

Rapport du Laboratoire Statistique et Modélisation

Année 2010

Responsable : Valentin PATILEA, Professeur des universités, INSA Rennes

Membres :

BENINEL Farid	Maître de conférence Université de Poitiers (CMC Ensai)
BIDAULT Alain	Enseignant-chercheur (CMC Ensai)
CAUSEUR David	Professeur des Universités Agrocampus Ouest
CLAUSS Pierre	Enseignant-chercheur (CMC Ensai)
COQUET François	Professeur des Universités INSEE
DELATTRE Eric	Maître de Conférences Université Cergy-Pontoise (CMC Ensai)
DIAYE Marc-Arthur	Maître de Conférences Université d'Evry (CMC Ensai)
DUVAL Laurence	Maître de Conférences INSEE
FROMONT RENOIR Magalie	Maître de Conférences Université Rennes 2 (CMC Ensai)
HRISTACHE Marian	Maître de Conférences INSEE
PATILEA Valentin	Professeur des universités INSA de Rennes
ROUVIERE Laurent	Maître de Conférences Université Rennes 2 (CMC Ensai)
TRUQUET Lionel	Maître de Conférences Université Rennes 1
VIAL Céline	Maître de Conférences Paris X Nanterre (CMC Ensai)
VILLA Christophe	Professeur Audencia Nantes
VIMOND Myriam	Maître de Conférences INSEE

Doctorants :

MAISTRE Samuel

1) Visiteurs

BLANCKE Délphine :	Université d'Avignon (France)
COMBES Jean-Baptiste :	University of Aberdeen (Royaume Uni)
FOULKES Andrea :	University of Massachusetts (Etats Unis)
GONZALEZ MANTEIGA Wenceslao :	Universidad de Santiago de Compostela (Espagne)
GRIEVE Andy :	King's College-London (Royaume Uni)
HIRUKAWA Junichi :	Niigata University (Japon)
LUTA Gheorghe :	Georgetown University (Etats-Unis)
SHEU Ching-Fan :	National Cheng-Kung University, Tainan (Taiwan)
STOREY John:	Princeton University (Etats Unis)
MACKENZIE Gilbert :	University of Limerick (Irlande)
NECIR Abdelhakim :	Universités de Biskra (Algérie)
SANCHEZ-SELLERO César :	Universidad de Santiago de Compostela (Espagne)
XU Jing :	University of Limerick (Irlande)

Participation à des Comités de rédaction de revues et éditeurs associés.

BENINEL F.:

Coordination d'un numéro spécial de la RNTI, consacré au « *Data mining et apprentissage statistique* ». Parution prévue en 2011.

COQUET F. :

Editeur associé de « *Statistics and Probability Letters* ».

2) LIVRES

CORNILLON P.A., GUYADER A., HUSSON F., JEGOU N., JOSSE, J., KLOAREG M., MATZNER-LOBER, E., ROUVIERE L. :

« *Statistiques avec R* », Presses Universitaire de Rennes, 2^{ème} édition.

3) ARTICLES ET RAPPORTS DE RECHERCHE

ARTICLES PUBLIES OU EN COURS DE PUBLICATION

Econométrie de la Finance et de l'Assurance

CLAUSS P.:

« Liquidity Risk Integration in Portfolio Choice: the Bid Efficient Frontier », *Journal of Modern Accounting and Auditing*, 6(7), 1-10.

Divers

DIAYE M.A., LAPIDUS A.:

« Pleasure and Belief in Hume's Decision Process », *European Journal of the History of Economic Thought*, à paraître.

Microéconométrie de l'entreprise, R et D, innovation

CRIFO P., DIAYE M.A.:

« The Composition of Compensation Policy: From Cash to Fringe Benefits », *Annales d'Economie et de Statistique*, à paraître.

Statistique

BIAU G., CADRE B., ROUVIERE L. :

« Statistical analysis of k-nearest neighbor collaborative recommendation », *The Annals of Statistics*, Vol. 38, 1568-1592.

BIGOT, J., LOUBES, J.M., VIMOND, M. :

« Semiparametric estimation of shifts on compact Lie groups for image registration », à paraître dans *Probability Theory and Related Fields*.

BLANKE D., VIAL C:

« Estimating the order of mean-square derivatives with quadratic variations », à paraître dans *Statistical Inference and Stochastic Processes*.

BLUM Y., LE MIGNON G., LAGARRIGUE S., CAUSEUR, D. :

« A factor model to analyse heterogeneity in gene expressions », *BMC Bioinformatics*. 11:368.

BOHNING D., HOLLING H., PATILEA V. :

« A limitation of the diagnostic-odds ratio in determining an optimal cut-off value for a continuous diagnostic test », à paraître dans *Statistical Methods in Medical Research*.

BONNERY D., BREIDT J., COQUET F.:

« Almost sure convergence of the sample cumulative distribution function under informative selection mechanism », à paraître dans *Bernoulli*.

BOUAGUEL W., BENINEL F., BEL MUFTI G. :

« From generalized discrimination to logistic discrimination: Logistic sub-models definition for populations mixture in credit scoring », Proceedings of *The Second Meeting on Statistics and Data Mining MSDM'2010*, Hammamet, pp. 51-60.

BURR T., HENGARTNER N., MATZNER-LOBER E., MYERS S., ROUVIERE L. :

« Smoothing low resolution gamma spectra », *IEEE Transactions on Nuclear Science*, Vol. 57, 2831-2840.

CAFFIER V., DIDELOT F. , PUMO B., CAUSEUR D. , DUREL C.E., PARISI L. :

« Aggressiveness of eight *Venturia inaequalis* isolates virulent or avirulent to the major resistance gene Rvi6 on a non-Rvi6 apple cultivar », *Plant pathology*. 59(6): 1072-1080.

CREMASCO C.P., LOUZADA F., MACKENZIE G. :

Sampling Based Inference for the Generalized Time Dependent Logistic Hazard Model, «*Journal of Statistical Theory and Applications* », Vol 9, No 2, 169-184.

FROMONT-RENOIR M., LAURENT B., REYNAUD-BOURET P.:

« Adaptive tests of homogeneity for a Poisson process », à paraître dans *Annales de l'IHP*.

HA, I.D., MACKENZIE G. :

« Robust Frailty Modelling using Non-Proportional Hazards Models », *Statistical Modelling*, Vol 10, No. 3, 315-332.

HA I.D., SYLVESTER R., LEGRAND C., MACKENZIE G. :

« Frailty Modelling for Survival Data from Multi-Centre Clinical Trials », à paraître dans *Statistics in Medicine*.

HERVE L., LEDOUX J., PATILEA V. :

« The Berry-Esseen bound of M-estimators for geometrically ergodic Markov chains », à paraître dans *Bernoulli*.

KACHOUR M., TRUQUET L. :

« A p-order signed integer-valued autoregressive (SINAR(p)) model », à paraître dans *Journal of Time Series Analysis*.

LAVERGNE P., PATILEA V. :

« One for all and all for one: Dimension reduction for regression checks », à paraître dans *Journal of Business and Economic Statistics*.

TRUQUET L.:

« A moment inequality of the Marcinkiewicz-Zygmund type for some weakly dependent random fields », *Statistics and Probability Letters* 80, 1673-1679.

M. VIMOND :

« Efficient estimation for a subclass of shape invariant models », *Annals of Statistics*, Vol. 38, No. 3, 1885-1912.

5) DOCUMENTS DE TRAVAIL ET TEXTES EN REVISION.

Divers

DELATTRE E., SAMSON, A.L. :

« Stratégies de localisation des médecins généralistes français : mécanismes économiques ou hédonistes ? », soumis à *Annales d'Economie et Statistique*.

DIAYE M.A., GREENAN, N., URDANIVIA, M. :

« Individual evaluation interview as a practice for subjective evaluation of performance. »

DIAYE M.A., GREENAN, N., PEKOVIC, S.:

« Sharing the fame of "being" ISO certified : Quality Supply chain Effect from the French Employer Survey. »

DIAYE M.A., GREENAN, N.:

« The Economics of Performance Appraisals. »

DIAYE M.A., SCHOCH, D.:

« Desires under uncertainty. »

LMADANI F.B., DIAYE M.A., URDANIVIA M. :

« Un test de la discrimination intersectionnelle sur le marché du travail en France. »

SHIRAISHI H., OGATA H., AMANO T., PATILEA V., VEREDAS D., TANIGUGHI M.:

« Optimal portfolios with end-of-period target », soumis à *The European Journal of Finance*.

VILLA C., YUSUPOV N. :

« Modern Microfinance : Cross sector partnership with motivated agents », cahiers de recherche, chaire Banque Populaire.

Statistique

AL-TAWARAH, Y., MACKENZIE G.:

« Two new Failure-time Regression Models from the GTDL family », en revision pour *Statistical Modelling*.

F. BENINEL :

« Using Ensemble methods in credit scoring ».

BOUAGUEL W., BENINEL F., BEL MUFTI G. :

« Mise à jour de règle d'affectation en credit-scoring », soumis à *RNTI*.

BOUAMOUD, M., DIAYE, M.A., WALKOWIAK, E. :

« Informal help in the workplace: workers' free-agency by-product or firms' organizational design by-product ? »

CAUSEUR, D., FRIGUET, C. HOUEE-BIGOT, M., KLOAREG, M.:

« Factor Analysis for Multiple Testing (FAMT): an R package for large-scale significance testing under dependence », en revision pour *Journal of statistical Software*.

FRIGUET, C., CAUSEUR, D. :

« Estimation of the proportion of true null hypotheses in high-dimensional data under dependence », en revision pour *Computational Statistics and Data Analysis*.

HENGARTNER, E., MATZNER-LOBER, E., ROUVIERE, L., BURR, T. :

« Multiplicative bias corrected nonparametric smoothers with application to nuclear energy spectrum estimation », en revision pour *Canadian Journal of Statistics*.

LAVERGNE P., PATILEA V. :

« Smooth Minimum Distance Estimation and Testing in Conditional Moment Restrictions Models: Uniform in Bandwidth Theory », en revision pour *Journal of Econometrics*.

LOPEZ O., PATILEA V., VAN KEILEGOM I.:

« A single index model with randomly right censored responses », en revision à *Bernoulli*.

MACKENZIE G, PENG D. :

« Precision of estimators in interval censored PH survival models », soumis à *JRSS B*.

PATILEA V., RAISSI H. :

« Adaptive estimation of vector autoregressive models with time-varying variance: application to testing linear causality in mean », en revision pour *Econometric Theory*.

PATILEA V., RAISSI H. :

« Corrected Portmanteau tests for multivariate autoregressive processes time-varying variance », soumis à *Journal of the American Statistical Association*.

PETIT C., BLANGIARDO M., RICHARDSON S., CHEVRIER C., CORDIER S., COQUET F. :

« A Bayesian model including multiple sources of exposure: effect of environmental insecticide exposure on fetal growth – the PELAGIE mother-child cohort », soumis.

XU, J., MACKENZIE, G. :

« Modelling covariance structure in bivariate marginal models for longitudinal data », en revision pour *Biometrika*.

XU J., MACKENZIE G. :

« On joint modelling of constrained mean and covariance structures for longitudinal data », soumis à *Biometrics*.

6) HABILITATIONS ET THESES DE DOCTORAT

HABILITATIONS

THESES DE DOCTORAT

Aucune thèse ou HDR soutenue par les membres du laboratoire. Plusieurs thèses encadrées, jurys de thèse/HDR, rapporteurs. Si besoin on peut inclure cette information.

6) CONTRATS DE RECHERCHE

Projet *Intersearch*, programme « PME » du pôle Image et Réseaux. Participants : Ensaï (Alain Bidault), l'ENSSAT de Lannion ainsi que les sociétés Swid et Semsoft. Durée 18 mois, à partir de la fin 2010.

Contrat ANR *Datalift*. Partenaires : INRIA (équipes EXMO et Edelweiss), Eurécom, Mondeca, Atos Origin Integration, IGN, INSEE (A. BIDAULT, LSM), FING. Contrat démarré en 10/2010 pour une durée de trois ans.

Contrat de recherche « *Stratégies de localisation des médecins libéraux* » dans le cadre de la chaire Santé Dauphine-Allianz. Participants E. DELATTRE (LSM), A.L. SAMSON. Contrat démarré en mai 2010 pour un an.

Allocation d'Installation Scientifique Rennes Métropole. Contrat démarré en 2009, continué en 2010, il prendra fin en aout 2011. Participants M.A. DIAYE et E. DELATTRE (LSM).

Contrat de recherche *Analyse des inégalités croisées santé-travail* auprès de l'Observatoire des Inégalités. Participants E. DELATTRE (LSM), M. SABATIER. Contrat démarré en mars 2010 pour un an.

Contrat ANR jeunes chercheurs *ATLAS*. Projet porté par P. REYNAUD-BOURET et G. STOLTZ. Parmi les participants M. FROMONT (LSM). Le contrat a pris fin en novembre 2010.

Contrat ANR blanc *CLARA (Clustering in High Dimension : Algorithms and Applications)*. Projet porté par B. PELLETIER. Parmi les participants L. ROUVIERE (LSM).

8) COMMUNICATIONS A DES SEMINAIRES ET CONGRES

BENINEL F.

« Implementing ensemble methods in credit scoring »

- ISBIS Meeting, Portoroz, Slovénie (4-9 juillet).

« From generalized discrimination to logistic discrimination: Logistic sub-models definition for populations mixture in credit-scoring » (avec W. Bouagueul et G. Bel Mufti).

- MSDM Meeting, Hamamet, Tunisie (11-12 Mars, 2010).

CAUSEUR D.

« Large scale significance testing in gene expression studies under dependence »

- Séminaire du Groupe de recherche SSB, Université d'Evry.

« A factor model to analyze heterogeneity in gene expression in a context of QTL characterization » (avec Y. Blum et S. Lagarrigue)

- 8th workshop Statistical Methods for Post-Genomic Data, Luminy, Marseille, 14-15 janvier.
- ISAG 2010 (International Society for Animal Genetics), Edinburgh, 26-30 juillet.

« Inférence sur réseaux géniques par Analyse en Facteurs » (avec Y. Blum, C. Friguet et S. Lagarrigue)

- Journées de Statistique de la SFdS, Marseille, 25-28 mai.

DELATTRE E.

« The determinants of physicians' choices for location : a discrete choice analysis for French General Practitioners » (avec A.L. Samson)

- 10th Journées LAGV, Marseille, juin.
- 8th European Conference on Health Economics, Helsinki, juillet.
- Second French Econometrics Conference, ENSAE-Paris, décembre 2010

« Stratégies de localisation des médecins généralistes français : mécanismes économiques ou hédonistes ? », (avec A.L. Samson)

- Health Economics Seminar, Paris School of Economics, october 2010
- 32^{èmes} Journées des économistes de la santé français, Lyon, décembre.

« Santé et trajectoires professionnelles : l'impact des inégalités sociales » (avec M. Sabatier)

- Colloque *Inégalité et Discriminations*, ENSAI, décembre.

DIAYE M.A.

« The Economics of Performance Appraisals »

- 3^e Conférence Euro-Africaine en Finance et Economie, Université de Paris 1 Panthéon-Sorbonne, Juin.

« Preference, Transitivity axiom and Rational Choice »

- Séminaire MASE (Modélisation et Analyse Statistique et Economie), Ecole polytechnique de Tunis, Septembre.

FROMONT-RENOIR M.

« Tests adaptatifs d'homogénéité pour un processus de Poisson »

- Séminaire du laboratoire JA Dieudonné, Univ. de Nice–Sophia Antipolis, octobre.

MACKENZIE G.

« Modelling covariance structure in bivariate marginal models for longitudinal data » (avec J. Xu)

- 25th IWSM, Glasgow, Juillet.
- Stochastic Modeling Techniques and Data Analysis International Conference, Chania, Crete, juin.

« Precision of estimators in interval censored PH survival models » (avec D. Peng)

- 25th IWSM, Glasgow, Juillet.

« A decade of covariance modelling »

- 12^e Rencontre Math-Industrie de la SMAI, « Les industriels et les mathématiciens se parlent » ENSAI, Mai.
- Stochastic Modeling Techniques and Data Analysis International Conference, Chania, Crete, juin.

PATILEA V.

« Inference with conditional moment restrictions in the presence of right-censoring on the dependent variable » (avec P. Lavergne et O. Lopez)

- Séminaire de Statistique, Université Toulouse 3, février.
- Econometric Seminar, Queen Mary University, mars.
- Statistical Seminar, Universidad de Santiago de Compostela, avril.
- Journées de Statistique de la SFdS, Marseille, mai.
- Weierstrass Institute of Applied Stochastics, Berlin, juin.
- 28th European Meeting of Statistician, Pyraeus, août.

« Adaptive estimation of VAR with Time-Varying variance: application to testing Causality in mean and VAR order »

- Statistical Seminar, Toulouse School of Economics, novembre.
- « A uniform Berry-Esseen theorem on M -estimators for geometrically ergodic Markov chains »
- 10ème Colloque Franco-Roumain de Mathématiques Appliquées, Poitiers, août.

6. FORMATION PAR LA RECHERCHE

COURS SUR LE SITE RENNAIS

PIERRE NEUVIAL

(Laboratoire Statistique et Génome, Génopôle, Evry)

« Méthodes statistiques pour l'analyse de données génomiques »

REMI GRIBONVAL

(Centre de Recherche INRIA Rennes - Bretagne Atlantique)

« Problèmes inverses et parcimonie »

9. VISITES A L'ETRANGER DES MEMBRES DU CREST

BENINEL F.

Office National de Statistiques, Alger, octobre 2010.

Université des Sciences et de la Technologie Houari Boumediene (USTHB), Alger, novembre 2010.

DIAYE M.A.

Center for the Study of Rationality, The Hebrew University of Jerusalem, Jerusalem, 17-30 juillet 2010.

PATILEA V.

Departamento de Estadística e Investigación Operativa, Facultad de Matemáticas, Universidad de Santiago de Compostela (1 semaine en avril)

ROUVIERE L.

Séjour de recherche à l'Institut d'océanographie de l'Université de San Diego (15 jours en février) dans le cadre du projet ANR CLARA.

Séminaire de Statistique IRMAR – ENSAI

Ensaï et l'équipe de statistique de l'IRMAR organise un séminaire hebdomadaire commun qui tourne sur trois sites (Ensaï, Université Rennes 1 et Université Rennes 2). Ci-dessous la liste des exposés qui ont eu lieu à l'Ensaï.

15 janvier

« Estimation des modèles VARMA structurels avec innovations linéaires non corrélées mais non indépendantes »

Boubacar Mainassara Yacouba (Université de Lille 3)

2 avril

« Sur Quelques Problèmes d'Apprentissage Supervisé et Non Supervisé »

Thomas Laloe (Université Lyon I)

7 mai

« Détection d'agrégats temporels et / ou spatiaux d'évènements ponctuels »

Christophe Demattei (CHU de Nimes)

25 juin

« Résidus et tests d'adéquation pour les processus ponctuels de Gibbs marqués »

Frederic Lavancier (Universite de Nantes)

24 septembre

« Space-time models for moving fields with an application to significant wave height fields »

Valérie Monbet (Université Rennes 1)

5 novembre

« Inférence statistique dans les modèles de Markov cachés à effets mixtes »

Maud Delattre (Université Paris 11)

3 décembre

« Mixture modeling for discovering genotype-trait associations »

Andrea Foulkes (University of Massachusets)

Le Laboratoire de Statistique et Modélisation

Le Laboratoire de Statistique et Modélisation (LSM) regroupe une grande partie des enseignants-chercheurs (EC) en poste permanent ou en poste chargé de mission à l'Ensaï, ainsi que un petit nombre des EC statisticiens du bassin rennais qui ne sont pas en poste à l'Ensaï. Le spectre des intérêts de recherche des membres du LSM est très large, allant de la modélisation statistique (aspects théoriques et pratiques), à l'économie quantitative (microéconométrie, économétrie avec applications au domaine du marché du travail, de la santé et de la finance), et l'informatique.

Les statisticiens du LSM collaborent étroitement avec l'équipe de statistique de l'IRMAR (Institut Rennais de Mathématiques ; UMR CNRS évalué A+ par l'Aeres dans la vague B en 2010), ils co-organisent un séminaire tournant, un colloque annuel de statistique (Journées de Statistique à Rennes ; JSTAR), ils développent des actions (papiers de recherche en commun, groupe de travail, ...). Depuis 2010, l'Ensaï a initié un séminaire des doctorants en statistique ouvert aux thésards rennais et aux anciens élèves/étudiants de l'Ensaï ou des établissements de Rennes qui ont poursuivi avec une thèse en statistique dans un autre centre universitaire. Les économistes du LSM ont bénéficié durant l'année 2010 d'une allocation d'installation scientifique de Rennes Métropole pour développer des projets de recherche de micro-économétrie appliquée. Les informaticiens du LSM sont impliqués dans des projets de recherche appliquée dans l'environnement breton ainsi que dans des projets nationaux au nom de l'Insee.

Les thématiques de recherche des membres du LSM pourraient se regrouper en quatre grands axes : biostatistique, économie et finance quantitatives, modélisation des structures des données complexes, statistique semi et non paramétrique. Au sein du LSM ces axes ne sont pas complètement cloisonnés, il existe des interactions entre les axes, plusieurs chercheurs appartient à plusieurs axes. A titre d'exemple, des méthodologies conçues pour les structures des données complexes et des procédures semi paramétriques sont utilisées en biostatistique, les méthodes semi et non paramétriques servent en économie et finance quantitative,...

Biostatistique

Les travaux de David Causeur (Agrocampus Ouest), Gilbert MacKenzie (Visiting professor Ensaï) et Valentin Patilea (Insa de Rennes et Ensaï) s'inscrivent dans cet axe.

Depuis 2009, D. Causeur co-dirige (50 %) la thèse de Yuna Blum (école doctorale VAS) portant sur l'inférence de réseaux de régulation génique par modélisation de composantes d'interaction entre gènes. Plus généralement, ce travail de recherche en biologie des systèmes a donné lieu à la création d'un groupe de travail fédérant les acteurs locaux de la recherche en statistique pour la biologie intégrative (<http://sibgroup.wordpress.com>). Un package R (<http://famt.free.fr>) est disponible pour l'ensemble des méthodes issues de ce travail collaboratif. Par ailleurs, D. Causeur dirige actuellement la thèse d'Anne Lehebel (2010-1013, école doctorale Matisse) sur la modélisation d'une dynamique épidémiologique animale non-observée à partir d'informations multi-sources. Cette thèse vise à mettre en place un système de surveillance épidémiologique en explorant la possibilité de transposer à un contexte épidémiologique les modèles espace-état étudiés par Pierre Tandeo dans une thèse que D. Causeur a dirigée, thèse soutenue en octobre 2010 (école doctorale VAS) sur la modélisation spatio-temporelle des variations de température à la surface des océans. Enfin, D. Causeur développe une collaboration avec le département de psychologie cognitive de National Cheng-Kung University (Tainan, Taiwan) sur la modélisation de l'activité cérébrale à partir de données d'électro-encéphalogrammes. Sur le plan méthodologique, l'objectif est de tenir compte de la dimension spatio-temporelle de l'organisation de cette

activité cérébrale dans les procédures de tests multiples visant, par exemple, à quantifier la réapparition de fonctions du cerveau après un AVC.

Gilbert MacKenzie a rejoint l'Ensaï en 2010. Ses intérêts en matière de recherche se portent sur la construction de modèles statistiques paramétriques et semi paramétriques orientés vers les applications, notamment en biostatistiques. G. MacKenzie a profité de sa présence au CREST-Ensaï pour continuer ses recherches en modélisation de la structure de covariance en présence des données longitudinales, domaine dans lequel il est un des spécialistes au niveau international. Ce type des modèles ont des applications immédiates, par exemples dans la pharmacovigilance. Il a également commencé un projet de recherche sur l'approche de modélisation conjointe de l'espérance et de la structure de covariance pour des données fonctionnelles avec Myriam Vimond (LSM). G. MacKenzie s'est également intéressé à la modélisation en analyse de survie, notamment aux approches alternatives à la modélisation classique de type Cox, comme la h-vraisemblance, pour des données multivariées. Ses modèles en analyse de survie sont motivés par des applications aux essais cliniques randomisés longitudinaux et les données censurées par intervalles. Enfin, il s'est intéressé à la modélisation des tables de contingences en grande dimension et creuses, comme c'est le cas des données d'analyse de sûreté en phase IV des essais cliniques, ou des données de micro-puces en génétique. Les travaux de G. MacKenzie ont donné lieu à plusieurs publications, articles soumis et conférences.

La recherche de Valentin Patilea avec applications à la biostatistique concerne principalement les modèles semi paramétriques définis par des équations estimantes conditionnelles (conditions de moments conditionnels) en présence des observations subissant un mécanisme de censure. Les modèles de régression, la régression quantile, la régression par variables instrumentales, ..., sont des cas particuliers de ces modèles. Les mécanismes de censure sont ceux habituellement rencontrés dans la pratique. Les problématiques abordées sont l'estimation des paramètres (travail en commun avec P. Lavergne et O. Lopez) le test d'adéquation (travail en commun avec O. Lopez). V. Patilea s'est également intéressé aux procédures habituelles de dépistage d'une maladie à l'aide d'un test avec un résultat une variable continue. Le problème consiste alors de choisir d'une manière « optimale » le seuil à partir duquel on déclare la maladie. Une analyse approfondie (tant théorique qu'à l'aide des exemples réels) est proposée dans un article en collaboration avec D. Böhning et H. Holling.

Economie et finance quantitatives

Les travaux de Pierre Clauss (Ensaï), E. Delattre (Ensaï), Marc-Arthur Diaye (Ensaï) se portent sur des thématiques d'économie et finance quantitative.

Pierre CLAUSS a poursuivi ses travaux en finance, plus précisément dans la problématique de gestion de portefeuille. Il s'intéresse au risque de liquidité et à une modélisation simple permettant de le prendre en compte dans le processus de sélection du portefeuille, notamment dans le cadre d'une approche frontière efficiente au sens du critère moyenne-variance. P. Clauss a également proposé des nouveaux indicateurs de performances pour les fonds d'investissement, en particulier les hedge-funds. Ces indicateurs tiennent compte du risque de liquidité du marché financier.

Dans le cadre de sa participation à la Chaire Santé Dauphine-Allianz, Eric Delattre (en collaboration avec Anne-Laure Samson) a développé une modélisation en termes de choix discrets multi-alternatives qui permet d'expliquer les choix de localisation des médecins français. Ce type d'approche permet en outre de simuler l'impact de politiques incitatives destinées à corriger les hétérogénéités spatiales en terme d'offre de soins. A l'aide des données SIP (Santé et Itinéraire Professionnel), E. Delattre et Mareva Sabatier ont proposé une approche tobit multivariée permettant d'expliquer simultanément les caractéristiques des

trajectoires individuelles de santé et sur le marché du travail tout en tenant compte des interactions entre ces trajectoires. Les premiers résultats montrent que les inégalités de niveau de vie produisent des effets négatifs et cumulatifs en affectant non seulement la santé mais aussi le parcours professionnel, alors que les inégalités de genre ou de nationalité ne produisent des effets directs que sur la vie professionnelle. E. Delattre développe également des activités de collaboration avec des anciens étudiants de l'Ensaï actuellement en thèse à l'étranger. Ainsi le séjour de Jean-Baptiste Combes (Univ. d'Aberdeen) a permis de mettre en place un projet de recherche sur les dynamiques salariales au sein des hôpitaux français, à l'aide des données DADS.

Enfin, dans le prolongement du colloque sur les inégalités et la discrimination qui s'est tenu à l'ENSAI en décembre 2010 sous l'impulsion de Marc-Arthur Diaye et Eric Delattre, Eric Delattre coordonne un numéro spécial de la revue *Economie et Statistique* consacré aux problèmes de discrimination et d'inégalités.

La recherche menée en 2010 par Marc-Arthur Diaye comporte deux volets. Un volet théorique sur les préférences individuelles et un volet empirique sur les pratiques incitatives dans les entreprises. En ce qui concerne le volet théorique, l'article avec Daniel Schoch (University of Nottingham in Malaysia) s'est poursuivi. Cet article est une généralisation au cas incertain d'un article de Diaye et Schoch (2009, *Journal of Mathematical Psychology*). L'article en cours avec Gleb Koshevoy (Laboratoire Poncelet, Moscou) s'est aussi poursuivi. Il porte d'une part sur une modélisation de la maximisation d'utilités espérées dans lesquelles les probabilités classiques sont remplacées par des intervalles (random sets) et d'autre part sur un test expérimental (effectué en 2008 à Paris 1). Enfin un travail théorique est en cours avec Nathalie Greenan (Centre d'Etudes de l'Emploi) sur la modélisation des entretiens d'évaluation dans le cas du travail individuel. En ce qui concerne le volet empirique, trois articles ont été poursuivis. Le premier co-écrit avec Nathalie Greenan et Sanka Pekovic (Centre d'études de l'Emploi) porte sur l'analyse, à partir du volet entreprise de l'enquête COI 2006, du comportement d'adoption des normes ISO dans le cadre d'une relation client-fournisseur (supply chain). En effet certaines entreprises ne sont pas certifiées mais exigent que leurs fournisseurs le soient. L'article vise à voir si ces entreprises non certifiées profitent (en termes monétaires) de l'effet signal engendré par leurs fournisseurs certifiés. Le deuxième article co-écrit avec Nathalie Greenan et Michal Urdanivia (Centre d'Etudes de l'Emploi) porte sur l'estimation (à partir de l'enquête COI 2006 appariée avec les DADS) de l'effet de l'évaluation des salariés sur leurs salaires. Il est une extension d'un article précédent (Diaye, Greenan, Urdanivia 2007, document de travail 12979 NBER). Son originalité porte sur la méthode utilisée (régression quantile avec variable instrumentale). Le troisième article co-écrit avec Fatima Lmadani (Urmis) et Michal Urdanivia porte sur une méthode pour tester la discrimination intersectionnelle sur le marché du travail (enquête FQP 2003).

Dans le cadre de cet axe, une conférence (110 participants) sur le thème de la mesure des inégalités a été organisée les 9 et 10 décembre 2010 à l'Ensaï par le CREST et l'Observatoire des Inégalités. E. Delattre et M.A. Diaye ont coordonné l'organisation de cette conférence.

Modélisation des structures des données complexes

Les travaux de Farid Beninel (Ensaï), Alain Bidault (Ensaï), D. Causeur (Agrocampus Ouest), Samuel Maistre (allocataire recherche Ensaï), Valentin Patilea (Ensaï) et Laurent Rouvière (Ensaï) s'inscrivent dans cet axe.

L'activité de recherche de Farid Beninel est centrée est autour de l'apprentissage supervisé et du datamining. La partie datamining est surtout pour parvenir aux informations sur les données, pouvant aider dans les ultimes étapes de l'apprentissage et pour nettoyer les

données et aboutir aux « bonnes données » à soumettre à l'algorithme d'apprentissage choisi. En matière d'apprentissage supervisé, F. Beninel s'intéresse à l'amélioration des règles d'affectation et leur mise à jour. En effet ces règles sont souvent estimées sous des hypothèses peu réalistes (l'indépendance des observations et l'homogénéité de la population ayant fourni l'échantillon...). Cette approche qui part des données et qui peut être assimilée à de l'ingénierie statistique, permet par moment d'identifier des pistes plus formalisées, en terme mathématique. Parmi les voies d'amélioration des règles explorées par F. Beninel citons : 1) l'apprentissage sur une sous population, pour prédire sur une autre ; 2) La combinaison des *classifieurs*.

Les problématiques de recherche d'Alain Bidault s'inscrivent dans le large domaine de recherche lié au web sémantique. Il cherche à faciliter l'accès à de larges volumes de données en s'appuyant sur une description sémantique de ces données, adaptée à chaque contexte de consultation des données. Dans le cadre du projet *Intelsearch* (projet financé par la Région et impliquant des PME du bassin rennais), il cherche à donner les moyens aux utilisateurs du système d'exploiter de façon naturelle, conviviale et ludique les informations présentes au sein de cet univers de l'audiovisuel. Ces travaux de recherche visent à organiser ce large volume d'information pour la rendre disponible et exploitable par un utilisateur. Dans le cadre du projet Datalift financé par l'ANR (l'INSEE en étant partenaire), A. Bidault s'est positionné sur différents niveaux du projet aussi bien sur la définition de l'architecture à retenir pour cet ascenseur de données que sur le langage à utiliser pour décrire les données et faciliter un accès à cette description.

Une partie des activités de recherche de D. Causeur ont été centrées sur les propriétés de méthodes d'inférence en grande dimension, et plus particulièrement les problèmes de stabilité de sélection de modèles que génèrent certaines structures de dépendance forte. Cette thématique prolonge celle abordée par Chloé Friguet dans sa thèse, que D. Causeur a dirigée, thèse soutenue en septembre 2010 (école doctorale Matisse), concernant l'impact de la dépendance sur les procédures de tests multiples en grande dimension.

En prolongement de son mémoire de M2 recherche et en collaboration avec une équipe de Technicolor, Samuel Maistre a proposé des nouvelles techniques pour la construction des intervalles confiance pour les recommandations collaboratives lorsque ces recommandations (estimations de préférences) sont obtenues par une approche de décomposition de la matrice d'observations (les choix déjà faits) à l'aide d'une factorisation en matrices des facteurs de taille réduite. Les intervalles sont obtenus par une technique de bootstrap qui utilise la décomposition matricielle. La nouvelle méthodologie a été appliquée sur les bases de données habituellement utilisées par les chercheurs du domaine.

Valentin Patilea, en collaboration avec son doctorant Matthieu Saumard de l'Insa de Rennes, ont étudié des modèles définis par des équations estimantes conditionnelles lorsque le paramètre et la variable de conditionnement prennent valeurs dans un espace de fonctions (e.g. l'espace des fonctions réelles de carré intégrable définies sur $[0,1]$). Deux procédures d'estimations ont été proposées : une en prolongement de la méthode de moments généralisés, outil classique en économétrie ; l'autre comme extension de l'approche *smooth minimum distance* de Lavergne et Patilea. Le problème de l'adéquation du modèle est actuellement à l'étude.

Laurent Rouvière s'intéresse aux problématiques très actuelles de systèmes de recommandation. En collaboration avec Gérard Biau (Université Paris 6) et Benoît Cadre (ENS Cachan, antenne de Bretagne), ils ont proposé un modèle stochastique permettant l'étude (convergence, vitesse de convergence) des systèmes dits de "recommandation collaborative". Ces systèmes conçus surtout par des chercheurs du monde de l'informatique, sont utilisés par de nombreux sites internet dans un but de marketing. Ils proposent des suggestions personnalisées à des utilisateurs (livres, films, musique) évaluées en fonction de

la proximité entre leurs goûts. Le travail en collaboration de L. Rouvière a été une des premières approches par modélisation probabiliste et statistique de techniques de recommandation collaborative.

Statistique semi et non paramétrique

Les travaux de François Coquet, Magalie Fromont, Marian Hristache, Samuel Maistre, Valentin Patilea, Laurent Rouvière, Céline Vial et Myriam Vimond (tous de l'Ensay) s'inscrivent dans cet axe.

François Coquet poursuit son travail avec Myriam Vimond et Jay Breidt (Colorado State University) concernant l'étude des données SAXS (Small Angle X-ray Scattering). Le problème considéré est un problème inverse du type Tikhonov. À partir de l'intensité de diffraction mesurée, on cherche à estimer la distribution de distance de la molécule. Cette distribution est à support compact, et elle est assez régulière : elle apporte des informations sur la géométrie de la molécule et sur ses interactions possibles. L'intensité et la distribution de distance sont reliées par la transformation de Fourier indirecte (Indirect Fourier Transform) qui est non inversible. F. Coquet, en collaboration avec Daniel Bonnéry et Jay Breidt, étudie les propriétés asymptotiques de l'estimateur de Horwitz-Thomson, notamment dans le cadre de plans de sondages raisonnablement stratifiés.

Magalie Fromont Renoir développe des travaux portant sur le rééchantillonnage et les tests non paramétriques adaptatifs au sens du minimax. Faisant suite à un article écrit en collaboration avec Béatrice Laurent et Patricia Reynaud-Bouret sur des tests adaptatifs d'homogénéité pour un processus de Poisson, dans le cadre de l'ANR Atlas, un nouvel article sur des tests adaptatifs de comparaison de processus de Poisson est en cours de rédaction. Les tests construits ici sont basés sur des méthodes de rééchantillonnage fines pour les processus de Poisson marqués, et font appel à des méthodes à noyaux utilisées en théorie de l'approximation et en apprentissage statistique. Des applications en neurosciences et à des données de l'Insee sur la répartition de services sont à l'étude.

Marian Hristache a poursuivi son étude des modèles semi paramétriques définis par des conditions de moments conditionnels. Le problème abordé consiste à considérer deux ou plusieurs restrictions de moments conditionnels avec des ensembles de variables de conditionnement différentes, pas forcément dans une relation d'inclusion. Les conditions de moment contiennent un paramètre d'intérêt de dimension finie et un paramètre de nuisance de dimension infinie. Les applications sont nombreuses en statistique et en économétrie. L'objectif de la recherche est de trouver la borne d'efficacité semi paramétrique d'un tel modèle général et d'en proposer des estimateurs efficaces. Cette année M. Hristache s'est concentré sur les applications de son étude générale à la modélisation des données incomplètes.

Samuel Maistre s'intéresse à des tests de significativité des variables en régression non paramétrique. Il est bien connu que lorsque le nombre de variables augmente, l'estimation d'une régression non paramétrique a des mauvaises propriétés statistiques. Pour diminuer l'effet de ce problème qui se pose également dans un test de significativité des variables en régression non paramétrique, S. Maistre a adapté et étendu une idée de réduction de la dimension par projections linéaires déjà utilisée dans le cadre des tests d'adéquation non paramétriques pour les modèles paramétriques. Ceci représente une première étape dans un projet de construction de tests pour la validation non paramétrique des modèles à direction révélatrices (index regressions).

Valentin Patilea, en collaboration avec Hamdi Raïssi (Insa de Rennes), a utilisé un estimateur à noyau de la variance dans un modèle VAR (Vector Autoregressive) avec un bruit de variance variable dans le temps afin de définir des meilleurs estimateurs des

coefficients du VAR. A l'aide de ces estimateurs, ils ont mis en place une technique de test de causalité ainsi que des corrections du test du portmanteau pour le choix de l'ordre du modèle. Les propriétés statistiques (en particulier des résultats asymptotiques uniformes dans la fenêtre de l'estimateur à noyau) des nouvelles procédures ont été étudiées et des exemples réels ont été traités.

Sur le thème de la statistique non paramétrique, Laurent Rouvière, en collaboration avec Tom Burr, Steve Myers, Nicolas Hengartner (chercheurs à Los Alamos) et Eric Matzner-Lober (Professeur à l'Université Rennes 2), a étudié les propriétés théoriques (vitesse de convergence) et pratiques (sur des données spectrales) d'une méthode permettant de réduire le biais d'estimateurs non paramétrique de la fonction de régression sans affecter la variance asymptotique.

Le travail de recherche de Céline Vial a essentiellement porté sur l'estimation de la régularité des processus. Ce travail a visé à estimer un paramètre de régularité d'un processus. En général ce (ou ces) paramètre(s) est (sont) inconnu(s) alors que de fortes hypothèses de régularité sont nécessaires dans un grand nombre de problèmes : approximation, intégration, prédiction et estimation. On a considéré ici un processus de classe Hölder $C(r, \beta)$ en moyenne quadratique, où les deux paramètres sont inconnus. Récemment, C. Vial et D. Blanke ont proposé un nouvel estimateur fortement convergent de r basé sur les variations quadratiques du processus. Une étude des performances numériques de cet estimateur et de celui proposé dans un travail antérieur de Blanke et Vial est en cours ainsi qu'une estimation du deuxième paramètre d'intérêt β .

Myriam VIMOND Myriam a effectué un travail en collaboration avec Jérémie Bigot et Jean-Michel Loubes (Université de Toulouse 3) qui porte sur l'estimation des déformations qui peut exister entre des images bruitées lorsque l'image de référence est inconnue. Les déformations sont modélisées comme des paramètres de dimension finie appartenant à un groupe de Lie. L'image de référence est modélisée par une fonction, i.e. c'est un paramètre infini-dimensionnel. Le modèle étudié est donc semi paramétrique. Nous avons travaillé à la construction d'un estimateur consistant pour les déformations. Cet estimateur est défini comme le minimum d'un critère qui est fonctions de la transformée de Fourier des images. Nous proposons également une étude sur la vitesse de convergence et la normalité asymptotique de cet estimateur via la carte exponentielle du groupe de Lie. Ce travail s'ouvre ensuite sur une étude de l'efficacité asymptotique pour des paramètres appartenant à un groupe de Lie ou une variété différentiable.