

# Rapport d'activités 2003-2007 du groupe ANALYSE DES DONNEES, PROBABILITES ET STATISTIQUES

## Composition du groupe (avec pourcentage de recherche dans ce groupe)

### Permanents :

Fadoua BALABDAOUI (MCF) (*depuis Septembre 2006*)  
Alain BENSOUSSAN (PR, émérite) (*2003-2004*) (30%)  
François BOLLEY (MCF) (40%, *depuis Septembre 2006*)  
Pierre CAZES (PR)  
Edwin DIDAY (PR)  
Halim DOSS (PR) (*depuis Février 2006*)  
Danièle FLORENS (PR) (30%)  
Ivan GENTIL (MCF) (50%)  
Mireille GETTLER SUMMA (MCF)  
Bernard GOLDFARB (MCF)  
Françoise GOUPIL (MDC)  
Arnaud GUILLIN (MCF) (*départ en Septembre 2006*)  
Bénédicte HAAS (MCF) (*depuis Septembre 2005*)  
Alessandra IACOBUCCI (IE) (*depuis Juin 2005*)  
Jean-Michel MARIN (MCF) (*départ en Septembre 2006*)  
Stefano OLLA (PR)  
Catherine PARDOUX (MCF)  
Christian ROBERT (PR)  
Judith ROUSSEAU (PR) (*depuis Septembre 2004*)  
Myriam TOUATI (IE)  
José TRASHORRAS (MCF) (*depuis Septembre 2004*)

### Doctorants (avec année de soutenance en cas de thèse déjà soutenue)

Billy AMZAL (*12/2004*), Abdourahamane BALDE (*05/2007*), Gaël BENABOU (*12/2005*), Cédric BERNARDIN (*06/2005*), Xavier BRY (*10/2004*), Roberto CASARIN, Christian DERQUENNE (*02/2006*), Aicha EL GOLLI (*06/2004*), Gholam Hossein GHOLAMI ZORABAD, Afonso FILIPE (*12/2005*), Basile GIADA (*06/2007*), Alzenny GOMES DA SILVA, Aude GRELAUD, Freddy Romero HERNANDEZ (co-tutelle avec l'IMPA), Meryem IMAKOR, Baba dit Malley KANE, Soleiman KHAZAEI, Gabriel KISSITA (*10/2003*), Fazlollah LAK, Mohamed LIMAM (*07/2005*), Chérif MBALLO (*12/2005*), Omar MERROUN, Katalin NAGY (co-tutelle avec l'Université de Budapest) (*06/2006*), Kutluhan PAK (*07/2005*), Mohamed Cherif RAHAL, Sophie RAINERO (*05/2006*), Djamal SECK, Arafat TAYEB (*10/2006*), Ibrahima WAGNE

Arnaud GUILLIN (*HDR, 11/2004*), Fabrice ROSSI (*HDR, 09/2006*)

## Présentation générale du groupe

### 1. Probabilités

En 2003 l'équipe de probabilité était constituée par Stefano Olla (professeur), Arnaud Guillin (MdC) et Ivan Gentil (MdC partagé à 60% avec l'équipe d'analyse). Elle s'est développée dans les dernières années avec le recrutement de José Trashorras (MdC en 2004), Bénédicte Haas (MdC en 2005) et François Bolley (MdC en 2006, partagé avec l'équipe d'Analyse). Notons que Arnaud Guillin a passé une HDR et a été recruté professeur à Marseille en 2006. En 2006 le professeur Halim Doss a rejoint le Ceremade. Enfin en 2007, les recrutements de Rémi Rhodes (MdC) et de Vincent Vargas (CR CNRS) ont continué à enrichir l'équipe des permanents. Les recherches développées ont des liens importants avec les équipes d'Analyse et de Statistique du CEREMADE. De plus tous les membres ont une collaboration active avec des chercheurs externes au laboratoire, en France comme à l'étranger.

Bénédicte Haas travaille sur les processus de fragmentations aléatoires et leur représentation généalogique sous forme d'arbres continus. José Trashorras s'intéresse aux modèles de verres de spin, à la modélisation multi-échelle et aux grandes déviations pour les permutations aléatoires. François Bolley et Ivan Gentil travaillent sur les inégalités fonctionnelles de types Poincaré ou Sobolev logarithmique pour des mesures de probabilités. Halim Doss étudie les grandes déviations pour les équations différentielles stochastiques rétrogrades à horizon aléatoire. Enfin Stefano Olla travaille sur des modèles microscopiques pour la mécanique statistique du non équilibre et sur l'homogénéisation des diffusions en environnement aléatoire.

### 2. Statistique mathématique

Depuis 2003, l'équipe de Statistique mathématique a évolué, du fait du départ de Jean-Michel Marin et d'Arnaud Guillin, respectivement pour l'INRIA et l'École Centrale de Marseille, et du fait du recrutement de Fadoua Balabdaoui. Cette modification a induit un déplacement des thèmes depuis la Statistique bayésienne paramétrique et les méthodes de simulation vers la Statistique nonparamétrique, puisque Fadoua Balabdaoui comme Judith Rousseau travaillent dans cette direction, en partie sous un angle bayésien. Le recrutement récent de Sophie Donnet, qui travaille également sur les méthodes de calcul statistique, va ré-équilibrer l'équipe.

Les travaux de Christian Robert, comme de Jean-Michel Marin et d'Arnaud Guillin (pour sa composante statistique), portent principalement sur l'analyse d'algorithmes de Monte Carlo pour l'approximation d'intégrales probabilistes. Ils ont en particulier introduit la notion de "population Monte Carlo", forme adaptative de la méthode d'importance, et participent de ce fait à l'ANR "Adap'MC" depuis 2005, en collaboration avec des chercheurs de Telecom, des Ponts et de Polytechnique. Christian Robert et Jean-Michel Marin contribuent également à deux autres ANR en collaboration avec des généticiens (Modgepop) et des cosmologues (Ecosstat), toujours sous l'angle des méthodes de calcul intensives. Par ailleurs, Judith Rousseau vient d'obtenir le financement d'une ANR SP Bayes qui est consacrée à l'étude des méthodes bayésiennes non- et semi-paramétriques et de leurs applications, en particulier en théorie du risque. Judith Rousseau et Fadoua Balabdaoui s'intéressent aux propriétés asymptotiques de type fréquentistes de

procédures statistiques (estimation et test). Judith Rousseau s'intéresse principalement aux méthodes bayésiennes alors que Fadoua Balbdaoui étudie les méthodes de maximum de vraisemblance, les deux ayant souvent des propriétés similaires.

Au cours de la période 2003-2007, les membres de l'équipe de Statistique mathématique ont contribué à l'organisation de plusieurs congrès et workshops, dont le 4ième workshop sur les méthodes bayésiennes objectives à Aussois en juin 2003, une demi-journée sur les problèmes inverses à la Royal Statistical Society à Londres en décembre 2003, les workshops Adap'ski et Adap'skii à Bormio (Italie) en janvier 2005 et 2008 respectivement, et le workshop "Bioinformatics, Genetics and Stochastic Computation: Bridging the Gap", à BIRS, Banff (Canada), en juillet 2007. Ils participent également activement à des comités scientifiques lors d'organisation de congrès internationaux (Valencia, O'Bayes, ISBA et autres). Tous les membres de l'équipe ont participé activement à des congrès internationaux et à des écoles d'été, en particulier autour de "Bayesian Core", rédigé par Jean-Michel Marin et Christian Robert et publié en janvier 2007 (Finlande, Espagne, Nouvelle Zélande, Australie, Venezuela). Christian Robert a également été nommé éditeur en chef de la revue de la Royal Statistical Society, JRSS B, pour la période 2006-2009.

### **3. Analyse des Données**

Pour le groupe d'Analyse des Données, un des thèmes de Mireille Gettler Summa, Bernard Goldfarb et Catherine Pardoux est celui de l'exploration et de la modélisation des chroniques multiples et multidimensionnelles. Des résultats ont été obtenus sur le codage, la définition de distances et l'élaboration d'algorithmes de classification de courbes. En particulier dans le cadre de la valorisation contractuelle des recherches du CEREMADE avec la société ISTHMA, des solutions originales de visualisation des séries multiples et multidimensionnelles ont été trouvées avec une approche fonctionnelle, et développées dans un logiciel qui a fait l'objet de deux brevets déposés initialement par l'Université Paris Dauphine.

Edwin Diday a proposé un nouveau paradigme intitulé Analyse des Données Symboliques (ADS) qui étend la fouille de données à de nouvelles unités statistiques exprimant un niveau de connaissances plus élevé que celui de simples individus (caractérisés par un vecteur de variables numériques ou qualitatives habituelles). Des méthodes de classification ont également été développées. Ces recherches ont donné lieu à trois contrats : i) ANR SEVEN pilotée par EDF Clamart; ii) Contrat S3 pour l'Ile-de-France piloté par le LCPC (laboratoire Central Des Ponts et Chaussées) pour l'étude de la détérioration des ouvrages publics; iii) Contrat avec le CHU de Dijon pour l'étude de trajectoires de patients dans les services des hôpitaux.

En analyse des données numériques, de nouvelles approches ont été développées dans les modèles complexes d'équations structurelles, avec en particulier : a) la méthode PML (Partial Maximum Likelihood) qui généralise le modèle PLS au cas de données qualitatives ou mixtes; b) la mise au point de nouvelles méthodes pour traiter la multicollinéarité, tout en gardant un bon pouvoir explicatif, avec des algorithmes originaux. Des logiciels nouveaux ont été mis au point, avec en particulier le développement et la maintenance du logiciel SODAS.

[Lien vers les fiches individuelles des membres du groupe](#)