Translations of Abstracts

BIOMETRIC METHODOLOGY

Dans plusieurs domaines d'application notamment en biologie, on cherche à décomposer les effets d'une exposition sur une réponse en des effets agissant par des voies différentes. Par exemple, on peut souhaiter décomposer l'effet d'une consommation importante d'alcool sur la pression artérielle systolique (PAS) en des effets s'exprimant sur l'indice de masse corporelle (IMC) via la gamma-glutamyl transpeptidase (GGT) et des effets s'exprimant par d'autres voies. Grâce principalement à des contributions provenant du champ de l'inférence causale, des progrès importants ont été accomplis concernant la compréhension de la nature précise des quantités statistiques à estimer qui rendent compte de tels effets intuitifs, les hypothèses nécessaires à leur identification et les méthodes statistiques requises. Ces contributions se sont concentrées presque uniquement sur des contextes avec un seul médiateur ou un ensemble de médiateurs considérés en un seul bloc. Dans beaucoup d'applications cependant, les chercheurs visent une décomposition bien plus ambitieuse en des effets spécifiques de voies particulières à partir de beaucoup de médiateurs. Dans cet article, nous donnons des définitions contrefactuelles de telles quantités à estimer dans des contextes avec des médiateurs multiples quand des médiateurs intervenant précocement peuvent avoir des effets sur des médiateurs intervenant plus tardivement et montrons qu'il y a de nombreuses décompositions possibles. Nous discutons d'hypothèses fortes permettant l'identification des effets et suggérons une approche basée sur des analyses de sensibilité quand certaines de ces hypothèses ne peuvent pas être justifiées. Ces idées sont illustrées sur des données de consommation d'alcool ainsi que de PAS, IMC et GGT tirées de l'étude dite « Izhevsk Family Study ». Nous cherchons à combler le fossé existant entre une théorie basé sur un médiateur unique et une théorie basé sur des médiateurs multiples, ce qui nous permet de souligner la nature ambitieuse de cette entreprise et de formuler des suggestions pratiques sur la meilleure façon de procéder.

Du fait de l'augmentation des coûts du développement thérapeutique, des critères de substitution pour évaluer des nouvelles molécules dans des essais cliniques sont de plus en plus demandés. Cependant, il est devenu évident que les critères de substitution requièrent une évaluation et une validation statistique avant de pouvoir être utilisés à la place du critère final dans les essais cliniques. Actuellement, deux paradigmes fondés sur l'inférence causale et sur la méta-analyse, dominent ce champ de recherche. Bien qu'une

littérature abondante concerne ces deux paradigmes, leur interrelation n'a pas été explorée jusqu'à présent. Dans ce travail, nous discutons le cadre conceptuel de ces deux approches et nous étudions le lien les unissant à partir d'éléments théoriques et d'études de cas. De plus, nous montrons d'une part que l'approche méta-analytique peut être incluse dans celle de l'inférence causale et d'autre part qu'il existe un argument heuristique qui explique pourquoi les critères de substitution évalués avec succès par l'approche méta-analytique ont fréquemment des propriétés intéressantes du point de vue de l'inférence causale. Un nouveau package, R *Surrogate*, est fourni pour réaliser de telles évaluations.

Y. Choai and S. Matsui **25**
Estimation of Treatment Effects in All-Comers Randomized Clinical Trials with a Predictive Marker

Les progrès récents en génomique et dans les biotechnologies ont accéléré le développement de traitements ciblés sur des marqueurs prédictifs de la réponse au traitement. Cependant, il est habituel qu'au début d'un essai de phase III il n'y ait pas de base biologique établie ou de données d'essais préliminaires sur un candidat marqueur concernant sa capacité à prédire les effets du traitement. Dans ce cas, il est raisonnable d'inclure et de randomiser tous les patients éligibles mais de planifier de façon prospective des analyses de sous-groupes basés sur ce marqueur. Un plan d'analyse dans ces schémas du tout-venant est l'approche dite de repli qui consiste à tester tout d'abord l'efficacité thérapeutique globalement chez tous les patients inclus puis dans le sous-groupe des patients positifs au biomarqueur si ce premier test n'est pas significatif. Avec cette approche, du fait de la nature adaptative de l'analyse et de la corrélation entre les deux tests, l'estimation de l'effet du traitement dans le sous-groupe à biomarqueur positif après un premier test global non significatif sera biaisée. Dans cet article, nous formulons la fonction de biais sur un domaine borné du paramètre d'efficacité en utilisant des fonctions polynomiales. Nous fournissons également une méthode d'estimation par intervalle basée sur une distribution normale bivariée doublement tronquée. Des simulations montrent que cette approche conduit à une réduction du biais. Un exemple d'application à un essai de phase III dans le cancer du poumon est présenté.

A. S. Hedayat, J. Wang, and T. Xu **33**
Minimum Clinically Important Difference in Medical Studies

Dans les essais cliniques, la différence minimale cliniquement importante (MCID) a fait l'objet d'un intérêt croissant en tant qu'outil majeur de support à l'inférence clinique et statistique. Beaucoup de méthodes d'estimation du MCID ont été développées, à partir d'intuitions variées, avec peu de justifications théoriques. Cet article propose un nouveau cadre d'estimation du MCID, intégrant à la fois des mesures diagnostiques objectives et des mesures subjectives (*patient-reported outcomes*, PRO). Il s'agit d'abord de formuler la recherche d'un MCID de population comme un problème de classification à forte marge puis de l'étendre à un MCID personnalisé permettant d'individualiser la valeur du seuil pour les patients dont les profils cliniques pourraient affecter les mesures des PRO. Deux propriétés sont particulièrement importantes : d'une part, nous démontrons que le cadre d'estimation proposé est asymptotiquement consistant et, d'autre part, nous déterminons une borne supérieure de l'écart entre le MCID estimé sur un échantillon fini

et le MCID idéal. L'avantage de la méthode que nous proposons est illustré au travers d'une variété de simulations et de deux essais cliniques de phase III.

O. Saarela and J. A. Hanley **42**
Case-Base Methods for Studying Vaccination Safety

La mise en commun des témoins dans des schémas cas-témoins nichés peut produire des gains d'efficacité substantiels comparée à l'analyse standard appariée sur le temps utilisant la méthode de Mantel-Haenszel ou la régression logistique conditionnelle. Dans le contexte des effets indésirables possibles des vaccinations de la petite enfance, nous proposons de mettre en commun l'information des témoins pour estimer la prévalence de l'exposition dans la population comme une fonction paramétrique ou non paramétrique du temps et éventuellement d'autres facteurs. Cette fonction peut à son tour être utilisée comme quantité estimée à insérer pour contrôler la confusion dans l'estimation ultérieure de rapports de taux. Nous déduisons les écarts-types pour les estimateurs en deux étapes qui en résultent, démontrons par des simulations les gains d'efficacité par rapport à une analyse appariée standard et proposons une nouvelle présentation graphique des données de vaccination et des temps avant événement indésirable. Nous formulons les méthodes dans le cadre général d'un échantillonnage basé sur les cas qui englobe les différentes méthodes cas-témoins et séries de cas.

Z. Geng, S. Wang, M. Yu, P. O. Monahan, V. Champion, and G. Wahba **53**
Group Variable Selection Via Convex Log-Exp-Sum Penalty with Application to a Breast Cancer Survivor Study

Dans de nombreuses applications scientifiques et d'ingénierie, des covariables sont naturellement regroupées. Lorsque les structures de groupe sont disponibles parmi les covariables, on s'intéresse généralement à l'identification à la fois des groupes importants et des variables importantes au sein des groupes sélectionnés. Parmi les méthodes existantes de sélection de variables groupe, certaines ne parviennent pas à conduire la sélection au sein du groupe. D'autres méthodes permettent la sélection à la fois du groupe et au sein du groupe mais les fonctions objectives correspondantes sont non convexes. Une telle non-convexité peut nécessiter un effort numérique supplémentaire. Dans cet article, nous proposons une nouvelle pénalité dite « Log-Exp-Sum » (LES) pour la sélection de la variable groupe. La pénalité LES est strictement convexe. Elle permet d'identifier des groupes importants ainsi que de sélectionner des variables importantes au sein du groupe. Nous développons un algorithme descendant coordonné au niveau du groupe pour ajuster le modèle. Nous dérivons également les marges d'erreurs non asymptotiques et la convergence asymptotique de la sélection du groupe pour notre méthode dans le cas de données à grandes dimensions où le nombre de covariables peut être beaucoup plus important que le nombre de sujets. Les résultats numériques mettent en évidence la bonne performance de notre méthode à la fois pour la sélection de variables et pour la prédiction. Nous avons appliqué la méthode proposée à un jeu de données de l'American Cancer Society sur les survivantes du cancer du sein. Les résultats sont cliniquement significatifs et peuvent aider à la conception de programmes d'intervention pour améliorer la qualité de vie des survivantes du cancer du sein.

Y. Shen, K. P. Liao, and T. Cai **63**

Sparse Kernel Machine Regression for Ordinal Outcomes

Des variables ordinales surviennent fréquemment dans les études cliniques où chaque sujet est affecté à une catégorie, les catégories étant ordonnées. Des règles de classification de résultats ordinaux peuvent être développées avec des modèles de régression couramment utilisés, tels que le modèle (fCR) du rapport complet de continuation (CR) qui permet aux effets de covariables de différer pour tous les rapports de continuation et le modèle CR avec une structure de risques proportionnels (pCR) qui suppose que les effets des covariables sont constants pour tous les rapports de continuation. Pour les situations où les effets des covariables diffèrent entre certains rapports de continuation, mais pas pour tous, l'ajustement par fCR ou par pCR peut conduire à une performance de prédiction non optimale. En outre, ces modèles standard ne permettent pas de prendre en compte des effets non-linéaires des covariables. Dans cet article, nous proposons une méthode de régression parcimonieuse à noyau de CR (KM) pour des résultats ordinaux où nous utilisons le cadre KM pour intégrer la non-linéarité et imposer la parcimonie sur les différences globales entre les effets des covariables de rapports de continuation pour contrôler un sur-ajustement. En outre, nous fournissons des règles basées sur les données pour sélectionner un noyau optimal afin de maximiser la précision de la prédiction. Les résultats des simulations montrent que les procédures que nous proposons fonctionnent bien dans les deux situations linéaires et nonlinéaires, en particulier lorsque le vrai modèle sous-jacent est entre deux modèles fCR et pCR. Nous appliquons nos procédures afin de développer un modèle de prédiction des niveaux d'anticorps anti-CCP chez les patients atteints de polyarthrite rhumatoïde et de démontrer l'avantage de notre méthode sur d'autres méthodes couramment utilisées.

L. Zhu, H. Zhao, J. Sun, W. Leisenring, and L. L. Robison **71**
Regression Analysis of Mixed Recurrent-Event and Panel-Count Data with Additive Rate Models

Les événements récurrents sont étudiées dans de nombreux domaines tels que la démographie, l'épidémiologie, la médecine et les sciences sociales (Cook et Lawless, 2007; Zhao et coll., 2011). Pour de telles analyses, deux types de données ont été étudiés : les données d'événements récurrents et les données de comptage en panel. Cependant, dans la pratique, il existe un troisième type de données, des données mixtes d'événements récurrents et de comptage en panel ou des données mixtes d'histoires d'événements. De telles données sont rencontrées lorsque certains sujets de l'étude sont suivis ou observés de façon continue dans le temps et fournissent ainsi des données d'événements récurrents tandis que les autres sujets sont observés uniquement à des temps discrets et ne donnent donc que des données de comptage en panel. Une situation plus générale est celle où chaque sujet est observé de façon continue sur certaines périodes de temps mais seulement à des temps discrets sur d'autres périodes de temps. Il existe peu de littérature sur l'analyse de ces données mixtes à l'exception de celle publiée par Zhu et coll. (2013). Dans cet article, nous considérons l'analyse de régression des données mixtes en utilisant le modèle de taux additif et développons des approches basées sur des équations d'estimation pour estimer les paramètres de régression d'intérêt. Aussi bien les propriétés en échantillon fini que les propriétés asymptotiques des estimateurs obtenus sont établies et les études numériques montrent que la méthode proposée fonctionne bien pour des situations concrètes. L'approche est appliquée à l'étude « Childhood Cancer Survivor

Study » qui a motivé cette étude.

F. Bartolucci and A. Farcomeni **80**
A Discrete Time Event-History Approach to Informative Drop-Out in Mixed Latent Markov Models with Covariates

Les modèles mixtes markoviens latents (MLM) forment un outil important de l'étude de données longitudinales lorsque les variables de réponse sont sous l'emprise d'hétérogénéités non observées fixe dans le temps et variable dans le temps, la dernière étant prise en compte par une chaîne de Markov cachée. Pour écarter ce biais en utilisant un modèle de ce type en présence de sorties d'étude informatives, nous proposons une extension de type événement-historique (EH) de l'approche de Markov latente, pouvant être utilisée avec des données longitudinales multivariées, pour lesquelles un ou plusieurs résultats de natures différentes sont observés à chaque date d'étape. La composante EH du modèle qui en résulte est relatée aux sorties d'études censurées par intervalle, et le biais dans la modélisation MLM est évité par des effets aléatoires corrélés, inclus dans les différentes composantes du modèle, et qui suivent des distributions latentes communes. Pour réaliser une estimation au maximum de vraisemblance du modèle proposé, à l'aide de l'algorithme espérance-maximisation (EM), nous étendons les récursions backward-forward usuelles de Baum et Welch. L'algorithme a la même complexité que celui choisi dans les cas de sorties non informatives. Nous illustrons l'approche proposée au travers de simulations, et d'une application fondée sur des données provenant d'une étude clinique sur la cirrhose biliaire primitive dans laquelle deux critères d'intérêt existent, l'un étant continu et l'autre binaire.

J. M. Lange, R. A. Hubbard, L. Y. T. Inoue, and V. N. Minin **90**
A Joint Model for Multistate Disease Processes and Random Informative Observation Times, with Applications to Electronic Medical Records Data

Les modèles multi-états sont utilisés pour caractériser les évolutions naturelles des individus au cours de maladies à états discrets. Les ressources en termes de données observationnelles issues d'enregistrements médicaux électroniques offrent de nouvelles opportunités pour étudier de telles maladies. Néanmoins, ces données sont des observations du processus à des temps particuliers et non en continu, ces temps pouvant être ou bien pré-déterminés et non informatifs ou bien déterminés par la survenues de symptômes et donc informatifs quant au stade sous-jacent de la maladie chez l'individu. C'est pourquoi nous avons développé un nouveau modèle de maladie, modèle conjoint des temps d'observation et des transitions entre les stades de la maladie. Le processus de la maladie est modélisé selon une chaîne de Markov à temps continu latent ; le processus d'observation, quant à lui, est modélisé selon un processus de Poisson modulé de façon markovienne où la fréquence des observations dépend des stades sous-jacents de la maladie chez les individus. Le processus de la maladie est observé à des temps dont certains sont informatifs et d'autres non, sachant qu'il est possible que certains états s'avèrent mal classifiés. Nous démontrons que le modèle est tout-à-fait accessible du point de vue calculatoire et concevons un algorithme d'estimation-maximisation (EM) pour l'estimation des paramètres. Sur données simulées, nous montrons combien les estimateurs issus de notre modèle conjoint des temps d'observation et des transitions inter-états conduisent à des estimateurs à la fois moins biaisés et plus précis des

paramètres décrivant la maladie. Nous appliquons le modèle à une étude réelle des événements secondaires du cancer du sein, en utilisant les enregistrements de mammographie et de biopsie d'un échantillon de femmes avec antécédent de cancer du sein primaire.

P. Blanche, C. Proust-Lima, L. Loubère, C. Berr, J.-F. Dartigues, and H. Jacqmin-Gadda **102**

Quantifying and Comparing Dynamic Predictive Accuracy of Joint Models for Longitudinal Marker and Time-to-Event in Presence of Censoring and Competing Risks

En raison de l'intérêt croissant pour la médecine personnalisée, les modèles conjoints pour marqueurs longitudinaux et données de survie ont récemment commencé à être utilisés pour calculer des prédictions de risques individuels dynamiques. Ces prédictions sont dites dynamiques car elles sont actualisées au fur et à mesure que l'information sur le profil de santé d'un sujet change au cours de son suivi. Dans ce travail, nous nous intéressons aux méthodes statistiques pour quantifier et comparer les capacités pronostiques de ce type de modèles pronostiques, en tenant compte de la censure à droite et d'éventuels risques concurrents. L'aire sous la courbe ROC (AUC) et le score de Brier (BS) sont utilisés pour quantifier les capacités pronostiques dynamiques. Une méthode non paramétrique de pondération par l'inverse de la probabilité de censure est utilisée pour estimer les courbes dynamiques de l'AUC et du BS en fonction du moment où les prédictions sont réalisées. Des résultats asymptotiques sont établis et des méthodes de calcul d'intervalles de confiance ponctuels et simultanés en sont dérivées. Des tests sont également proposés pour comparer les courbes de capacités pronostiques dynamiques de deux modèles. Les performances des procédures d'inférence sont évaluées par des simulations. Les méthodes proposées sont appliquées pour comparer différents modèles pronostiques de la démence chez les personnes âgées, en tenant compte du risque concurrent de décès. Les modèles pronostiques sont basés sur les mesures répétées de deux tests psychométriques. Ils sont estimés sur la cohorte française Paquid et leurs capacités pronostiques sont évaluées et comparées sur la cohorte française des Trois-Cités.

X. Chang, R. Waagepetersen, H. Yu, X. Ma, T. R. Holford, R. Wang, and Y. Guan **114**

Disease Risk Estimation by Combining Case–Control Data with Aggregated Information on the Population at Risk

Nous proposons un cadre statistique innovant consistant à compléter des données cas-témoins par des résumés statistiques pour un ensemble de facteurs de risque sur la population à risque considérée. Notre approche consiste d'abord à écrire deux équations d'estimation non biaisées, l'une basée sur les données cas-témoins et l'autre sur les données sur les cas et les résumés statistiques puis à les combiner de façon optimale afin d'obtenir une autre équation d'estimation utilisable pour l'estimation. La méthode proposée est simple sur le plan calculatoire et plus efficace que les approches standard basées sur les seules données cas-témoins. Nous établissons également les propriétés asymptotiques de l'estimateur résultant et étudions ses propriétés sur des échantillons de taille finie par simulation. Nous illustrons la méthode proposée par l'étude des facteurs de risque du cancer de l'endomètre en utilisant des données d'une récente étude cas-témoins

basée sur une population et des résumés statistiques issus du programme « Behavioral Risk Factor Surveillance System » qui est le programme d'estimation populationnelle de l'institut de recensement des États-Unis (« US Census Bureau ») et du département des transports de l'état du Connecticut (« Connecticut Department of Transportation »)

Les méthodes par variable instrumentale (IV) sont populaires dans les études non expérimentales pour estimer les effets causaux des interventions médicales. Ces approches permettent l'estimation consistante des effets du traitement même si des facteurs de confusion importants ne sont pas observés. Malgré l'utilisation croissante de ces méthodes, il y a eu peu d'extensions des méthodes IV aux problèmes des données censurées. Dans cet article, nous discutons les difficultés de l'application des techniques IV au modèle à risques proportionnels et démontrons l'utilité de la formulation à risques additifs pour les analyses IV avec données censurées. Sous l'hypothèse de modèles d'équations structurelles linéaires pour la fonction de risque, nous développons une expression analytique d'un estimateur en deux étapes pour l'effet causal dans le modèle à risques additifs. Les méthodes autorisent à la fois des expositions continues et discrètes et permettent l'estimation de mesures causales de survie relative. Les propriétés asymptotiques des estimateurs sont déduites et il est montré que les inférences qui en résultent se comportent bien dans des études de simulation et dans une application à un jeu de données sur l'efficacité d'un nouvel agent chimiothérapeutique pour le cancer du côlon.

Nous considérons les problèmes de sélection de modèles et d'estimation dans un contexte où plusieurs modèles concurrents basés sur des équations différentielles ordinaires (ODE) existent et où tous ces modèles sont des cas particuliers d'un modèle « complet ». Nous proposons une approche peu coûteuse au plan calculatoire qui se base sur l'estimation statistique dans le modèle complet puis sur une combinaison de l'approximation des moindres carrés (LSA) et du LASSO adaptatif. Nous montrons que la méthode qui en résulte et que nous appelons la méthode LSA est (asymptotiquement) une méthode de sélection de modèles de type oracle. Les propriétés sur des échantillons finis de la méthode LSA proposée sont étudiées au moyen d'une étude de simulation. Cette étude examine le pourcentage de sélection des modèles ODE corrects, l'efficacité de l'estimation des paramètres par rapport à l'utilisation des modèles corrects et complets et les probabilités de recouvrement des intervalles de confiance estimés pour les paramètres ODE. Elle montre des performances satisfaisantes vis-à-vis de tous ces critères. Notre méthode est illustrée par la sélection du meilleur modèle ODE prédateur-proie pour la dynamique des populations de lynx et de lièvre parmi quelques modèles ODE bien connus et biologiquement interprétables.

Pocock et collègues (2012) ont proposé une approche dite de « win ratio » pour analyser les critères d'évaluation composites dont les composantes sont d'importances différentes du point de vue clinique. Dans cette communication, nous établissons un cadre statistique pour cette approche. Nous en dérivons une hypothèse nulle et proposons un estimateur de la variance de type fermé pour la statistique du « win ratio », applicable à toute situation d'appariement. Notre étude de simulation montre que l'estimateur proposé se comporte bien indépendamment de l'amplitude de l'effet du traitement et du type de la distribution jointe des événements.

M. P. Fay, M. A. Proschan, and E. Brittain **146**
Combining One-Sample Confidence Procedures for Inference in the Two-Sample Case

Nous présentons une méthode simple et générale pour combiner deux procédures de confiance pour un échantillon pour obtenir des inférences dans le problème de deux échantillons. Certaines applications donnent des connexions frappantes avec les méthodes établies ; par exemple, la combinaison de procédures exactes de confiance pour la loi binomiale donne de nouveaux intervalles de confiance sur la différence ou le rapport des proportions qui correspondent à des inférences en utilisant le test exact de Fisher et les études numériques montrent des intervalles de confiance associés liés au taux d'erreur de type I. La combinaison de procédures exactes de confiance pour un échantillon Poissonnien recrée les intervalles de confiance standard pour le rapport et en propose de nouveaux pour la différence. La combinaison de procédures de confiance associée aux tests t pour un échantillon recrée les intervalles de Behrens-Fisher. D'autres applications donnent de nouveaux intervalles de confiance avec moins de suppositions que précédemment nécessaires. Par exemple, la méthode crée de nouveaux intervalles de confiance sur la différence de médianes qui ne nécessitent pas de suppositions de décalage et de continuité. Nous créons un nouvel intervalle de confiance pour la différence entre deux distributions de survie à un point de temps fixe quand il y a censure indépendante en combinant la procédure récemment mise au point du produit bêta de confiance pour chaque échantillon séparé. L'intervalle résultant est conçu pour garantir une couverture indépendamment de la taille de l'échantillon ou de la censure de la distribution, et produit des conclusions équivalentes à un test exact de Fisher quand il n'y a pas de censure. Nous montrons théoriquement que lors de la combinaison d'intervalles asymptotiquement équivalents à des intervalles Normaux, notre méthode a une couverture asymptotiquement exacte. Toutes les situations étudiées suggèrent la garantie d'une couverture nominale pour notre nouvel intervalle chaque fois que les procédures de confiance originales elles-mêmes assurent cette couverture.

A. Touloumis, S. Tavaré, and J. C. Marioni **157**
Testing the Mean Matrix in High-Dimensional Transposable Data

Les informations structurelles des données transposables de grande dimension nous permettent d'écrire les données enregistrées pour chaque sujet dans une matrice de telle sorte que les lignes et les colonnes correspondent à des variables d'intérêt. Un problème important est de tester l'hypothèse nulle que la matrice des moyennes a une structure particulière sans ignorer la structure de dépendance parmi et/ou entre les variables de lignes et de colonnes. Pour y remédier, nous développons une procédure de test non-

paramétrique générique et peu coûteuse en terme de calcul pour évaluer l'hypothèse que, dans chaque sous-ensemble prédéfini de colonnes (lignes), le vecteur colonne (ligne) de la moyenne reste constant. Dans les études de simulation, la procédure de test proposée semble avoir de bonnes performances et, contrairement aux approches simples, elle conserve le risque α nominal et reste puissante même si les variables de ligne et/ou de colonne ne sont pas indépendantes. Nous illustrons l'utilisation de la méthodologie proposée par deux exemples empiriques sur des puces à ADN.

BIOMETRIC PRACTICE

Les études de séries temporelles multi-sites ont apporté la preuve d'une association entre l'exposition à court terme à des particules (PM) et les effets néfastes sur la santé mais la taille de l'effet varie à travers les États-Unis. La variabilité de l'effet peut être en partie due aux différents niveaux d'exposition et à des caractéristiques de santé au sein des communautés mais également à la composition chimique des particules qui peut varier considérablement selon le lieu et le temps. L'objectif de ce document est d'identifier les composants particulièrement nocifs de ces mélanges chimiques. En raison du grand nombre de composants hautement corrélés, nous devons intégrer certaines régularisations dans un modèle statistique. Nous supposons que, à chaque localisation spatiale, les coefficients de régression proviennent d'un modèle de mélange avec recherche stochastique pour la sélection de variables mais en utilisant un partage d'information entre l'inclusion des variables et la taille de l'effet dans les différents endroits. Le modèle diffère des techniques actuelles de sélection de variables spatiales en s'adaptant à la sélection des variables locales et globales. Le modèle est utilisé pour étudier l'association entre les particules fines (PM<2,5μm), mesurées dans 115 comtés à l'échelle nationale sur la période 2000-2008, et les admissions aux urgences cardiovasculaires chez les patients du régime Medicare.

Les modèles conjoints pour les critères d'évaluation longitudinaux et les critères de type survie sont de plus en plus utilisés dans les études cliniques. La spécification correcte des effets fixes et aléatoires est fondamentale en pratique pour l'analyse statistique de ces données. Une sélection simultanée de variables parmi les composantes longitudinales et de survie est nécessaire pour se prémunir contre une mauvaise spécification du modèle. Cependant, la sélection des variables dans de tels modèles n'a pas été étudiée et il n'existe à notre connaissance aucun outil logiciel disponible. Dans cet article, nous décrivons une méthode de vraisemblance pénalisée avec des fonctions de pénalité de type LASSO adaptatif pour permettre la sélection simultanée d'effets fixes et aléatoires dans les modèles conjoints. Afin de sélectionner les composantes de variance des effets aléatoires, nous reparamétrons les composantes de variance en utilisant la décomposition de Cholesky, ce qui aboutit à introduire une fonction de pénalité de type « shrinkage » de groupe. Afin de réduire le biais dans l'estimation résultant de la pénalisation, nous

proposons une procédure de sélection en deux étapes dans laquelle la seconde étape permet de réduire le biais. La vraisemblance pénalisée est approximée par une quadrature gaussienne et optimisée par l'utilisation d'un algorithme EM. Une étude de simulation montre que les résultats de sélection sont excellents lors de la première étape et que les biais d'estimation sont faibles dans la seconde étape. À titre d'illustration, nous analysons un marqueur clinique mesuré de façon longitudinale et la survie des patients dans une cohorte de patients atteints d'insuffisance cardiaque.

R. Graziani, M. Guindani, and P. F. Thall **188**
Bayesian Nonparametric Estimation of Targeted Agent Effects on Biomarker Change to Predict Clinical Outcome

L'effet d'un agent sur la réponse clinique d'un patient atteint de cancer peut être, de façon putative, prédit par l'intermédiaire de son effet sur des événements biologiques survenant plus précocement. Ce fait est motivé par des essais précliniques sur la cellule ou l'animal où des biomarqueurs dichotomiques ou quantitatifs permettent d'identifier de tels événements. Lors de l'évaluation d'agents ciblés chez l'homme, les questions centrales sont de savoir si la distribution d'un biomarqueur varie au cours du traitement, de connaître la nature et l'intensité d'une telle modification et si cette dernière est associée avec la réponse clinique.
La difficulté majeure dans l'estimation de ces effets réside dans le fait que la distribution d'un biomarqueur peut être très complexe, varie considérablement selon les patients et possède une relation compliquée avec celle de la réponse clinique. Nous présentons un cadre probabiliste cohérent en vue de la modélisation et de l'estimation de ces aspects au travers d'un modèle de mélange non-paramétrique et bayésien hiérarchique pour les biomarqueurs que nous utilisons pour définir un profil fonctionnel d'une variation de distribution pré versus post traitement du biomarqueur. La forme fonctionnelle est similaire à la caractéristique opérationnelle du récepteur utilisée dans les tests de diagnostic et nous utilisons le profil comme une covariable dans un modèle de régression de la réponse clinique. La méthodologie est illustrée par l'analyse de données d'un essai clinique dans le cancer de la prostate utilisant l'imatinib pour cibler un facteur de croissance dérivé des plaquettes avec pour but clinique l'amélioration du temps de survie sans progression.

X. J. Lee, C. C. Drovandi, and A. N. Pettitt **198**
Model Choice Problems Using Approximate Bayesian Computation with Applications to Pathogen Transmission Data Sets

Dans le cadre d'inférences complexes, il arrive que les fonctions de vraisemblance ne puissent pas être exprimées sur le plan analytique ou gérées sur le plan calculatoire, ce qui les rend inaccessibles aux méthodes bayésiennes standard. Les méthodes de Calcul Bayésien Approché (CBA) abordent un tel problème inférentiel en remplaçant les évaluations directes de la vraisemblance par celles issues d'un échantillonnage répété à partir du modèle. Les méthodes CBA ont été appliquées essentiellement à des problèmes d'estimation de paramètres et moins à des problèmes de choix de modèles et ce en raison du surcroît de difficulté lié à la gestion de la multiplicité des espaces des différents modèles. L'algorithme CBA proposé ici aborde les problèmes de choix de modèles en généralisant la méthode de Fearnhead et Prangle (2012) où les moyennes a posteriori des

paramètres des modèles estimés par régression sont utilisées comme statistiques de base pour la mesure de la discordance entre modèles. Une régression logistique multinomiale pas à pas additionnelle est effectuée sur la variable indicatrice du modèle et les probabilités estimées du modèle sont incorporées dans les statistiques clés en vue du choix du modèle. Une étape de saut réversible de type Chaine de Markov - Monte Carlo est également incluse dans l'algorithme afin d'explorer de façon minutieuse la diversité des espaces des modèles. Afin de prouver sa robustesse, cet algorithme a été appliqué, en vue de sa validation, à différents modèles dont les probabilités réelles s'étendent sur une gamme très large. Son application à des exemples de complexité variable portant sur la transmission de trois agents pathogènes illustre son utilité dans le choix d'un modèle de transmission privilégié de ces agents.

Z. C. Quiroz, M. O. Prates, and H. Rue **208**
A Bayesian Approach to Estimate the Biomass of Anchovies Off the Coast of Perú

Le courant de Humboldt nord (NHCS) est l'écosystème le plus productif en poissons au monde. En particulier, l'anchois péruvien (Engraulis ringens) constitue la proie majeure des principaux grands prédateurs tels que les oiseaux marins, les poissons, les humains et les autres mammifères. Dans ce contexte, il est important de comprendre la dynamique de la distribution de la population des anchois afin à la fois de les préserver et de d'exploiter cette ressource sur le plan économique. En utilisant les données recueillies par l'« Instituto del Mar del Perú » (IMARPE) pendant une enquête lors d'une expédition scientifique, nous présentons une analyse statistique dont les objectifs principaux sont les suivants : 1) s'adapter aux caractéristiques des données échantillonnées telles que la dépendance spatiale, les proportions élevées de zéros (mesures nulles) et la grande taille des échantillons ; 2) fournir des éléments sur la dynamique de la population d'anchois ; 3) proposer un modèle permettant l'estimation et la prédiction de la biomasse d'anchois dans la zone du NHCS au large du Pérou. Ces données ont été analysées dans un cadre bayésien en utilisant la méthode d'approximation nichée d'intégration de Laplace (INLA). Ensuite, afin de sélectionner le meilleur modèle et d'étudier la capacité prédictive de chaque modèle, nous avons comparé des modèles et réalisé des vérifications de leur caractère prédictif. Enfin, nous avons mené des diagnostics d'influence spatiale bayésienne pour le modèle retenu.

S. Ventz and L. Trippa **218**
Bayesian Designs and the Control of Frequentist Characteristics: A Practical Solution

Les concepts fréquentistes tels que le contrôle du risque de première espèce ou le taux de fausses découvertes sont bien établis dans la littérature médicale et sont souvent exigés par les autorités de régulation. La plupart des schémas bayésiens sont définis sans considération explicite de ces caractéristiques fréquentistes Une fois le schéma bayésien établi, les statisticiens utilisent des simulations et des paramètres de réglage pour se conformer à un ensemble de caractéristiques opérationnelles fixées. Ces ajustements affectent l'utilisation de l'information préalable et des fonctions d'utilité. Dans cet article, nous considérons une approche bayésienne décisionnelle pour des schémas expérimentaux permettant de satisfaire des caractéristiques fréquentistes fixées. Nous définissons des schémas optimaux sous un ensemble de contraintes pouvant être requises par les autorités de régulation. Notre approche associe l'utilisation de fonctions d'utilité

interprétables et de critères fréquentistes et permet de sélectionner un schéma optimal qui satisfait un ensemble de caractéristiques opérationnelles fixées. Nous illustrons cette approche sur un essai de phase II à plusieurs bras conduit avec une approche séquentielle groupée et sur un essai de transition.

Les études écologiques qui utilisent des données recueillies sur des groupes d'individus et non sur les individus eux-mêmes sont sujettes à de nombreux biais qui ne peuvent pas être contrôlés sans un recours à des données individuelles. Dans le contexte d'une réponse rare, le schéma d'étude hybride pour l'inférence écologique combine de façon efficiente des données obtenues au niveau de groupes d'individus et des données cas-témoins obtenues au niveau des individus. En dehors de cas relativement simples, l'utilisation de ce schéma est malheureusement limitée en pratique car l'évaluation de la vraisemblance hybride est d'une difficulté calculatoire prohibitive. Dans cet article, nous commençons par proposer et développer une représentation alternative de la vraisemblance hybride. Nous proposons ensuite, à partir de cette nouvelle représentation, plusieurs approximations, ce qui conduit à une réduction considérable des difficultés calculatoires. Une étude de simulation approfondie montre que, pour de nombreux scénarios, les estimateurs basés sur la vraisemblance hybride approchée ont les mêmes caractéristiques opérationnelles que ceux basés sur la vraisemblance hybride exacte sans conduire à une augmentation du biais ni à une réduction de l'efficacité. Puis, pour les cas où les approximations peuvent être invalides, nous développons une estimation pragmatique et une stratégie d'inférence qui utilisent la forme approchée de quelques contributions à la vraisemblance et la forme exacte des autres contributions. Cette stratégie permet de trouver un compromis entre difficulté calculatoire et précision selon les problèmes considérés. Enfin, une retombée de ce travail est de nous permettre de proposer, pour la première fois, une caractérisation explicite du schéma de données hybrides agrégées qui combine des données agrégées (Prentice & Sheppard, 1995) et des données cas-témoins. Ces méthodes sont illustrées sur des données portant sur les naissances survenues en Caroline du Nord (États-Unis) entre 2007 et 2009.

Le modèle de N-mélange est très utilisé pour estimer l'abondance d'une population quand la probabilité de détection est inconnue et à partir d'un ensemble de dénombrements pouvant être répétés dans le temps et dans l'espace (Royle, 2004, *Biometrics* **60**, 105-115). Nous expliquons et nous tirons parti de l'équivalence des modèles de N-mélange et des modèles multivariés poissonniens et basés sur la loi binomiale négative, ce qui conduit à de nouvelles approches puissantes pour ajuster ces modèles. Nous montrons en particulier que, quand la probabilité de détection et le nombre d'occasions d'échantillonnage sont faibles, des estimations infinies de l'abondance peuvent être obtenues. Nous proposons une covariance de l'échantillon pour diagnostiquer cet événement et montrons sa bonne performance pour le modèle de Poisson. Des estimations infinies peuvent être manquées en pratique du fait des procédures numériques d'optimisation qui s'arrêtent à des valeurs arbitrairement élevées.

Nous montrons que l’utilisation d’une borne K pour une sommation infinie dans la vraisemblance de N-mélange peut conduire à une sous-estimation de l’abondance ce qui implique que l’utilisation de valeurs par défaut de K faite dans des logiciels informatiques est déconseillée. Nous proposons une approche alternative simple et automatique pour choisir K. Les méthodes sont illustrées par l’analyse de données sur la tortue d’Hermann (Testudo *hermanni*)

En vue d’applications aux études observationnelles contemporaines, nous présentons une classe de modèles fonctionnels qui étendent les schémas nichés et croisés. Ces modèles tiennent compte de la transmission naturelle des structures de corrélation provenant des schémas d’échantillonnage dans les études où les unités fondamentales d’observation sont des fonctions ou des images. L’inférence est basée sur des formes quadratiques fonctionnelles et leur relation avec la structure sous-jacente des processus latents. Une procédure d’estimation rapide sur le plan calculatoire est développée pour des données de grande dimension. Les méthodes sont utilisées dans des applications qui comprennent des données d’accéléromètres de haute fréquence pour l’étude de l’activité quotidienne, des données de tonalité vocale pour l’analyse phonétique et des données d’EEG pour l’étude de l’activité électrique du cerveau pendant le sommeil.

L'étude des populations difficiles d’accès ou cachées pose des défis importants. En règle générale, aucune base de sondage n'est disponible et les membres de la population sont difficiles à identifier ou à recruter à partir de bases de sondage plus larges. Ceci est particulièrement vrai pour des populations à haut risque pour le VIH/sida. L’échantillonnage conduit par les répondants (ECR) est souvent utilisé dans ce contexte pour l'estimation de la prévalence de l'infection. Dans ces populations, le nombre de personnes à risque d'infection et le nombre de personnes infectées sont d'une importance fondamentale. Cet article présente une étude de cas de l'estimation de la taille d’une population difficile d’accès sur la base des données recueillies par ECR. Nous étudions deux populations de prostituées et d’hommes ayant des relations sexuelles avec des hommes au Salvador. L'approche est bayésienne et nous considérons différentes formes d’information a priori, y compris l’utilisation des directives de l’ONUSIDA pour la taille de la population dans la région. Nous montrons que la méthode est capable de quantifier la masse d'informations sur la taille de la population disponible dans les échantillons ECR. À titre de validation séparée, nous comparons nos résultats à ceux estimés par extrapolation d'une étude de capture-recapture des villes salvadoriennes. Les résultats de notre étude de cas sont largement comparables à ceux de l'étude de la capture-recapture alors qu’ils diffèrent de ceux des directives de l'ONUSIDA. Notre méthode est généralement applicable à des données provenant d'études par ECR et nous fournissons un logiciel permettant de la mettre en œuvre.

Reader Reaction to "A Robust Method for Estimating Optimal Treatment Regimes" by Zhang et al. (2012)

Une publication récente compare des méthodes d'estimation d'un régime de traitement optimal basées sur la régression et sur la probabilité inverse (Zhang et coll., 2012) et montre que, dans le cas d'un nombre restreint de covariables, les méthodes pondérées selon la probabilité inverse sont plus robustes en cas d'erreur de spécification du modèle que les méthodes de régression. Nous étudions si l'utilisation de modèles qui s'ajustent mieux aux données réduit les inquiétudes concernant la non-robustesse des méthodes de régression. Nous étendons l'étude de simulation de Zhang et coll. (2012) en considérant un plus grand nombre de covariables et nous incluons des modèles paramétriques et non-paramétriques mieux ajustés dans la méthode de régression. Nous montrons que, dans la mesure où les modèles deviennent plus flexibles et s'ajustent mieux aux données, les préoccupations concernant le manque de robustesse de la méthode de régression sont réduites. L'incorporation de l'algorithme des forêts aléatoires (« Random Forest ») dans la méthode pondérée selon la probabilité inverse augmentée améliore ses propriétés et la rend tout à fait similaire à la méthode de régression qui utilise également les forêts aléatoires.