

Université Jean Monnet

**MÉMOIRE D'HABILITATION À DIRIGER DES
RECHERCHES**

DISCIPLINE : MATHÉMATIQUES ET LEURS INTERACTIONS

présenté le 14 mai 2007

par Laurent CARRARO

Professeur à l'Ecole des mines de Saint-Etienne

Modèles probabilistes spatiaux pour l'ingénierie

Membres du jury :

M. Mario AHUES, Professeur à l'Université Jean Monnet, examinateur

M. Anestis ANTONIADIS, Professeur à l'Université Joseph Fourier, rapporteur

M. Georges OPPENHEIM, Professeur à l'Université de Marne-la-Vallée, rapporteur

M. Grigory PANASENKO, Professeur à l'Université Jean Monnet, examinateur

M. Michel SCHMITT, Professeur à l'Ecole des mines de Paris, rapporteur

TABLE DES MATIÈRES

Table des matières.....	2
Introduction.....	3
Des mathématiques pures à la création d'un laboratoire de mathématiques appliquées dans une école d'ingénieurs.....	4
Modèles probabilistes spatiaux pour l'ingénierie.....	9
A) Des percées applicatives.....	9
1. Techniques probabilistes et statistiques en synthèse d'images.....	9
2. Modèles probabilistes pour les aciers moulés.....	10
3. Modèles probabilistes en cinétique chimique hétérogène.....	11
B) Des avancées méthodologiques.....	13
1. Les solutions statistiques intrinsèques de l'équation de Burgers.....	13
2. L'équation de Bachelier et le problème de Skorokhod.....	15
3. Propagation d'incertitudes – la MCV-optimalité.....	16
C) Le thème « computer experiments ».....	17
1. Génèse.....	17
2. Evaluation de plans d'expériences numériques.....	18
3. Construction de métamodèles - l'approche géostatistique.....	19
Conclusion et perspectives.....	22
1.La MCV-optimalité.....	22
2.Valorisation d'informations a priori.....	24
Bibliographie.....	26

INTRODUCTION

Le présent document constitue mon mémoire d'habilitation à diriger des recherches. Il retrace mon activité scientifique, bilan structuré et perspectives, mais s'attache également à décrire mon parcours depuis le début des années 1990 suite à mon intégration à l'Ecole Nationale Supérieure des Mines de Saint-Etienne (ENSM.SE). Cela peut sembler étonnant dans un mémoire d'habilitation à diriger des recherches, mais mon goût pour les questions de qualité me fait donner une place privilégiée aux questions de processus, même si les produits en tant que tels gardent une importance essentielle. Et mes travaux de recherche ne sont évidemment pas indépendants du contexte dans lequel ils ont été conduits.

Ce document est structuré en 2 sections. La première est consacrée à ma trajectoire personnelle, à mon passage du statut de mathématicien à celui de mathématicien appliqué, et décrit la manière dont j'ai mis sur pied, puis développé, une équipe d'enseignement et de recherche à l'ENSM.SE, actuellement département 3MI (Méthodes et Modèles Mathématiques pour l'Industrie).

La deuxième section constitue le coeur scientifique de ce mémoire. Elle est elle-même structurée en 3 parties, la première consacrée à mes travaux à forte connotation applicative, la seconde à mes activités scientifiques situées en amont des applications précitées, et la troisième au thème de recherche que j'ai fait émerger au sein du département 3MI, qui est devenu depuis bientôt 2 ans le thème principal d'activité de l'équipe.

Le document s'achève sur une conclusion générale et des perspectives, avant une bibliographie sélective.

SECTION I

DES MATHÉMATIQUES PURES À LA CRÉATION D'UN LABORATOIRE DE MATHÉMATIQUES APPLIQUÉES DANS UNE ÉCOLE D'INGÉNIEURS

Il est assez fréquent dans la communauté mathématique de nier la différence entre mathématiques fondamentales et mathématiques appliquées. Et il est vrai que lorsque l'on réfléchit par exemple aux propriétés mathématiques d'un modèle probabiliste utilisé en synthèse d'images, l'activité développée est de nature strictement mathématique et aucune différence de fond ne peut raisonnablement être mise en évidence alors entre ce que font le mathématicien appliqué et le mathématicien dit fondamental. Pour autant, l'activité d'un mathématicien appliqué ne se limite pas, loin s'en faut, à la résolution de problèmes mathématiques déjà formulés ou dont l'énoncé est assez largement précisé puisque, déjà, avant de résoudre un problème, il faut s'attacher à le poser. Et mon expérience m'a montré que la réalité de cet exercice est fort loin d'être une simple figure de rhétorique, car cette phase de compréhension et de mise en forme d'un problème est souvent longue, complexe. D'ailleurs, pour être tout à fait honnête, il n'est pas rare lors d'une action liée à un travail de doctorat, que la compréhension en profondeur des questions à résoudre doit attendre la fin de la thèse, voire au-delà ! Et les seules qualités scientifiques - essentielles - ne sont pas suffisantes et ne peuvent exercer leur pleine efficacité que par le biais de qualités humaines (dialogue, respect de l'interlocuteur, curiosité...) qui agissent alors comme des catalyseurs¹.

Ma formation est celle d'un mathématicien, sans aucun souci applicatif². Pourtant, ma thèse porte sur les processus stochastiques qui semblent à première vue un sujet d'étude assez proche d'applications. Mais la problématique est alors celle de l'étude d'objets définis par ailleurs dont, comme à l'habitude, de nombreuses propriétés restent à éclaircir et dont l'étude me conduit d'ailleurs vers l'analyse harmonique sur les groupes. J'aperçois alors quelques liens avec la statistique non paramétrique mais n'ai ni le temps ni les bases pour les comprendre, et encore davantage pour les exploiter. On notera sur ce point que je découvre le calcul des probabilités très tardivement à travers un seul cours de

¹Les collègues chimistes savent bien que dans la plupart des cas, sans catalyseur, il n'y a pas de réaction.

²Sur ce point, je me souviens du plaisir et de l'émotion intenses procurés par les aspects totalement abstraits de mes études de mathématiques, marquées du sceau profond de l'école de pensée Bourbaki.

DEA consacré aux processus de Lévy, avec une vision très fonctionnelle (décomposition spectrale des semi-groupes).

Sans même imaginer ce que peut être une application réelle des mathématiques, j'ai à cette époque le goût, que j'ai conservé, pour la confrontation de disciplines variées. Il s'agit alors de disciplines ressortant toutes du domaine mathématique mais mon plaisir est grand lorsque je m'aperçois que telle approche topologique fournit des propriétés algébriques pour un sous-groupe du groupe linéaire, ou que la transformée de Young d'une fonction convexe – qui mène directement aux inégalités de Chernoff, germes des théorèmes de grandes déviations – consiste simplement à paramétrer une fonction par les hyperplans tangents à son graphe, et donc par les points de l'espace projectif à travers la dualité bien connue dans ce contexte. Ces moments de connexion entre des points de vue différents correspondent à des instants où je perçois profondément l'activité mathématique comme un art où la notion de beauté est prépondérante.

Le monde mathématique est vaste, immensément vaste. Pourtant, après avoir exercé en tant qu'ATER entre 1987 et 1989 à l'Université Claude Bernard, puis en tant qu'agrégé préparateur en 1989-1990 à l'ENS Lyon, je ressens une certaine frustration du fait que mes interlocuteurs parlent le même langage, et ont des façons de penser qui se ressemblent, ce qui ne satisfait pas ma curiosité naturelle. Par ailleurs, je suis passionné par l'enseignement et les lieux où j'exerce pendant ces années me semblent peu propices pour assumer cette passion. Sans revenir sur le tristement célèbre statut des enseignants-chercheurs décrit par le décret de 1984³, je me contenterai de noter que l'on parle constamment de *charge* d'enseignement comme si cette activité pesait d'un poids trop important sur les épaules des enseignants-chercheurs. On comprendra donc que je cherche, et ceci dès 1988, un lieu où je puisse d'une part étendre mes contacts scientifiques hors de la sphère strictement mathématique et où d'autre part la formation semble davantage valorisée. C'est ainsi qu'à la seconde tentative, je rejoins l'École des mines de Saint-Etienne, à partir de l'automne 1990 en y occupant un poste de maître-assistant. Ceci sans vraiment savoir ce que peut être une école d'ingénieurs, milieu dont j'ignore tout, mais avec l'espoir – non déçu par la suite – que cet environnement entre davantage en résonance avec mes goûts.

L'intégration n'est pas des plus faciles pour un individu comme moi qui ignorait presque, deux années plus tôt, l'existence même de ce type d'écoles, et tous mes repères sont ainsi altérés. Tout d'abord, je fais partie d'une « équipe » qui ne comporte que deux mathématiciens et mon seul collègue J.M. Pla, professeur, a cessé toute activité de recherche depuis longtemps. De ce fait, la grande majorité de mes interlocuteurs n'est pas constituée de mathématiciens et je comprends à grand peine la nature de leurs activités, d'une variété extrême au regard de mon passé (géologues, gestionnaires, informaticiens, chimistes, métallurgistes...). Par ailleurs, mes débuts en enseignement ne sont pas très concluants. Je choisis en effet de présenter aux élèves-ingénieurs le cours de calcul

³Je garde l'espoir et milite pour que cette erreur stratégique qui enferme les Universités et leurs personnels dans une activité uniforme, fort éloignée des besoins de flexibilité du temps présent, soit corrigée suffisamment tôt.

intégral que j'ai assuré l'année précédente aux normaliens, partant du principe que ces deux publics ont passé un concours analogue... Autant dire que le succès n'est pas au rendez-vous alors que je leur décris par exemple de manière très précise les problèmes de prolongement de mesures et que mon discours est entièrement centré sur la maîtrise des concepts de la théorie de l'intégration !

Mais cette situation inconfortable me satisfait et je me donne à l'époque comme défi de comprendre au moins superficiellement l'activité de mes collègues, et surtout de rencontrer un succès raisonnable pour ce qui concerne mes enseignements. La route est longue mais mes deux objectifs principaux sont convergents puisqu'ils sont tous deux liés à la compréhension de la place des mathématiques dans les sciences de l'ingénieur. C'est la tâche à laquelle je m'attelle avec enthousiasme dès 1990. Pour ce faire, je multiplie les contacts avec les collègues de l'école et encadre le plus grand nombre possible d'étudiants en stage dans les entreprises. Après un an d'efforts, un premier bilan s'impose, très contrasté. D'une part les mathématiques nécessaires en sciences de l'ingénieur recouvrent un champ extrêmement vaste, très proche de celui occupé par les mathématiques pures. A contrario, les mathématiques utilisées par les ingénieurs, ou même les scientifiques rencontrés à l'école, sont souvent d'un niveau très élémentaire alors que quelques outils d'actualité leur permettraient de faire des progrès substantiels. Compte tenu de mon passé scientifique, je décide d'approfondir mes connaissances en processus aléatoires, dont j'ai pu apercevoir qu'ils s'appliquaient dans des secteurs très variés (synthèse d'images, traitement d'images, matériaux, géologie...), en étudiant particulièrement les aspects inférentiels, dont j'ignore encore tout. Je rencontre là le monde statistique, découverte essentielle, et profite de toutes les occasions pour progresser. Je relis par exemple en détail le projet de livre « Régression non linéaire » de A. Antoniadis, J. Berruyer, R. Carmona, publié ensuite chez Economica [Antoniadis 1992].

Des premières collaborations se concrétisent : je co-encadre le stage de DEA de J.L. Maillot en synthèse d'images en 1992. Il s'inscrit en thèse à l'automne de cette même année et je co-encadre son travail. Par ailleurs des contacts pris avec EDF Septen dès 1992, avec le premier contrat auquel je participe, se transforment à l'automne 1993 en proposition de thèse sur la modélisation et la caractérisation des aciers austéno-ferritiques. L'équipe de mathématiques appliquées de l'école s'adjoint donc son premier doctorant, Laure Cadet-Messiaen, financée par EDF, inscrite en thèse de morphologie mathématique à l'École des mines de Paris. Je parviens également cette même année à recruter sur un poste d'attaché de recherche agrégé vacant un jeune agrégé, Xavier Bay, qui s'inscrit en thèse à l'Université Joseph Fourier sous la direction de A. Antoniadis. De ce fait, l'équipe est alors constituée dès l'automne 1993 d'un professeur, un maître-assistant, et deux doctorants, et j'en prends la responsabilité avec le soutien total de J.M. Pla.

Les années qui suivent me voient poursuivre mes efforts pour accroître et stabiliser les activités de l'équipe, tâche délicate dans la mesure où les besoins en enseignement sont importants puisque tous les enseignements de mathématiques réalisés à l'école le sont peu à peu par les membres de l'équipe. Nous adoptons pour ce faire une démarche pragmatique, dégagée autant que nos passés le permettent d'a priori sur ce qu'il convient ou non d'enseigner. Et les enseignements changent ainsi profondément à la fois au niveau

fond (disparition progressive des enseignements de tronc commun consacrés à l'intégration ou à l'analyse fonctionnelle) mais aussi et surtout au niveau de la forme avec l'utilisation intensive de l'ordinateur comme moyen expérimental pour comprendre les mathématiques de l'ingénieur ou l'utilisation de pédagogies inductives (mises en situation, apprentissage par problèmes...). Dans ces conditions, je parviens à obtenir un poste de maître-assistant, occupé à l'automne 1994 par F. Beaucoup qui prend en charge toute la partie analyse numérique du cursus. Pour autant, ces activités sont très lourdes et il est difficile en l'état de répondre convenablement aux besoins. Pour illustrer ce point, on peut noter que pendant l'année 1995/1996, je totalise 307 h équivalent TD, alors que F. Beaucoup en effectue 254, J.M. Pla 179, X. Bay 132 et L. Messiaen 85⁴. Paradoxalement, les renforts obtenus par l'équipe se traduisent par une activité d'enseignement par individu qui augmente. Ceci du fait que les collègues d'autres disciplines souhaitent que les mathématiciens de l'école soient davantage présents dans les enseignements dispensés ; succès indéniable dont les conséquences sont difficiles à maîtriser au quotidien. Il faut dire que notre enthousiasme nous fait prendre des risques peu maîtrisés. Nous lançons par exemple les projets de mathématiques réalisés par la promotion entière, en relations avec les centres de recherche de l'école, alors que nous sommes seulement 3 permanents !

Nous parvenons à stabiliser la position de X. Bay lorsqu'il termine sa thèse en 1997 et recrutons également sur ressources propres L. Messiaen. Cette période voit le département assurer un chiffre d'affaires contractuel conséquent et assez constant lui permettant de profiter d'opportunités en termes de recrutement de personnel⁵, et cet aspect ne changera plus depuis. Mais il est difficile sous la pression des activités de formation de stabiliser scientifiquement les thèmes abordés. C'est ainsi que X. Bay se réoriente sur des aspects mathématiques financières du fait de la demande des élèves et que L. Messiaen prend en charge un nouvel axe de deuxième année consacré aux techniques statistiques. Je continue pour autant mes activités de recherche tout en démarrant de nouvelles collaborations industrielles, avec l'IPSN et Elf notamment, qui s'avéreront très fructueuses par la suite. Et l'identité de l'équipe est maintenant stabilisée autour du tryptique formation, industrie, recherche, même si l'origine universitaire majoritaire des membres mènent à insister particulièrement sur les aspects industriels.

Il faut attendre l'automne 1999 pour que les effectifs commencent à être à la hauteur des activités déployées. On compte en effet à cette date 4 permanents dans l'équipe (dont un sur ressources propres) et deux post-docs. Cette situation permet de relancer l'activité de recherche de l'équipe. Nous nous attachons en effet les services de O. Roustant qui prépare en septembre 1999 une thèse de l'Université Claude Bernard sous la direction de J.P. Laurent, puis nous recrutons un nouveau doctorant en la personne de C. Helbert, en lien avec le projet Infotherm qui regroupe universitaires et industriels ; elle prépare une thèse de l'ENSM.SE sous ma direction à partir de l'automne 2000⁶.

⁴Rappelons que X. Bay et L. Messiaen sont doctorants.

⁵Nous avons ainsi en 1998 2 CDD et 1 CDI sur Armines, pour 3 permanents.

⁶Je suis nommé professeur en mai 2000 et peut à ce titre diriger une thèse de doctorat.

Mes responsabilités administratives (directeur du centre SIMMO en 2000-2001, puis directeur adjoint de l'école entre 2001 et 2005) m'obligent à ralentir mon activité scientifique. Néanmoins, l'équipe est lancée avec des activités d'enseignement plébiscitées par tous, administration de l'école, collègues, étudiants. Les activités de recherche commencent elles aussi à être reconnues, tout en ayant une tendance à la dispersion compte tenu de la multiplicité des interlocuteurs, en particulier au niveau industriel. Avant de retrouver à temps plein l'équipe courant 2005, je fais émerger le thème de recherche « computer experiments » et négocie dans ce cadre le détachement à l'école de B. Corre, ingénieur senior de Total, à partir d'avril 2005. Nous recrutons alors 3 nouveaux doctorants, permettant à l'équipe de refocaliser ses activités scientifiques sur un thème commun, et d'atteindre ainsi une masse critique et une visibilité nationale sur ce thème.

SECTION II

MODÈLES PROBABILISTES SPATIAUX POUR L'INGÉNIERIE

A) DES PERCÉES APPLICATIVES

Les trois exemples qui suivent sont relatifs à des thèses de doctorat que j'ai encadrées, qui sont bi-disciplinaires et dont, par suite, le jury et les rapporteurs ont rassemblé mathématiciens et spécialistes du domaine concerné.

1. Techniques probabilistes et statistiques en synthèse d'images

Ce travail de thèse correspond à mon premier co-encadrement, et concerne l'utilisation des techniques de tracé de rayons pour la création d'images de synthèse. Deux contributions principales sont apportées dans ce travail.

L'équation donnant la luminance en chaque point d'une scène, appelée équation de rendue, n'est explicitée qu'en 1986 [Kajiya 1986]. Cette équation est hautement non linéaire et l'une des techniques de résolution, le tracé de rayons, consiste à lancer des rayons lumineux dans la scène et à observer leurs rebonds sur les objets présents. Cette technique est assez populaire en 1992 lorsque J. L. Maillot débute sa thèse. Par contre, aucune écriture probabiliste formelle n'est à cette époque proposée. La première avancée réalisée consiste à expliciter l'algorithme sous la forme d'une chaîne de Markov dont l'espace d'état est $S \times \Omega$, où S représente la surface de l'ensemble des objets et Ω l'ensemble des directions orientées de rayons. Dans cette modélisation, l'état de la chaîne au « temps n » n'est autre que la luminance et la direction d'un rayon lumineux après n rebonds sur des objets. Par ce biais, J.L. Maillot montre que la luminance solution de l'équation de rendu s'explique simplement en fonction de cette chaîne qui est naturellement arrêtée (le rayon peut en effet être absorbé, tomber sur une source lumineuse, ou disparaître sans toucher d'objet). L'expression obtenue est alors la première à être explicitée, et une stratégie de simulations de Monte Carlo est construite pour aborder sur des bases solides le tracé de rayons. Les techniques de réduction de variance, et en premier lieu l'importance sampling adaptatif, y jouent un rôle crucial.

La seconde contribution concerne le tracé de rayons progressif, introduit par J. L. Maillot. Il s'agit de mettre au point un algorithme de détection de zones homogènes dans une image en construction de manière à concentrer les efforts de calcul sur des zones comportant de très fortes variations de luminance. Compte tenu du contexte, nous proposons une méthode basée sur la statistique séquentielle de Wald (voir [Wald 1948] et

[Siegmond 1985]). Cette approche est particulièrement adaptée ici puisque qu'elle consiste à fixer par avance les niveaux des risques de première et de seconde espèce et à continuer l'échantillonnage, et donc le lancer de rayons, tant que la décision n'a pas pu être prise. De manière plus précise, le test utilisé est celui du rapport séquentiel de vraisemblances, avec des lois hypergéométriques qui ne peuvent être approximées dans ce cadre par des lois binomiales.

La seconde partie des résultats obtenu a été publiée dans [Maillot 1992], alors que la première n'a malheureusement pas pu l'être, J. L. Maillot ayant terminé son travail de doctorat [Maillot 1996] tout en étant ingénieur chez Dassault Systèmes. C'est la raison pour laquelle la seconde partie de ses travaux a eu un impact beaucoup plus important dans la littérature du domaine que la première, pourtant à mes yeux et aux yeux de son jury de thèse plus innovante.

2. Modèles probabilistes pour les aciers moulés

L'objet initial de la thèse consiste à confirmer, ou infirmer, le lien entre certaines propriétés mécaniques d'aciers biphasés et les propriétés morphologiques des deux phases. La question est d'importance pour EDF à une époque où se pose la question de la prolongation de la durée de vie des centrales nucléaires. Les pièces observées sont les coudes moulés du circuit primaire. Compte tenu de la complexité topologique des deux phases, la démarche consiste à définir un modèle probabiliste décrivant la morphologie de l'acier, à en inférer les paramètres au moyen de techniques d'analyse d'images, et à examiner le lien entre les paramètres du modèle et les propriétés mécaniques considérées (essentiellement résilience).

Outre la démarche, rendue presque indispensable par le fait supplémentaire que les images proviennent au mieux de coupes successives très difficiles à recoller, l'apport essentiel du travail consiste en l'introduction d'un modèle probabiliste simple mais pour autant réaliste qui donne des résultats proches des observations, pourtant très variables dans leur morphologie.

Le modèle est construit par analogie avec le processus de refroidissement après moulage, qui voit l'apparition à l'état solide d'austénite dans une matrice de ferrite. Et les observations faites montrent qu'un effet de répulsion existe entre les deux phases. Le modèle est construit à partir d'une chaîne de Markov homogène dont l'espace d'état est $E = \{0,1\}^{\mathbb{Z}^3}$ et dont la dynamique est formée des évolutions indépendantes de chaque site de \mathbb{Z}^3 (ici, l'état 1 pour le site x représente la phase austénitique). En d'autres termes, conditionnellement au passé, chaque site évolue indépendamment et la dynamique de la chaîne est définie par la seule fonction :

$$(1) \quad e \in E \rightarrow p_e(x) = P(X_{n+1}(x) = 1 / X_n = e) \in [0,1]$$

Et on suppose évidemment que la fonction qui précède est invariante par translation.

L'application des critères classiques d'ergodicité (par exemple [Templeman 1972]) permet alors, en lien avec des techniques de classes monotones, d'établir que la dynamique choisie transforme, en toute généralité, un processus stationnaire et ergodique en un

processus stationnaire et ergodique. On déduit de tout cela que pour tout instant n , le processus X_n indexé par \mathbb{Z}^3 est stationnaire et ergodique. Par ailleurs, il n'est pas difficile de voir que la chaîne converge vers le processus constant égal à 1 (austénite en chaque site). C'est la raison pour laquelle on arrête la chaîne (analogue de la trempe des métallurgistes) au premier instant N où une proportion p d'austénite est atteinte. Ainsi, un modèle de l'acier biphasé n'est autre que le processus $X = X_N$. Les propriétés d'ergodicité précédentes montrent que le temps d'arrêt N est en fait déterministe, et donc que le processus X lui-même est également stationnaire et ergodique. L'inférence est donc possible et a été réalisée comme indiqué ci-avant à partir d'analyses d'images bi-dimensionnelles.

On notera pour terminer que ces résultats auraient pu découler de résultats généraux sur les systèmes de particules ou d'automates cellulaires (par exemple [Holley 1972]) si la probabilité $p_e(x)$ dépendait de e par un voisinage de x . Ce n'a pas été notre choix du fait que l'effet de répulsion a été construit à partir de la distance entre x et le plus proche voisin austénitique, distance non bornée.

Je citerai simplement, pour terminer sur ce sujet, le commentaire du rapporteur métallurgiste, Y. Bréchet. « Le travail de Mme Messiaen est novateur et bien conduit et ouvre des perspectives intéressantes pour la modélisation de la mécanique et de la rupture de structures à morphologie complexe. ».

3. Modèles probabilistes en cinétique chimique hétérogène

Ce sujet a émergé à partir de discussions avec le département PROCESS (Procédés et Evolution des Systèmes avec Solides) de l'ENSM.SE qui concernaient la transformation thermique de solides. Le point de départ a été une réflexion sur l'interprétation de résultats expérimentaux concernant le temps de latence avant germination pour des grains solides. Un fait notable est qu'un gros grain se comporte de manière différente d'une population de petits grains de même surface. Les germes sont conséquences de lacunes dans le réseau cristallographique, dues aux conditions ambiantes (détachement de molécules du fait de la température par exemple), et impliquent des changements locaux de phase. De ce fait, un modèle probabiliste m'a semblé adapté et des discussions nombreuses avec mes interlocuteurs m'ont mené à proposer un modèle basé sur une distribution poissonnienne des germes à la surface du grain. Par une loi des grands nombres adaptée (il s'agit ici de max et non de moyenne, d'où l'apparition de lois max-stables), je montre alors que le temps de latence d'un gros grain doit suivre une loi de Weibull. Et la confrontation avec les expériences disponibles confirme, ou plutôt n'infirme pas⁷ comme dans toute décision statistique, le modèle.

Dans ces conditions, les collègues chimistes souhaitent approfondir la réflexion et me signalent qu'un modèle probabiliste qui ressemble à celui que j'ai construit existe dans leur communauté, le modèle de Mampel [Mampel 1940]. Le raisonnement développé par Mampel à l'époque, et jamais vraiment repris par les chimistes depuis, est très difficile à

⁷Cette remarque fait écho à la faible puissance du test due au très petit nombre d'observations.

suivre du fait d'hypothèses probabilistes peu explicites et d'un raisonnement sur un grain moyen peu formalisé. Ce modèle ne suffit plus aux besoins de mes interlocuteurs car il suppose des grains sphériques et des conditions isothermes et isobares. Or, des expériences menées depuis plusieurs années montrent que les géométries des grains sont très variables (cylindres, plaquettes...). De plus, la stabilité des conditions stoechiométriques rend impossible l'intégration de ce modèle pour des simulations de fours industriels où pression et température varient dans le temps et l'espace. Car il s'agit de coupler, pour de tels développements informatiques, les aspects thermiques et cinétiques.

Je reprends intégralement la modélisation pour me convaincre qu'il est possible de retrouver les résultats de Mampel par une approche probabiliste explicite et lance une thèse de doctorat sur le sujet, thèse réalisée par C. Helbert. Elle obtient ainsi une modélisation d'une population de grains dont les hypothèses sont les suivantes.

- Les grains évoluent de façon indépendante les uns des autres. Ils passent d'une phase A à une phase B par apparition d'un germe de la phase B à la surface de A puis croissance de ce germe.

- La germination qui se produit à la surface S du grain est stochastique, selon un processus spatio-temporel de Poisson $(N(F))_F$ d'intensité :

$$(2) \quad n(F) = \iint_{(t,\sigma) \in F} \gamma(t) dt d\sigma$$

où F est un borélien de l'espace $[0, +\infty[\times S$ et $d\sigma$ la mesure d'aire sur S. La fonction γ est alors appelée fréquence surfacique de germination.

- La croissance d'un germe après son apparition est linéaire dans le temps, selon l'équation :

$$(3) \quad \frac{dV}{dt} = \varphi(t) V_{m_A} S$$

où V_{m_A} est le volume molaire de la phase initiale A, V est le volume occupé par la phase B au temps t et $\varphi(t)$ est la réactivité de croissance.

On définit alors pour tout temps t, tout point x du grain, et tout point σ de la surface S du grain, l'instant $\tau(x, \sigma, t)$ comme étant l'instant de naissance auquel un germe né en σ atteint le point x au temps t. De cette façon, en posant :

$$(4) \quad S_{t,x} = \{(u, \sigma) \in [0, T[\times S, u \leq \tau(x, \sigma, t)\}$$

On obtient aisément la formule qui suit, donnant la proportion $\beta(t)$ d'un grain transformée en phase B au temps t :

$$(5) \quad \beta(t) = 1 - \frac{1}{V} \int_V 1_{N(S_{t,x})=0} dx$$

Sous les hypothèses d'indépendance entre grains, cette expression mène très simplement au volume d'une poudre transformée, et ceci quelle que soit la géométrie des grains, leur variation de forme (on intègre facilement des distributions granulométriques), et les conditions stoechiométriques. Les expressions obtenues se prêtent évidemment à des évaluations numériques par des méthodes de Monte Carlo et permettent surtout de construire un code générique permettant la modélisation de transformations variées. Ces travaux réalisés par C. Helbert, comprenant le traitement d'expériences de laboratoire via

une analyse détaillée de la structure des bruits sous-jacents, ont été publiés dans [Helbert 2004]. Ce travail a été également présenté aux journées annuelles de cinétique hétérogène [Helbert 2002] et a reçu le prix Jean Besson, qui récompense le meilleur travail de doctorat de l'année. Mais son apport dans le cadre de sa thèse [Helbert 2005] ne s'est pas réduit à cette modélisation puisqu'elle a ouvert de nouvelles voies en revisitant les hypothèses faites pour le processus de germination via l'introduction d'une chaîne de Markov d'espace d'état l'ensemble des fonctions de la surface du grain à valeurs dans un espace discret, modèle analogue à celui introduit dans [Messiaen 1997].

Je me permets de citer ici un extrait du commentaire du rapporteur statisticien, D. Serant. « Ce modèle présente une avancée importante relativement aux travaux antérieurs [...]. L'élégance du modèle [...] est d'oublier le caractère physiquement markovien du phénomène de germination. » Et je ne résiste pas au commentaire final du rapporteur chimiste, F. Berthier, qui comble mes convictions sur la pluridisciplinarité. « Il est important de noter que ces thèses multidisciplinaires sont exigeantes. Peut-être est-ce la raison pour laquelle elles sont si rares... Il faut espérer que ce système de collaboration se généralise au sein de la recherche française. Une telle thèse réclame de la doctorante une grande ouverture d'esprit, beaucoup de culture et de recul, ce dont Céline Helbert ne manque clairement pas. ».

B) DES AVANCÉES MÉTHODOLOGIQUES

1. Les solutions statistiques intrinsèques de l'équation de Burgers

L'équation de Burgers visqueuse est l'équation suivante :

$$(6) \quad \frac{\partial u}{\partial t}(x, t) + u(x, t) \frac{\partial u}{\partial x}(x, t) = \mu \frac{\partial^2 u}{\partial x^2}(x, t)$$

$$(7) \quad u(x, 0) = f(x)$$

Cette équation a été étudiée de manière intensive depuis les travaux pionniers de [Hopf 1950] et [Cole 1951]. Dans sa version la plus courante, $u(.,t)$ représente le champ de vitesse au temps t d'un fluide visqueux de viscosité μ , avec $\mu > 0$. Cette équation a été introduite initialement comme un modèle mono-dimensionnel simple de turbulence, qui décrit la formation et l'évolution de chocs dans un fluide compressible. Elle admet une solution explicite via la transformation de Hopf-Cole :

$$(8) \quad u = -2\mu \frac{1}{\theta} \frac{\partial \theta}{\partial x}$$

La fonction θ est alors solution de l'équation de la chaleur, et la solution fait donc intervenir la densité du mouvement brownien, fonction de Green du problème.

Les solutions de l'équation de Burgers non visqueuse, i.e. pour $\mu = 0$, s'obtiennent alors par passage à la limite, qui est effectué par une méthode de Laplace. La solution obtenue ainsi est solution d'un problème d'optimisation, et fait intervenir la transformée de Legendre de la primitive de la fonction f , à laquelle est ajoutée une dérive linéaire.

Au début des années 1990, de nombreux travaux de physiciens théoriciens visent à obtenir des solutions de l'équation de Burgers non visqueuse avec une condition initiale f aléatoire. Citons par exemple [She 1992] et [Sinai 1992]. Leur approche consiste à utiliser la transformation de Hopf-Cole et à tenter d'identifier, vis des passages à la limite adéquats, le processus résultant.

Nous abordons avec J. Duchon la question sous un angle radicalement différent. Raisonnant sur la loi au temps t du processus solution, nous montrons directement que le flot de l'équation de Burgers conserve la classe des processus de Lévy à sauts négatifs ; cette dernière condition sur les sauts provenant des conditions d'entropie de Lax. De manière plus précise, seule la loi des accroissements du processus est considérée, c'est la raison pour laquelle nous choisissons de qualifier les solutions de solutions statistiques intrinsèques, en référence aux travaux des géostatisticiens [Matheron 1973]. C'est pour moi la première occasion où je rencontre des lois de probabilités sur des espaces fonctionnels quotient, mais je reviendrai plus tard sur ces questions avec le krigeage. Les résultats obtenus sont annoncés dans [~ 1994], puis publiés de manière complète dans [~ 1998].

Le résultat principal obtenu est particulièrement simple puisque qu'il peut s'énoncer comme suit (voir [Bertoin 1996] pour les notions concernant les processus de Lévy).

Théorème [~ 1998]

Soit $(X(t); t \geq 0)$ une famille de processus de Lévy homogènes de variance finie et à sauts négatifs. On note μ_t la loi des accroissements de $X(t)$ et $\varphi(t, \cdot)$ son exposant de Laplace.

On suppose que pour tout $w \geq 0$, $t \mapsto \varphi(t, w)$ est dérivable de dérivée localement bornée.

Alors la famille $(\mu_t; t \geq 0)$ est solution statistique intrinsèque de l'équation de Burgers si et seulement si son exposant de Laplace satisfait l'équation de Burgers non visqueuse pour $t > 0$ et $w > 0$.

On voit donc que les solutions statistiques intrinsèques de l'équation de Burgers sont obtenues vis des solutions usuelles, non stochastiques, de cette même équation. Il est de ce fait facile de donner ces solutions statistiques explicites à l'équation de Burgers non visqueuse, et nos travaux sont les premiers à le faire. Ils sont très vite utilisés par J. Bertoin, de l'Université de Paris VI, spécialiste des processus de Lévy qui s'intéresse à l'équation de Burgers. Il redémontre [Bertoin 1998] nos résultats dans le cas particulier d'une condition initiale brownienne par la méthode classique découlant de la transformation de Hopf-Cole, avant de lancer deux thèses sur des sujets connexes ([Giraud 2001], [Winkel 2001]). S. Jaffard [Jaffard 1999] intègre nos résultats dans le cadre de ses réflexions sur les relations entre la turbulence et les fonctions multifractales. Son article montre en effet que les processus de Lévy sont multifractals et notre contribution établit ainsi que certaines solutions de l'équation de Burgers le sont aussi.

2. L'équation de Bachelier et le problème de Skorokhod

Suite à des contacts avec P. Crépel et B. Bru, tous deux probabilistes et historiens de sciences, j'entame en 1991 la lecture de l'oeuvre de L. Bachelier. Je débute par sa thèse [Bachelier 1900] dans laquelle il est maintenant bien connu qu'il y définit le mouvement brownien (de trois manières différentes) et utilise le processus ainsi construit pour modéliser le cours de la rente à 3%. Ce travail a été très tardivement reconnu par les communautés mathématique et financière, mais c'est maintenant chose faite puisque le séminaire qui fait référence en France sur les questions de mathématiques financières s'intitule « séminaire Bachelier ». Mais je poursuis mes lectures, assez difficilement, car le formalisme de Bachelier diffère très sensiblement de celui que nous utilisons aujourd'hui et la plupart de ses résultats sont établis au moyen de raisonnements essentiellement littéraires. Pourtant, je découvre des résultats remarquables, en particulier dans ce qui me semble être son chef d'oeuvre, son article sur « les probabilités continues » [Bachelier 1906]. On trouve en effet dans cet article une définition générale des diffusions, et la résolution d'un certain nombre d'équations différentielles stochastiques (EDS), en particulier l'équation de Langevin dont la solution est le célèbre processus d'Ornstein-Uhlenbeck. Pour ce faire, il quitte le monde de la finance et utilise une analogie avec des situations de jeux pour mener à bien ses raisonnements. Dans son vocabulaire, les « conditions du jeu » représentent les coefficients de l'EDS considérée.

Mais Louis Bachelier considère également une EDS qui est demeurée apparemment inconnue de la communauté probabiliste jusqu'à ce que je la rencontre. Il s'agit de l'équation :

$$(9) \quad dY_t = \varphi(\bar{Y}_t) dB_t$$

où $B = (B_t)_t$ est le mouvement brownien standard, φ est une fonction borélienne positive, et $\bar{Y}_t = \sup \{Y_s, 0 \leq s \leq t\}$.

Je résous cette équation en 1993 au moyen du calcul du premier ordre de Azema-Yor (voir par exemple [Revuz 1991]). Je note au passage qu'en remplaçant le mouvement brownien par une martingale locale continue $X = (X_t)_t$ on obtient en résolvant l'équation de Bachelier une formule de dualité entre martingales, analogue à une transformation de Legendre. Je discute de ces résultats avec M. Yor qui les cite dans les éditions suivantes de son livre avec D. Revuz [Revuz 1991]. Je continue dans cette voie et montre que la formule de dualité entre martingales permet de retrouver très simplement la solution de Azema-Yor au problème de plongement de Skorokhod (voir [Azema 1979]) ; je publie ces résultats dans un rapport interne [~ 1995] avant de les oublier, préoccupé alors par d'autres questions (cf. par exemple mes activités d'enseignement évoquées plus haut !). Je retrouve par hasard ce manuscrit au début de l'année 2006 et l'envoie à M. Yor, devenu membre de l'Académie des Sciences, qui le transmet à un de ses élèves, I. Obloj, qui vient de soutenir sa thèse fin 2005 sur le plongement de Skorokhod ([Obloj 2005]). Celui-ci est très intéressé par mon approche et nous décidons de publier ensemble mes résultats, sa contribution consistant essentiellement à tenir compte des nombreuses avancées réalisées depuis 1995. Car des hypothèses faites dans mon article de 1995

peuvent être relâchées ; en particulier, la continuité des trajectoires des martingales considérées n'est pas du tout nécessaire. N. El Karoui et A. Meziou, motivés par des questions de gestion de portefeuille utilisent les résultats obtenus et les précisent en établissant des propriétés d'uniforme intégrabilité dans le cas de certaines transformées de martingales du type précédent. Ces résultats devraient paraître au Séminaire de probabilités [~ 2007-1].

3. Propagation d'incertitudes – la MCV-optimalité

La théorie moderne des plans d'expériences a été forgée comme on le sait par Sir R. Fisher, avec la première publication d'un ouvrage sur le sujet ([Fisher 1935]), basé sur ses travaux de statisticien à la station expérimentale de Rothamsted, l'une des plus anciennes institutions de recherche en agronomie. Cette théorie permet de planifier des expériences de façon à recueillir le maximum d'information et d'interpréter au mieux les résultats obtenus. Mais tout ceci concerne des expériences réelles pour lesquelles les notions de bruit de mesure et de répétabilité permettent de fonder une théorie statistique. Et la théorie s'est considérablement développée depuis, avec en particulier les travaux sur l'optimalité des plans (voir [Fedorov 1972]). Dans le cas d'expériences numériques, c'est-à-dire lorsque une expérience correspond à l'évaluation d'une quantité au moyen d'un outil de calcul numérique, théorie et problèmes changent radicalement. Nous évoquerons plus loin comment les hypothèses habituelles, qui sont celles de la régression, ne peuvent plus être satisfaites. Sans considérer dans un premier temps ce problème fondamental, je me suis appliqué à donner des critères de choix de plans d'expériences numériques qui soient adaptés à un objectif précis, celui de la propagation d'incertitudes.

Le problème de propagation d'incertitudes est facile à comprendre. On utilise un simulateur numérique afin de prévoir le comportement d'un système. Supposons que l'on caractérise ce comportement par une variable, une « réponse » notée y , et que celui-ci soit influencé par d'autres variables résumées dans un vecteur x . On connaît mal dans la réalité la valeur de x , et on caractérise cette méconnaissance par une loi de probabilité pour x . Enfin, y est liée à x par le simulateur :

$$(10) \quad y = f(x)$$

Il s'agit de déterminer la loi de y , connaissant celle de x . Ce problème est très facile à résoudre en pratique si l'on s'autorise un grand nombre d'évaluations de la fonction f , via une méthode de type Monte Carlo. Si ce n'est pas possible, on va chercher à construire une fonction approchée f_{app} à partir d'un petit nombre d'évaluations de la fonction f pour des valeurs de x « bien choisies ». Cette approximation étant construite, on effectuera des simulations de Monte Carlo avec cette fonction approchée pour obtenir une approximation de la loi de y . Un instant de réflexion permet de se convaincre que ce type de problème ne se rencontre pas dans le cas d'expériences réelles, c'est la raison pour laquelle une étude spécifique a dû être menée. Partant d'un modèle linéaire pour la fonction f_{app} , on aboutit à l'équation :

$$(11) \quad f(x) = X(x)\beta + \epsilon(x)$$

où $\epsilon(x)$ représente l'écart au modèle, dont on suppose qu'il s'agit d'un bruit blanc gaussien.

Les paramètres sont estimés à partir d'un plan d'expériences noté \mathbf{X} . Dans ces conditions, on a :

$$(12) \quad f_{app}(x) = X(x)\hat{\beta} + \eta(x)$$

où η est un bruit blanc gaussien d'écart-type celui estimé à partir du plan \mathbf{X} , noté $\hat{\sigma}$.

On vérifie que, sous les hypothèses précédentes⁸, les lois des variables aléatoires⁹ $f(x)$ et $f_{app}(x)$ ont la même espérance et que la variance de $f(x)$ est inférieure à celle de $f_{app}(x)$. Le critère d'optimalité, désigné par MCV, pour Monte Carlo Variance, n'est autre que la différence des deux variances qui précèdent. Il s'interprète comme une distance entre les deux lois de probabilité qui ne dépend que de certains moments de la loi du vecteur x , et est donc plus robuste à des écarts à la loi que des critères plus précis basés par exemple sur des distances de Kullback ou de Hellinger. Des calculs probabilistes montrent que la minimisation du critère équivaut à celle d'un critère de type IMSE, mais l'intérêt de l'approche est ici de construire un critère adapté à l'étude visée qui est celle de la propagation d'incertitudes. Tout ceci a été exposé dans les deux articles [~ 2005] et [~ 2007-2], qui envisagent des algorithmes de recherche de plans optimaux, notamment en restreignant la recherche à la classe des hypercubes latins. On reviendra sur ces questions dans la partie « conclusion et perspectives ».

C) LE THÈME « COMPUTER EXPERIMENTS »

1. Génèse

A l'instar de tout thème de recherche, les « computer experiments » recouvrent un ensemble de problématiques (propagation d'incertitudes, sensibilité, optimisation, calage...) dont les méthodes de résolution font appel à des techniques variées. Je me suis confronté à certains de ces problèmes au cours de mes activités contractuelles dès 1997. Le premier problème auquel j'ai été confronté était posé par le CEA, centre de Cadarache. Il s'agit d'estimer des quantités nucléaires dites différentielles, les sections efficaces, à partir d'expériences intégrales qui correspondent à l'observation de variables mesurées sur un système nucléaire complexe. Mathématiquement, ce processus d'estimation peut se voir simplement sous la forme d'une régression non linéaire :

$$(13) \quad y = f(\theta) + \epsilon \quad ,$$

où y représente le vecteur des réponses mesurées, θ le vecteur des sections efficaces et ϵ l'erreur de mesure supposée aléatoire.

Mais la fonction f qui précède est représentée informatiquement par un code de calcul complexe dont l'évaluation est coûteuse en temps. La question que j'ai dû traiter à l'époque n'était pas liée à cette spécificité mais était placée à un autre niveau. Il s'agissait

⁸Le bruit blanc n'en est pas tout à fait un pour rester une véritable fonction.

⁹Bien noter que l'alea provient à la fois des observations faites sur le plan d'expérience \mathbf{X} et du caractère aléatoire de la variable x .

d'aider les physiciens nucléaires à situer précisément le problème, et en particulier à faire la lumière sur l'utilisation dans ce cadre du formalisme bayésien, qui ne faisait pas l'unanimité parmi eux. J'ai ensuite travaillé à partir de 1998 sur ces questions analogues avec B. Corre, ingénieur de la société Elf EP, devenue plus tard Total. Le problème est connu dans la littérature pétrolière sous le nom de « history matching ». Il s'agit mathématiquement du même problème que le précédent puisque l'on mesure ici des productions de pétrole au cours du temps et qu'il s'agit alors d'utiliser ces informations pour obtenir une nouvelle estimation du sous-sol. La spécificité de ce problème est double. Tout d'abord l'évaluation de la fonction f est beaucoup plus coûteuse que dans le cas nucléaire, et la connaissance avant exploitation d'un champ pétrolier est essentielle dans la procédure d'estimation. En d'autres termes, un cadre bayésien semble particulièrement adéquat. Les méthodes utilisées à l'époque par B. Corre faisaient une utilisation intensive de la théorie des plans d'expériences. Mais cette théorie était utilisée dans un cadre très particulier, puisque les expériences étaient numériques. Après avoir pris connaissance de la bibliographie disponible, et en particulier les travaux de Sacks, Welch and al. ([Sacks 1989-1], [Sacks 1989-2], [Welch 1992]), qui définissent un cadre formel dans lequel la théorie des plans d'expériences numériques trouve sa place, je lance de premiers travaux. Le premier est réalisé par C. Helbert lors de son stage de fin d'études en 2001 qui teste les méthodes de krigeage sur des sorties d'un simulateur d'écoulement, ainsi que de premières idées de critères de choix de plans d'expériences numériques qui vont par la suite mener à la définition de la MCV-optimalité. Y. Richet réalise également une partie de son stage de fin d'études en 2002 chez Total, et met au point une méthode de calage semi-automatique basée sur le krigeage. L'organisation par l'équipe du premier séminaire français sur les plans d'expériences numériques en juin 2003 entame la mise en place de cette thématique de recherche au sein de l'équipe. L'arrivée en détachement de B. Corre à l'école à partir d'avril 2005, lié à mon retour à temps plein dans l'équipe, donne de nouveaux moyens pour aborder ce thème de recherche.

Les lignes qui précèdent montrent comment la thématique « computer experiments » a émergé à partir de problèmes rencontrés au niveau des applications. Mais les outils nécessaires pour aborder ces questions m'étaient déjà familiers, et recouvraient la quasi-totalité de mes activités passées. On y retrouve en effet les techniques de régression, paramétriques ou semi-paramétriques, le krigeage dans ses variantes les plus courantes (krigeages simple, ordinaire, universel) mais aussi plus avancées (krigeage bayésien, anisotropies géométrique et zonale), mais aussi la considération des processus aléatoires dont la loi n'est connue que sur un espace fonctionnel quotient (comme les FAI-k des géostatisticiens) que j'ai rencontré avec l'équation de Burgers, ou les techniques bayésiennes et les méthodes de Monte Carlo.

2. Evaluation de plans d'expériences numériques

On a vu avec le thème de la MCV-optimalité de quelle manière avait été introduit explicitement l'objectif de propagation d'incertitudes dans un critère d'optimalité de plans. Mais les questions de planification d'expériences numériques ne se posent évidemment pas

dans ce seul cas d'application. De nombreux auteurs utilisent de façon générale des space-filling design (SFD) dans la phase initiale d'exploration d'un code de calcul, l'idée étant qu'en l'absence d'information sur le lien entre les variables d'entrée x du code et les variables de sortie y , on cherche à explorer le plus largement possible le domaine de variation des x . On trouve dans [Koehler 1996] une revue des plans communément utilisés, hypercubes latins, tableaux orthogonaux, suites à faible discrédance.

De nombreux critères de qualité des plans introduits sont définis ; citons par exemple l'entropie, l'IMSE, les critères minimax ou maximin, etc... Mais tous ces critères ne tiennent pas compte d'une propriété essentielle que doivent satisfaire les plans d'expériences numériques, à savoir la bonne répartition spatiale des points du plan, mais aussi la bonne répartition des projections des points sur des espaces de dimension plus faible. Ceci découle d'au moins deux considérations. D'une part le fait que les réponses y peuvent ne dépendre effectivement que d'une partie des facteurs x , auquel cas, les points du plan doivent être bien répartis dans le sous-espace concerné. Mais on peut également évoquer les questions d'aliasing bien connues en traitement du signal ou de l'image qui font se détourner de plans dotés de certaines régularités (voir sur ce sujet le très intéressant article de Yellot sur les cellules photoréceptrices du singe Rhésus [Yellot 1983]).

C'est la raison pour laquelle j'ai orienté, en collaboration avec O. Roustant, une partie des travaux de doctorat de J. Franco sur ces questions. Nous avons ainsi défini un outil à la fois numérique et graphique qui permette de vérifier si un plan donné possède des projections sur les droites qui soient bien réparties. Celui-ci est basé sur la discrédance du plan projeté et est donc lié de près à la statistique de Kolmogorov-Smirnov. Il a permis de détecter très simplement des comportements inadéquats pour une suite de Halton de 250 points en dimension 15 ou un tableau orthogonal linéaire de 49 points en dimension 3, comportements jamais relevés à ma connaissance dans la littérature des plans d'expériences numériques. Cet outil est basé sur le fait que la statistique considérée est libre, et construit à partir de simulations de Monte Carlo effectuées pour chaque valeur n du nombre de points du plan. La loi asymptotique de la statistique peut s'obtenir via le principe d'invariance de Donsker [Donsker 1952]. Je retrouve là des notions que j'ai utilisées pendant ma thèse en 1985, avec la question de la convergence des processus empiriques multi-dimensionnels et des classes de Vapnik-Chervonenkis. Les premiers résultats ont été annoncés dans [~ 2006-1] et [Franco 2007-1] et l'ensemble est sur le point d'être soumis à publication [Franco 2007-2].

3. Construction de métamodèles - l'approche géostatistique

Depuis les articles fondateurs de Sacks, Welch and al. précités ([Sacks 1989-1], [Sacks 1989-2], [Welch 1992]), la théorie des plans d'expériences numériques utilise de façon courante les modèles géostatistiques et le krigeage. Le point faible de cette approche est celui de l'inférence. Les géostatisticiens réalisent l'inférence du variogramme dans les dimensions de l'espace ambiant, c'est-à-dire 1, 2 ou 3, dans un contexte où les données sont relativement abondantes et où l'erreur d'estimation du variogramme n'a pas de conséquences majeures si elle reste modérée. Mais les notions de variographie gagnent

singulièrement en complexité lorsque la dimension augmente. C'est sans doute la raison pour laquelle la plupart des auteurs proposent un modèle très simple, analogue à celui de la formule (11) :

$$(14) \quad y(x) = \beta + Z(x)$$

où Z est un processus aléatoire gaussien centré de variance σ^2 et de fonction de corrélation R . La fonction de corrélation est le plus souvent gaussienne, de la forme suivante :

$$(15) \quad R(h) = \exp \left\{ - \sum_{i=1}^n (h_i / p_i)^2 \right\}$$

Les quantités p_i apparaissent alors comme étant des facteurs d'échelle, égaux à une constante multiplicative près à la portée des géostatisticiens. Un tel modèle souffre d'au moins deux défauts. Le premier vient du fait qu'aucune tendance n'est en général introduite, sous prétexte que les résultats sont peu différents si l'on en introduit une. Pourtant, le simulateur représenté par la fonction f possède souvent des comportements vis-à-vis de certains des facteurs, par exemple monotonie, peu compatibles avec cette hypothèse. Il est vrai que toute fonction, même croissante, peut être considérée comme la réalisation d'un processus stationnaire, mais le prix à payer se situe alors au niveau de l'ergodicité puisque l'inférence devient très difficile. Le second, lié au premier, concerne les portées qui sont très souvent dans la pratique beaucoup plus grandes que la largeur de l'intervalle d'observation, et les questions d'ergodicité reviennent par ce biais.

Face à de telles insuffisances, on peut envisager de nombreuses alternatives. L'une d'entre elles, particulièrement populaire, est celle développée par Oakley et O'Hagan ([Oakley 2002], [Oakley 2004]) qui s'est concrétisée par un logiciel intitulé BACCO (Bayesian Analysis of Computer Code Outputs). Cette démarche a également été utilisée [Goria 2004], dans le contexte historique de la géostatistique, pour l'évaluation financière de projets miniers. Elle consiste à se protéger contre l'erreur de modèle en les mélangeant, en d'autres termes en utilisant une approche bayésienne. C'est la voie que suit D. Ginsbourger qui a débuté sa thèse en novembre 2005. Il travaille à une extension de l'algorithme d'optimisation EGO [Jones 1998], intitulée AMEGO (Agregation of Models for Efficient Global Optimization), basée sur une approche bayésienne. Les premiers résultats sont encourageants et une première publication doit voir le jour rapidement. On notera que son travail s'est nourri de réflexions fondamentales sur l'estimation de processus gaussiens pour les computer experiments [Ginsbourger 2007].

Mais l'approche précédente se contente de mélanger des modèles du même type. C'est la raison pour laquelle un autre thème de travail est celui de la complexification raisonnée du modèle décrit par les équations (14) et (15). Ainsi des travaux sont en cours, en collaboration avec A. Badea et D. Dupuy, pour introduire une tendance non triviale, par exemple de type GAM, et des modèles géostatistiques distincts de (15). L'introduction de modèles comprenant des anisotropies zonales permet notamment de prendre en compte simplement les facteurs peu influents (voir [Dupuy 2006]). Cette approche, pour l'instant effectuée dans le cadre inférentiel classique, doit être développée dans un cadre bayésien. On notera d'ailleurs pour terminer sur ce sujet que des notions aussi élémentaires que le

krigeage ordinaire ou le krigeage universel ne trouvent un véritable fondement probabiliste que dans un cadre bayésien (voir [Goria 2004] ou [~2006-2]).

CONCLUSION ET PERSPECTIVES

Mon activité de recherche s'est déployée au cours des années passées dans des directions variées, avec des domaines d'applications eux-mêmes extrêmement différents. Et mes activités ont poursuivi deux objectifs, l'un applicatif visant à proposer des approches originales dans le domaine concerné, l'autre consacré à la mise au point de méthodes visant à la résolution de problèmes de nature mathématique. J'ai depuis 2005 utilisé ces diverses expériences pour faire émerger, me concernant, mais aussi et surtout pour le département 3MI, la thématique « computer experiments ». Le laboratoire commence à avoir sur ce thème un bon niveau de reconnaissance. Il est en effet leader du consortium DICE (Deep Inside Computer Experiments) qui regroupe des partenaires industriels (Total, Renault, EDF), des instituts de recherche (ONERA, IRSN, IFP) et des universitaires (U. P. Cézanne, Paris-Sud, J. Fourier). Un PPF a été également attribué au laboratoire LAMUSE de l'Université J. Monnet, sous le titre *Amélioration de la méthode d'interpolation par l'analyse numérique et asymptotique*, auquel nous contribuons de manière essentielle. Il s'agit d'examiner la façon dont les méthodes asymptotiques peuvent être utilisées pour la construction de métamodèles (tendances dans des modèles de krigeage par exemple). Enfin, le département 3MI participe activement au projet de GDR intitulé MASCOT NUM (Méthodes d'Analyse Stochastiques des COdes et Traitements NUMériques) qui regroupe la quasi totalité des scientifiques français travaillant sur ces questions. Au total, nous devrions dans les deux ans qui viennent nous situer au meilleur niveau français sur cette thématique avec un double ancrage académique et industriel. Sur ce dernier plan, des contacts avec l'IFP, EADS et Bosch devraient déboucher dans des délais raisonnables, augmentant d'autant notre assise applicative. Sur le plan académique, je résume ci-après des idées que j'ai commencé à développer.

1. La MCV-optimalité

Les travaux sur la MCV-optimalité peuvent être étendus dans au moins deux directions. Déjà, on a vu que le modèle utilisé dans notre approche tel que résumé par la formule (11) est basé sur un bruit blanc dont la justification pour des expériences numériques n'est pas claire¹⁰. C'est la raison pour laquelle une extension des résultats au

¹⁰Ceci étant, on prendra garde au fait que la répétabilité parfaite d'un code numérique n'est jamais satisfaite. Quelquefois, le code est structurellement stochastique alors que dans d'autres cas, il comporte des composantes chaotiques. Hormis ces cas extrêmes, la réponse de tout code dépend de variables dites de simulation (architecture des ordinateurs utilisés, système d'exploitation, critères d'arrêt, variables de stabilisation numérique, choix de la discrétisation...) qui conditionnent le résultat et sont dans leur plus grande part non explicites.

cas de processus gaussiens de structure de covariance non triviale est entamée, ce qui ne pose pas de problème de fond mais tout au plus quelques questions numériques. Sur un autre plan, l'utilisation des techniques de RKHS (Reproducing Kernel Hilbert Spaces) permet de compléter de manière simple et élégante le modèle donné par la formule (11). Supposons en effet que s'ajoute à la composante linéaire un facteur h^{11} , fonction appartenant à un RKHS noté H , de noyau reproduisant K et de norme $\|\cdot\|$. On obtient ainsi :

$$(16) \quad f(x) = X(x)\beta + h(x) + \epsilon(x)$$

Ce modèle est introduit par [Yue 1999] dans un contexte d'expériences réelles sensiblement différent du nôtre. Ne connaissant pas, par principe, la fonction h , la procédure d'estimation de la partie linéaire est la même que dans l'approche MCV-optimalité précédente, c'est-à-dire basée sur les moindres carrés. Dans ces conditions, le modèle approché est le même que précédemment et est donc fourni par la formule (12). On s'intéresse alors naturellement à l'écart-quadratique moyen¹² :

$$(17) \quad C = E\left[\left(f(x) - f_{app}(x)\right)^2\right]$$

Des techniques assez standard de calcul dans un RKHS permettent alors de majorer cette quantité par une somme pondérée de deux termes, un terme de biais et un terme de variance, chacun ne dépendant que du seul plan d'expériences. Seuls les poids affectés à ces deux termes mettent en jeu les tailles du biais $\|h\|^2$ et du bruit σ^2 . De sorte que la confiance ou la défiance dans le modèle linéaire, mesurée par le rapport des tailles précitées, suffit à déterminer un plan d'expériences pour lequel la quantité C sera la plus faible possible. Ces premières réflexions ont été présentées dans [~ 2007-3] et font l'objet de travaux en cours. On notera que ces questions se rapprochent de celles de l'apprentissage d'une fonction appartenant à un RKHS, abordées dans les années 1980 à travers les fonctions splines, et peu exploitées depuis, alors que leur pertinence pour l'exploration des codes numériques semble assez claire. On notera d'ailleurs le travail récent de [Amato 2006] sur le sujet, en lien avec les décompositions en bases d'ondelettes, qui montre l'actualité de ce type d'approche.

A moyen terme, la construction de plans approchant l'optimum des critères de type MCV-optimalité constitue un véritable challenge. Une des idées possibles consiste à lier les procédés de construction de plans mis au point actuellement par J. Franco en lien avec X. Bay, basés sur des processus ponctuels soumis à une dynamique markovienne, à des critères à optimiser, dont celui de MCV. Et cette piste semble très raisonnable dans la mesure où les processus ponctuels introduits par J. Franco sont des champs markoviens et que le lien entre champs markoviens et critères à minimiser sont maintenant bien connus.

¹¹Pour assurer l'identifiabilité du modèle, la fonction h doit satisfaire des conditions d'orthogonalité.

¹²Dans le cas où h est nulle, cette quantité est égale à une constante près au critère de MCV-optimalité.

2. Valorisation d'informations a priori

Par ailleurs, les modèles qui sont utilisés de façon assez courante pour approximer les codes numériques sont des processus gaussiens stationnaires et les travaux les plus souvent cités, comme indiqué précédemment, sont ceux de Oakley et O'Hagan ([Oakley 2002] et [Oakley 2004]) qui proposent une méthodologie de type bayésien, et ont développés un outil informatique, intitulé BACCO (Bayesian Analysis of Computer Code Outputs). Néanmoins, l'utilisation des outils géostatistiques en grande dimension est pour l'instant largement en friches et la célèbre malédiction de la dimension semble assez peu évoquée dans ces travaux. Avant d'aller plus loin, il me semble important d'exprimer une conviction personnelle ; l'utilisation de techniques bayésiennes, qu'elles soient liées à des modèles gaussiens ou à d'autres types de modèles¹³, me semble être l'un des éléments clés des recherches à venir pour au moins une raison simple. Mon expérience m'a montré que les ingénieurs qui faisaient des expériences numériques incorporaient toujours, et de façon essentielle, de l'information métier dans leur réflexion. C'est la raison pour laquelle toute méthode mathématique devra permettre à l'utilisateur d'intégrer au moins partiellement cette connaissance *a priori* dans sa démarche. Et l'analyse bayésienne est un excellent outil pour ce faire. Mais dans ces conditions, l'utilisation faite intensivement dans la littérature de lois a priori non informatives me paraît à proscrire, sauf pour des exemples de nature purement mathématique. D'ailleurs, cette affirmation est de notoriété publique dans un domaine comme celui de l'analyse d'images où les informations a priori sont traduites par exemple de manière particulièrement simple à travers le formalisme des champs markoviens.

Pour incorporer de telles informations, deux pistes complémentaires émergent naturellement. La première consiste à donner à l'ingénieur qui réalise une étude à partir d'un code numérique des outils d'aide à la décision lui permettant d'utiliser sa connaissance métier dans le processus. C'est la démarche qui a été suivie lors de son stage de fin d'études par Yann Richet [Richet 2001] pour le calage d'historiques de production. Son travail, basé sur la visualisation de certains profils de vraisemblance, a d'ailleurs été mis depuis en exploitation par la société Total pour ses besoins propres. Il y a là de nombreuses voies à explorer dans la continuation des travaux de [Ginsbourger 2007].

Mais on peut également adopter une démarche analogue à celle utilisée en finance mathématique. L'une des hypothèse faite le plus souvent est celle de la complétude des marchés qui affirme en substance que les cours des différents actifs donnent la totalité de l'information disponible. De ce point de vue, il est inutile de chercher à acquérir d'autres types d'informations, car elles ont toutes un effet sur le marché. Un exemple bien connu est celui de la hausse récente des cours du pétrole, reflet des anticipations des acteurs sur les futures tensions du fait de la diminution des réserves disponibles. Dans notre contexte, il est assez clair que le code de calcul lui-même résume l'ensemble des informations génériques, auxquelles s'ajoutent des informations particulières relatives au phénomène étudié. Par exemple, en neutronique, le code utilisé intègre l'ensemble des connaissances

¹³On pense en particulier aux SVM (support vector machine) qui trouvent d'ailleurs naturellement leur place dans le cadre des RKHS..

relatives à la propagation et l'évolution des neutrons alors que l'ingénieur faisant une étude de criticité aura en tête des notions métier lui faisant dire par exemple que l'effet de telle enceinte de protection ne peut dépasser un certain niveau. Mais si le code résume à lui seul l'ensemble des informations génériques disponibles, à quoi bon parler d'informations a priori ? Le problème vient du fait que le code est d'une complexité et d'un temps calcul tels qu'il n'est pas possible de l'exploiter de très nombreuses fois. C'est bien là tout le problème des « computer experiments ». Il est néanmoins possible de dégrader la qualité du code en le simplifiant ou en utilisant un autre code plus simpliste pour obtenir de premières informations. Ceci peut se faire en modifiant le modèle approché ou en considérant que le code simplifié donne une image floue de la « réalité ». C'est la voie que nous commençons à suivre avec le premier travail de [Helbert 2007].

De manière plus prospective, je réfléchis à l'heure actuelle à l'intégration de processus intrinsèques dans la démarche, en me basant notamment sur les travaux en cours de J. Duchon [Duchon 2007] et sur la thèse de E. Vasquez [Vasquez 2005]. L'un des intérêts que je perçois de ces notions étant de permettre la considération de familles de « covariances » plus générales que celles rencontrées dans le cadre standard, qui doivent permettre de contraindre de manière efficace les modèles introduits. Exemple parmi d'autres, les splines plaques minces ont connu un succès indéniable depuis les années 70. Et ce type de fonction n'a pu être considéré qu'à travers des noyaux de type positif conditionnel qui sont la « covariance » de fonctions aléatoires connues à un polynôme de degré 1 près.

BIBLIOGRAPHIE

- [Amato 2006] Amato U., Antoniadis A., Pensky M., *Wavelet kernel penalized estimation for non-equispaced design regression*, *Statistics and Computing*, **16**, 1, p. 37-56.
- [Antoniadis 1992] Antoniadis A., Berruyer J., Carmona R., *Régression non linéaire*, *Economica*.
- [Azema 1979] Azema J., Yor M. : *Une solution simple au problème de Skorokhod*, Séminaire de probabilités, XIII, *Lecture Notes in Math.*, **721**, p. 90-115, Springer.
- [Bachelier 1900] Bachelier L. : *Théorie de la spéculation*, *Ann. Sci. de l'Ecole Normale Sup.*, 3e série, **17**, p. 21-86.
- [Bachelier 1906] Bachelier L. : *Théorie des probabilités continues*, *J. de Math. Pures et Appl.*, **2**, p. 259-327.
- [Bertoin 1996] Bertoin J. : *Lévy processes*, Cambridge University Press, Cambridge.
- [Bertoin 1998] Bertoin J. : *The inviscid Burgers equation with Brownian initial velocity*, *Comm. Math. Phys.*, **193**, p. 397-406.
- [Cole 1951] Cole J. D. : *On a quasi-linear parabolic equation occurring in aerodynamics*, *Quart. Appl. Math.*, **9** (3), pp. 225–236.
- [Donsker 1952] Donsker M. D. : *Justification and extension of Doob's heuristic approach to the Kolmogorov-Smirnov theorems*, *Annals of Mathematical Statistics*, **23**, p. 277-281.
- [Duchon 2007] Duchon J., *Modèles intrinsèques et probabilités conditionnelles*, en préparation, communication personnelle.
- [Dupuy 2006] Dupuy D., *Méthode de krigeage avec anisotropie zonale – le cas I33I*, rapport interne ENSM.SE.
- [Fedorov 1972] Fedorov V.V., *Theory of Optimal Experiment Design*, Academic Press, New York.
- [Fisher 1935] : *The Design of Experiments*, Oliver & Boyd, Edinburgh.
- [Franco 2007-1] Franco J., Carraro L., Roustant O., Jourdan A., *A radar-shaped statistic for testing and visualizing uniformity properties in computer experiments*, ENBIS Conference on Computer experiments versus Physical experiments, Turin (avril).
- [Franco 2007-2] Franco J., Carraro L., Roustant O., Jourdan A., *A radar-shaped statistic for testing and visualizing uniformity properties in computer experiments*, en préparation.
- [Ginsbourger 2007] Ginsbourger D., Dupuy D., Badea A., Carraro L., Roustant O., *A note on the choice and the estimation of kriging models for the analysis of deterministic computer experiments*, ENBIS Conference on Computer experiments versus Physical experiments, Turin (avril).

- [Giraud 2001] Giraud C. : *Turbulence de Burgers et agrégation de particules lorsque l'état initial est aléatoire*, thèse de doctorat, Univ. Paris VI.
- [Goria 2004] Goria S. : *Evaluation d'un projet minier : approche bayésienne et options réelles*, thèse de doctorat, Ecole des mines de Paris.
- [Helbert 2002] Helbert C., Carraro L., Perrin S., Valdivieso F., Pijolat M. : *Le modèle de Mampel par Monte-Carlo*, 33ièmes Journées d'Etude de Cinétique Hétérogène, Compiègne. (Prix Jean Besson)
- [Helbert 2004] Helbert C., Touboul E., Perrin S., Carraro L., Pijolat M. : *Stochastic and deterministic models for nucleation and growth in non-isothermal and/or non-isobaric powder transformation*, Chemical Engineering Science, **59** (7), p. 1393-1401.
- [Helbert 2005] Helbert C. : *Modèles probabilistes de germination/croissance pour la transformation des poudres*, thèse de doctorat, Ecole des mines de Saint-Etienne.
- [Helbert 2007] Helbert C., Dupuy D., Carraro L., *Assessment of uncertainty in computer experiments, from Universal Kriging to Bayesian Kriging*. ENBIS Conference on Computer experiments versus Physical experiments, Turin (avril).
- [Holley 1972] Holley R. : *Markovian interaction processes with finite range interactions*, Ann. Math. Statistics, **43**, p. 1961-1967.
- [Jaffard 1999] Jaffard S. : *The multifractal nature of Lévy processes*, Probab. Theory Related Fields, **114**(2), p. 207-227.
- [Jones 1998] Jones D. R., Schonlau M., Welch W. J. : *Efficient global optimization of expensive black-box functions*, Journal of Global Optimization, **13**, p. 455–492.
- [Kajiyi 1986] Kajiyi J. T. : *The rendering equation*, Comm. Of the ACM, **20** (4), p. 143-149.
- [Koehler 1996] Koehler J. R., Owen A. B. : *Computer experiments*, Handbook of statistics, **13**, p. 261-308.
- [Hopf 1950] Hopf E. : *The partial differential equation $ut + uux = \mu uxx$* , Comm. Pure and Appl. Math., **3**, pp. 201–230.
- [Maillot 1992] Maillot J. L., Carraro L., Peroche B. : *A progressive ray tracing*, Proceedings of the third eurographics workshop on rendering, p. 9-19.
- [Maillot 1996] Maillot J. L. : *Pseudo-réalisme et progressivité pour le tracé de rayons*, thèse de doctorat, Université Jean Monnet.
- [Mampel 1940] Mampel K. L. : *Die Zeitemsatzformeln für ein Pulver aus kugelförmigen Teichen*, Z. Phys. Chem. A, **187**, p. 43 et 235.
- [Matheron 1973] Matheron G. : *The intrinsic random functions and their applications*, Adv. in Applied Prob., **5**, p. 439-468.
- [Messiaen 1997] Messiaen L. : *Recherche des paramètres morphologiques influents pour la prévision des caractéristiques mécaniques d'un acier austénoferritique*, thèse de doctorat, Ecole des mines de Paris
- [Obloj 2005] Obloj J. : *The Skorokhod embedding problem and some families of Brownian martingales*, thèse de doctorat, Univ. Paris VI.

- [Oakley 2002] Oakley J., O'Hagan A. : *Bayesian inference for the uncertainty distribution of computer model outputs*. Biometrika, **89**, p. 769-784.
- [Oakley 2004] Oakley J., O'Hagan A. : *Probabilistic sensitivity analysis of complex models: a Bayesian approach*. J. of the Royal Stat. Soc. B, **66**, p. 751-769.
- [Revuz 1991] Revuz D., Yor M. : *Continuous martingales and brownian motion*, Springer, first edition.
- [Sacks 1989-1] Sacks J., Welch W.J., Mitchell T.J., Wynn H.P. : *Design and Analysis of Computer Experiments*, Statistical Science, **4** (4), p. 409-435.
- [Sacks 1989-2] Sacks J., Schiller S.B., Welch W.J. : *Designs for Computer Experiments*, Technometrics, **31** (1), p. 41-47.
- [She 1992] She Z. S., Aurell E., Frisch U., *The inviscid Burgers equation with initial data of Brownian type*, Comm. Math. Phys., **148**, p. 632-641.
- [Siegmund 1985] Siegmund D. : *Sequential analysis*, Springer, Berlin.
- [Sinai 1992] Sinai Y., *Statistics of shocks in solution of inviscid Burgers equation*, Comm. Math. Phys., **148**, p. 601-621.
- [Templeman 1972] Templeman A. A. : *Ergodic theorems for general dynamical systems*, Trans. Moscow Math. Society, **26**, p. 94-132.
- [Vasquez 2005] Vasquez E., *Modélisation comportementale de systèmes non-linéaires multivariés par méthodes à noyaux et applications*, thèse de doctorat, Université Paris XI.
- [Wald 1948] Wald A. : *Sequential analysis*, J. Wiley & sons, New York.
- [Welch 1992] Welch W. J., Buck R. J., Sacks J., Wynn H. P., Mitchell T. J., Morris, M. D. : *Screening, Predicting, and Computer Experiments*, Technometrics, **34** (1), p. 15-25.
- [Winkel 2001] Winkel M. : *Quelques contributions à la théorie des processus de Lévy, et des applications en turbulence et en économétrie*, thèse de doctorat, Univ. Paris VI.
- [Yellott 1983] Yellott J. I. : *Spectral consequences of photoreceptor sampling in the rhesus retina*, Science, **221**, p. 382-385.
- [Yue 1999] Yue R., Hickernell F.J., *Robust designs for fitting linear models with misspecification*, Stat. Sinica, **9**, p. 1053-1069.

- [~ 2007-1] L. Carraro, N. El Karoui, A. Meziou, J. Obloj : *On Azema-Yor martingales : further properties and applications*, soumis au séminaire de probabilités (Springer).
- [~ 2007-2] L. Carraro, B. Corre, C. Helbert, O. Roustant, S. Josserand : *Optimal designs for propagation of uncertainty in computer experiments*, à paraître dans Chemometrics and Intelligent Laboratory Systems.

- [~ 2007-3] L. Carraro : *Mini-cours : RKHS, métamodèles et plans d'expériences*. Journées plans d'expériences numériques, Solaize (mars).
- [~ 2006-1] L. Carraro, J. Franco, O. Roustant : *Un nouveau critère statistique en forme de radar pour la mesure de l'uniformité des plans d'expériences et de leurs projections sur les sous-espaces*, actes des XXXVIIIèmes journées de statistiques, Clamart, juin 2006.
- [~ 2006-2] L. Carraro, *Bayesian kriging for dummies*, note interne ENSM.SE.
- [~ 2005] L. Carraro, B. Corre, C. Helbert, O. Roustant : Construction d'un critère d'optimalité pour plans d'expériences numériques dans le cadre de la quantification d'incertitudes, *Revue de Statistique Appliquée*, **3** (4), p. 87-103.
- [~ 1998] Carraro L., Duchon J. : *Equation de Burgers avec conditions initiales à accroissements indépendants et homogènes*, *Annales de l'Institut Henri Poincaré, Analyse Non Linéaire*, **15** (4) p. 431-458.
- [~ 1995] Carraro L. : *The Bachelier equation*, Note interne centre Simade, Ecole des mines de St Etienne.
- [~ 1994] Carraro L., Duchon J. : *Solutions statistiques intrinsèques de l'équation de Burgers et processus de Lévy*, *C.R. Acad. Sci. Paris, Série I*, **319**, p. 855-858.