de ce variogramme. En pratique, il est très rare de connaître ce variogramme. Il faut donc l'estimer à partir d'observations de la sortie. Sachant que la définition de la structure de dépendance peut avoir une grande influence sur la précision des prédictions du modèle de Krigeage, l'étape de l'estimation du variogramme ou de la fonction de covariance conditionne la qualité du modèle. Dans l'approche dite variographique classique, on recourt à ce que l'on appelle la nuée variographique pour estimer le variogramme. La nuée variographique est en fait un ensemble d'écartés entre des points de l'espace d'entrée associés aux carrés des différences entre les sorties correspondantes. Considérons un ensemble de  $N$  points d'entrée associés à leur image par le modèle physique :  $\{(\mathbf{x}^{(1)}, f(\mathbf{x}^{(1)})) \dots (\mathbf{x}^{(N)}, f(\mathbf{x}^{(N)}))\}$ . On a alors la nuée variographique correspondante :

$$\nu = \left\{ \left( \|\mathbf{x}^{(i)} - \mathbf{x}^{(j)}\|, \frac{1}{2}(f(\mathbf{x}^{(i)}) - f(\mathbf{x}^{(j)}))^2 \right), \{i, j\} \in \llbracket 1, N \rrbracket^2 \right\} \quad (4.33)$$

On peut théoriquement estimer empiriquement le variogramme à partir d'un jeu de données, mais cela nécessite un grand nombre de points qui ne sont pas forcément à disposition. La solution généralement privilégiée est donc de partir d'un modèle paramétrique construit à partir d'une expertise métier dont on estimera les paramètres à l'aide de la nuée variographique. Les techniques utilisées pour estimer les paramètres variographiques sont

multiples mais nécessitent un nombre de points significatif qui devient vraiment très important au fur et à mesure que l'on augmente le nombre de paramètres d'entrée ; problème qui ne se rencontre pas en géostatistique où la majeure partie des problèmes sont en dimension 2 ou 3 [Ginsbourger 2009]. Dans les problématiques développées dans ces travaux de thèse, on veut pouvoir traiter de manière automatique des problèmes très différents ayant potentiellement un nombre important de paramètres d'entrée. On ne privilégiera donc pas l'approche du variogramme pour l'emploi du Krigeage dans ces travaux de thèse. Il est cependant intéressant de citer les travaux de [Jawad 2014] intégrant une application de la dosimétrie numérique utilisant l'approche variographique. Le cas d'application en 2 dimensions vise à évaluer l'exposition d'un modèle anatomique induite par une onde plane ayant ses angles d'azimut et d'élévation variables.

Une seconde approche également utilisée pour modéliser la structure de dépendance du processus Gaussien est de remplacer l'estimation du variogramme par l'estimation de la fonction de covariance. Comme pour l'approche du variogramme, on va ici aussi définir a priori une structure paramétrique pour la fonction de covariance. L'estimation des paramètres de covariance est faite ici par la maximisation d'un critère d'adéquation entre les données observées et les prédictions fournies par le processus muni de la fonction de covariance paramétrique en question. Typiquement, des méthodes de maximisation de vraisemblance ou de validation croisée peuvent être utilisées comme on le verra par la suite. C'est cette approche qui sera utilisée dans les travaux présentés dans ce mémoire.

Comme on l'a vu dans l'Eq. (4.30), la covariance s'exprime dans le cas de processus Gaussiens stationnaires en fonction d'une fonction d'autocorrélation (ou du noyau de covariance qui est proportionnel à l'autocorrélation dans le cas stationnaire) définie a priori. Cet a priori est généralement choisi en fonction de la connaissance de certaines caractéristiques de la sortie du modèle physique. On utilisera ici des fonctions d'autocorrélation anisotropes. L'utilisation de telles fonctions est pertinente quand les paramètres d'entrée ont des influences différentes sur la variable de sortie. On présente plusieurs exemples de fonctions d'autocorrélation qui seront utilisées par la suite.

**La fonction d'autocorrélation Nugget** La fonction d'autocorrélation nugget se définit de la façon suivante :

$$R(\mathbf{x} - \mathbf{x}') = \mathbf{1}_{\mathbf{x}=\mathbf{x}'}(\mathbf{x} - \mathbf{x}') \quad (4.34)$$

Dans cette équation  $\delta$  représente la fonction de Dirac. Un processus Gaussien muni de ce noyau aura donc toutes ses observations indépendantes et identiquement distribuées. Cette fonction d'autocorrélation sert donc à modéliser un bruit blanc. En pratique, ce noyau de covariance est utilisé en plus d'un autre noyau quand la sortie du modèle physique est intrinsèquement incertaine à cause d'une incertitude de mesure par exemple.

**La fonction d'autocorrélation Gaussienne** Un autre noyau couramment utilisé est la fonction d'autocorrélation Gaussienne. Elle se définit comme suit :

$$R(\mathbf{x} - \mathbf{x}', \boldsymbol{\theta}) = \exp \left( -\frac{1}{2} \sum_{i=1}^M \left( \frac{|x_i - x'_i|}{\theta_i} \right)^2 \right) \quad (4.35)$$

Dans cette équation,  $\boldsymbol{\theta} = \{\theta_1, \dots, \theta_M\}$  représente les paramètres de covariance modélisés ici par des longueurs de corrélation que l'on devra estimer en fonction d'observations de la sortie. Ce noyau de covariance confère au processus de sortie un degré infini de

différentiabilité. Il est donc utilisé lorsque l'on a cet a priori de caractère  $C^\infty$  sur la sortie de son modèle physique.

**La fonction d'autocorrélation de Matérn** Le noyau de Matérn est également très utilisé dans le domaine de l'apprentissage statistique [Santner et al. 2003]. Il est défini par l'expression :

$$R(\mathbf{x} - \mathbf{x}', \boldsymbol{\theta}, \nu) = \prod_{i=1}^M \frac{1}{2^{\nu-1} \Gamma(\nu)} \left( 2\sqrt{\nu} \frac{|x_i - x'_i|}{\theta_i} \right)^\nu \mathcal{K}_\nu \left( 2\sqrt{\nu} \frac{|x_i - x'_i|}{\theta_i} \right) \quad (4.36)$$

Dans l'équation ci-dessus,  $\boldsymbol{\theta} = \{\theta_1, \dots, \theta_M\}$  est toujours le vecteur des paramètres de covariance (longueurs de corrélation) à déterminer.  $\nu \geq 1/2$  est ce que l'on appelle le paramètre de forme,  $\Gamma$  est la fonction Gamma et  $\mathcal{K}_\nu$  est la fonction de Bessel de seconde espèce [Abramovitz and Stegun 1972]. Cette fonction d'autocorrélation de Matérn possède des caractéristiques intéressantes car en fonction de la valeur du paramètre de forme  $\nu$ , le processus Gaussien sera  $\lceil \nu - 1 \rceil$  différentiable ( $\lceil \cdot \rceil$  correspond ici à la fonction partie entière par excès). En pratique la valeur de  $\nu$  utilisée est égale à  $3/2$  pour un processus 1 fois différentiable et égale à  $5/2$  pour un processus 2 fois différentiable.

Ayant affaire à des phénomènes réguliers dans la problématique de l'exposition et de la dosimétrie numérique, les choix du noyau Gaussien ou du noyau de Matérn avec  $\nu = 5/2$  apparaît pertinent.

#### 4.2.4 Le meilleur prédicteur linéaire sans biais (BLUP)

Soit un plan d'expériences de  $N$  points  $\mathcal{X}^N = \{\mathbf{x}^{(1)} \dots \mathbf{x}^{(N)}\}$ . Sous les hypothèses du Krigeage universel, on définit le meilleur estimateur linéaire non biaisé de la sortie du modèle physique (BLUP for *best linear unbiased predictor*) de la manière suivante :

$$\hat{m}_{\hat{\Psi}}(\mathbf{x}) = \boldsymbol{\psi}(\mathbf{x})^T \hat{\boldsymbol{\beta}} + \mathbf{r}(\mathbf{x})^T \mathbf{R}^{-1}(\mathbf{y} - \boldsymbol{\Psi} \hat{\boldsymbol{\beta}}) \quad (4.37)$$

où  $\boldsymbol{\psi}(\mathbf{x}) = (\psi_0(\mathbf{x}), \dots, \psi_{P-1}(\mathbf{x}))^T$  et  $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \dots, \hat{\beta}_{P-1})^T$  sont respectivement les vecteurs des fonctions de régression et des estimateurs des coefficients de régression,  $\mathbf{r}(\mathbf{x}) = (R(\mathbf{x} - \mathbf{x}^{(1)}, \boldsymbol{\theta}), \dots, R(\mathbf{x} - \mathbf{x}^{(N)}, \boldsymbol{\theta}))^T$  est le vecteur de corrélation entre  $\mathbf{x}$  et les points du plan d'expériences,  $\mathbf{R} = (R(\mathbf{x}^{(i)} - \mathbf{x}^{(j)}, \boldsymbol{\theta}))_{1 \leq i, j \leq N}$  est la matrice de corrélation du plan d'expériences,  $\mathbf{y} = (f(\mathbf{x}^{(1)}), \dots, f(\mathbf{x}^{(N)}))^T$  est la réponse correspondante aux points du plan d'expériences et  $\boldsymbol{\Psi} = (\boldsymbol{\psi}(\mathbf{x}^{(1)}), \dots, \boldsymbol{\psi}(\mathbf{x}^{(N)}))^T$  est la matrice des fonctions de régression calculées aux points du plan d'expériences. Dans la formulation de l'Eq. (4.37),  $\hat{\boldsymbol{\beta}}$  est l'estimateur des moindres carrés généralisés :

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{\Psi}^T \mathbf{R}^{-1} \boldsymbol{\Psi})^{-1} \boldsymbol{\Psi}^T \mathbf{R}^{-1} \mathbf{y} \quad (4.38)$$

Il est important de remarquer que si l'ensemble des fonctions de régression se réduit au polynôme constant égal à 1, le modèle de Krigeage est appelé *Krigeage ordinaire*. La partie déterministe du modèle de Krigeage ordinaire se limite donc à la moyenne estimée de la sortie. En plus de cet estimateur sans biais, le Krigeage fournit un intervalle d'incertitude autour de cet estimateur. En effet, l'erreur du modèle étant modélisée dans l'Eq. (4.28) par un processus Gaussien, le BLUP est lui même l'espérance conditionnelle aux observation d'un processus Gaussien de variance  $s_{\hat{\Psi}}^2(\mathbf{x})$  :

$$s_{\hat{\Psi}}^2(\mathbf{x}) = \sigma^2 \left( 1 - \mathbf{r}(\mathbf{x})^T \mathbf{R}^{-1} \mathbf{r}(\mathbf{x}) + \mathbf{u}(\mathbf{x})^T (\boldsymbol{\Psi}^T \mathbf{R}^{-1} \boldsymbol{\Psi})^{-1} \mathbf{u}(\mathbf{x}) \right) \quad (4.39)$$



avec

$$\mathbf{u}(\mathbf{x}) = \mathbf{\Psi}^T \mathbf{R}^{-1} \mathbf{r}(\mathbf{x}) - \psi(\mathbf{x}) \quad (4.40)$$

La précision d'un modèle de Krigeage est fortement dépendante de la précision avec laquelle on infère les longueurs de corrélation  $\boldsymbol{\theta}$  de la fonction d'autocorrélation. Parmi les méthodes classiquement utilisées pour l'inférence de ces paramètres il y a premièrement les méthodes basées sur la maximisation d'un critère d'adéquation entre les données d'observation et le modèle de Krigeage comme la validation croisée ou le maximum de vraisemblance [Bachoc 2013]. L'estimation de ces hyperparamètres peut également se faire à travers une approche bayésienne [Santner et al. 2003]. Dans ces travaux, les longueurs de corrélation sont inférées par la méthode du maximum de vraisemblance. On rappelle l'expression de la vraisemblance des observations  $\mathbf{y}$  respectant la distribution normale multivariée :

$$L(\mathbf{y}|\boldsymbol{\beta}, \sigma^2, \boldsymbol{\theta}) = \frac{1}{((2\pi\sigma^2)^N [\det \mathbf{R}(\boldsymbol{\theta})])^{1/2}} \exp \left[ -\frac{1}{2\sigma^2} (\mathbf{y} - \mathbf{\Psi}\boldsymbol{\beta})^T \mathbf{R}(\boldsymbol{\theta})^{-1} (\mathbf{y} - \mathbf{\Psi}\boldsymbol{\beta}) \right] \quad (4.41)$$

La simplification de la vraisemblance en prenant l'opposé du logarithme  $-\log(L)$  et en utilisant l'expression de l'Eq. (4.38) aboutit à une expression qui dépend seulement du vecteur des hyperparamètres. La maximisation de la vraisemblance revient alors à minimiser une expression plus simple appelée *fonction de vraisemblance concentrée* :

$$L_r(\boldsymbol{\theta}) = \hat{\sigma}^2(\boldsymbol{\theta}) [\det \mathbf{R}(\boldsymbol{\theta})]^{1/N} \quad (4.42)$$

où :

$$\hat{\sigma}^2(\boldsymbol{\theta}) = \frac{1}{N} [\mathbf{y} - \mathbf{\Psi}\hat{\boldsymbol{\beta}}(\boldsymbol{\theta})]^T \mathbf{R}^{-1}(\boldsymbol{\theta}) [\mathbf{y} - \mathbf{\Psi}\hat{\boldsymbol{\beta}}(\boldsymbol{\theta})] \quad (4.43)$$

et  $N$  est le nombre de points d'observation. Ainsi l'estimation de  $\boldsymbol{\theta}$  est faite par la minimisation globale de cette fonction de vraisemblance concentrée. Dans ces travaux, le problème d'optimisation globale est traité à l'aide de l'algorithme BOXMIN de la boîte à outil Matlab *DACE* [Lophaven et al. 2002] qui est basé sur un algorithme de dichotomie multivariée. Cet algorithme assez simple peut voir ses performances se dégrader avec une augmentation de la dimension du problème. Il existe d'autres algorithmes d'optimisation plus complexes et plus performants comme celui de la boîte à outil R *DiceKriging* [Roustant et al. 2010]. Néanmoins celui de la boîte *DACE* suffit pour les exemples et les méthodes présentées dans ces travaux.

Le méta-modèle basé sur le Krigeage est interpolant. Il est également asymptotiquement consistant quand la fonction de covariance du processus sous-jacent est continue sur sa diagonale [Vazquez 2005](pp. 132-156). On entend par consistance asymptotique le fait que la variance de la différence entre le processus Gaussien associé au meilleur prédicteur linéaire sans biais et le processus exact sous-jacent modélisé tend vers 0 quand  $N$  tend vers l'infini [Vazquez 2005](pp. 132-156). De plus, le processus de sortie est Gaussien conditionnellement aux observations en chaque point du domaine d'entrée :

$$Y(\mathbf{x})|\mathbf{y} \sim \mathcal{N}(\hat{m}_{\hat{\mathbf{y}}}(\mathbf{x}), s_{\hat{\mathbf{y}}}^2(\mathbf{x})) \quad (4.44)$$

On peut voir sur la Figure 4.14 un exemple de meilleur estimateur linéaire sans biais de la fonction analytique en une dimension  $y = x \sin(x)$  construit à partir des points d'observation marqués par une croix rouge. Cette figure présente également l'intervalle de confiance à 95% conditionnel aux observations du modèle de Krigeage obtenu à partir



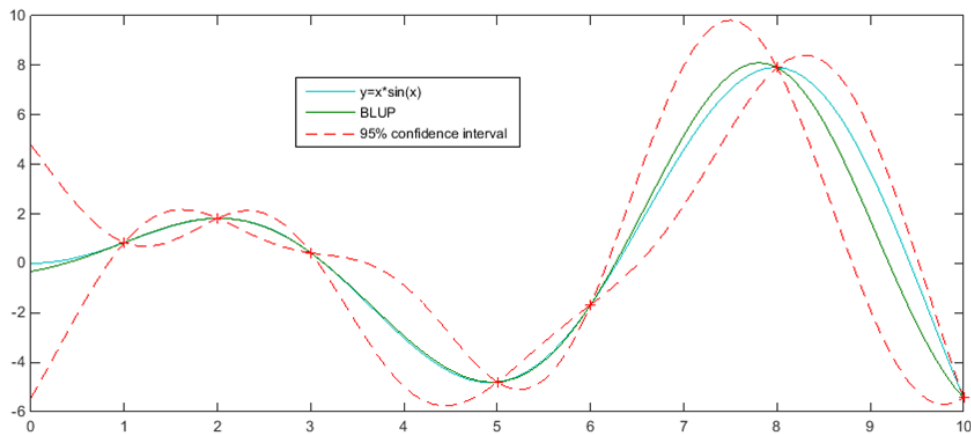


FIGURE 4.14 : Exemple de modèle de Krigeage pour la fonction  $y = x \sin(x)$

de la variance locale de l'Eq. (4.39). Un indicateur de précision d'un modèle de Krigeage universel classiquement utilisé est l'*erreur moyenne quadratique intégrée* (IMSE) [Picheny et al. 2010, Sacks et al. 1989a] définie par :

$$IMSE = \int_{\mathbf{x}} s_y^2(\mathbf{x}) p_{\mathbf{x}}(\mathbf{x}) d\mathbf{x} \quad (4.45)$$

Néanmoins, l'IMSE ne reflète pas la capacité de généralisation du méta-modèle. Dans un but d'analyse de la capacité de généralisation du modèle et de comparaison de performances avec les développements du chaos polynomial, la précision du modèle de Krigeage sera estimée dans mes travaux par validation croisée *leave-one-out* (cf Section 4.1.5). Comme pour les développements en séries polynomiales, l'erreur *leave-one-out* définie dans l'Eq. (4.13) peut être calculée analytiquement par la méthode décrite dans [Dubrule 1983].

Les fonctions de régression sont souvent déterminées à partir de la connaissance a priori de l'évolution de la sortie en fonction des entrées [Oakley 2004]. Un modèle de Krigeage universel peut donc être difficile à construire dans le sens où l'on ne bénéficie que rarement de connaissances a priori sur la sortie d'un modèle physique. Sans cet apriori, la solution généralement privilégiée est de considérer uniquement le polynôme constant égal à 1 comme fonction de régression, ce qui revient à un *Krigeage ordinaire*. Une des contributions de mes travaux a donc été d'élaborer une méthodologie pour construire un modèle de Krigeage universel à l'aide des polynômes d'un développement de chaos polynomial.

### 4.3 Hybridation du chaos polynomial et du Krigeage universel

La méthodologie proposée par ces travaux est basée sur l'utilisation des polynômes sélectionnés par l'algorithme LARS dans un développement de chaos polynomial comme fonctions de régression dans la formulation du Krigeage universel de l'Eq. (4.28). Nous proposons donc d'utiliser la capacité de généralisation des développements en séries polynomiales ainsi que la capacité du LARS à sélectionner l'ensemble creux de polynômes

le plus pertinent pour construire l'approximation du Krigeage universel. Le but est de sélectionner les fonctions qui apportent le plus d'information utile au modèle de Krigeage. En faisant cela, le processus Gaussien n'a plus pour rôle que de modéliser le résidu entre la sortie et la partie déterministe de la formulation constituée par ces fonctions de régression. L'algorithme utilisé pour construire un tel modèle de Krigeage se présente comme suit :

1. Construire un plan d'expériences initial d'une taille donnée ;
2. Appliquer l'algorithme LARS à un ensemble candidat de polynômes issu d'un développement de chaos polynomial ;
3. A chaque étape de sélection du LARS, construire le modèle de Krigeage universel correspondant à l'ensemble courant de polynômes sélectionnés ;
4. Appliquer une LOOCV sur chaque métamodèle ainsi construit pour évaluer leur précision ;
5. Sélectionner alors le modèle avec la plus grande capacité de généralisation, i.e. celui produisant l'erreur *leave-one-out* la plus faible ;
6. Si la capacité de généralisation obtenue n'est pas satisfaisante, augmenter la taille du plan d'expériences en ajoutant de nouveaux points et reprendre l'algorithme à la seconde étape.

Il est intéressant de noter que le premier polynôme choisi par le LARS est toujours le polynôme constant, le Krigeage ordinaire est donc toujours le premier modèle dont la précision est évaluée. En conséquence si aucun autre polynôme n'apporte une amélioration de précision au modèle de Krigeage, le modèle choisi par l'algorithme sera le Krigeage ordinaire. On se situe donc dans une démarche de sélection du modèle de Krigeage représentant au mieux les données. On appellera cette méthode hybride le LARS-Kriging-PC. Cette méthodologie de construction d'un modèle de Krigeage universel à l'aide de polynômes issus d'un développement de chaos polynomial a fait l'objet d'une publication d'article dans la revue *Journal of Computational Physics* [Kersaudy et al. 2015]. Il existe des points communs entre cette approche et une autre méthode appelée *blind-Kriging* développée dans [Joseph et al. 2008] mais notre méthode diffère de cette dernière par l'algorithme de sélection des fonctions de régression et l'utilisation d'une base de chaos polynomial.

### 4.3.1 Exemple d'application analytique : la fonction d'Ishigami

Cette section est dédiée à la validation du LARS-Kriging-PC avec l'exemple analytique de la fonction d'Ishigami présentée dans l'Eq. (4.26). On construit donc un méta-modèle de chaos polynomial LARS classique, un méta-modèle de Krigeage ordinaire et un méta-modèle de LARS-Kriging-PC sur des plans d'expériences LHS de différentes tailles. On part d'un plan initial de 40 points augmenté par la suite via la technique NLHS présentée plus haut dans la Section 4.1.7. L'erreur moyenne quadratique relative présentée dans l'Eq. (4.11) de ces méta-modèles est ensuite calculée à partir de 100000 simulations de Monte Carlo (la fonction d'Ishigami étant rapide à calculer). Cette démarche est répétée 50 fois (à partir de 50 LHS initiaux différents) dans le but d'observer la variabilité des résultats. Les résultats sont présentés en Figure 4.15 sous forme de boxplots affichant la

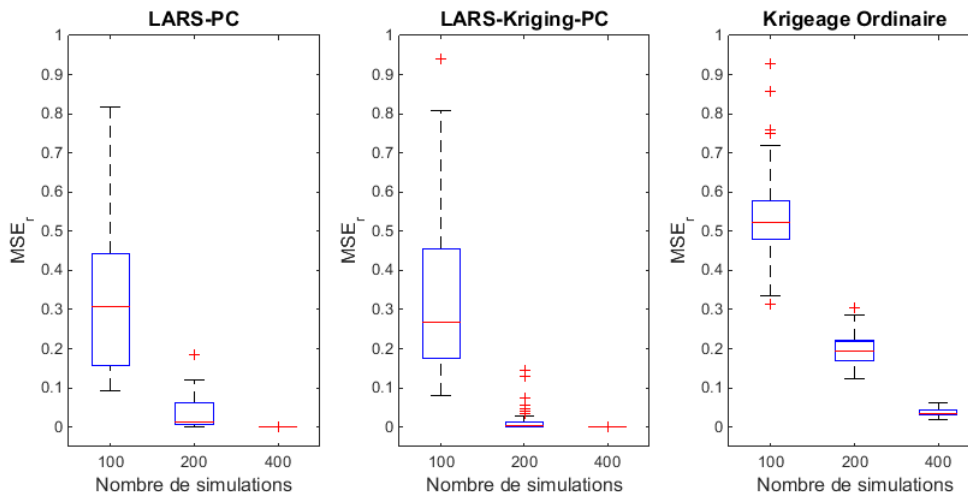


FIGURE 4.15 : Comparaison des méthodes de LARS-PC, de LARS-Kriging-PC et de Krigage ordinaire pour 50 LHS initiaux différents de 40 points augmentés jusqu'à 160 points par la technique NLHS pour la fonction d'Ishigami.

médiane de l'erreur, les quartiles et les valeurs maximales et minimales de l'erreur sur les 50 répétitions indépendantes. La Figure 4.15 montre que le LARS-Kriging-PC obtient une meilleure précision que le chaos polynomial LARS classique (LARS PC) et que le Krigage ordinaire en terme de valeur médiane. L'écart de précision semble plus significatif entre le LARS-Kriging-PC et le Krigage ordinaire qu'entre le LARS-Kriging-PC et le chaos polynomial LARS classique. Enfin, pour un faible nombre de points, le Krigage ordinaire semble avoir les résultats les moins variables d'entre les trois méthodes.

### 4.3.2 Exemple d'application analytique : la fonction de Borehole

On applique maintenant la même démarche que dans la section précédente mais pour une autre fonction test analytique : la fonction de Borehole. La fonction de Borehole est une fonction test utilisée pour des problèmes de prédiction [Morris et al. 1993, Xiong et al. 2013]. Cette fonction comporte 8 paramètres d'entrée et modélise le débit d'eau à travers un trou de forage par l'équation suivante :

$$f(\mathbf{x}) = \frac{2\pi x_3(x_4 - x_6)}{\ln(x_2/x_1) \left(1 + \frac{2x_7x_3}{\ln(x_2/x_1)x_1^2x_8} + \frac{x_3}{x_5}\right)} \quad (4.46)$$

Tous les paramètres d'entrée sont modélisés par des lois uniformes indépendantes dont les intervalles de variation sont listés dans la Table 4.3.

On effectue la même comparaison de performances que dans le cas de la fonction d'Ishigami et les résultats sont présentés en Figure 4.16 sur le même modèle que précédemment. Comme pour la fonction d'Ishigami, il apparaît que le LARS-Kriging-PC a une précision plus élevée que le Krigage ordinaire et le chaos polynomial classique en terme de valeur médiane. Néanmoins, pour un nombre de point égal à 50, il y a un nombre significatif de points considérés comme aberrants qui donnent des erreurs plus élevées que pour le Krigage ordinaire. Cela peut être dû à une mauvaise optimisation de la boîte à outil DACE

Paramètre d'entrée	Intervalle
$x_1$	[0.05; 0.15]
$x_2$	[100; 50000]
$x_3$	[63070; 115600]
$x_4$	[990; 1110]
$x_5$	[63.1; 116]
$x_6$	[700; 820]
$x_7$	[1120; 1680]
$x_8$	[9855; 12045]

TABLE 4.3 : Domaine de variation des paramètres d'entrée pour la fonction de Borehole

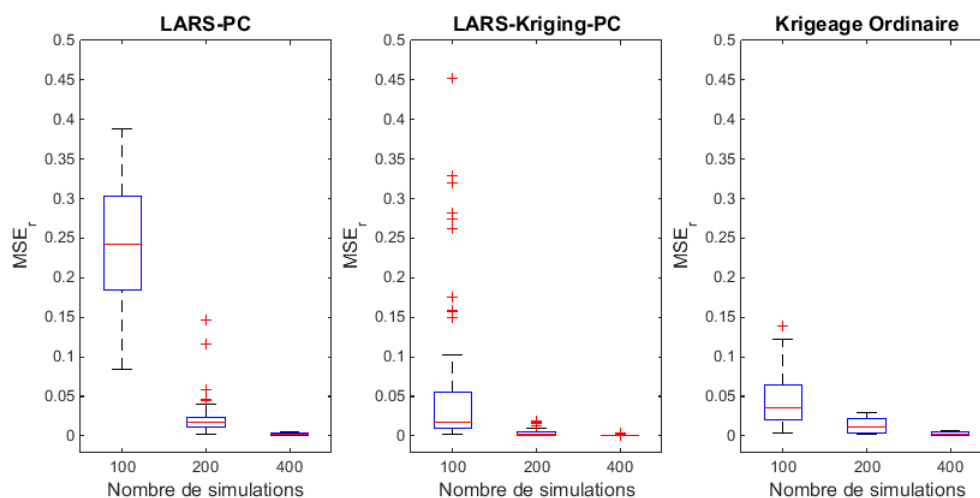


FIGURE 4.16 : Comparaison des méthodes de LARS-PC, de LARS-Kriging-PC et de Krigage ordinaire pour 50 LHS initiaux différents de 50 points augmentés jusqu'à 200 points par la technique NLHS pour la fonction de Borehole.

utilisée ici [Lophaven et al. 2002] qui fournit un algorithme d'optimisation des hyperparamètres basique et qui peut générer des solutions sous-optimale en dimension élevée. Ce phénomène n'apparaît pas pour un nombre de points plus élevé où le LARS-Kriging-PC obtient toujours de meilleurs résultats. Comme pour la fonction d'Ishigami, le Krigage ordinaire reste la méthode avec les performances les moins variables pour un faible nombre de points.

### 4.3.3 Exemple d'application analytique : la fonction de Sobol

On veut maintenant tester les performance du LARS-Kriging-PC avec un exemple en grande dimension. On utilise la fonction analytique de Sobol' en dimension 20 dont l'expression est :

$$f(\mathbf{x}) = \prod_{i=1}^{20} \frac{|4x_i - 2| + c_i}{1 + c_i} \quad \mathbf{x} \in [0, 1]^{20} \quad (4.47)$$

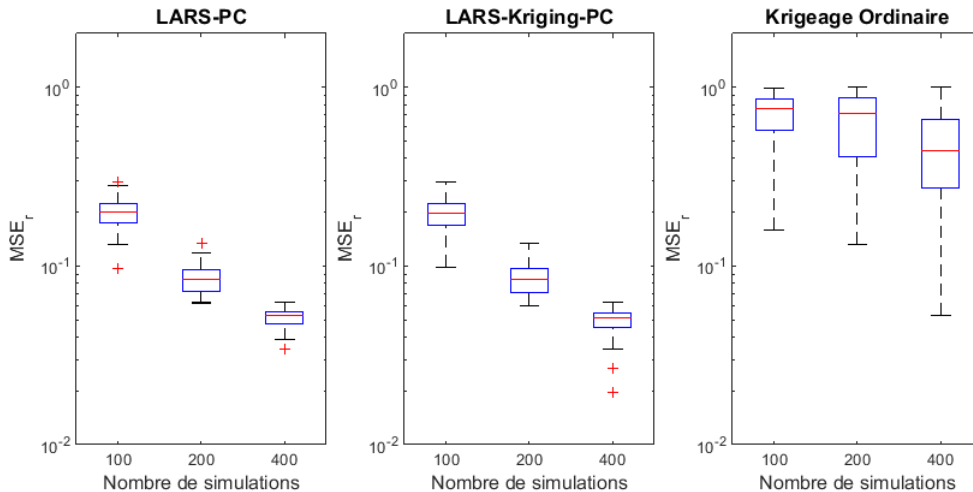


FIGURE 4.17 : Comparaison des méthodes de LARS-PC, de LARS-Kriging-PC et de Krigage ordinaire pour 50 LHS initiaux différents de 50 points augmentés jusqu'à 200 points par la technique NLHS pour la fonction de Sobol.

avec

$$\mathbf{c} = (1, 2, 5, 10, 20, 50, 100, 500, 500, 500, 500, 500, 500, 500, 500, 500, 500, 500, 500, 500) \quad (4.48)$$

La même procédure que pour les deux précédents exemples est répétée pour cette fonction. Les résultats sont présentés en Figure 4.17. Ces résultats confirment le caractère optimal de la méthode LARS-Kriging-PC. En effet, les performances du LARS-PC-Kriging sont significativement meilleures que pour le Krigage ordinaire et sont au moins égales à celles du LARS-PC classique. Le caractère creux de la fonction de Sobol' qui a quelques variables plus influentes que les autres semble bien capté par l'algorithme du LARS. Par conséquent, les modèles de LARS-PC-Kriging ont une meilleure précision que ceux du Krigage ordinaire car ils bénéficient des polynômes LARS captant ce caractère creux de la fonction. A travers les trois exemples analytiques que l'on vient de présenter, il apparaît que les performances du LARS-Kriging-PC sont dépendantes du problème et du modèle physique étudié. En présence d'un problème où le LARS-PC et le Krigage ordinaire ont des performances médiocres, le LARS-Kriging PC pourra aussi avoir des performances limitées et le manque de précision devra être compensé par l'addition de points supplémentaires. Mais d'un point de vue applicatif, cette approche nous assure d'avoir des performances au moins aussi bonnes que les deux autres méthodes. Et du fait que le Krigage ordinaire et le LARS-PC sont deux techniques de méta-modélisation très différentes qui peuvent avoir des performances totalement différentes selon le problème, l'optimalité du LARS-Kriging-PC a un intérêt non négligeable.

#### 4.3.4 Application à l'exposition d'un fœtus à un système femto-cell

Cette approche a été utilisée pour étudier l'influence du positionnement d'un système femtocell sur l'exposition corps entier d'un fœtus. Le modèle anatomique que l'on utilise

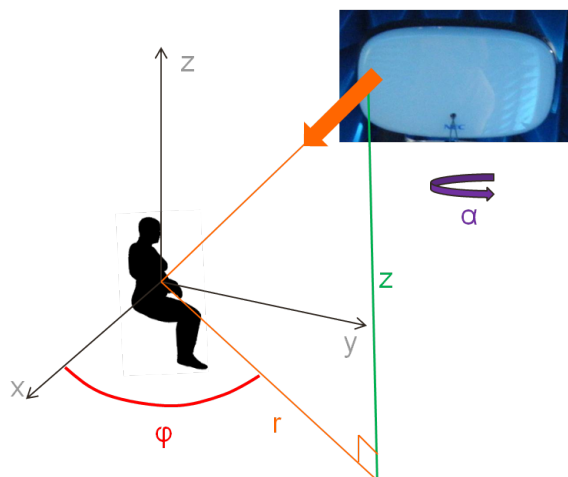


FIGURE 4.18 : Représentation géométrique des paramètres d'entrée pour le problème de l'exposition du fœtus à une femtocell

Paramètre d'entrée	Intervalle
$r$	$[1; 3]$ mètres
$\phi$	$[0, 360]$ degrés
$z$	$[-0.75, 1.75]$ mètres
$\alpha$	$[0, 360]$ degrés

TABLE 4.4 : Domaine de variation des paramètres d'entrée pour le problème de l'exposition du fœtus à une station de bases de femtocell

est ici un modèle de femme enceinte à la 26<sup>ème</sup> semaine d'aménorrhée. Ce modèle est le modèle de Victoria enceinte que l'on a utilisé dans la Section 2.5. Pour cette étude, ce modèle de femme enceinte est déformé en position assise avec les bras levés comme si elle tapait sur un clavier d'ordinateur. Une station de base de femtocell est ensuite déplacée (cf Section 2.4) autour de ce modèle de femme enceinte et on évalue le DAS corps entier du fœtus par l'hybridation modes sphériques/FDTD présentée en Section 2.4. Le DAS corps entier du fœtus est défini comme le rapport de la puissance électromagnétique absorbée du fœtus sur le poids total du fœtus. Pour étudier l'influence de la position de la station de base placée dans une pièce sur l'exposition du fœtus, on fait l'hypothèse que tous les positionnements relatifs de la femme enceinte et du système femtocell ont des chances égales d'exister. On définit alors quatre paramètres d'entrée incertains et indépendants pouvant influencer le DAS corps entier du fœtus sur le modèle du cas-test de la Section 4.6 (cf Figure 4.18) :

- Les trois coordonnées cylindriques  $r$ ,  $\phi$ ,  $z$  qui décrivent la position spatiale de la station de base dans un repère centré sur le corps de la femme enceinte.
- L'angle de rotation de la station de base par rapport à l'axe vertical  $Oz$ .

Les quatre paramètres d'entrée sont représentés par des lois uniformes indépendantes dont les domaines de variation sont donnés dans la Table 4.4. On constitue un plan d'expériences LHS initial de 50 points qui est augmenté par la technique NLHS jusqu'à 100 et 200

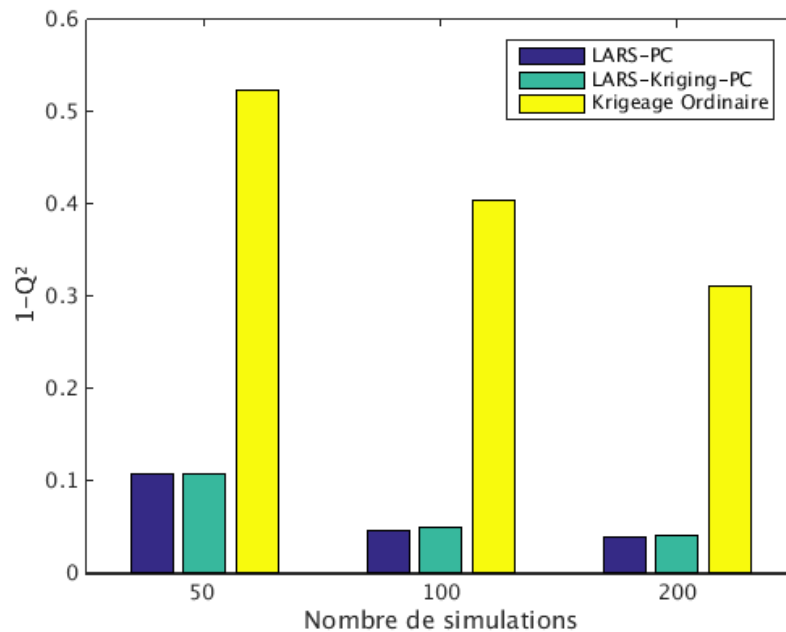


FIGURE 4.19 : Comparaison des méta-modèles LARS PC, LARS-Kriging-PC et Krigeage ordinaire pour le problème de l'exposition du fœtus à la station de base en utilisant l'erreur relative de la validation croisée *leave-one-out* en fonction du nombre de points.

points pour des besoins de précision. A partir de ces données les mêmes méta-modèles que précédemment sont construits : un chaos polynomial LARS classique, un LARS-Kriging-PC et un Krigeage ordinaire. Pour estimer la précision des méta-modèles générés, on se sert de l'erreur relative tiré de la validation croisée *leave-one-out* présentée dans la Section 4.1.5 :

$$1 - Q^2 = \frac{Err_{LOO}}{\hat{\sigma}_y^2} \quad (4.49)$$

Cette erreur est calculée pour les différents modèles et pour les différents nombres de points du plan d'expériences. Les résultats présentés en Figure 4.19 montrent que le LARS-Kriging-PC a une précision significativement supérieure au Krigeage ordinaire mais obtient des performances similaires au chaos polynomial classique quel que soit le nombre de points et génère des erreurs relatives inférieures à 0.05 pour 100 et 200 points. Pour évaluer l'influence relative des paramètres d'entrée sur la variance du DAS corps entier du fœtus, les indices totaux de Sobol' sont estimés ainsi que les indices de premier ordre [Sobol' 2001] à partir du modèle de LARS-Kriging-PC de 200 points en utilisant 1 000 000 simulations de Monte Carlo. Ces indices de Sobol' sont présentés dans la Table 4.5. Cette table montre que la variance du DAS corps entier du fœtus est principalement expliquée par les paramètres  $r$  et  $\phi$ . En particulier, le paramètre le plus influent est l'angle d'azimut  $\phi$  de la station de base de femtocell autour du corps de la femme enceinte. Cette prédominance s'explique principalement par le fait que l'épaisseur de tissu de la femme enceinte séparant le fœtus de l'air est fortement variable avec l'angle  $\phi$ . L'absorption des ondes par le corps de la femme enceinte avant qu'elles arrivent jusqu'au fœtus varie donc en conséquence. De plus, la Table 4.5 montre une différence significative entre les indices de premier ordre et les

Paramètre d'entrée	Indice de premier ordre	Indice total
$r$	0.171	0.508
$\phi$	0.389	0.764
$z$	0.014	0.261
$\alpha$	0.002	0.131

TABLE 4.5 : Indices de Sobol totaux et de premier ordre des paramètres d'entrée pour le problème de l'exposition du fœtus à la station de base obtenus à partir de 1 000 000 simulations de Monte Carlo du BLUP du métamodèle LARS-Kriging-PC pour 200 points.

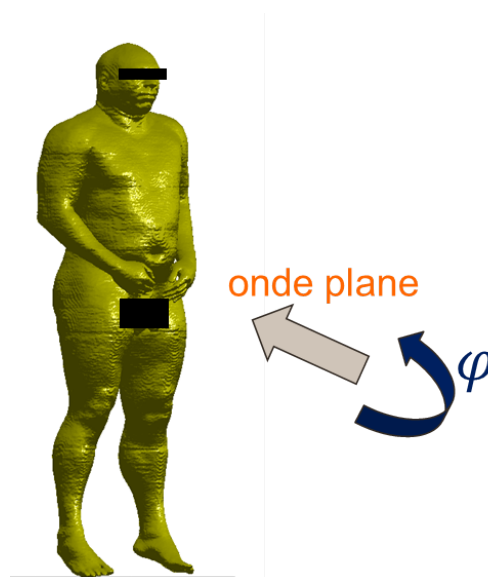


FIGURE 4.20 : Schéma du cas test d'application de l'incertitude morphologique.

indices totaux. Cela signifie qu'il y a de fortes interactions dans le phénomène étudié. Le calcul des indices d'ordre plus élevé montre que pour les indices de second ordre, seul l'indice d'interaction entre  $r$  et  $\phi$  est significatif et que la variance expliquée par les paramètres d'entrée  $z$  et  $\alpha$  est constituée principalement des indices d'ordre 3 et 4. Ces variables sont alors toujours en interaction avec  $r$  et/ou  $\phi$ .

#### 4.3.5 Application à l'étude de l'influence de l'incertitude morphologique.

L'utilisation des polynômes de développement du chaos polynomial dans la partie déterministe du modèle de Krigeage universel est maintenant testée sur un autre cas-test issu de la dosimétrie. Comme présenté dans la Section 3.2.3, on cherche à étudier l'influence de l'incertitude morphologique du corps humain sur la puissance absorbée par le corps. On considère donc le cas d'exposition présenté dans la Figure 4.20. Dans ce cas d'exposition le modèle anatomique étudié est Visible Human [Ackerman 1995] que l'on va déformer à l'aide des techniques de morphing présentée dans la section 3.2.3. Deux types de paramètres d'entrée sont ici retenus : d'un côté les paramètres morphologiques présentés dans la section 3.2.3 et de l'autre côté une incertitude sur la source qui est ici une onde plane.



Paramètre d'entrée	Loi	$\theta_1$	$\theta_2$
Stature	Lognormale	5.175	0.044
Tour de poitrine	Gumbel	96.51	7.919
Tour de ceinture	Gumbel	83.68	8.098
Carrure	Lognormale	3.736	0.094
Hauteur de l'entrejambe	Lognormale	4.370	0.066
$\phi$	Uniforme	0	360

TABLE 4.6 : Lois marginales des paramètres d'entrée.

L'angle d'azimut  $\psi$  de l'onde plane est un paramètre d'entrée muni d'une loi uniforme sur  $[0, 360^\circ]$ . La liste des paramètres d'entrée ainsi que les lois associées sont présentées dans la Table 4.6. Pour les 5 premiers paramètres, la structure de dépendance est celle présentée dans la section 3.2.3 et le paramètre d'azimut est défini comme indépendant des 5 autres paramètres morphologiques. La variable de sortie est ici le DAS corps entier du modèle anatomique. A l'aide de la transformée de Nataf, on se ramène à un espace standardisé pour pouvoir construire les méta-modèles de chaos polynomial et de Krigeage tels que présentés dans les exemples précédents. A partir d'un LHS initial de 100 points que l'on augmente jusqu'à 300 points pour la technique du NLHS, on construit les trois modèles de LARS-PC, de LARS-Kriging-PC et de Krigeage ordinaire. La précision des modèles construits est comme dans la section précédente évaluée à l'aide de l'erreur relative de la validation croisée de l'Eq. (4.49). La Figure 4.21 présente les résultats sous le même format que la Figure 4.19. Cette figure confirme à nouveau le caractère optimal du LARS-PC-Kriging. Pour un nombre de 100 points, on observe que la méthode d'hybridation obtient une meilleure précision que les deux autres méthodes, en particulier si l'on s'intéresse à la différence de précision avec le Krigeage ordinaire. Il apparaît par la suite qu'avec l'ajout de points, le modèle de LARS-PC-Kriging voit sa précision améliorée jusqu'à obtenir une erreur relative inférieure à 1% et que les autres méthodes suivent la même tendance en venant in fine approcher ses performances. Il est alors possible de visualiser la distribution statistique du DAS corps entier par la propagation de l'incertitude morphologique et de l'incertitude de l'orientation de la source par simulations de Monte Carlo à travers le méta-modèle de LARS-Kriging-PC le plus précis (300 points). Cette distribution est présentée en Figure 4.22. On observe une bimodalité sur cette distribution vraisemblablement due au comportement oscillant de la sortie en fonction de l'angle d'azimut. Le coefficient de variation de cette distribution est égal à  $CV = 0.245$ . A partir du modèle optimal obtenu avec 300 points, une analyse de sensibilité est effectuée qui prend en compte la dépendance entre les 5 premiers paramètres. On utilise pour cela les indices de Kucherenko présentés dans la Section 3.3.4. Les indices de premier ordre et les indices totaux pour chaque variable sont calculés en utilisant 1000000 simulations de Monte Carlo propagées à travers le méta-modèle. Les résultats sont présentés dans la Table 4.7. Ce tableau montre clairement que le paramètre le plus influent dans ce problème d'exposition est celui de l'orientation de la source. En effet, les indices de premier ordre et totaux de Kucherenko de l'angle d'azimut de l'onde plane (qui correspondent ici aux indices de Sobol' puisque ce paramètre est indépendant des autres) sont respectivement égaux à 0.71 et 0.74. On remarque au passage la faible différence entre les valeurs de ces indices, ce qui nous indique qu'il y a peu d'interaction entre l'azimut et les paramètres morphologiques. Il

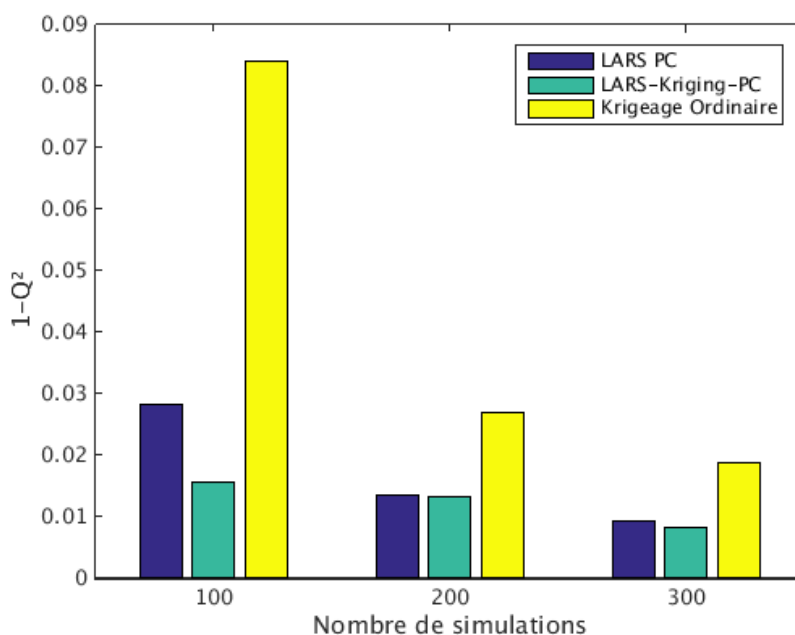


FIGURE 4.21 : Comparaison des métamodèles LARS PC, LARS-Kriging-PC et Krigeage ordinaire pour l'incertitude morphologique en utilisant l'erreur relative de la validation croisée *leave-one-out* en fonction du nombre de points.

Paramètre d'entrée	Indice de premier ordre	Indice total
Stature	0.034	0.006
Tour de poitrine	0.22	0.033
Tour de ceinture	0.23	0.046
Carrure	0.075	0.025
Hauteur de l'entrejambe	0.022	0.006
$\phi$	0.713	0.748

TABLE 4.7 : Indices de Kucherenko totaux et de premier ordre des paramètres d'entrée pour le problème de l'incertitude morphologique obtenus à partir de 1 000 000 simulations de Monte Carlo du BLUP du méta-modèle LARS-Kriging-PC pour 300 points.

pourra éventuellement être envisagé dans de futures études de cas similaires le traitement séparé de l'incertitude de la source et de l'incertitude morphologique à travers deux métamodèles de dimension réduite. Pour ce qui est des indices des paramètres morphologiques, on remarque tout d'abord que les indices totaux sont tous très faibles. Ce résultat n'est pas très surprenant étant donné la dépendance assez élevée entre certains paramètres d'entrée comme expliqué dans la Section 3.3.4. Il est donc intéressant de considérer les indices totaux faisant intervenir plusieurs paramètres à la fois pour avoir accès à l'influence de l'incertitude morphologique dans sa globalité. Ce qui s'avère pertinent dans ce cas est de considérer l'indice total de l'ensemble des 5 paramètres morphologiques qui est égal ici à 1 moins l'indice de premier ordre de l'angle d'azimut. On voit que l'on obtient dans ce cas un indice total de l'ensemble des paramètres morphologiques égal à 0.287. Si l'influence

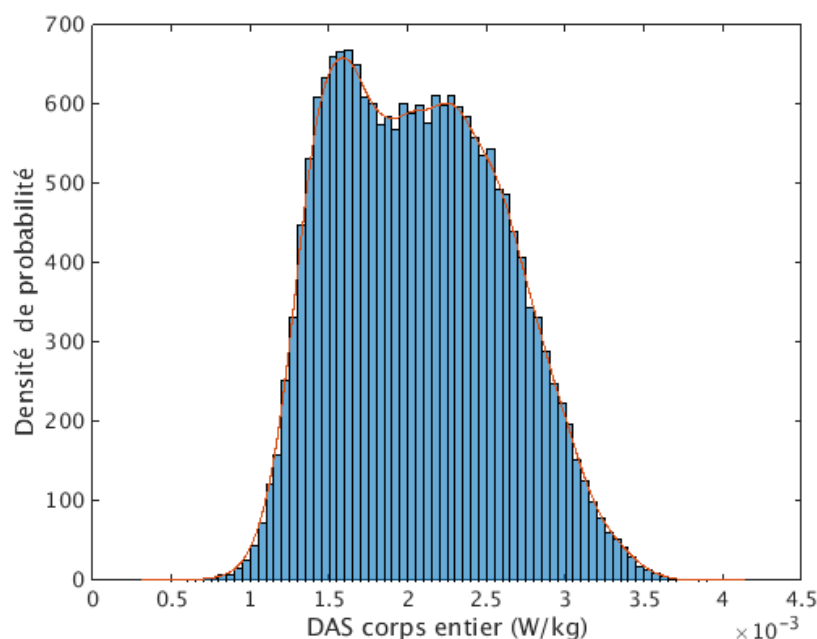


FIGURE 4.22 : Distribution du DAS corps entier obtenu par propagation de l’incertitude morphologique et de l’incertitude de la source à travers le méta-modèle de LARS-Kriging-PC construit à partir du LHS de 300 points.

des paramètres morphologiques est moindre que celle de l’azimut elle reste cependant non négligeable. En regardant les influences séparées des paramètres morphologiques à l’aide des indices du premier ordre, il apparaît que seuls le tour de poitrine et le tour de ceinture ont une influence significative sur le DAS corps entier du modèle anatomique. On rappelle que ces deux paramètres sont très dépendants ( $\rho_S = 0.8$ , cf Table 3.2), ils ont donc une grande partie de leur influence qui est commune à cause de cette corrélation. En conclusion, il ressort alors que la variation du DAS corps entier du modèle anatomique est en majeure partie due à l’angle d’azimut de l’onde plane source de l’exposition et de l’influence conjointe du tour de poitrine et du tour de ceinture. Ces résultats d’analyse de sensibilité sur les paramètres dépendants sont néanmoins à prendre avec une certaine retenue car comme expliqué dans la Section 3.2.3, la communauté scientifique propose plusieurs méthodes concernant l’analyse de sensibilité de paramètres dépendants qui donnent parfois des résultats contradictoires et le consensus sur les méthodes à employer n’a pas encore été trouvé.

## 4.4 Conclusion

Pour conclure, ce chapitre a présenté les constructions de méta-modèles pour la propagation de l’incertitude à travers l’utilisation de plusieurs méthodes connues ou nouvelles. Premièrement, l’approche de régression pour la construction de développement tronqués de chaos polynomial a été introduite et testée sur des cas-test qui ont pour certains fait l’objet d’une publication d’article [Kersaudy et al. 2014]. Ensuite, le Krigeage universel a été introduit pour pouvoir justifier d’une nouvelle méthode *hybride* qui utilise des poly-

nômes sélectionnés par le LARS pour des développements de chaos polynomial en tant que fonctions de régression dans la partie déterministe du modèle de Krigeage universel. Cette nouvelle méthode appelée *LARS-Kriging-PC*, qui a fait l'objet d'une publication d'article [Kersaudy et al. 2015], est testée à l'aide de plusieurs exemples analytiques et de cas-test concrets issus de la dosimétrie numérique. Cette méthode montre sur ces exemples son caractère optimal par rapport au chaos polynomial LARS classique et au Krigeage ordinaire.

Toutes ces méthodes se révèlent efficaces pour construire des méta-modèles avec une bonne précision globale et peuvent donc être utilisées pour une analyse globale des distributions de sortie. En revanche, si on veut conduire une analyse locale de ces distributions, par exemple à travers l'étude des quantiles de queue de distribution, la précision globale du modèle ne sera pas suffisante. En se restreignant au cas des quantiles, il faut trouver des moyens d'avoir accès à la précision de l'estimation de ces quantiles. Le Chapitre 5 s'attachera donc au développement de méthodes pour estimer la précision des quantiles calculés à partir de méta-modèles, en particulier de méta-modèles de chaos polynomial. De plus, pour satisfaire à l'exigence de parcimonie dans l'ajout de points au plan d'expériences, ce chapitre développera une méthodologie de planification orientée vers l'estimation efficace et parcimonieuse des quantiles à 95% des distributions de sortie.



# Chapitre 5

## Planification d'expérience adaptative pour l'estimation du quantile à 95% de la distribution de sortie

### Sommaire

---

<b>5.1</b>	<b>La méthode BOAS (<i>Bootstrapped Oriented Adaptive Sampling</i>)</b>	<b>96</b>
5.1.1	La théorie du bootstrap et les intervalles de confiance . . . . .	96
5.1.2	Stratégie adaptative de planification d'expériences . . . . .	99
5.1.3	Critères d'arrêt et estimateurs de précision . . . . .	103
5.1.4	Exemple analytique : la fonction d'Ishigami . . . . .	103
5.1.5	Exemple analytique : la fonction de Borehole . . . . .	106
5.1.6	Exemple analytique : la fonction de Sobol . . . . .	108
5.1.7	Exemple d'application : exposition d'un modèle d'enfant à une station de base de femtocell . . . . .	111
<b>5.2</b>	<b>Comparaison de performances avec la méthode GPS (<i>Gaussian Process Shrunk</i>)</b>	<b>112</b>
5.2.1	La méthode GPS . . . . .	113
5.2.2	Comparaison de performances avec la fonction d'Ishigami . . . . .	114
5.2.3	Comparaison de performances avec la fonction de Borehole . . . . .	115
<b>5.3</b>	<b>Conclusion</b>	<b>117</b>

---

Comme expliqué dans la Section 3.3.2, au delà de l'analyse de sensibilité globale, les quantiles de la distribution de sortie sont aussi une information qui nous intéresse. En particulier, on s'intéresse dans ces travaux de thèse au quantile à 95% de la distribution de sortie avec un contexte qui reste le même que précédemment : on veut une estimation du quantile à 95% dont on sait estimer la précision (que l'on veut la plus haute possible) et que l'on peut obtenir à partir d'un nombre de points limité. On rappelle l'expression du quantile à 95% en transcrivant l'expression de l'Eq. (3.33), le quantile  $q_{95}$  est ainsi la valeur de la variable de sortie telle que :

$$\mathbb{P}(Y \leq q_{95}) = F_Y(q_{95}) = 0.95 \quad (5.1)$$

Classiquement, pour estimer la valeur de ce quantile, on utilise son approximation par Monte Carlo que ce soit à l'aide du modèle physique ou bien à l'aide d'un méta-modèle. Soit un échantillon de Monte Carlo de la sortie  $Y$  de taille  $n$  :  $\{y_1, \dots, y_n\}$ . On rappelle l'approximation par Monte Carlo de la fonction de répartition de la variable aléatoire  $Y$  :

$$\hat{F}_Y(y) = \frac{1}{n} \sum_{i=1}^n \mathbf{1}_{y_i \leq y} \quad (5.2)$$

L'estimateur Monte Carlo du quantile à 95% de la variable de sortie est donc défini de la manière suivante :

$$\hat{q}_{95} = \inf \left\{ y, \hat{F}_Y(y) \geq 0.95 \right\} = y_{(\lceil 0.95n \rceil)} \quad (5.3)$$

Dans l'Eq. (5.3),  $[\cdot]$  représente toujours la fonction valeur entière supérieure et  $y_{(i)}$  qui est la  $i$ -ème statistique d'ordre. Théoriquement la précision de cette approximation du quantile est connue. En effet, on connaît le comportement limite de l'estimation quand  $n$  tend vers l'infini ; l'incertitude sur son estimation est sans biais et tend vers une loi normale dont on connaît la variance [Cannamela et al. 2008] :

$$\sqrt{n}(\hat{q}_{95} - q_{95}) \xrightarrow[n \rightarrow \infty]{} \mathcal{N}(0, \sigma_{\hat{q}_{95}}^2) \quad \text{avec} \quad \sigma_{\hat{q}_{95}}^2 = \frac{0.95(1 - 0.95)}{p_Y^2(q_{95})} \quad (5.4)$$

On voit avec cette expression qui est généralisable pour n'importe quel quantile que plus on cherche un quantile élevé, plus la variance d'estimation sera grande. Ainsi avec le quantile à 95%, il n'est donc pas possible d'utiliser les simulations de Monte Carlo de façon directe dans le cas du modèle physique de la FDTD pour obtenir une estimation de cette grandeur d'intérêt. On peut bien sûr comme précédemment utiliser un méta-modèle à travers lequel on pourra propager l'incertitude des paramètres d'entrée. Néanmoins, il faut que ce méta-modèle soit très précis dans la région du quantile si l'on veut pouvoir faire confiance à cette estimation. Dans la méthodologie développée dans le chapitre précédent, on voit que pour estimer la précision des méta-modèles, des estimateurs de précision généraux qui calculaient une précision moyenne sur tout le domaine d'entrée étaient utilisés. On n'a donc pas nécessairement l'assurance que les méta-modèles en question soient précis dans la région du quantile. Il nous faut donc trouver un estimateur de la précision du quantile ainsi qu'un moyen d'augmenter cette précision avec un nombre de points le plus réduit possible.

Dans la littérature, on rencontre plusieurs méthodes d'estimations de quantiles élevés d'une distribution de sortie. Tout d'abord, on peut citer les méthodes s'attachant à réduire

la variance des estimations faites à partir d'échantillons Monte Carlo. Il y a par exemple les méthodes utilisant des *variables de contrôle* ou *control variate* en anglais [Hesterberg and Nelson 1998, Nelson 1987]. Leur principe est de considérer une variable aléatoire additionnelle dont on connaît parfaitement la distribution ou du moins la partie d'intérêt et dont on sait qu'elle est dépendante statistiquement de celle dont on veut estimer le quantile. Il est alors possible d'utiliser cette variable comme variable de contrôle pour réduire la variance de l'estimation du quantile. Sur le même principe, on trouve la *stratification contrôlée* [Cannamela et al. 2008]. Le principe est ici d'estimer la fonction de répartition par partie à l'aide d'un modèle que l'on a à disposition et dont on se sert comme variable de contrôle. Avec cette estimation par partie, il est possible de générer un nombre de simulations Monte Carlo plus important dans la zone intéressante pour l'estimation du quantile et par là même de réduire la variance de l'estimation. Enfin, on peut aussi citer les méthodes d'*échantillonnage préférentiel* ou *importance sampling* [Glynn 1996] où l'idée est d'échantillonner la variable aléatoire dans la zone intéressante pour l'estimation du quantile à l'aide d'une variable additionnelle dite *instrumentale* qui permet cet échantillonnage préférentiel. Dans notre cas, il paraît compliqué d'utiliser de telles méthodes pour l'estimation de quantiles car elles nécessiteraient un temps de calcul trop conséquent.

Une alternative se situe dans l'utilisation de méta-modèles et dans la conception de méthodologies de planification d'expériences adaptatives orientée vers l'estimation de quantile. Il existe beaucoup de ces méthodes dans le cadre de l'utilisation de modèle de Krigeage. Ces méthodes développées à l'origine pour l'analyse de fiabilité ou l'optimisation globale peuvent être facilement adaptées pour l'estimation de quantile. On retrouve par exemple la méthode de l'*expected improvement* [Jones et al. 1998, Vazquez and Bect 2010] où les points sont ajoutés au plan d'expériences de manière à ce qu'ils permettent d'améliorer l'estimation du maximum. On trouve également la méthode SUR pour *Stepwise Uncertainty Reduction* [Bect et al. 2012] qui est une méthode orientée vers l'estimation d'une probabilité de défaillance et qui a été adaptée pour l'estimation de quantile dans [Arnaud et al. 2010, Jala 2013]. On peut aussi citer une méthode d'analyse de fiabilité développée dans [Picheny et al. 2010] qui utilise un critère d'erreur pondérée d'un modèle de Krigeage pour définir des zones d'exploration pour l'ajout de nouveaux points. Ces méthodes font partie des nombreuses initiatives pour le développement de stratégies adaptatives dans le domaine de l'analyse de fiabilité [Bichon et al. 2011, Echard et al. 2011, Dubourg et al. 2011]. Toujours dans le cadre de modèles de Krigeage, il existe des méthodes spécifiquement développées pour l'estimation de quantiles. Par exemple, [Oakley 2004], utilise la normalité des modèles de Krigeage conditionnelle aux plans d'expériences pour définir une zone d'exploration enrichie à l'aide de plans LHS tronqués. Enfin, on peut citer la méthode GPS pour *Gaussian Process Shrunk* [Jala 2013, Jala et al. 2014] développée au sein d'Orange Labs pour la résolution de problèmes issus de la dosimétrie. Cette méthode utilise des fonctions d'encadrement pour définir les nouvelles zones d'exploration. On développera plus en détail cette méthode dans la suite de ce chapitre. Pour résumer, toutes ces méthodes utilisent des modèles de Krigeage et peuvent bien entendu être utilisées avec les modèles de LARS-Kriging-PC présentés dans le chapitre précédent. En revanche, il n'existe pas à notre connaissance dans la littérature de telle méthode de planification orientée pour les modèles de chaos polynomial. Une approche basée sur des rééchantillonnages bootstrap a certes été développée dans Notin [2010] pour estimer l'incertitude sur les quantiles mais elle n'intégrait pas de méthode de planification spécifique pour cette



estimation ce qui est l'objectif principal de ce chapitre.

Dans ce chapitre on s'attache donc à développer une stratégie de planification adaptative pour estimer le quantile dont on estime la précision à l'aide de rééchantillonnages bootstrap du plan d'expériences. Cette méthode que l'on appelle le BOAS pour *Bootstrapped Oriented Adaptive Sampling* s'inspire de la méthode GPS et est à destination des utilisateurs de la théorie du chaos polynomial (mais est aussi applicable à toute technique de méta-modèle) pour laquelle il n'a pas été développé à notre connaissance de telle stratégie adaptative. Cette méthode du BOAS fait l'objet d'une soumission d'article dans l'*International Journal on Uncertainty Quantification*.

## 5.1 La méthode BOAS (*Bootstrapped Oriented Adaptive Sampling*)

### 5.1.1 La théorie du bootstrap et les intervalles de confiance

Une méthode couramment utilisée pour extraire un intervalle de confiance d'une estimation d'une quantité d'intérêt est la méthode du *bootstrap*. Cette méthode a originellement été introduite par Efron [Efron 1979, Efron and Gong 1983]. L'origine sémantique du nom de la méthode provient de la philosophie de son utilisation. L'idée est d'extraire un maximum d'information et de précision à partir d'un unique échantillon de faible taille dont on dispose. Considérons un échantillon d'observation  $\mathbf{z} = \{z_i, i \in \llbracket 1, N \rrbracket\}$  indépendant et identiquement distribué d'une variable aléatoire  $Z \in \mathbb{R}$  dont la fonction de répartition est notée  $F$ . Considérons également une quantité statistique d'intérêt  $T(F)$  de cette variable aléatoire. Le principe du bootstrap est de rééchantillonner  $\mathbf{z}$  avec remise pour obtenir une estimation de la distribution de l'incertitude associée à l'estimation  $\hat{T}$  de la quantité d'intérêt  $T(F)$ . Le rééchantillonnage de l'échantillon d'origine  $\mathbf{z}$  se fait suivant une loi multinomiale où chaque point d'observation est muni d'un poids  $1/N$ . Le résultat de ce rééchantillonnage est donc un nouvel échantillon  $\mathbf{z}^*$  de la même taille que  $\mathbf{z}$  et constitué des observations de  $\mathbf{z}$  avec remise, c'est-à-dire que des éléments de l'échantillon d'origine  $\mathbf{z}$  pourront être répétés plusieurs fois dans  $\mathbf{z}^*$ . Si on compte le nombre d'observations différentes dans le vecteur rééchantillonné, il sera donc inférieur ou égal à  $N$ . Dans [Gazut 2007], il est montré qu'en moyenne le nombre d'observations différentes au sein d'un vecteur rééchantillonné est de  $0.63N$ .

Considérons maintenant un ensemble de  $B$  rééchantillonnages bootstrap de  $\mathbf{z} : \{\mathbf{z}_k^*, k \in \llbracket 1, B \rrbracket\}$ , ainsi que les  $B$  estimations associées de  $T(F) : \{\hat{T}_k^* = \hat{T}(\mathbf{z}_k^*), k \in \llbracket 1, B \rrbracket\}$ . Des approximations de la moyenne et de la variance de l'estimateur  $\hat{T}$  sont obtenues par les expressions suivantes :

$$\hat{\mu}_{\hat{T}} = \frac{1}{B} \sum_{k=1}^B \hat{T}_k^* \quad (5.5)$$

$$\hat{\sigma}_{\hat{T}}^2 = \frac{1}{B-1} \sum_{k=1}^B (\hat{T}_k^* - \hat{\mu}_{\hat{T}})^2 \quad (5.6)$$

De manière générale, l'ensemble  $\{\hat{T}_k^* = \hat{T}(\mathbf{z}_k^*), k \in \llbracket 1, B \rrbracket\}$  représente une estimation empirique de la distribution de l'incertitude associée à l'estimateur  $\hat{T}$ . En conséquence des intervalles de confiance peuvent également être déduits de ces rééchantillonnages bootstrap

en prenant les quantiles empiriques de la distribution de l'incertitude de  $\hat{T}$ . En effet en extrayant les quantiles  $\alpha/2$  et  $(1-\alpha/2)$  de la distribution bootstrap, on obtient une estimation d'un intervalle de confiance de degré  $\alpha$  :

$$T(F) \in [\hat{T}_{([\alpha B/2])}^* ; \hat{T}_{([(1-\alpha/2)B])}^*] \quad (5.7)$$

Cette méthode d'obtention d'intervalles de confiance est appelée la méthode des *pourcentiles simple* et requiert environ  $B = 1000$  rééchantillonnages bootstrap pour avoir une bonne approximation des quantiles de la distribution bootstrap [Efron 1979]. D'autres méthodes existent pour définir des intervalles de confiance à partir de rééchantillonnages bootstrap [Efron and Tibshirani 1981]. On peut par exemple citer la méthode de *l'erreur standard* qui consiste à considérer la distribution bootstrap de l'incertitude de l'estimateur  $\hat{T}$  comme suivant une loi normale. Dans ce cas, il n'y a besoin que de calculer la variance et la moyenne de la distribution bootstrap à l'aide de l'Eq. (5.6) et de considérer comme bornes de l'intervalle de confiance les quantiles correspondants de la loi normale de moyenne  $\hat{\mu}_{\hat{T}}$  et de variance  $\hat{\sigma}_{\hat{T}}^2$ . Cette méthode a l'avantage de ne nécessiter que de  $B = 100$  rééchantillonnages [Efron 1979]. En revanche, les distributions bootstrap peuvent rarement être assimilées à des lois normales et en l'absence de ce type d'a priori pour nos cas d'application on n'utilisera pas cette méthode ici. Il existe également d'autres méthodes pour obtenir des intervalles de confiance : la méthode des *pourcentiles avec correction pour le biais* et celle des *pourcentiles avec correction pour le biais et accélération*. Ces méthodes permettent toutes deux de corriger le biais éventuel de la distribution bootstrap et la deuxième permet d'insérer un terme constant dans le calcul de l'intervalle qui compense un éventuel changement de la variance de l'estimateur  $\hat{T}$  quand  $T(F)$  varie. Ces méthodes fournissent des intervalles plus précis mais nécessitent plus de rééchantillonnages bootstrap que la méthode des pourcentiles simples. Le coût de ces rééchantillonnages supplémentaires devient alors significatif pour la méthode de planification adaptative d'expériences développée dans ce chapitre. Comme compromis entre la précision et le temps de calcul, le choix est fait d'utiliser la méthode des pourcentiles simples.

Dans le contexte de ces travaux de thèse, c'est l'application de la théorie du bootstrap pour l'estimation des résidus de modèles de régression [Efron 1979, Freedman 1981] qui nous intéresse, ceci afin d'obtenir un intervalle de confiance sur les variations des méta-modèles construits. Plus précisément, on veut ici l'appliquer aux modèles de chaos polynomial présentés dans le Chapitre 4. La méthodologie employée s'inspire des principes de l'utilisation du bootstrap avec des méthodes d'apprentissage statistique. On peut citer le *bagging* pour *Bootstrap Aggregating* [Breiman 1996] ou encore la méthode LDR pour *Learner Disagreement by experiment Resampling*. Dans le cas particulier des méta-modèle de chaos polynomial par régression, considérons une troncature fixée dont les multi-indices des polynômes forment l'ensemble  $\mathbb{A}$  selon les notations utilisées dans la section 4.1.2. La variation locale de la prédiction des méta-modèles de régression issus du chaos polynomial sera alors donnée par l'incertitude faite sur l'estimation des coefficients  $\beta = \{\beta_{\alpha}, \alpha \in \mathbb{A}\}$  du modèle de régression sous-jacent que l'on note ici  $f_{reg}$  :

$$f_{reg} = \sum_{\alpha \in \mathbb{A}} \beta_{\alpha} \psi_{\alpha}(\mathbf{X}) \quad (5.8)$$

Afin d'estimer la variation des coefficient du méta-modèle  $f_{reg}$ , le principe est d'utiliser les rééchantillonnages bootstrap comme autant de plans d'expériences à partir desquels

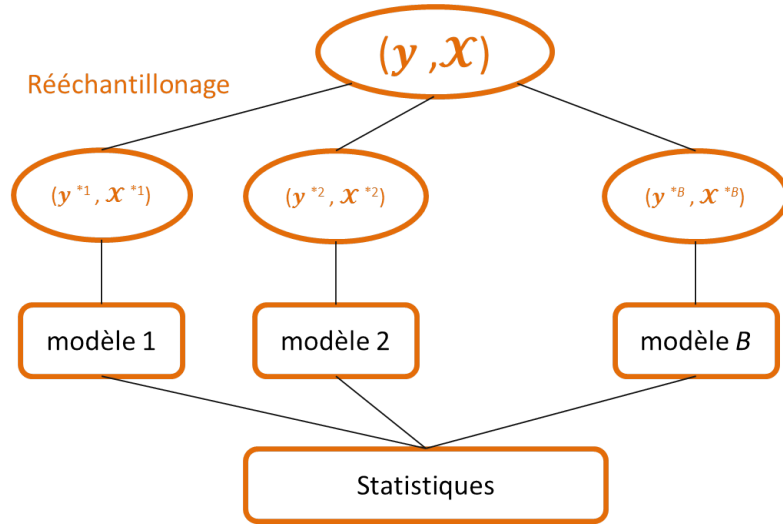


FIGURE 5.1 : Schéma de principe de l'utilisation du bootstrap pour un modèle de régression.

le méta-modèle est construit. Pour un  $\mathbf{x} \in \mathbb{X}$  donné, on a alors autant de prédictions de  $y$  possibles que de rééchantillonnages. L'estimation de l'incertitude sur les coefficients nous fournit donc une estimation de l'incertitude sur la prédiction locale du méta-modèle. L'utilisation du bootstrap est donc une façon de générer cette incertitude locale qui est intrinsèque dans le cas d'une modélisation par processus Gaussiens. Un schéma explicatif du principe de l'utilisation du bootstrap en régression est donné en Figure 5.1. Pour entrer plus dans le détail, on considère le plan d'expériences muni de l'ensemble de points  $\mathcal{X}^N = \{\mathbf{x}^{(i)}, i \in \llbracket 1, N \rrbracket\}$  ainsi que  $B$  rééchantillonnages bootstrap  $\{\mathcal{X}_k^{*N} = \{\mathbf{x}_k^{*(1)}, \dots, \mathbf{x}_k^{*(N)}\}, k \in \llbracket 1, B \rrbracket\}$  de cet ensemble. Le calcul des coefficients bootstrap des différents modèles de chaos polynomial s'effectue toujours par moindres carrés ordinaires :

$$\hat{\beta}_k^* = (\Psi_k^{*T} \Psi_k^*)^{-1} \Psi_k^{*T} \mathbf{y}_k^* \quad k \in \llbracket 1, B \rrbracket \quad (5.9)$$

Ici, en reprenant les notation de la Section 4.1,  $\hat{\beta}_k^* = (\hat{\beta}_{k_{\alpha(0)}}^*, \dots, \hat{\beta}_{k_{\alpha(P-1)}}^*)^T$  est le vecteur des coefficients scalaires du  $k$ -ième modèle bootstrap,  $\Psi = (\psi_{\alpha(j)}(\mathbf{x}_k^{*(i)}))_{1 \leq i \leq N, 0 \leq j \leq P-1}$  est la matrice des polynômes calculés au points du plan d'expériences rééchantillonné et  $\mathbf{y}_k^* = (f(\mathbf{x}_k^{*(1)}), \dots, f(\mathbf{x}_k^{*(N)}))^T$ . On remarque que dans notre démarche, les polynômes de la matrice  $\Psi$  sont ceux sélectionnés par le LARS à partir du plan d'expériences d'origine  $\mathcal{X}^N$  et que ceux-ci ne diffèrent pas avec les différents échantillons bootstrap. Ainsi, on estime l'incertitude liée à l'estimation des coefficients associés à ces polynômes [Shao 1996, Gazut 2007]. On peut effectués les analyses statistiques de cette incertitude d'estimation des coefficient à partir des Eq. (5.5), Eq. (5.6) et Eq. (5.7). Mais la propagation de l'incertitude de l'estimation du coefficient à travers le méta-modèle de régression peut également être utilisée pour fournir une incertitude locale sur la prédiction de  $f_{reg}(\mathbf{x})$ . Si on considère une réalisation  $\mathbf{x} \in \mathbb{X}$  du vecteur aléatoire d'entrée  $X$  et  $(\hat{f}_1^*(\mathbf{x}), \dots, \hat{f}_B^*(\mathbf{x}))^T$  le vecteur des  $B$  estimations de la sortie par les différents méta-modèles bootstrap. Ce vecteur est donc un échantillonnage empirique de la distribution bootstrap de l'estimation locale  $\hat{f}(\mathbf{x})$  ( $\hat{f}$  étant le méta-modèle de chaos polynomial généré à partir du plan d'expériences  $\mathcal{X}^N$ ). De la même manière que précédemment, il est alors possible de définir un intervalle de

confiance approché par la méthode des pourcentiles simples en prenant les estimateurs empiriques des quantiles de la distribution bootstrap de la prédiction de  $f_{reg}(\mathbf{x})$  :

$$f_{reg}(\mathbf{x}) \in [\hat{f}_{(\lceil \alpha B/2 \rceil)}^*(\mathbf{x}) ; \hat{f}_{(\lceil (1-\alpha/2)B \rceil)}^*(\mathbf{x})] \quad (5.10)$$

Cet intervalle de confiance est la base de la stratégie adaptative d'enrichissement du plan d'expériences orientée vers l'estimation efficace du quantile à 95%.

### 5.1.2 Stratégie adaptative de planification d'expériences

Le principe de la stratégie adaptative présentée dans cette section est d'encadrer le méta-modèle exact sous-jacent  $f_{reg}$  censé approcher le modèle physique  $f$  avec deux fonctions d'encadrement. A partir de ces fonctions d'encadrement, on veut tirer des conclusions sur l'incertitude liée à l'estimation du quantile à 95% de la distribution de sortie à travers une démarche heuristique en réduisant l'intervalle entre ces deux fonctions dans un sous-ensemble de l'espace de sortie que l'on suppose contenir la valeur du quantile à 95% estimé de la distribution de sortie qui sera défini plus tard. L'approche adoptée pour réduire cet intervalle est l'ajout de points dans le sous-ensemble de l'espace d'entrée dont le sous ensemble de sortie est l'image par  $f$ . On définit tout d'abord les deux fonctions d'encadrement qui à  $\mathbf{x}$  associent les bornes respectives de l'intervalle de l'Eq. (5.10) :

$$\begin{aligned} L_\alpha : \mathbb{X} &\rightarrow \mathbb{R} & U_\alpha : \mathbb{X} &\rightarrow \mathbb{R} \\ \mathbf{x} &\mapsto \hat{f}_{(\lceil \alpha B/2 \rceil)}^*(\mathbf{x}) & \mathbf{x} &\mapsto \hat{f}_{(\lceil (1-\alpha/2)B \rceil)}^*(\mathbf{x}) \end{aligned} \quad (5.11)$$

On précise bien ici que  $L_\alpha$  et  $U_\alpha$  ne sont pas égales à un méta-modèle bootstrap en particulier. En effet suivant la valeur de  $\mathbf{x}$ , ce ne sera pas forcément les mêmes méta-modèles bootstrap qui interviendront dans la définition de l'intervalle de confiance de l'Eq. (5.10).

De l'Eq. (5.11), il est clair que l'on peut déduire l'intervalle de confiance approché sur  $f_{reg}(\mathbf{x})$  suivant qui sera valable pour tout  $\mathbf{x} \in \mathbb{X}$  :

$$f_{reg}(\mathbf{x}) \in [L_\alpha(\mathbf{x}) ; U_\alpha(\mathbf{x})] \quad \forall \mathbf{x} \in \mathbb{X} \quad (5.12)$$

On définit  $q_{95L_\alpha}$  et  $q_{95U_\alpha}$  comme les quantiles à 95% respectifs des distributions des variables aléatoires  $L_\alpha(\mathbf{X})$  et  $U_\alpha(\mathbf{X})$  :

$$\mathbb{P}(L_\alpha(\mathbf{X}) \leq q_{95L_\alpha}) = 0.95 \quad (5.13)$$

$$\mathbb{P}(U_\alpha(\mathbf{X}) \leq q_{95U_\alpha}) = 0.95 \quad (5.14)$$

Dans une approche heuristique, en prenant  $\alpha = 0.05$  afin d'avoir des intervalles de confiance approchés à 95% pour l'Eq. (5.12), ces quantiles se veulent être des indicateurs représentatifs de l'incertitude faite sur l'estimation du quantile à 95% de la sortie. La pertinence de ces indicateurs sera analysée par la suite sur des exemples analytiques. Dans de futur travaux et afin d'aller plus loin que le caractère simplement heuristique de cette approche, une analyse mathématique de ces quantiles serait intéressante pour éventuellement pouvoir les utiliser en tant qu'intervalles de confiance avec un degré de confiance à définir. L'objectif est ici de réduire l'écart entre les deux quantiles définis en Eq. (5.13) et Eq. (5.14) afin de réduire l'incertitude sur l'estimation du quantile. Pour réduire cette incertitude, des points sont ajoutés au plan d'expériences dans la région de ce

quantile pour réduire l'incertitude entre les deux fonctions d'encadrement  $L_\alpha$  et  $U_\alpha$  dans cette région. Ainsi on définit la région  $\mathcal{D}^{N+1}$  de sélection du point additionnel du nouveau plan augmenté  $\mathcal{X}^{N+1}$  :

$$\mathcal{D}^{N+1} = \{\mathbf{x} \in \mathbb{X}, U_\alpha(\mathbf{x}) > q_{95_{L_\alpha}}\} \quad (5.15)$$

Pour justifier de l'intérêt de l'ajout de points dans cette zone  $\mathcal{D}^{N+1}$  on démontre la proposition suivante :

**Proposition 5.1.** Soit les fonctions  $L_\alpha, U_\alpha$  définies précédemment dans l'Eq. (5.11) et  $q_{95_{L_\alpha}}$  et  $q_{95_{U_\alpha}}$  les quantiles à 95% associés. Alors on a l'inégalité suivante :

$$q_{95_{U_\alpha}} - q_{95_{L_\alpha}} \leq \sup_{\mathbf{x} \in \mathcal{D}^{N+1}} (U_\alpha(\mathbf{x}) - L_\alpha(\mathbf{x})) \quad (5.16)$$

*Démonstration.* Considérons l'ensemble  $\mathcal{C}_1 = \{\mathbf{x} \in \mathbb{X}, L_\alpha(\mathbf{x}) > q_{95_{L_\alpha}}\}$ . On rappelle qu'avec la formulation du quantile de l'Eq. (5.1), on a :

$$\mathbb{P}(\mathbf{X} \in \mathcal{C}_1) = 0,05$$

Soit  $\mathbf{x} \in \mathcal{C}_1$ , on sait par définition des fonctions d'encadrement de l'Eq. (5.11) que  $U_\alpha(\mathbf{x}) \geq L_\alpha(\mathbf{x}) > q_{95_{L_\alpha}}$ . On peut en déduire que  $\mathbf{x} \in \mathcal{D}^{N+1}$  et ceci étant vrai pour tout  $\mathbf{x} \in \mathcal{C}_1$ , il vient :

$$\mathcal{C}_1 = \{\mathbf{x} \in \mathcal{D}^{N+1}, L_\alpha(\mathbf{x}) > q_{95_{L_\alpha}}\}$$

En réécrivant  $L_\alpha$  sous la forme  $L_\alpha = U_\alpha - (U_\alpha - L_\alpha)$ , on a :

$$\mathcal{C}_1 = \{\mathbf{x} \in \mathcal{D}^{N+1}, U_\alpha(\mathbf{x}) > q_{95_{L_\alpha}} + U_\alpha(\mathbf{x}) - L_\alpha(\mathbf{x})\}$$

Or, par définition, pour tout  $\mathbf{x} \in \mathcal{D}^{N+1}$ , il est clair que :

$$q_{95_{L_\alpha}} + \sup_{\mathbf{x} \in \mathcal{D}^{N+1}} (U_\alpha(\mathbf{x}) - L_\alpha(\mathbf{x})) \geq q_{95_{L_\alpha}} + U_\alpha(\mathbf{x}) - L_\alpha(\mathbf{x})$$

En définissant  $\mathcal{C}_2 = \{\mathbf{x} \in \mathcal{D}^{N+1}, U_\alpha(\mathbf{x}) > q_{95_{L_\alpha}} + \sup_{\mathbf{x} \in \mathcal{D}^{N+1}} (U_\alpha(\mathbf{x}) - L_\alpha(\mathbf{x}))\}$ , en prenant  $\mathbf{y} \in \mathcal{C}_2$ , on a :

$$U_\alpha(\mathbf{y}) > q_{95_{L_\alpha}} + \sup_{\mathbf{x} \in \mathcal{D}^{N+1}} (U_\alpha(\mathbf{x}) - L_\alpha(\mathbf{x})) \geq q_{95_{L_\alpha}} + U_\alpha(\mathbf{y}) - L_\alpha(\mathbf{y})$$

On en déduit que  $\mathbf{y} \in \mathcal{C}_1$  et ceci étant vrai pour tout  $\mathbf{y} \in \mathcal{C}_2$ , il vient :

$$\mathcal{C}_2 \subseteq \mathcal{C}_1 \quad (5.17)$$

Définissons maintenant  $\mathcal{C}_3 = \{\mathbf{x} \in \mathbb{X}, U_\alpha(\mathbf{x}) > q_{95_{L_\alpha}} + \sup_{\mathbf{x} \in \mathcal{D}^{N+1}} (U_\alpha(\mathbf{x}) - L_\alpha(\mathbf{x}))\}$ . D'une part, il est assez évident que  $\mathcal{C}_2 \subseteq \mathcal{C}_3$ . D'autre part, en prenant  $\mathbf{z} \in \mathcal{C}_3$ , il vient :

$$U_\alpha(\mathbf{z}) - \sup_{\mathbf{x} \in \mathcal{D}^{N+1}} (U_\alpha(\mathbf{x}) - L_\alpha(\mathbf{x})) > q_{95_{L_\alpha}}$$

Comme par définition,  $\sup_{\mathbf{x} \in \mathcal{D}^{N+1}} (U_\alpha(\mathbf{x}) - L_\alpha(\mathbf{x})) \geq 0$  (puisque  $U_\alpha \geq L_\alpha$ ), on peut en déduire que  $U_\alpha(\mathbf{z}) > q_{95_{L_\alpha}}$  et donc que  $\mathbf{z} \in \mathcal{D}^{N+1}$ . Ceci étant vrai pour tout  $\mathbf{z} \in \mathcal{C}_3$ , on en déduit que  $\mathcal{C}_3 \subseteq \mathcal{C}_2$  et par là même :

$$\mathcal{C}_3 = \mathcal{C}_2 \quad (5.18)$$

Par combinaison des Eq. (5.17) et (5.18), il est possible d'écrire que  $\mathcal{C}_3 \subseteq \mathcal{C}_1$  et donc, d'après la formulation du quantile, il vient :

$$\mathbb{P}(\mathbf{X} \in \mathcal{C}_3) \leq \mathbb{P}(\mathbf{X} \in \mathcal{C}_1) = 0.05$$

Toujours d'après la formulation du quantile, on en déduit l'inégalité suivante :

$$q_{95_{U_\alpha}} \leq q_{95_{L_\alpha}} + \sup_{\mathbf{x} \in \mathcal{D}^{N+1}} (U_\alpha(\mathbf{x}) - L_\alpha(\mathbf{x})) \quad (5.19)$$

Ce qui conclut la Proposition 5.2.1. □

Ainsi avec la Proposition 5.2.1, il apparaît qu'en réduisant l'écart maximal entre  $U_\alpha$  et  $L_\alpha$  sur la zone  $\mathcal{D}^{N+1}$ , il y a également réduction de l'écart entre les quantile à 95% des fonctions d'encadrement. L'expérience nous montre qu'avec les rééchantillonnages bootstrap qui ont permis d'obtenir  $U_\alpha$  et  $L_\alpha$ , ajouter un point à l'endroit de l'écart maximal entre ces deux fonctions ne contribue pas forcément à réduire cet écart. En revanche, on observe qu'ajouter un point selon une méthode classique de *space-filling* restreinte à la région d'intérêt de l'espace d'entrée  $\mathcal{D}^{N+1}$  contribue au fur et à mesure des itérations à réduire cet écart maximal. C'est la solution envisagée ici. Le choix est donc fait d'ajouter le nouveau point d'expérience  $\mathbf{x}^{(N+1)}$  suivant un critère maxmin restreint à la région  $\mathcal{D}^{N+1}$  :

$$\mathbf{x}^{(N+1)} = \max_{\mathbf{x} \in \mathcal{D}^{N+1}} \min_{\mathbf{x}^{(i)} \in \mathcal{X}^N} \|\mathbf{x} - \mathbf{x}^{(i)}\| \quad (5.20)$$

Les points sont ainsi ajoutés de façon itérative en recalculant les modèles à chaque fois jusqu'à ce que l'incertitude sur le quantile soit réduite en dessous d'un seuil fixé. On présente un exemple en 1D pour illustrer la progression de la méthode itération après itération. Cet exemple s'attache à l'estimation du quantile de la fonction  $f : x \mapsto x * \sin(x)$  pour  $x \in [0, 10]$  et suivant une loi uniforme sur son domaine de définition. La Figure 5.2 présente les résultats pour cet exemple 1D avec quelques itérations représentatives. Sur cette figure, il y a un graphe pour chaque itération. Chaque courbe en pointillés représente soit l'évolution de la sortie du modèle physique, soit l'une des deux fonctions d'encadrement. Chacune de ces fonctions est associée à une ligne horizontale de la même couleur dont la hauteur indique la valeur de son quantile à 95%. On a par conséquent en bleu nuit  $f$  (tirets) et  $q_{95}$  (trait plein), en bleu clair  $U_\alpha$  (tirets) et  $\hat{q}_{95_{U_\alpha}}$  (trait plein) et en noir  $L_\alpha$  (tirets) et  $\hat{q}_{95_{L_\alpha}}$  (trait plein). En trait plein rouge figure le modèle de chaos polynomial  $\hat{f}$  obtenu à partir des points de  $\mathcal{X}^N$  qui eux figurent en points rouges sur la courbe bleue pointillée de  $f$ . Sur l'axe des abscisses figurent en vert les points d'échantillonnages de  $\mathcal{D}^{N+1}$  sur lesquels est appliqué le critère maxmin pour choisir le nouveau point de simulation  $\mathbf{x}^{(N+1)}$ . Ce dernier est également indiqué en vert sur la fonction  $f$ . Cette figure montre d'abord que les quantiles  $\hat{q}_{95_{U_\alpha}}$  et  $\hat{q}_{95_{L_\alpha}}$  encadrent la valeur du quantile et représentent bien l'incertitude d'estimation faite sur le quantile. On observe ensuite que l'ajout de points dans la région du quantile contribue bien à réduire l'écart entre les deux fonctions d'encadrement dans cette région et par là même, la largeur de l'intervalle de confiance sur le quantile à 95%; à la dixième itération on a bien convergé vers la valeur  $q_{95}$ . Il apparaît également que si le méta-modèle de chaos polynomial est trop imprécis en certains endroits et que les rééchantillonnages bootstrap génèrent des régions  $\mathcal{D}^{N+1}$  qui ne correspondent pas à la région du quantile (exemple sur les itérations 2, 3 et 5 dans la zone gauche des courbes), l'ajout de points au "mauvais endroit" ne fera que préciser le modèle dans cette région pour définir avec plus de précision  $\mathcal{D}^{N+1}$  à l'itération suivante. En conclusion, on a une validation de principe de la démarche avec cet exemple 1D.



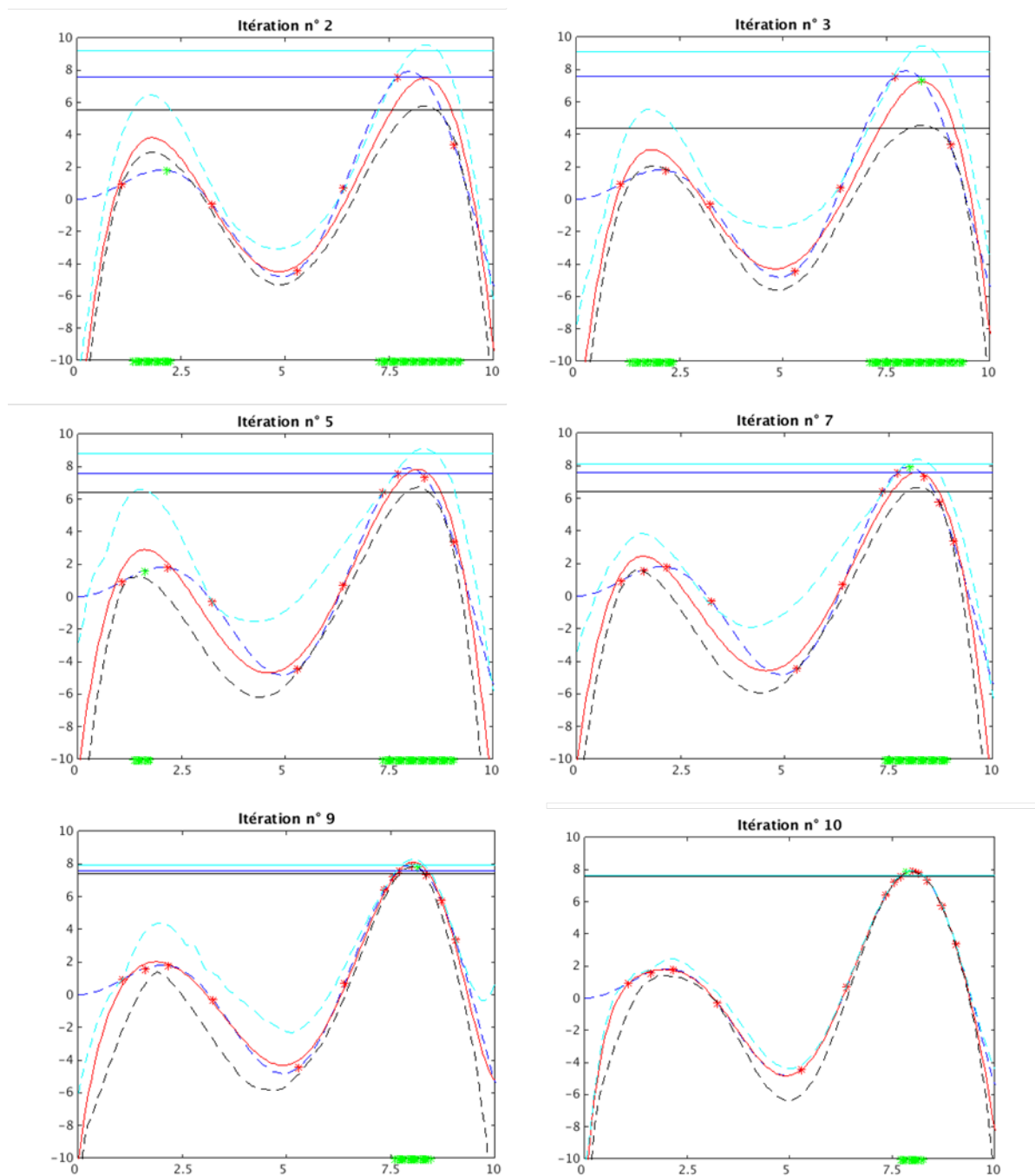


FIGURE 5.2 : Illustration pas à pas de l'application de la méthode BOAS sur la fonction test  $f : x \mapsto x \sin(x)$ . En bleu nuit  $f$  (tirets) et  $q_{95}$  (trait plein), en rouge  $\hat{f}$  (trait plein) et les points de  $\mathcal{X}^N$ , en bleu clair  $U_\alpha$  (tirets) et  $\hat{q}_{95U_\alpha}$  (trait plein), en noir  $L_\alpha$  (tirets) et  $\hat{q}_{95L_\alpha}$  (trait plein) et en vert  $\mathbf{x}^{(N+1)}$  (point sur  $f$ ) et les points d'échantillonnage Monte Carlo de  $\mathcal{D}^{N+1}$  (axe des abscisses)

### 5.1.3 Critères d'arrêt et estimateurs de précision

Si l'on est parvenu, à l'aide des rééchantillonnages, à obtenir un indicateur de précision sur l'estimation du quantile à 95% de la distribution de sortie modélisée par un méta-modèle de chaos polynomial, il faut maintenant définir des critères d'arrêt qui utiliseront des estimateurs de la précision de l'estimation du quantile. Un critère intuitif pour juger de la précision de l'estimation de quantile à 95% pourrait être l'erreur relative faite sur cette estimation :

$$\epsilon_r = \frac{|\hat{q}_{95} - q_{95}|}{|q_{95}|} \quad (5.21)$$

Néanmoins, dans notre démarche applicative, la valeur réelle de  $q_{95}$  n'est bien sûr pas accessible. Il est donc nécessaire d'avoir des critères de précision accessibles à partir des quelques expériences disponibles et du modèle construit. Dans cette section deux critères sont définis et sont voués à être utilisés comme critères d'arrêt dans les problèmes d'application. Pour le premier critère, une démarche naturelle est d'utiliser les quantiles des fonctions d'encadrement définis dans les Eq. (5.13) et (5.14). Considérons donc l'écart relatif entre ces deux quantiles. Ce critère est appelé *l'incertitude relative* sur l'estimation du quantile à 95%.

$$U_r = \frac{|\hat{q}_{95_{U_\alpha}} - \hat{q}_{95_{L_\alpha}}|}{|\hat{q}_{95_f}|} \quad (5.22)$$

Le second critère est une version pondérée du coefficient de détermination  $Q^2$  issu de la validation croisée leave-one-out (cf Eq. (4.19)). L'idée est ici d'effectuer une validation croisée uniquement sur les points d'expérience situés dans la zone d'exploration  $\mathcal{D}^{N+1}$ . Ainsi l'expression de l'Eq. (4.13) est pondérée par la fonction suivante :

$$\begin{aligned} w_N &: \mathbb{X} \rightarrow \{0, 1\} \\ \mathbf{x} &\mapsto \mathbf{1}_{\mathcal{D}^{N+1}}(\mathbf{x}) \end{aligned} \quad (5.23)$$

Dans cette équation,  $\mathbf{1}_{\mathcal{D}}$  est la fonction indicatrice de  $\mathcal{D}$ . On exprime ainsi le nouveau *coefficient de détermination* pondéré comme suit :

$$Q_w^2 = 1 - \frac{Err_{LOO_w}}{\hat{\sigma}_y^2} \quad (5.24)$$

Où

$$Err_{LOO_w} = \frac{1}{N} \sum_{i=1}^N w_N(\mathbf{x}^{(i)}) \left( f(\mathbf{x}^{(i)}) - \hat{f}^{(-i)}(\mathbf{x}^{(i)}) \right)^2 \quad (5.25)$$

Ce coefficient de détermination leave-one-out pondéré vise à estimer la précision du méta-modèle dans la zone d'exploration qui est la région d'intérêt pour l'estimation du quantile. Il fournit donc indirectement une indication sur la précision de l'estimation du quantile elle-même. Au delà de la simple estimation de précision sur le quantile à 95%, on utilise ces deux critères comme critère d'arrêt dans l'ajout de points dans les problèmes applicatifs que l'on présentera par la suite.

### 5.1.4 Exemple analytique : la fonction d'Ishigami

On veut maintenant valider l'approche BOAS avec des exemples analytiques de la même manière que dans la Section 4.3.1. La méthode BOAS est validée avec des méta-modèles de



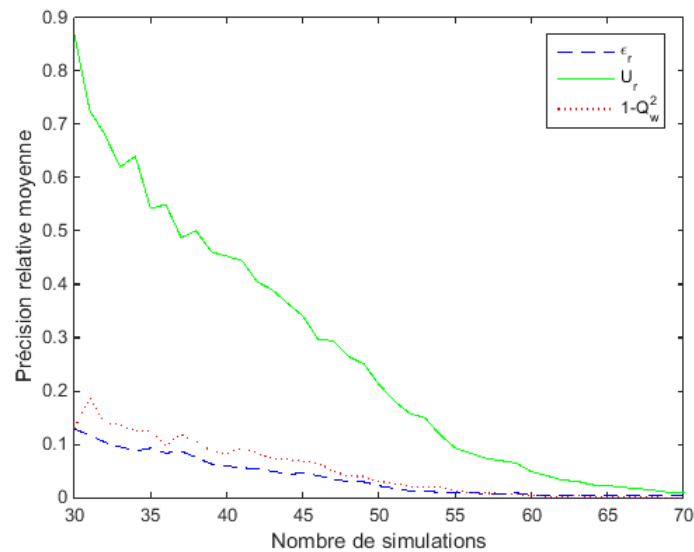


FIGURE 5.3 : Valeurs moyennes de  $\epsilon_r$ ,  $U_r$  et  $1 - Q_w^2$  en fonction du nombre de simulations sur les 50 répétitions du LHS initial pour la fonction d'Ishigami.

chaos polynomial LARS. Premièrement, la fonction d'Ishigami que l'on a présentée dans l'Eq. (4.26), est utilisée. L'approche BOAS est appliquée sur cette fonction à partir d'un LHS initial de 40 points que l'on augmentera jusqu'à 70 points. Pour commencer on veut observer la convergence de l'estimation du quantile à 95% à l'aide de la méthode BOAS avec l'erreur relative de l'Eq. (5.21). Ici la valeur de référence de  $q_{95}$  est estimée à l'aide de 1000000 simulations de Monte Carlo. Pour illustrer leurs performances, sont affichés sur le même graphe que l'erreur relative les estimateurs de précision des Eq. (5.22) et Eq. (5.24). La procédure complète est répétée 50 fois avec 50 LHS initiaux différents dans l'objectif d'estimer l'incertitude des performances des différentes quantités affichées due au caractère aléatoire du LHS initial. Ainsi, la Figure 5.3 montre l'évolution de l'erreur relative  $\epsilon_r$  et des estimateurs de précision  $U_r$  et  $1 - Q_w^2$  moyennés sur les 50 répétitions. Il apparaît tout d'abord sur cette figure que l'erreur relative de l'estimation du quantile converge bien vers 0 avec l'ajout de points. Comme pour l'exemple 1D, le principe de l'approche BOAS est ici validé. Ensuite, on remarque que les estimateurs de précision  $U_r$  et  $1 - Q_w^2$  convergent également vers 0. Il est également remarquable que l'incertitude relative  $U_r$  surestime la valeur réelle de l'erreur relative ce qui est une caractéristique importante dans une démarche applicative. En ce qui concerne l'erreur relative leave-one-out pondérée  $1 - Q_w^2$ , on peut difficilement comparer sa valeur avec celle de l'erreur relative faite sur le quantile. En effet, elle représente la précision globale dans la région d'intérêt pour l'enrichissement du plan d'expériences et pas exactement l'erreur faite sur le quantile. Néanmoins c'est un indicateur supplémentaire qui peut s'avérer utile dans une démarche applicative pour arrêter l'enrichissement du plan d'expérience si on considère par exemple que pour un cas d'application donné, l'incertitude relative a un temps de convergence trop long.

Il est maintenant intéressant de vérifier si la vitesse de convergence de la méthode BOAS est effectivement plus rapide qu'une méthode classique d'enrichissement global comme de NLHS présenté dans la Section 4.1.7. Ainsi on effectue de même 50 répétitions d'un enrichissement NLHS à par d'un plan LHS initial de 40 points. On utilise la même méthode

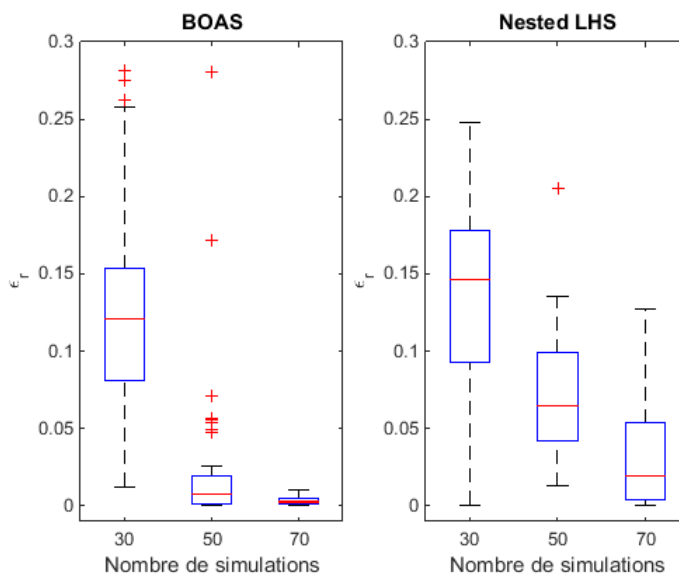


FIGURE 5.4 : Comparaison des performances des méthodes BOAS et NLHS vis-à-vis de  $\epsilon_r$  avec 50 plans LHS initiaux différents pour la fonction d'Ishigami.

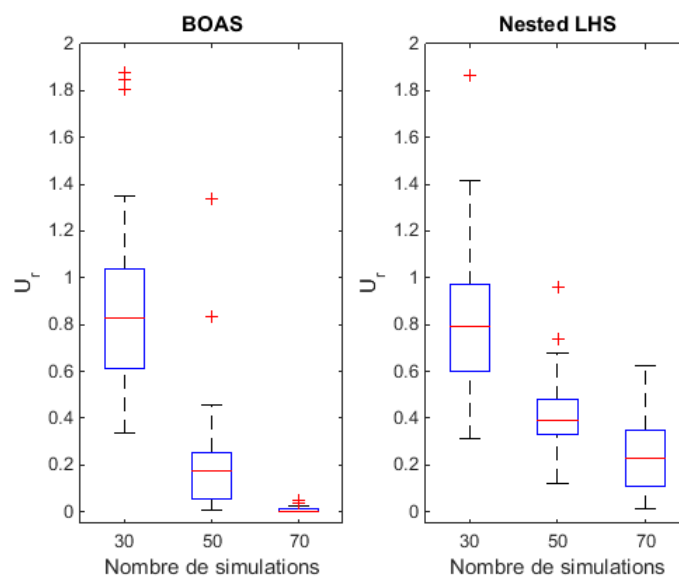


FIGURE 5.5 : Comparaison des performances des méthodes BOAS et NLHS vis-à-vis de  $U_r$  avec 50 plans LHS initiaux différents pour la fonction d'Ishigami.

de méta-modélisation de chaos polynomial LARS que pour le BOAS et l'estimation du quantile est faite avec les mêmes méthodes de rééchantillonnages bootstrap. Sur la Figure 5.4 sont présentés les résultats avec des boxplots représentant la valeur médiane, les quartiles et les valeurs extrêmes de l'erreur relative  $\epsilon_r$  des 50 répétitions indépendantes. Les croix rouges représentent des valeurs que l'algorithme de représentation des boxplots juge aberrantes. Il apparaît sur cette figure que le BOAS obtient de meilleures performances en terme de vitesse de convergence que le NLHS. Pour la médiane (et même pour les écart

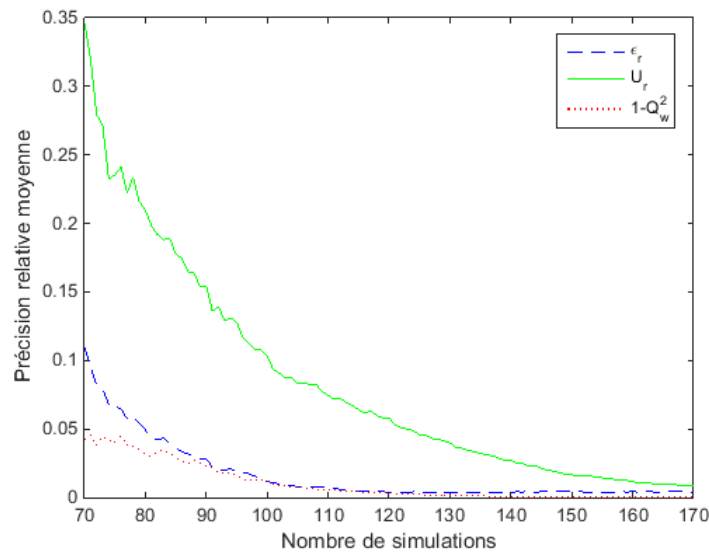


FIGURE 5.6 : Valeurs moyennes de  $\epsilon_r$ ,  $U_r$  et  $1 - Q_w^2$  en fonction du nombre de simulations sur les 50 répétitions du LHS initial pour la fonction de Borehole.

entre valeurs extrêmes et entre quartiles) les deux méthodes d'enrichissement partent évidemment de performances similaires au nombre de points initial et ensuite le BOAS réduit l'erreur relative faite sur le quantile significativement plus rapidement que le NLHS ainsi que l'incertitude sur l'erreur relative due au plan initial. S'il est possible d'observer des valeurs aberrantes à 50 points, l'amélioration en terme de vitesse de convergence se confirme totalement à 70 points.

D'un point de vue applicatif, il est également intéressant de vérifier si la vitesse de convergence de l'incertitude relative  $U_r$  est également plus rapide pour le BOAS que pour le NLHS. En effet ce critère d'arrêt est la seule information disponible sur la précision quand on considère un problème d'application où l'on n'a pas accès à la valeur de  $q_{95}$ . On présente en Figure 5.5 la comparaison du BOAS et du NLHS pour l'incertitude relative  $U_r$  avec la même présentation que pour la Figure 5.4. Les conclusions sont les mêmes que pour l'erreur relative. On valide ainsi le caractère optimal du remplissage adaptatif orienté du BOAS.

### 5.1.5 Exemple analytique : la fonction de Borehole

Une analyse similaire à celle précédente est menée pour une autre fonction de test analytique, la fonction de Borehole, dont l'expression est donnée dans l'Eq. (4.46). L'analyse conduite pour la fonction d'Ishigami est ici reproduite à l'identique pour confirmer les résultats en dimension plus grande (8 paramètres d'entrée). La Figure 5.6 montre l'évolution de  $\epsilon_r$ ,  $U_r$  et  $1 - Q_w^2$  moyennés sur les 50 répétitions partant de plans LHS initiaux de 70 points enrichis à l'aide de la méthode BOAS jusqu'à 170 points. On observe ici encore la convergence de la méthode BOAS pour l'estimation du quantile à 95%. Comme précédemment l'incertitude relative  $U_r$  surestime l'erreur relative réelle. Cela confirme que  $U_r$  est basé sur un intervalle de confiance fiable pour un problème applicatif. Enfin on observe également la convergence de  $1 - Q_w^2$ . Avec la même démarche que pour la fonction d'Ishi-

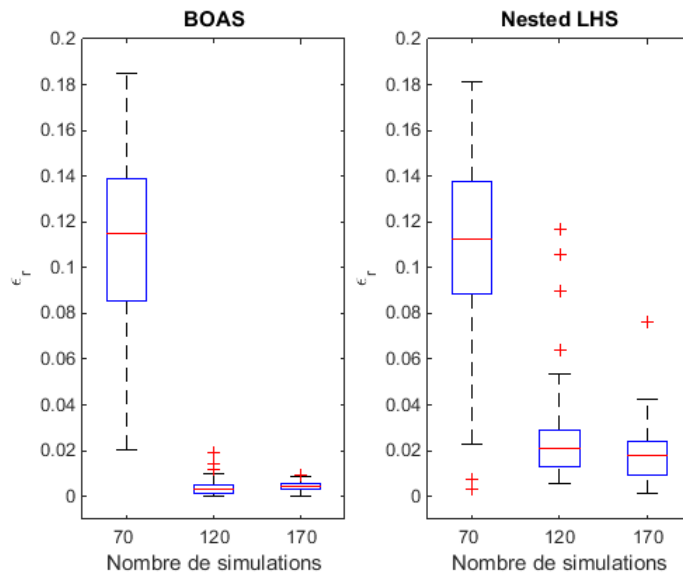


FIGURE 5.7 : Comparaison des performances des méthodes BOAS et NLHS vis-à-vis de  $\epsilon_r$  avec 50 plans LHS initiaux différents pour la fonction de Borehole.

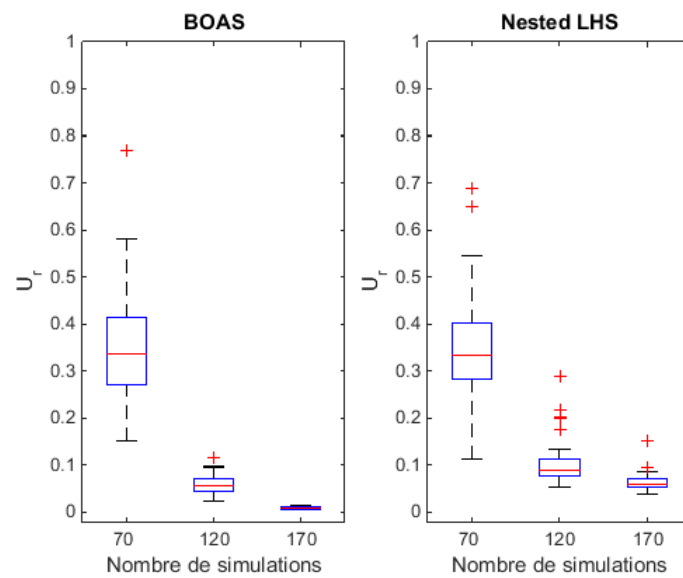


FIGURE 5.8 : Comparaison des performances des méthodes BOAS et NLHS vis-à-vis de  $U_r$  avec 50 plans LHS initiaux différents pour la fonction de Borehole.

gami, les vitesses de convergence des méthodes BOAS et NLHS sont comparées pour  $\epsilon_r$  et  $U_r$ . Les résultats sont présentés en Figure 5.7 pour l'erreur relative et en Figure 5.8 pour l'incertitude relative. De nouveau, on observe qu'à partir de performances équivalentes pour les plans initiaux, l'approche BOAS montre une vitesse de convergence plus grande que le NLHS pour l'erreur relative. On observe également une réduction significative de la dispersion des performances en fonction du plan initial pour 120 et 170 points. En ce qui concerne la comparaison pour  $U_r$ , à 120 points l'amélioration est moins significative même si le BOAS obtient de meilleures performances en terme de valeur médiane. Néanmoins, si on prend en compte les valeurs jugées aberrantes pour le NLHS la réduction de la dispersion est plus remarquable. La différence de performance devient ensuite plus claire à 170 points.

### 5.1.6 Exemple analytique : la fonction de Sobol

Dans le but d'analyser les performances de la méthode BOAS sur un exemple en grande dimension, la même démarche est appliquée sur la fonction de Sobol' en dimension 15. L'expression de cette fonction est la même que dans l'Eq. (5.26) mais en supprimant les 5 dernières dimensions. On considère 5 dimensions de moins pour que le temps de calcul des 50 répétitions du processus entier de la méthode BOAS soit raisonnable. On obtient donc toujours l'expression suivante :

$$f(\mathbf{x}) = \prod_{i=1}^{15} \frac{|4x_i - 2| + c_i}{1 + c_i} \quad \mathbf{x} \in [0, 1]^{15} \quad (5.26)$$

mais cette fois-ci avec :

$$\mathbf{c} = (1, 2, 5, 10, 20, 50, 100, 500, 500, 500, 500, 500, 500, 500, 500) \quad (5.27)$$

La fonction de Sobol' est un exemple intéressant pour le test de la méthode BOAS parce qu'étant potentiellement une limitation pour ses performances. En effet, il est attendu que cet exemple limite les performances du BOAS non pas en terme de vitesse de convergence de l'estimation du quantile elle-même mais plutôt en terme de vitesse de convergence de  $U_r$ . En effet, la fonction de Sobol' est une fonction dont les valeurs maximales – et donc la région du quantile à 95% – se situent sur les sommets de l'hypercube du domaine d'entrée. Or il est expliqué dans Gazut [2007] que les rééchantillonnages bootstrap peuvent être instables sur les bords du domaine dans le cas d'utilisation de méthodes de régression, ce qui aboutit à des intervalles de confiances qui ont du mal à se réduire avec l'ajout de points. Sont en cause les effets leviers des modèles de régressions. L'intervalle de confiance donné par le bootstrap aura donc plus de difficultés à être réduit dans ces zones. On teste donc la méthode BOAS sur cette fonction en reprenant la même démarche que pour les deux précédents exemples. La Figure 5.9 montre l'évolution de  $\epsilon_r$ ,  $U_r$  et  $1 - Q_w^2$  moyennés sur les 50 répétitions partant de plans LHS initiaux de 100 points enrichis à l'aide de la méthode BOAS jusqu'à 400 points. Pour ce qui est de l'erreur relative  $\epsilon_r$  faite sur le quantile on remarque qu'elle passe assez rapidement sous la barre des 1% d'erreur – ce qui est une erreur largement acceptable pour nos cas d'application – ainsi que l'erreur relative leave-one-out pondérée  $1 - Q_w^2$ . Pour ce qui est de l'incertitude relative  $U_r$ , la Figure 5.9 confirme ce que l'on attendait, on observe que cet estimateur a du mal à converger. Des expériences additionnelles nous montrent que  $U_r$  finit par rejoindre l'erreur relative mais au

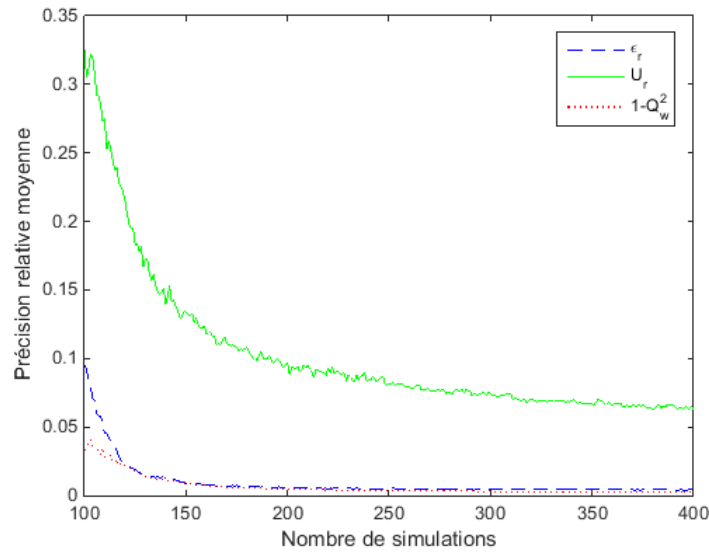


FIGURE 5.9 : Valeurs moyennes de  $\epsilon_r$ ,  $U_r$  et  $1 - Q_w^2$  en fonction du nombre de simulations sur les 50 répétitions du LHS initial pour la fonction de Sobol.

prix de nombreuses simulations supplémentaires. L'intérêt d'avoir plusieurs critères pour juger de la précision de l'estimation apparaît clairement ici. Dans un cas comme celui-là, en observant la décroissance lente de  $U_r$  mais la convergence très rapide de  $1 - Q_w^2$ , on peut d'abord observer la zone d'exploration. Et si on constate que celle-ci se situe sur les bords du domaine, il est possible d'arrêter l'enrichissement uniquement en fonction du critère  $1 - Q_w^2$ . Il est important de noter que pour d'autres techniques de construction de méta-modèles, ce comportement d'instabilité des rééchantillonnages bootstrap sur les bords du domaine ne s'observe pas nécessairement. Par exemple, pour une approche étudiée dans [Gazut 2007, Gazut et al. 2008] basée sur le bootstrap et utilisant des réseaux de neurones, ce comportement ne s'observe pas. Le BOAS utilisé avec cette technique de construction de méta-modèles ne rencontrerait donc pas cette limitation. L'efficacité de la méthode BOAS en serait sans doute renforcée.

Comme précédemment, les performances de la méthode BOAS sont comparées aux performances du NLHS en suivant la même démarche et les résultats sont présentés dans les Figures 5.10 et 5.11. L'intérêt du BOAS apparaît clairement sur la Figure 5.10. En effet, quand sur la fonction de Sobol, le NLHS produit des erreurs relatives sur les estimations du quantile à peu près constantes autour de 10%, on voit clairement que le BOAS contribue à fortement réduire l'erreur par l'ajout de points. On observe le même comportement quant à l'évolution de la dispersion des résultats par rapport au plan d'expérience initial. Pour la Figure 5.11, les mêmes conclusions peuvent être tirées. Même si on observe une vitesse de convergence de  $U_r$  plus lente pour cet exemple que pour les précédents, le BOAS contribue à significativement réduire l'incertitude  $U_r$  par rapport au NLHS. Il en est de même pour la dispersion des résultats par rapport au plan initial. A travers ces résultats, il apparaît que même avec un exemple théoriquement limitant par rapport à la méthode, on obtient des estimations de quantiles avec des erreurs faibles.

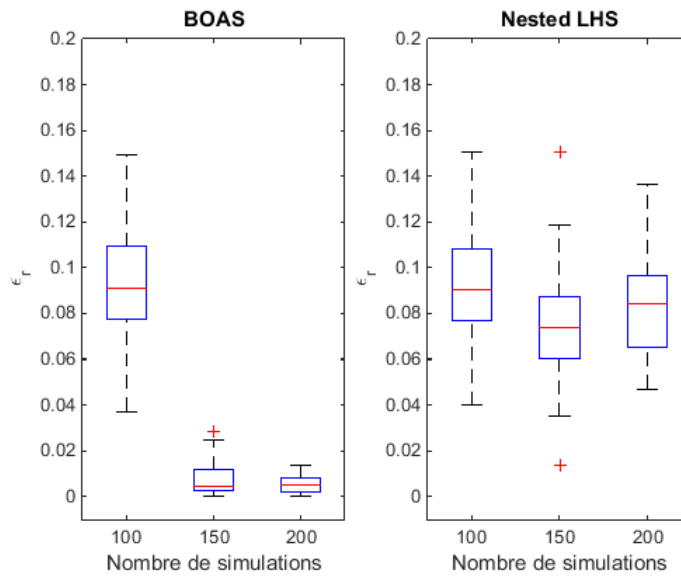


FIGURE 5.10 : Comparaison des performances des méthodes BOAS et NLHS vis-à-vis de  $\epsilon_r$  avec 50 plans LHS initiaux différents pour la fonction de Sobol.

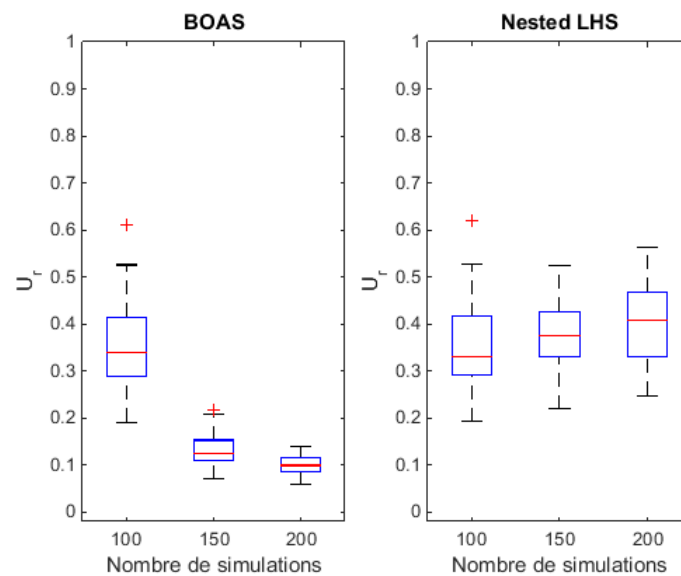


FIGURE 5.11 : Comparaison des performances des méthodes BOAS et NLHS vis-à-vis de  $U_r$  avec 50 plans LHS initiaux différents pour la fonction de Sobol.

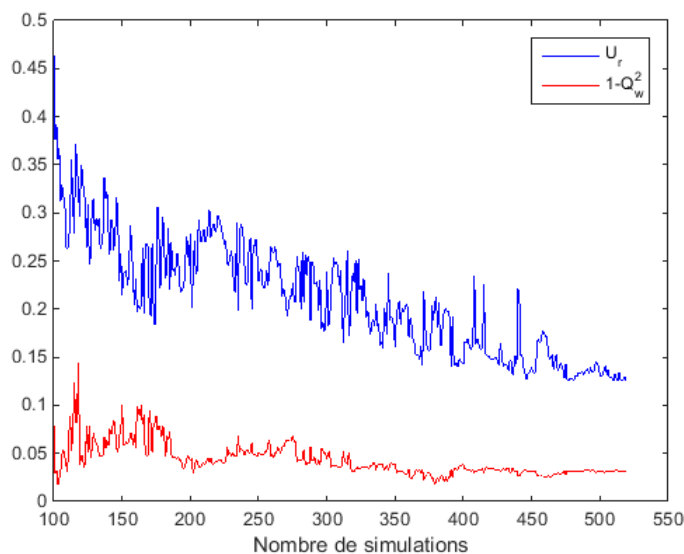


FIGURE 5.12 : Evolution de  $U_r$  et  $1 - Q_w^2$  en fonction du nombre de points dans le plan d'expériences pour l'exemple d'application de l'exposition d'un modèle d'enfant.

### 5.1.7 Exemple d'application : exposition d'un modèle d'enfant à une station de base de femtocell

La capacité de la méthode BOAS à réduire l'incertitude sur l'estimation du quantile à 95% de la distribution du DAS corps entier va maintenant être illustrée sur un exemple issu de la dosimétrie numérique. L'exemple du modèle anatomique d'enfant exposé à une station de base de femtocell de la Section 4.1.8 est repris dans cette section. En partant du LHS de 100 points initial du traitement de cet exemple dans la Section 4.1.8, on ajoute des points au plan d'expériences à l'aide de la méthode BOAS. La Figure 5.12 présente l'évolution des indicateurs de précision  $U_r$  et  $1 - Q_w^2$  en fonction du nombre de points ajoutés au plan d'expériences. La Figure 5.12 montre qu'à partir d'une incertitude relative supérieure à 45% l'ajout de point contribue à la réduire significativement jusqu'à 12%. Dans cet exemple, la zone d'intérêt pour l'estimation des quantiles se situe sur les bords du domaine comme dans la fonction de Sobol'. On observe donc une diminution assez lente de l'incertitude. Il apparaît également que la valeur de  $1 - Q_w^2$  se stabilise à une valeur faible de 0.03. Ne pouvant pas ajouter un grand nombre de simulations supplémentaires car étant limité par le temps de calcul, le choix est fait d'arrêter l'ajout de points et de conserver la valeur finale de l'estimation de quantile associée à une incertitude relative de 12%, ce qui est jugé suffisant pour ce cas-test au regard des contraintes au niveau du temps de calcul. La stabilisation de  $1 - Q_w^2$  à une valeur de 3% nous conforte également sur la confiance que l'on peut accorder à la valeur du quantile estimé. Pour une meilleure visualisation du résultat, on représente sur la Figure 5.13 la valeur du quantile et les indicateurs de précision correspondants sur la figure de l'estimation de la distribution totale à laquelle on était parvenu à la fin de la section 4.1.8. Pour conclure, on a ici montré la capacité de la méthode BOAS à réduire significativement l'incertitude sur l'estimation du quantile à 95% d'une distribution de sortie sur un problème concret de dosimétrie numérique.



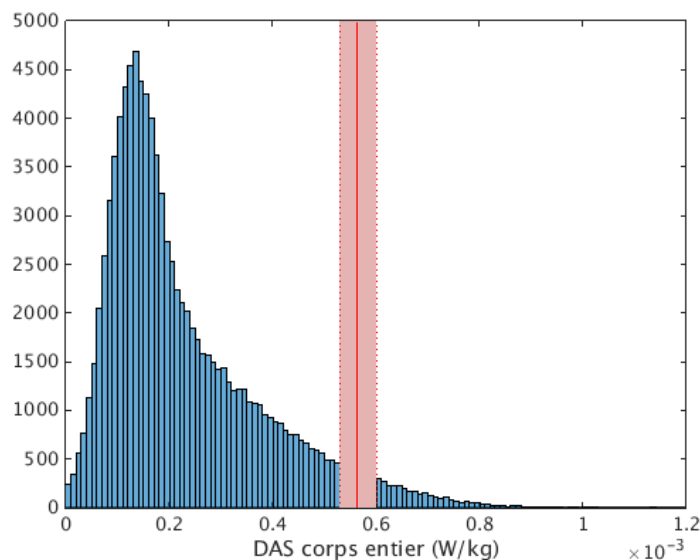


FIGURE 5.13 : Représentation de la valeur du quantile à 95% de la distribution du DAS corps entier avec ses bornes de confiance.

## 5.2 Comparaison de performances avec la méthode GPS (*Gaussian Process Shrunk*)

Pour finir, il est intéressant de comparer les performances du BOAS avec d'autres méthodes de planification d'expériences adaptative. A notre connaissance, il n'existe pas d'autres méthodes de planification adaptative pour les méta-modèle de chaos polynomial de régression. En revanche, pour les méta-modèles de Krigage, il existe plusieurs méthodes dont la méthode GPS que l'on a introduit au début de ce chapitre et dont le principe a servi de base à la méthode BOAS. On choisit donc de comparer les performances de la méthode BOAS avec cette méthode qui a prouvé son efficacité dans [Jala et al. 2014] pour l'estimation du quantile à 95% de distributions de sortie de codes de calcul nécessitant des temps de calcul importants. Avant toute chose, il est important de souligner que cette comparaison ne pourra pas amener de conclusion définitive du type : telle méthode réussit mieux que l'autre. En effet on compare ici des méthodes de planification adaptative sur deux types de méta-modèle différents : d'une part le chaos polynomial LARS et d'autre part le Krigage universel vu dans la section 4.3 (LARS-Kriging-PC). Les performances seront alors dépendantes de la capacité qu'a le méta-modèle à réagir à l'ajout de points sur un problème particulier. C'est également pour cela que l'on présentera moins d'exemples de validation de performances que dans les précédentes sections car on s'attend à avoir autant de conclusions différentes que d'exemples proposés. Cette section a donc uniquement vocation à donner des éléments sur les comportements respectifs du chaos polynomial LARS classique et du LARS-Kriging-PC sur des approches d'enrichissement orienté basées sur des principes similaires. On commence d'abord par présenter la méthode GPS que l'on peut retrouver en détail dans [Jala 2013].

### 5.2.1 La méthode GPS

La méthode GPS est une méthode de planification d'expériences adaptative orientée vers l'estimation de quantiles de la distribution de sortie qui utilise des modèles de Krigeage. Il est donc supposé ici que l'on a modélisé la sortie  $Y$  à l'aide d'un modèle de Krigeage universel dont on a calculé le *BLUP*  $\hat{m}_{\hat{Y}}$  (cf Eq. (4.37)) et la variance locale d'estimation  $s_{\hat{Y}}^2$  (cf Eq. (4.39)). Étant donné que cette méthode a inspiré celle du BOAS, le processus de sélection des points est assez similaire. Le principe est de partir d'un ensemble de points que l'on note  $\mathcal{S}$  de cardinal élevé par rapport au nombre de points du plan d'expériences. D'après la méthode proposée par [Srinivas et al. 2012] et à partir des points de  $\mathcal{S}$ , il est possible de proposer des fonctions d'encadrement  $m_N^L$  et  $m_N^U$  qui fournissent un intervalle de confiance de degré au moins  $1 - \alpha$  pour les valeurs prises par le modèle physique  $f$  et qui dépendent du nombre de points  $N$  de  $\mathcal{S}$  ajoutés au plan d'expériences  $\mathcal{X}^N$ . On précise que l'on a toujours  $\mathcal{X}^N \subset \mathcal{S}$ . Les fonctions d'encadrement sont définies comme suit :

$$\begin{aligned} m_N^L &: \mathbb{X} \rightarrow \mathbb{R} & m_N^U &: \mathbb{X} \rightarrow \mathbb{R} \\ \mathbf{x} &\mapsto \hat{m}_{\hat{Y}}(\mathbf{x}) - \sqrt{\gamma_N} s_{\hat{Y}}(\mathbf{x}) & \mathbf{x} &\mapsto \hat{m}_{\hat{Y}}(\mathbf{x}) + \sqrt{\gamma_N} s_{\hat{Y}}(\mathbf{x}) \end{aligned} \quad (5.28)$$

avec

$$\gamma_N = 2 \ln \left( \frac{\pi^2 N^2}{6} \right) + 2 \ln \left( \frac{N_m}{\alpha} \right) \quad (5.29)$$

Pour tout  $\mathbf{x} \in \mathcal{S}$ , on a donc l'intervalle de confiance :

$$f(\mathbf{x}) \in [m_N^L(\mathbf{x}) ; m_N^U(\mathbf{x})] \quad \forall \mathbf{x} \in \mathcal{S} \quad (5.30)$$

Avec une démonstration disponible dans [Jala 2013], on en déduit l'intervalle de confiance sur le quantile de la distribution de sortie :

$$q_{95} \in [q_{95_N}^L ; q_{95_N}^U] \quad (5.31)$$

Avec  $q_{95_N}^L$  et  $q_{95_N}^U$  qui sont les quantiles respectifs des distributions des variables aléatoires  $m_N^L(\mathbf{X})$  et  $m_N^U(\mathbf{X})$ . La zone d'intérêt pour l'enrichissement du plan d'expériences est définie de la même façon que pour la méthode BOAS :

$$\mathcal{D}^{N+1} = \{ \mathbf{x} \in \mathbb{X}, m_N^U(\mathbf{x}) \geq \hat{q}_{95_N}^L \} \quad (5.32)$$

Ainsi, par une démonstration similaire à celle de la Proposition , on montre que :

$$|q_{95} - q_{95_f}| \leq \sup_{\mathbf{x} \in \mathcal{D}^{N+1}} s_{\hat{Y}}(\mathbf{x}) \quad (5.33)$$

Enfin, il faut déterminer la règle avec laquelle le point est ajouté dans la zone d'enrichissement  $\mathcal{D}^{N+1}$ . Étant donné que l'on utilise ici des modèles de Krigeage, une façon judicieuse de diminuer la valeur de  $\sup_{\mathbf{x} \in \mathcal{D}^{N+1}} s_{\hat{Y}}(\mathbf{x})$  et donc de l'erreur faite sur l'estimation du quantile est d'ajouter le nouveau point à l'endroit où la variance du Krigeage est maximale :

$$\mathbf{x}^{(N+1)} = \arg \max_{\mathbf{x} \in \mathcal{D}^{N+1}} s_{\hat{Y}}(\mathbf{x}) \quad (5.34)$$

En effet, le BLUP étant interpolant, l'ajout de points aux maxima de variance diminue dans le même temps la borne supérieure des valeurs prises par cette variance locale dans la zone  $\mathcal{D}^{N+1}$ . Toutes les démonstrations relatives à la méthode GPS sont disponibles

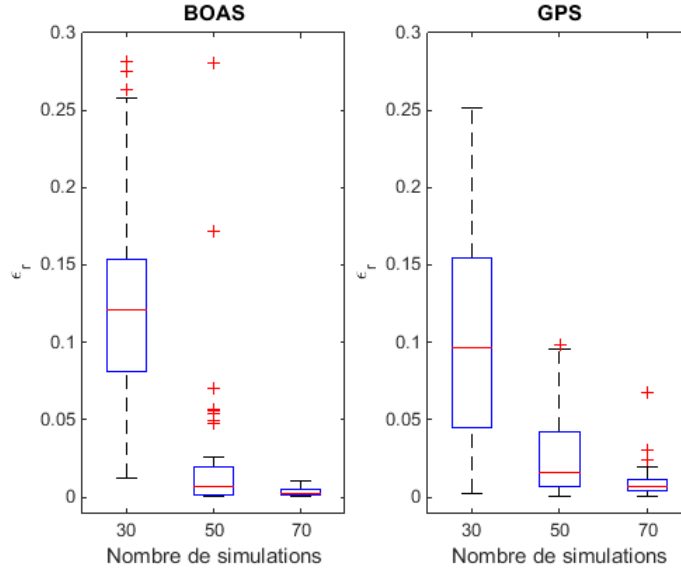


FIGURE 5.14 : Comparaison des performances des méthodes BOAS et GPS vis-à-vis de  $\epsilon_r$  avec 50 plan LHS initiaux différents pour la fonction d'Ishigami.

dans [Jala 2013]. Pour évaluer les performances de la méthode GPS et pour pouvoir les comparer aux performances de la méthode BOAS, on utilisera d'une part l'erreur relative faite sur le quantile de l'Eq. (5.21) et d'autre part l'incertitude relative de l'Eq. (5.22) qui avec la méthode GPS s'exprime :

$$U_r = \frac{|\hat{q}_{95N}^U - \hat{q}_{95N}^L|}{|\hat{q}_{95f}|} \quad (5.35)$$

## 5.2.2 Comparaison de performances avec la fonction d'Ishigami

Les performances des deux méthodes vont être comparées sur des exemples analytiques, premièrement la fonction d'Ishigami (cf Eq. (4.26)) et deuxièmement la fonction de Borehole (cf Eq. (4.46)). Pour cela, la même procédure que dans la section 5.1.4 est appliquée mais en remplaçant la comparaison avec le NLHS par une comparaison avec la méthode GPS. En ce qui concerne les modèles de Krigeage utilisés pour la méthode GPS, les modèles hybrides de Krigeage universel présentés dans la Section 4.3 utilisant des polynômes de la base du chaos polynomial LARS comme fonctions de régression seront utilisés. Ainsi, 50 LHS initiaux générés indépendamment sont enrichis séparément avec la méthode GPS et avec la méthode BOAS. Les évolutions et la dispersion des résultats des indicateurs de performance  $\epsilon_r$  et  $U_r$  sont ensuite affichés sous forme de boxplots. Les résultats sont présentés en Figure 5.14 et 5.15. Pour ce qui est de l'erreur relative réelle  $\epsilon_r$  sur la Figure 5.14, on remarque que partant de performances à peu près comparables pour le nombre de points initiaux, la méthode BOAS a une réduction plus rapide de l'erreur commise sur le quantile en valeur médiane même si la différence est légère. De plus, à 50 points où la différence est la plus importante, les meilleures performances en valeur médiane et en étalement du BOAS sont pondérées par le nombre plus important de valeurs considérées comme aberrantes par l'algorithme des boxplots. On peut donc conclure qu'au niveau de

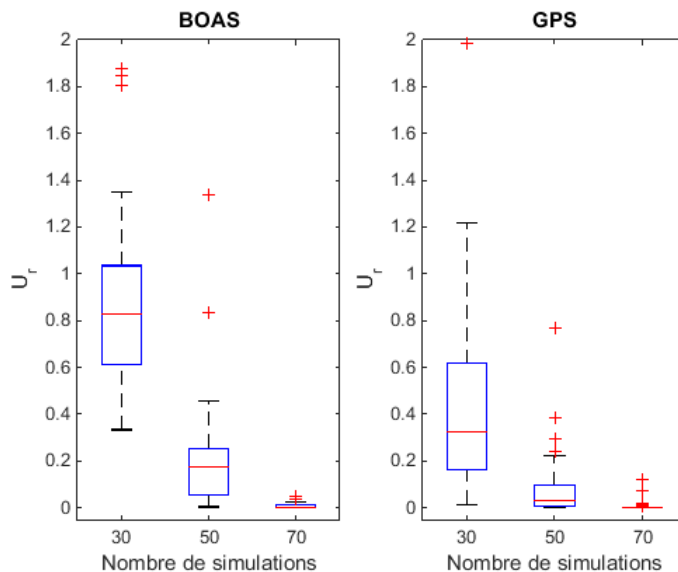


FIGURE 5.15 : Comparaison des performances des méthodes BOAS et GPS vis-à-vis de  $U_r$  avec 50 plan LHS initiaux différents pour la fonction d'Ishigami.

la vitesse de réduction de l'erreur avec l'ajout de point, la méthode BOAS et la méthode GPS ont des performances comparables. En ce qui concerne l'estimateur de précision  $U_r$  qui est celui auquel on a accès dans un contexte applicatif, il apparaît sur la Figure 5.15 que c'est le GPS qui donne les incertitudes les plus faibles pour 30 et 50 points et qui surestime néanmoins l'erreur relative réelle. En revanche pour 70 points, quand l'erreur relative réelle devient faible, la Figure 5.15 montre qu'il est possible dans certain cas que l'incertitude  $U_r$  sous-estime légèrement l'erreur relative réelle ce qui n'est pas le cas pour la méthode BOAS. Cette sous-estimation trouve son origine dans l'incertitude dans l'estimation des paramètres du modèle de Krigage qui induit un léger biais au modèle et rend les fonctions d'encadrements légèrement moins efficaces concernant l'encadrement du modèle physique. Les performances des deux méthodes peuvent donc être qualifiées de comparables excepté l'inconvénient d'une légère sous-estimation de l'erreur réelle quand cette erreur réelle devient faible pour le GPS.

### 5.2.3 Comparaison de performances avec la fonction de Borehole

On effectue maintenant la même analyse sur la fonction de Borehole (cf Eq. (4.46)). Les résultats sont présentés sur les Figures 5.16 et 5.17. Pour l'erreur relative  $\epsilon_r$ , il apparaît sur la Figure 5.16, que les modèles de Krigage universel sur les plans d'expériences initiaux obtiennent des performances significativement meilleures que les modèles de chaos polynomial LARS. Malgré cette "avance" des modèles de Krigage, l'ajout de points suivant la méthode BOAS contribue rapidement à rattraper le "retard" sur la méthode GPS. On observe une réduction très rapide de l'erreur relative pour le BOAS pour cet exemple analytique. Finalement, pour 170 points, l'erreur du BOAS est équivalente au GPS pour la médiane mais avec une dispersion des résultats plus réduite. En ce qui concerne l'incertitude relative  $U_r$ , la situation inverse de celle de la fonction d'Ishigami est observée. En effet, dans le cas de la fonction de Borehole, c'est la méthode BOAS qui fournit l'incertitude la

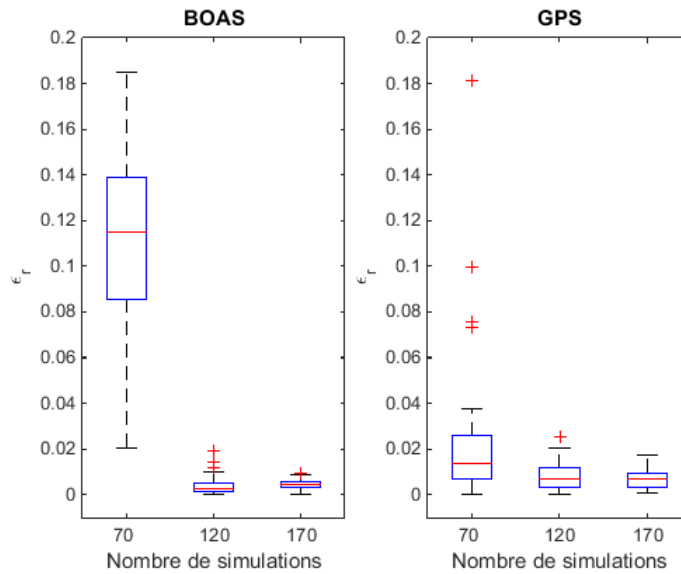


FIGURE 5.16 : Comparaison des performances des méthodes BOAS et GPS vis-à-vis de  $\epsilon_r$  avec 50 plans LHS initiaux différents pour la fonction de Borehole.

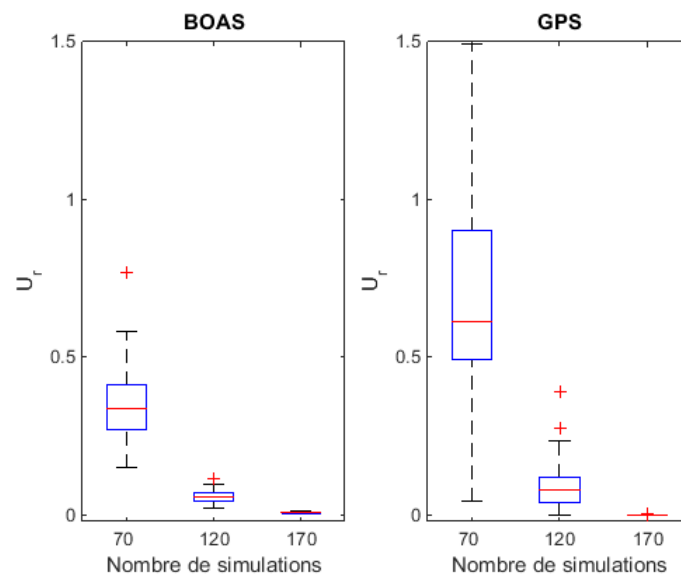


FIGURE 5.17 : Comparaison des performances des méthodes BOAS et GPS vis-à-vis de  $U_r$  avec 50 plans LHS initiaux différents pour la fonction de Borehole.

plus proche de l'erreur relative. Enfin on retrouve l'inconvénient de la sous-estimation de l'erreur de la méthode GPS pour 170 points quand l'erreur relative réelle est réduite. Pour conclure, dans le cas de la fonction de Borehole, l'ajout de points selon les deux méthodes conduit à des performances équivalentes malgré les différences de performances initiales dues aux modèles employés.

### 5.3 Conclusion

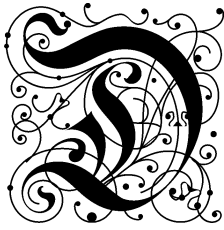
Pour conclure, ce chapitre a présenté une nouvelle méthode appelée BOAS permettant de construire itérativement un plan d'expérience orienté vers l'estimation efficace du quantile à 95% de la distribution de sortie. Cette méthode permet également de déduire des estimateurs de la précision du quantile grâce aux rééchantillonnages bootstrap utilisées pour définir la nouvelle zone de recherche. Le BOAS a été appliqué à plusieurs exemples analytiques et à un exemple issu de la dosimétrie numérique et a montré son efficacité pour l'estimation du quantile par rapport à une méthode plus classique de space filling (NLHS). Cette méthode fait l'objet d'une soumission d'article de revue scientifique. Si le BOAS a été initialement conçu pour des méta-modèles de chaos polynomial, il peut cependant être appliqué à d'autres méthodes telles que le Krigeage ou encore les réseaux de neurones. On peut également souligner que pour cette dernière technique, on ne sera pas confronté à la limitation que l'on a remarquée quand la zone d'intérêt pour la recherche de nouveaux points se situe sur les bords du domaine [Gazut 2007]. Enfin, ce chapitre a présenté une étude comparative de la méthode GPS utilisant des modèles de LARS-Kriging-PC et la méthode BOAS utilisant des modèles de chaos polynomial LARS. La comparaison sur les deux exemples analytiques proposés montre des performances similaires pour les deux méthodes de planification d'expériences adaptatives.



## Chapitre 6

### Conclusion générale et perspectives





ans ce rapport de thèse sont présentées la problématique générale de mes travaux ainsi que mes contributions. En ce qui concerne la modélisation électromagnétique, après une prise en main du logiciel FDTD que j'avais à disposition, mon travail a consisté en l'utilisation et en l'adaptation des modes sphériques pour la modélisation de sources en zone intermédiaire telles que les stations de base femtocell.

En ce qui concerne la modélisation statistique de la sortie du code de calcul en fonction de paramètres d'entrée variables, j'ai dans un premier temps travaillé à maîtriser les approches de modélisation qui me semblaient adaptées à notre problème, à savoir le développement en séries polynomiales du chaos polynomial avec une approche non intrusive par régression et le Krigeage universel. J'ai ensuite travaillé d'un côté sur les techniques de planification d'expériences à travers ma proposition de stratégie séquentielle de planification suivant un critère maximin. Cette stratégie s'avère efficace par rapport aux plans LHS classiques sur l'exemple de la fonction d'Ishigami pour augmenter la précision des modèles de chaos polynomial construits. D'un autre côté, j'ai travaillé à l'amélioration des modèles de Krigeage universel en particulier par rapport à sa formulation basique (le Krigeage ordinaire). Ma contribution a donc été de proposer une combinaison du chaos polynomial et du Krigeage universel à travers l'utilisation des polynômes sélectionnés par l'algorithme LARS comme fonctions de régression pour la partie déterministe du Krigeage universel. Cette méthode apparaît comme une méthode optimale qui construit un méta-modèle au moins aussi précis, sinon plus, que le chaos polynomial par régression classique et le Krigeage ordinaire. Cette méthode a été ensuite mise en oeuvre sur des exemples de fonctions analytiques et sur un exemple d'exposition d'un fœtus aux émissions électromagnétiques d'un système femtocell. Les résultats confirment son caractère optimal par rapport aux deux autres méthodes.

Enfin, pour répondre à l'objectif de parcimonie dans l'appel au modèle physique pour l'estimation du quantile à 95% des distributions de sortie, j'ai proposé une nouvelle méthode appelée BOAS pour *Bootstrapped Oriented Adaptive Sampling* permettant une planification d'expériences adaptative pour l'estimation du quantile à 95 % en utilisant des modèles de chaos polynomial régressifs intégrant des polynômes de base sélectionnés par le LARS. Cette méthode repose sur des rééchantillonnages bootstrap du plan d'expériences permettant de définir des intervalles de confiance locaux et par conséquent des fonctions d'encadrement du modèle physique à partir des points servant à la construction du méta-modèle de chaos polynomial. Ces fonctions d'encadrement servent alors à délimiter une zone d'exploration réduite pour l'ajout de nouveaux points tout en fournissant des indicateurs de précision sur l'estimation du quantile. La méthode BOAS prouve son efficacité pour l'estimation du quantile de façon parcimonieuse par rapport à d'autres stratégies d'enrichissement global. Elle montre également des performances équivalentes à la stratégie GPS (*Gaussian Process Shrunk*) basée sur le même principe mais utilisant des modèles de Krigeage. Une limitation des performances propres à l'utilisation de modèles de régression a également été observée pour le BOAS. Lorsque la zone d'exploration se situe sur les bords du domaine, les effets levier des points situés dans cette zone provoquent une forte instabilité des rééchantillonnages bootstrap. En conséquence, l'incertitude relative servant à estimer la précision de l'estimation du quantile a dans ce cas une vitesse de convergence assez lente. Néanmoins, sur l'exemple observé, cette limitation n'a pas d'effet sur la réduction de l'erreur réelle faite sur le quantile et il est possible d'utiliser le deuxième estimateur de précision de l'erreur leave-one-out pondérée pour juger de la précision de

l'estimation. Il est également à noter que cette limitation ne doit pas se rencontrer a priori pour des techniques de construction de méta-modèles différentes. On peut citer l'exemple des réseaux de neurones.

En ce qui concerne les perspectives de mes travaux, commençons par parler des polynômes de base du chaos polynomial. Dans les problèmes rencontrés en dosimétrie numérique, il est fréquent d'être confronté à des paramètres d'entrées périodiques (par exemple l'angle d'azimut). Dans ces travaux, ces paramètres sont traités comme des paramètres suivant une loi uniforme sur  $[-1, 1]$ . Par conséquent, l'utilisation de l'information correspondant à  $f(2+x) = f(x)$  (la fonction étant 2-périodique) n'est pas utilisée au voisinage de 1. Un objectif pourrait être de constituer une base de polynômes 2-périodiques afin d'utiliser la totalité de l'information disponible pour modéliser la sortie. Des polynômes construits à partir d'une base de Fourier par exemple pourraient être utiles pour modéliser des phénomènes périodiques.

Plus fondamentalement, concernant les modèles de Krigeage utilisant des polynômes de la base du chaos polynomial comme fonctions de régression, une perspective serait de construire des noyaux de covariance de sorte que la moyenne du processus Gaussien soit orthogonale à la partie déterministe du modèle de Krigeage. Ainsi, le processus Gaussien devrait avoir plus de facilités à modéliser précisément le résidu de la partie déterministe et de la sortie.

Quant à la méthode BOAS, une suite à donner à ces travaux pourrait être d'abord de pousser vers l'estimation de quantiles plus élevés que celui à 95% pour observer jusqu'où l'approche est efficace. En effet si toutes les démonstrations et raisonnements sont valables pour n'importe quel quantile, il n'est pas certain que l'ajout de points dans la zone réduite d'exploration réduise efficacement l'écart maximal entre les fonctions d'encadrement pour des quantiles très élevés. D'autre part, il pourrait également être intéressant de tester la méthode BOAS avec d'autres approches pour la construction des méta-modèles, en particulier pour observer ses réactions vis-à-vis de la limitation des effets levier pour les modèles de régression. Enfin, concernant ces effets leviers, il pourrait être intéressant d'intégrer un facteur de régularisation sur l'incertitude relative pour que cet estimateur soit plus proche de la réalité quand la zone d'exploration se situe sur les bords du domaine.

Plus largement, les applications qui ont été abordées dans ces travaux ne sont qu'un cas particulier d'une problématique plus générale de gestion de l'incertitude qui touche de nombreux secteurs de l'ingénierie utilisant la simulation numérique. En effet, dans de nombreux domaines, les outils de simulation sont devenus très précis et capables de modéliser et d'analyser des phénomènes avec souvent beaucoup de réalisme. Néanmoins, afin de caractériser et de prévoir les performances de systèmes divers, il est souvent nécessaire de prendre en compte la complexité et le caractère aléatoire de leur réalité d'utilisation et de fabrication. Ma participation dans le projet CHORUS m'a permis de me confronter au fait que de nombreux acteurs industriels de secteurs très différents tels que l'aéronautique, l'énergie, l'acoustique ou les télécommunications rencontraient le même besoin de prévoir et d'analyser les réponses de leurs simulateurs en intégrant les incertitudes liées à l'utilisation et à la réalisation de leurs systèmes. Cette tendance déjà très répandue dans le domaine de la mécanique, de la prospection minière ou encore de la météorologie trouve ainsi des échos de plus en plus nombreux dans les domaines de la conception d'équipements, de l'électromagnétisme et des télécommunications. Sur le sujet des métamatériaux, l'utilisation du chaos polynomial dans [Arya et al. 2015] pour évaluer la variabilité de la réponse électromagnétique d'une structure périodique millimétrique en est un exemple.

Les approches et démarches proposées dans ce mémoire s'inscrivent donc dans ce mouvement général et ont vocation à être appliquées dans ces domaines où la demande de prise en charge de la complexité est de plus en plus forte.

# Références

- Abramovitz, M. and Stegun, I. A. (1972). *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. Dover Publications, New-York, NY.
- Ackerman, M. (1995). Accessing the Visible Human project, D-Lib Mag. [http://www.nlm.nih.gov/research/visible/visible\\_human.html](http://www.nlm.nih.gov/research/visible/visible_human.html).
- ARCEP (2013). Observatoire des service mobiles. <http://www.arcep.fr/index.php?id=11677&L=0#c26979>.
- Arnaud, A., Bect, J., Couplet, M., Pasanisi, A., and Vazquez, E. (2010). Evaluation d'un risque d'inondation fluviale par planification séquentielle d'expériences. In *42ème Journées de Statistique*, pages 1–6.
- Arya, R., Kersaudy, P., Wiart, J., and Mittra, R. (2015). Statistical Analysis of Periodic Structures and Frequency Selective Surfaces using the Polynomial Chaos Expansions. *FERMAT*, 12.
- Bachoc, F. (2013). Cross Validation and Maximum Likelihood estimation of hyper-parameters of Gaussian processes with model misspecification. *Computational Statistics and Data Analysis*, 66 :55–69.
- Balanis, C. A. (1988). *Antenna Theory : Analysis and Design*. John Wiley and Sons, New-York, NY.
- Beard, B., Kainz, W., Onishi, T., Iyama, T., Watanabe, S., Fujiwara, O., Wang, J., Bit-Babik, G., Faraone, A., Wiart, J., Christ, A., Kuster, N., Kyoung Lee, A., Kroeze, H., Siegbahn, M., Keshvari, J., Abrishamkar, H., Simon, W., Manteuffel, D., and Nikolovski, N. (2006). Comparisons of computed mobile phone induced SAR in the SAM phantom to that in anatomically correct models of the human head. . *IEEE Transactions on Electromagnetic Compatibility*, 48(2) :397–407.
- Bect, J., Ginsbourger, D., Li, L., Picheny, V., and Vazquez, E. (2012). Sequential design of experiments for estimation of a probability of failure. *Statistics and Computing*, 22(3) :773–793.
- Berenger, J. (1994). A perfectly matched layer for the absorption of electromagnetic waves. *Journal of Computational Physics*, 114(2) :185–200.
- Berveiller, M., Sudret, B., and Lemaitre, M. (2006). Stochastic finite elements : a non intrusive approach by regression. *Eur. J. Comput. Mech.*, 15(1-3) :81–92.

- Bibin, L., Anquez, J., De la Plata Alcalde, J., Boubekeur, T., Angelini, E., and Bloch, I. (2010). Whole-body pregnant woman modeling by digital geometry processing with detailed uterofetal unit based on medical images. *IEEE Transactions on Biomedical Engineering*, 57(10) :2346–2358.
- Bichon, B., McFarland, J., and Mahadevan, S. (2011). Efficient surrogate models for reliability analysis of systems with multiple failure modes. *Reliability Engineering & System Safety*, 96(10) :1386–1395.
- Blatman, G. and Sudret, B. (2008). Sparse polynomial chaos expansions and adaptive stochastic finite elements using a regression approach. *Comptes Rendus Mécanique*, 336(6) :518–523.
- Blatman, G. and Sudret, B. (2010). An adaptive algorithm to build up sparse polynomial chaos expansions for stochastic finite element analysis. *Probabilistic Engineering Mechanics*, 25(2) :183–197.
- Blatman, G. and Sudret, B. (2011). Adaptive sparse polynomial chaos expansion based on least angle regression. *Journal of Computational Physics*, 230(6) :2345–2367.
- Borgonovo, E., Castaings, W., and Tarantola, S. (2007). Moment independent importance measures : New results and analytical test cases. *Risk Analysis*, 31 :404–428.
- Breiman, L. (1996). Bagging predictors. *Machine Learning*, 24(2) :123–140.
- Caniou, Y. (2012). *Analyse de sensibilité globale pour les modèles de simulation imbriqués et multi-échelles*. PhD thesis, Université Blaise Pascal - Clermont II.
- Cannamela, C., Garnier, J., and Iooss, B. (2008). Controlled Stratification for Quantile Estimation. *Annals of Applied Statistics*, 2(4) :1554–1580.
- CENELEC (2001). Norme européenne Basic Standard for Measurement of Specific Absorption Rate related to Human Exposure. CENELEC EN 50361. <http://www.cenelec.eu>.
- Chastaing, G., Gamboa, F., and Prieur, C. (2010). Generalized Hoeffding-Sobol decomposition for dependent variables - application to sensitivity analysis. *Technical Report, hal-00648404 GdR MASCOT-NUM*, 225.
- Christ, A., Kainz, W., Hahn, E., Honegger, K., Zefferer, M., Neufeld, E., Rascher, W., Janka, R., Bautz, W., Chen, J., Shen, J., Oberle, M., Szczerba, D., Kam, A., Guag, J., and Kuster, N. (2010). The Virtual Family –development of surface based anatomical models of two adults and two children for dosimetric simulations. *Physics in Medicine and Biology*, 55(2) :33–38.
- Christ, A. and Kuster, N. (2005). Differences in RF energy absorption in the heads of adults and children. *Bioelectromagnetics Supplement*, 7.
- Conil, E., Hadjem, A., Gati, A., Wong, M.-f., and Wiart, J. (2011). Influence of Plane-Wave Incidence Angle on Whole Body and Local Exposure at 2100 MHz. *IEEE Transactions on Electromagnetic Compatibility*, 53(1) :48–52.

- Conil, E., Hadjem, A., Lacroux, F., Wong, M. F., and Wiart, J. (2008). Variability analysis of SAR from 20MHz to 2.4 GHz for different adult and child models using finite-difference time-domain. *Physics in Medicine and Biology*, 53 :1511–1525.
- Dahdouh, S., Varsier, N., Serrurier, A., de le Plata Alcade, J., Anquez, J., Angelini, E. D., Wiart, J., and Bloch, I. (2014). A comprehensive tool for image-based generation of fetus and pregnant women mesh models for numerical dosimetry studies. *Physics in Medicine and Biology*, 89 :4583–4602.
- Dimbylow, P. (2005). Development of the female voxel phantom, NAOMI and its application to calculations of induced current densities and electric fields from applied low frequency magnetic and electric fields. *Physics in Medicine Biology*, 50(6) :1047–1070.
- Dubourg, V., Sudret, B., and Bourinet, J. (2011). Reliability-based design optimization using kriging surrogates and subset simulation. *Struct. Multidisc. Optim.*, 44(5) :673–690.
- Dubrule, O. (1983). Cross validation of kriging in a unique neighborhood. *Journal of the International Association for Mathematical Geology*, 15(6) :687–699.
- Echard, B., Gayton, N., and Lemaire, M. (2011). AK-MCS : An active learning reliability method combining Kriging and Monte Carlo Simulation. *Structural Safety*, 33(2) :145–154.
- Efron, B. (1979). Bootstrap Methods : Another Look at the Jackknife. *The Annals of Statistics*, 7(1) :1–26.
- Efron, B. and Gong, G. (1983). A Leisurely Look at the Bootstrap , the Jackknife , and Cross-Validation. *The American Statistician*, 37(1) :36–48.
- Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). LEAST ANGLE REGRESSION. *The Annals of Statistics*, 32(2) :407–499.
- Efron, B. and Tibshirani, R. J. (1981). *An introduction to the Bootstrap*. Chapman & Hall/CRC.
- EL Habachi, A. (2011). *Propagation de la variabilité de la morphologie humaine sur le Débit d’Absorption Spécifique en dosimétrie numérique*. PhD thesis, IES de l’Information des Télécommunications et des Systèmes.
- El Habachi, A., Conil, E., Hadjem, A., Vasquez, E., Wong, M. F., Gati, A., Fleury, G., and Wiart, J. (2010). Statistical analysis of whole-body absorption depending on anatomical human characteristics at a frequency of 2.1 GHz. *Physics in Medicine and Biology*, 55(7) :1875–1887.
- Freedman, D. A. (1981). Bootstrapping Regression Models. *The Annals of Statistics*, 9(6) :1218–1228.
- Gazut, S. (2007). *Conception et mise en oeuvre de nouvelles méthodes d’élaboration de plans d’expériences pour l’apprentissage de modèles non linéaires*. PhD thesis, Université Paris-Sud XI.

- Gazut, S., Martinez, J.-M., Dreyfus, G., and Oussar, Y. (2008). Towards the optimal design of numerical experiments. *IEEE transactions on neural networks*, 19(5) :874–882.
- Ghanem, R. G. and Spanos, P. D. (2003). *Stochastic Finite Elements : A Spectral Approach*. Courier Dover Publications.
- Ghanmi, A. (2012). *Analyse de l'exposition aux ondes électromagnétiques des enfants dans le cadre des nouveaux usages et des nouveaux réseaux*. PhD thesis, Université Paris-Est.
- Gilli, L., Lathouwers, D., Kloosterman, J., van der Hagen, T., a.J. Koning, and Rochman, D. (2013). Uncertainty quantification for criticality problems using non-intrusive and adaptive Polynomial Chaos techniques. *Annals of Nuclear Energy*, 56 :71–80.
- Ginsbourger, D. (2009). *Multiplés métamodèles pour l'approximation et l'optimisation de fonctions numériques multivariées*. PhD thesis, Ecole Nationale Supérieure des Mines de Saint-Etienne.
- Glynn, P. W. (1996). Importance Sampling for Monte Carlo Estimation of Quantiles. In *Proceedings of the Second International Workshop on Mathematical Methods in Stochastic Simulation and Experimental Design*, pages 180–185, St. Petersburg.
- Hansen, J. E. (1988). *Spherical Near-Field antenna measurements*. Peter Peregrinus Ltd, London, UK.
- Hesterberg, T. C. and Nelson, B. L. (1998). Control Variates for Probability and Quantile Estimation. *Management Science*, 44(9) :1295–1312.
- Hirata, A., Ito, N., and Fujiwara, O. (2009). Influence of electromagnetic polarization on the whole-body averaged SAR in children for plane-wave exposures. *Physics in Medicine and Biology*, 54 :59–65.
- Hirata, A. Kodera, S. W. J. and Fujiwara, O. (2007). Dominant factors for influencing whole-body average SAR due to far-field exposure in whole-body resonance frequency and GHz regions. *Bioelectromagnetics*, 28 :484–487.
- Holland, R. and Williams, J. (1983). Total-field versus scattered field finite difference codes : a comparative assessment. *IEEE Transactions on Nuclear Science*, 30(6) :4583–4588.
- ICNIRP (1998). Guidelines for limiting exposure to time-varying electric, magnetic and electromagnetic fields (up to 300 GHz). *Health Physics*, 44 :1630–1639.
- Iskander, M. and Yun, Z. (2002). Propagation prediction models for wireless communication systems. *IEEE Transactions on Antennas Propagation*, 50 :662–673.
- Isukapalli, S. (1999). *Uncertainty Analysis of Transport-Transformation Models*. PhD thesis, The State University of New Jersey.
- Jala, M. (2013). *Plans d'expériences adaptatifs pour le calcul de quantiles et application à la dosimétrie*. PhD thesis, Télécom Paris-Tech.

- Jala, M., Lévy-Leduc, C., Moulines, E., Conil, E., and Wiart, J. (2014). Sequential design of computer experiments for the assessment of fetus exposure to electromagnetic fields. *Technometrics*.
- Jawad, O. (2014). *Etude de l'exposition des personnes aux ondes électromagnétiques en environnement complexe*. PhD thesis, Université Libre de Bruxelles et Université Pierre et Marie Curie.
- Jensen, F., Frandsen, A., and Copenhagen, D. (2004). On the number of modes in spherical wave expansions. In *Proc. 26th AMTA*, volume 2, pages 489–494.
- Johnson, M., Moore, L., and Ylvisaker, D. (1990). Minimax and maximin distance designs. *Journal of Statistical Planning and Inference*, 26 :131–148.
- Jones, D., Schonlau, M., and Welch, W. (1998). Efficient Global Optimization of expensive black-box functions. *Journal of Global Optimization*, 13(4) :455–492.
- Joseph, V., Sudjianto, A., and Hung, Y. (2008). Blind Kriging : a new method for developing metamodels. *ASME Journal of Mechanical Design*, 130(3) :031102–1–8.
- Karhunen, K. (1947). Uber lineare Methoden in der Wahrscheinlichkeitsrechnung. *Annales Academiae scientiarum Fennicae : Mathematica-Physica*, 37 :1–79.
- Kersaudy, P., Mostarshedi, S., Sudret, B., Picon, O., and Wiart, J. (2014). Stochastic Analysis of Scattered Field by Building Facades Using Polynomial Chaos. *IEEE Transactions on Antenna Propagation*, 62(12) :6382–6393.
- Kersaudy, P., Sudret, B., Varsier, N., Picon, O., and Wiart, J. (2015). A new surrogate modeling technique combining Kriging and polynomial chaos expansions - Application to uncertainty analysis in computational dosimetry. *Journal of Computational Physics*, 286 :103–117.
- Kucherenko, S., tanrantola, S., and Annoni, P. (2012). Estimation of global sensitivity indices for models with dependent variables. *Computer Physics Communications*, 183 :937–946.
- Le Maitre, O. P., Reagan, M. T., Najm, H. N., Ghanem, R. G., and Knio, O. M. (2002). A Stochastic Projection Method for Fluid Flow. *Journal of Computational Physics*, 181(1) :9–44.
- Lebrun, R. and Dutfoy, A. (2009). A generalization of the Nataf transformation to distributions with elliptical copula. *Probabilistic Engineering Mechanics*, 24(2) :172–178.
- Li, G. and Rabitz, H. (2010). Global Sensitivity Analysis for Systems with Independent and/or Correlated Inputs. *Journal of Physical Chemistry*, 114 :6022–6032.
- Lophaven, S., Nielsen, H., and Sondergaard, J. (2002). *DACE, A Matlab Kriging Toolbox*. Technical University of Denmark.
- Lucor, D. and Karniadakis, G. (2004). Physical Systems with Random Uncertainties : Chaos Representations with Arbitrary Probability Measure. *SIAM Journal on Scientific Computing*, 26(2) :750–735.



- Matheron, G. (1969). *Les cahiers du Centre de morphologie mathématique de Fontainebleau, Fascicule 1 : Le krigeage universel*. Ecole de Mines de Paris, Fontainebleau.
- Matthies, H. and Keese, A. (2005). Galerkin methods for linear and nonlinear elliptic stochastic partial differential equations. *Computer Methods in Applied Mechanics and Engineering*, 194(12) :1295–1331.
- McKay, M., Conover, W., and Beckman, R. (1979). A comparison of three methods for selecting values of input variables in the analysis of an output from a computer code. *Technometrics*, 21 :239–245.
- Merewether, D., Fisher, R., and Smith, F. (1980). On implementing a numeric Huygens source scheme in a finite difference program to illuminate scattering bodies. *IEEE Transactions on Nuclear Science*, 27(6) :1829–1833.
- Molinaro, A. M., Simon, R., and Pfeiffer, R. M. (2005). Prediction error estimation : A comparison of resampling methods. *Bioinformatics*, 21(15) :3301–3307.
- Morris, M., Mitchell, T., and Ylvisaker, D. (1993). Bayesian design and analysis of computer experiments : use of derivatives in surface prediction. *Technometrics*, 35(3) :243–255.
- Mostarshedi, S., Richalot, E., Laheurte, J.-M., Wong, M., Wiart, J., and Picon, O. (2010). Fast and accurate calculation of scattered electromagnetic fields from building faces using Green’s functions of semi-infinite medium. *IET Microwaves, Antennas & Propagation*, 4 :72–82.
- Nagaoka, T., Togashi, T., Saito, K., Ito, K., and Watanabe, S. (2007). An anatomically realistic whole-body pregnant-woman model and specific absorption rates for pregnant-woman exposure to electromagnetic plane waves from 10 MHz to 2 GHz. *Phys. Med. Biol.*, 52(22) :6731–6745.
- Nataf, A. (1962). Détermination des distributions dont les marges sont données. *Comptes Rendus de l’Académie des Sciences de Paris*, 225 :42–43.
- Nelsen, R. (1999). *An introduction to Copulas*. Springer.
- Nelson, B. L. (1987). On Control Variate Estimators. *Computers & Operations Research*, 14(3) :219–225.
- Notin, A. (2010). *Evaluation à moindre coût de la fiabilité des structures sollicitées en fatigue*. PhD thesis, Université de Technologie de Compiègne.
- Oakley, J. (2004). Estimating percentiles of uncertain computer code outputs. *Journal of the Royal Statistical Society : Series C (Applied Statistics)*, 53(1) :83–93.
- Owen, A. (1994). Controlling correlations in Latin Hypercube samples. *Journal of American Statistical Association*, 89 :1517–1522.
- Park, J. S. (1994). Optimal Latin-Hypercube designs for computer experiments. *Journal of Statistical Planning and Inference*, 39 :95–111.

- Peyman, A., Gabriel, C., Benedickter, H., and Fröhlich, J. (2011). Dielectric properties of human placenta, umbilical cord and amniotic fluid. *Physics in Medicine and Biology*, 56 :564–612.
- Picheny, V., Ginsbourger, D., Roustant, O., Haftka, R. T., and Kim, N.-H. (2010). Adaptive Designs of Experiments for Accurate Approximation of a Target Region. *Journal of Mechanical Design*, 132(7).
- Rasmussen, C. E. and Williams, C. K. I. (2006). *Gaussian processes for machine learning*. MIT Press, Cambridge.
- Roustant, O., Ginsbourger, D., and Deville, Y. (2010). *DiceKriging : Kriging methods for computer experiments*, R package version 1.1.
- Sacks, J., Schiller, S. B., and Welch, W. J. (1989a). Design for Computer Experiments. *Technometrics*, 31(1) :41–47.
- Sacks, J., Welch, W. J., Mitchell, T. J., and Wynn, H. P. (1989b). Design and Analysis of Computer Experiments. *Statistical Science*, 4(4) :409–423.
- Saltelli, A., Chan, K., and Scott, E. M. (2000). *Sensitivity analysis*. J Wiley and Sons.
- Santner, T., Williams, B., and Notz, W. (2003). *The design and analysis of computer experiments*. Springer. Springer Series in Statistics.
- Schwarz, G. (1978). Estimation the dimension of a model. *The Annals of Statistics*, 6(2) :461–464.
- Shao, J. (1996). Bootstrap Model Selection. *Journal of the American Statistical Association*., 91(434) :655–665.
- Silly-Carette, J. (2008). *Modélisation avancée de l'absorption des ondes électromagnétiques dans les tissus biologiques*. PhD thesis, Université de Paris 6.
- Sklar, A. (1959). Fonctions de répartition à n dimensions et leurs marges. *Publications de l'Institut de Statistique de l'Université de Paris*, pages 229–231.
- Sklar, A. (1973). Random variables, joint distribution functions and copulas. *Kibernetika*., 9(6) :449–460.
- Sobol', I. M. (2001). Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates. *Mathematics and Computers in Simulation*, 55 :271–280.
- Soize, C. and Ghanem, R. (2004). Physical Systems with Random Uncertainties : Chaos Representations with Arbitrary Probability Measure. *SIAM Journal on Scientific Computing*, 26(2) :395–410.
- Srinivas, N., Krause, A., Kakade, S., and Seeger, M. (2012). Information-theoretic regret bounds for gaussian process optimization in the bandit setting. *IEEE Transactions on Information Theory*, 56(5) :3250–3265.
- Sudret, B. (2008). Global sensitivity analysis using polynomial chaos expansions. *Reliability Engineering & System Safety*, 93(7) :964–979.

- Taflove, A. and Hagness, S. C. (2005). *Computational Electrodynamics. The Finite-Difference Time Domain Method*. Artech House, Boston, 3rd edition.
- Tang, B. (1998). Selecting latin hypercubes using correlation criteria. *Statistica Sinica*, 8 :965–977.
- TNS-SOFRES (2014). Etude Connected Life. <http://www.tns-sofres.com/connectedlife>.
- Vazquez, E. (2005). *Modélisation comportementale de systèmes non-linéaires multivariables par méthodes à noyaux et applications*. PhD thesis, Université Paris XI - UFR scientifique d'Orsay.
- Vazquez, E. and Bect, J. (2010). Convergence properties of the expected improvement algorithm with fixed mean and covariance functions. *Journal of Statistical Planning and Inference*, 140(11) :3088–3095.
- Wang, G. (2003). Adaptive response surface method using inherited latin hypercube design points. *J. Mech. Des.*, 125 :210–220.
- Wiener, N. (1938). The Homogeneous Chaos. *American Journal of Mathematics*, 60(4) :897–936.
- Xiong, S., Qian, P., and Wu, C. (2013). Sequential design and analysis of high-accuracy and low-accuracy computer codes. *Technometrics*, 55(1) :37–46.
- Xiu, D. and Karniadakis, G. E. M. (2002). The wiener-asky polynomial chaos for stochastic differential equations. *SIAM Journal on Scientific Computing*, 24(2) :619–644.
- Xu, X. and Eckerman, K. (2010). *Handbook of Anatomical Models for Radiation Dosimetry*. Taylor and Francis Group LLC, Boca Raton, FL, USA.
- Yee, K. (1966). Numerical solution of initial boundary value problems involving Maxwell's equations in isotropic media. *IEEE Transactions on Antennas Propagation*, 14(3) :302–307.

# Publications personnelles

## Articles de revues avec comité de lecture

Kersaudy, P., Sudret, B., Varsier, N., Picon, O., and Wiart, J. (2015). A new surrogate modeling technique combining Kriging and polynomial chaos expansions – Application to uncertainty analysis in computational dosimetry. *Journal of Computational Physics*, 286 :103–117.

Kersaudy, Mostarshedi, S., P., Sudret, B., Picon, O., and Wiart, J. (2014). Stochastic Analysis of Scattered Field by Building Facades Using Polynomial Chaos. *IEEE Transactions on Antenna Propagation*, 62(12) : 6382–6393.

Arya, R.K., Kersaudy, P., Wiart, J., and Mittra, R. (2015). Statistical Analysis of Periodic Structures and Frequency Selective Surfaces using Polynomial Chaos Expansions. *FERMAT*, 12.

## Conférences

Kersaudy, P., Gati, A., Hadjem, A., Varsier, N., Conil, E., Picon, O., and Wiart, J. (2014). Equivalence principle use to assess the exposure induced by a Femto antenna at 2100MHz. *8th European Conference on Antenna and Propagation (EuCAP)*, pages 2462-2463.

Kersaudy, P., S., Sudret, B., Picon, O., and Wiart, J. (2014). Use of Polynomials of Chaos as Regression Functions for Universal Kriging Models and Application to Numerical Dosimetry. *SIAM Annual Meeting 2014*.

Kersaudy, P., Mostarshedi, S., Sudret, B., Picon, O., and Wiart, (2015). Statistical Analysis of Scattered Field by Building Facades. *URSI AT-RASC 2015*.

Kersaudy, P., Hadjem, A., Varsier, N., Mostarshedi, S., Picon, O., and Wiart, (2015). Statistical Analysis of the Influence of the Morphology onto the Whole-Body Exposure to Radio-Frequency Waves. *URSI AT-RASC 2015*.

Kersaudy, P., Hadjem, A., Mostarshedi, S., Picon, O., and Wiart, J. (2015). Modélisation de l'exposition induite par une source de type station de base de femtocell. *19èmes Journées Nationales Microondes*.





---

## Résumé

---

L'objectif de cette thèse est de traiter la problématique de la caractérisation et du traitement de la variabilité de l'exposition humaine aux ondes radio à travers l'utilisation de la dosimétrie numérique. En effet, si les progrès dans le domaine du calcul hautes performances ont contribué à significativement réduire les temps de simulation pour l'évaluation de l'exposition humaine, ce calcul du débit d'absorption spécifique reste un processus coûteux en temps. Avec la grande variabilité des usages, cette contrainte fait que la prise en compte de l'influence de paramètres d'entrée aléatoires sur l'exposition ne peut se faire par des méthodes classiques telles que les simulations de Monte Carlo.

Nous proposons dans ces travaux deux approches pour répondre à cette problématique. La première s'appuie sur l'utilisation et l'hybridation de méthodes de construction de modèles de substitution afin d'étudier l'influence globale des paramètres d'entrée. La deuxième vise à l'évaluation efficace et parcimonieuse des quantiles à 95% des distributions de sortie et s'appuie sur le développement d'une méthode de planification d'expériences adaptative et orientée couplée à la construction de modèles de substitution. Les méthodes proposées dans ce mémoire sont comparées et testées sur des exemples analytiques et ensuite appliquées à des problèmes concrets issus de la dosimétrie numérique.

---

## Abstract

---

The purpose of this thesis is to deal with the problem of the management and the characterization of the variability of the human exposure to radio-frequency waves through the use of the numerical dosimetry. As a matter of fact, if the recent advances in the high performance computing domain led to reduce significantly the simulation duration for the evaluation of the human exposure, this computation of the specific absorption rate remains a time-consuming process. With the variability of the usage, this constraint does not allow the analysis of the influence of random input parameters on the exposure to be achieved with classical approaches such as Monte Carlo simulations.

In this work, two approaches are proposed to address this problem. The first one is based on the use and the hybridization of construction methods of surrogate models in order to study the global influence of the input parameters. The second one aims at assessing efficiently the 95th-percentiles of the output distributions in a parsimonious way. It is based on the development of an adaptive and oriented methodology of design of experiments combined with the construction of surrogate models. In this manuscript, the proposed methods are compared and tested on analytical examples and then applied to full-scale problems from the numerical dosimetry.