Translations of Abstracts

Biometric Methodology

Cet article s'intéresse à l'association entre la suppression du VIH et la restauration de l'immunité, dans le but de fournir un outil d'évaluation de la thérapie antirétrovirale et de surveillance de la progression de la maladie. Les données d'une étude récente sur le SIDA sont utilisées à titre d'illustration. Nous modélisons conjointement la dynamique virale du VIH et le temps de diminution du ratio CD4/CD8 en présence d'un processus avec erreurs de mesure pour les CD4. L'estimation des paramètres du modèle se fait simultanément à travers une méthode basée sur une approximation de Laplace et l'algorithme Monte Carlo EM communément utilisé. Les approches et la plupart des points présentés dans cet article ont des applications plus générales.

Dans cet article, nous présentons une nouvelle méthodologie statistique pour les études longitudinales en foresterie où les arbres sont sujets à des infections récurrentes, et où le risque d'infection dépend de la croissance des arbres au cours du temps. Comprendre la nature de cette dépendance a des implications sur les programmes de reboisement et de sélection d'espèces. Dans ce contexte, l'analyse statistique est d'autant plus compliquée que les schémas d'observations conduisent à des données incomplètes, censurées par intervalle et présentant une grande variabilité spatiale. En outre, les données sont collectées sur un grand nombre de sites, ce qui pose des problèmes numériques pour la modélisation spatio-temporelle. Un modèle conjoint pour l'infection et la croissance est développé ; la récurrence des infections est modélisée à l'aide d'un processus de Poisson non-homogène tandis qu'un modèle non linéaire dynamique dans l'espace permet de représenter les trajectoires de croissance. Ces trajectoires de croissance sont basées sur le modèle de croissance de von Bertalanffy et une paramétrisation variant dans l'espace est utilisée. La variabilité spatiale des paramètres de croissance est modélisée à l'aide d'un processus spatial multivarié obtenu par convolution de noyaux. L'inférence est effectuée dans un cadre Bayésien avec une implémentation basée sur les méthodes hybrides Monte Carlo. La méthodologie est appliquée pour l'analyse des données sur onze années d'une étude sur l'infestation par le charançon de l'épinette blanche en Colombie-Britannique.

Nous étudions la forme et l'analyse des études longitudinales après échantillonnage cas-témoin, lorsque l'intérêt se porte sur la relation entre une variable binaire longitudinale de réponse liée à la variable (cas-témoin) d'échantillonnage, et un ensemble de covariables. Nous proposons un cadre de modélisation semi-paramétrique basé sur un modèle marginal de réponse binaire longitudinale et un modèle auxiliaire pour l'état « cas-témoin » des sujets. Dans cette approche, l'analyste doit postuler la prévalence des cas dans la population, qui est alors utilisée pour calculer un terme de compensation dans le modèle auxiliaire. Les estimations des paramètres de ce modèle sont utilisés pour calculer les termes d'équilibre dans le modèle longitudinal de réponse. En examinant l'impact de la prévalence et d'une mauvaise spécification du modèle auxiliaire, nous montrons que les estimateurs des paramètres des covariables ne dépendant pas du temps, autres que l'intercept, sont raisonnablement robustes, mais que l'intercept et les estimateurs des paramètres des covariables dépendant du temps peuvent être tr§s sensibles à de telles erreurs de spécification. Nous étudions les questions de forme et d'analyse affectant l'efficacité de l'étude, c'est-à-dire : le choix de la variable d'échantillonnage et l'intensité de sa relation à la variable de réponse, la stratification, le choix des poids de covariance, et le degré de flexibilité du modèle auxiliaire. Cette recherche résulte d'une étude longitudinale après échantillonnage cas-témoin du développement temporel des symptômes de l'ADHD.

M. Liu, W. Lu, and C.-h. Tseng **374**
Cox Regression in Nested Case–Control Studies with Auxiliary Covariates

Le dispositf cas-contrôle emboîté (CCE) est une méthode d'échantillonnage populaire dans de grandes études épidémiologiques, à cause de son rapport coût-efficacité, pour investiguer la relation temporelle entre des maladies et des expositions à l'environnement ou à des précurseurs biologiques. L'estimateur du maximum de vraisemblance partielle de Thomas est couramment utilisé pour estimer les paramètres de régression dans le modèle de Cox pour données CCE. Dans cet article, nous considérons la situation où une information échec/censure et quelques covariables brutes sont disponibles pour toute la cohorte en plus des données CCE, et nous proposons un estimateur amélioré qui est asymptotiquement plus efficace que l'estimateur de Thomas. Nous adoptons une approche de projection qui, jusqu'ici, a seulement été utilisée dans des situations d'échantillonnage de validation aléatoire, et montrons qu'elle peut être adaptée aux dispositifs CCE où le schéma d'échantillonnage est un processus dynamique et n'est pas indépendant des contrôles. Sous certaines conditions, la convergence et la normalité asymptotique de l'estimateur proposé sont établies, et un estimateur convergent est aussi introduit pour la variance. De plus, un estimateur simplifié approché est proposé quand la maladie est rare. Des simulations extensives sont réalisées pour évaluer la performance, pour des échantillons finis, des estimateurs proposés et pour comparer leur efficacité avec celle de l'estimateur de Thomas ou d'autres estimateurs concurrents. De plus, des analyses de sensibilité sont réalisées pour montrer le comportement des estimateurs proposés quand les hypothèses du modèle sont en défaut, et nous trouvons que les biais sont raisonnablement petits dans les situations réelles. Nous montrons aussi comment la méthode proposée fonctionne avec des données issues d'études de la tumeur de Wilm.

J. Qin and Y. Shen **382**
Statistical Methods for Analyzing Right-Censored Length-Biased Data under Cox Model

Les données mesurant des temps avant événement associées à un biais de longueur se rencontrent communément dans des applications allant des études de cohortes épidémiologiques aux études économiques sur le travail en passant par les essais de prévention des cancers. Un problème statistique récurrent est de savoir comment étudier l'association entre des facteurs de risque et la survie dans la population cible lorsque les observationssont associées à un biais de longueur. Dans cet article, nous montrons comment estimer ces effets avec le modèle semi-paramétrique à risque proportionnel de Cox. De façon générale, la structure du modèle de Cox est modifiée sous cet échantillonnage de durées biaisées. Bien que l'approche existante de la vraisemblance partielle puisse être utilisée pour étudier l'effet de covariables sous troncature à gauche, elle n'est pas efficace pour analyser des données de durées biaisées. Nous proposons deux approches par équations estimatrices pour estimer les coefficients de covariables sous le modèle de Cox. Nous utilisons des processus stochastiques modernes et la théorie des martingales pour dériver les propriétés asymptotiques des estimateurs. Nous évaluons la performance empirique et l'efficacité des 2 méthodes par des études de simulations complètes. Nous utilisons des données d'une étude de démence pour illustrer la méthodologie proposée et détaillons les algorithmes computationnels pour les estimations ponctuelles, ce qui permet de les intégrer directement aux fonctions Splus ou R existantes.

L. Xue, L. Wang, and A. Qu **393**
Incorporating Correlation for Multivariate Failure Time Data When Cluster Size Is Large

Nous proposons une nouvelle méthode d'estimation pour des données multivariées de temps de défaillance en utilisant l'approche de la fonction d'inférence quadratique (QIF). La méthode proposée incorpore efficacement les corrélations intra-grappes. C'est pourquoi, elle est plus efficace que les méthodes qui ignorent les corrélations intra-grappes. De plus, la méthode proposée est facile à implémenter. Au contraire des équations d'estimation pondérées dans Cai et Prentice (1995), il n'est pas nécessaire d'estimer explicitement les paramètres de corrélation. Cette simplification est particulièrement utile quand on analyse des grappes de grande taille où il est difficile d'estimer la corrélation intra-grappes. Sous certaines conditions de régularité, nous montrons la consistance et la normalité asymptotique des estimateurs QIF proposés. Un test du Chi-2 est aussi développé pour les tests d'hypothèse. Nous conduisons une simulation de Monte-Carlo étendue pour étudier la performance sur des échantillons de taille finie des méthodes proposées. Nous illustrons aussi les méthodes proposées en analysant des données de cirrhose biliaire primitive (BPC).

Dans les grandes études de cohorte, il arrive souvent que la mesure de certaines covariables soient couteuses et seulement réalisée sur un échantillon de validation. D'autre part, des mesures de covariables moins couteuses mais moins fiables sont disponibles pour tous les patients. Pour l'analyse de ce type de données, une méthode d'estimation de type régression par calibration (Prentice 1982) est une méthode couramment utilisée et a été appliquée avec un modèle de Cox par Wang, Hsu, Feng et Prentice (1997) sous des hypothèses de normalité pour la mesure de l'erreur et de maladies rares. Dans cet article, nous considérons la méthode d'estimation de type régression par calibration pour un modèle semi-paramétrique à risque accéléré avec des covariables sujettes à des erreurs de mesures. Des propriétés asymptotiques de la méthode proposée sont étudiées sous un schéma d'échantillonnage en deux étapes pour la validation des données qui sont sélectionnées via un simple échantillonnage aléatoire, résultant d'observations ni indépendantes ni identiquement distribuées. Nous montrons que les estimateurs convergent vers des paramètres bien définis. En particulier, une estimation non-biaisée est réalisable sous des modèles dont l'erreur de mesure est normale et additive pour des covariables normales et sous des modèles d'erreur de Berkson. La méthode proposée donne de bons résultats dans une étude de simulation à partir d'échantillons finis. Nous appliquons également la méthode proposée à une étude sur la mortalité due à la dépression.

L'utilisation naïve de covariables mal mesurées amène à des estimateurs non convergents des effets des covariables dans des modèles de régression. Diverses méthodes ont été proposées pour traiter ce problème incluant la vraisemblance, la pseudo-vraisemblance, les équations estimantes et les méthodes bayesiennes, toutes ces méthodes nécessitant typiquement soit des échantillons de validation internes ou externes ou des études répliquées. Nous considérons un problème rencontré dans une série d'études en orthopédie où l'intérêt est d'examiner l'effet d'une réponse sérologique à court-terme et d'autres covariables sur le risque de développer une complication à long-terme appelée thrombose veineuse profonde. La réponse sérologique est la production par le patient d'anticorps en réponse à l'administration d'un médicament anti-thrombotique, mais cette réponse ne peut être connue qu'à l'occasion d'une prise de sang faite à la fin de l'hospitalisation. Le délai de séroconversion est donc soumis à un processus de censure car les sujets testés avant la séroconversion sont classés à tort comme négatifs. Nous développons une approche fondée sur la vraisemblance pour l'ajustement de modèles de régression prenant en compte cette erreur de détermination de la séroconversion en utilisant des estimateurs paramétriques et non-paramétriques de la distribution du délai de séroconversion. On montre que cette méthode réduit le biais résultant d'une approche naïve sur des simulations et une application à des données d'études en orthopédie fournit une autre illustration.

J. Cao, A. Moosman, and V. E. Johnson **426**
A Bayesian Chi-Squared Goodness-of-Fit Test for Censored Data Models

Nous proposons un diagnostic de qualité d'ajustement bayésien de modèles par Khi-deux dans le cadre de l'analyse de données censurées. La statistique de test a la forme de la statistique de test du Khi-deux de Pearson et est facilement calculable à partir des algorithmes de Chaînes de Markov Monte Carlo. L'originalité de ce diagnostic repose sur le fait qu'il est défini uniquement sur les temps d'événements observés. Parce qu'il ne nécessite pas l'imputation des temps d'événements pour les observations censurées, nous montrons qu'en présence d'une grande proportion de données censurées, il peut avoir une puissance supérieure pour détecter l'inadéquation du modèle par rapport à un test construit à partir des données complètes. Dans une étude de simulation, nous montrons que les tests basés sur ce diagnostic ont une puissance comparable et des risques de première espèce meilleurs que ceux du test communément utilisé et proposé par Akritas (1988). Un avantage important du diagnostic proposé est qu'il peut être appliqué à différents types de modèles d'analyse de données censurées, incluant les modèles linéaires généralisés ainsi que les modèles avec erreurs non identiquement distribuées et non additives. Nous illustrons le diagnostic de modèle proposé en testant l'adéquation de deux modèles de survie paramétriques expliquant la survenue de pannes du moteur principal des navettes spatiales.

S. Zhang, P. Müller, and K.-A. Do **435**
A Bayesian Semiparametric Survival Model with Longitudinal Markers

Nous avons travaillé sur les données d'un essai clinique de traitements du cancer métastatique de la prostate, dans lequel les patients ont été inclus au terme d'historiques médicamenteux divers. L'hétérogénéité de la population en résultant soulève certains problèmes méthodologiques au niveau de l'inférence statistique, pour prédire le temps jusqu'à progression dans les différents bras de traitement. Ladite inférence est encore compliquée par la nécessité d'inclure dans le modèle un marqueur longitudinal comme covariable. Face à ces enjeux, nous avons développé un modèle semi-paramétrique d'inférence conjointe des données longitudinales et du temps jusqu'à événement. L'approche proposée couvre le cas d'une possible guérison chez certains patients. Nous postulons pour distribution du temps jusqu'à événement un processus non paramétrique d'arbre de Polya tandis que les données longitudinales sont prises en charge par un modèle à effets mixtes. Intégrer une régression sur covariables dans un modèle non paramétrique de temps jusqu'à événement représente généralement, et en particulier dans un modèle d'arbre de Polya, une gageure. En exploitant le fait que la covariable elle-même est une variable aléatoire, nous sommes parvenus à implémenter la régression désirée en factorisant le modèle conjoint du temps jusqu'à événement et de la variable longitudinale d'intérêt dans un modèle marginal du temps jusqu'à événement associé à une régression des données longitudinales sur cette variable de temps : nous modélisons ainsi implicitement la régression désirée par sa distribution conditionnelle inverse.

Dans la classification de données fonctionnelles, les observations fonctionnelles sont souvent contaminées par différents effets systématiques, tels que les effets lots aléatoires causés par des artéfacts du dispositif ou des effets fixes causés par des facteurs liés à l'échantillon. Ces effets peuvent mener à un biais de classification et ne devraient donc pas être négligés. Un autre problème est la sélection de fonctions quand les prédicteurs sont de multiples fonctions, certaines d'entre elles pouvant être redondantes. Les difficultés ci-dessus surviennent dans une application sur données réelles où nous utilisons la spectroscopie par fluorescence pour détecter un pré-cancer du col de l'utérus. Dans ce papier, nous proposons un modèle hiérarchique Bayésien qui prend en compte les effets lots aléatoires et sélectionne les fonctions qui ont un effet parmi de multiples prédicteurs fonctionnels. Des effets fixes ou prédicteurs sous forme non-fonctionnelle sont aussi inclus dans le modèle. La dimension des données fonctionnelles est réduite par une expansion sur une base orthonormée ou des composantes principales fonctionnelles. Pour l'échantillonnage *a posteriori*, nous utilisons un échantillonneur hybride Metropolis-Hastings/Gibbs, qui souffre d'un mélange lent. Un algorithme de Monte-Carlo Evolutionnaire est appliqué pour améliorer le mélange. Une simulation et une application sur données réelles montre que le modèle proposé fournis une sélection exacte des prédicteurs fonctionnels ainsi qu'une bonne classification.

Nous considérons une régression linéaire pénalisée, en particulier pour les problèmes « *p* grand, *n* petit » pour lesquels les relations parmi les prédicteurs sont décrites *a priori* par un réseau. Nous avons été motivés par une classe d'exemples incluant la modélisation d'un phénotype à travers des profils d'expression de gènes en prenant en compte le fonctionnement coordonné des gènes sous forme de chemins ou de réseaux biologiques. Pour incorporer la connaissance a priori de niveaux d'effet similaire de prédicteurs au voisinage dans le réseau, nous proposons une pénalité groupée, basée sur la norme $L_{\square}$ qui lisse les coefficients de régression des prédicteurs du réseau. La principale caractéristique de la méthode proposée est sa capacité à réaliser automatiquement une sélection groupée de variables et d'exploiter les effets de se regroupement. Nous discutons également des effets des choix de $\square$ et de certains poids dans la norme $L_{\square}$. Des études de simulation démontrent une meilleure performance de la méthode proposée en échantillon fini comparée au Lasso, réseau élastique et une méthode récente basée sur un réseau. La nouvelle méthode fonctionne mieux dans la sélection des variables pour toutes les simulations considérées. A des fins illustratives, la méthode est appliquée à une micropuce pour prédire les temps de survie pour des patients ayant un glioblastome, en utilisant un ensemble de données d'expression de gènes et un réseau de gènes compilé à partir de chemins KEGG (Kyoto Encyclopedia of Genes and Genomes).

L'étude par micropuce de l'expression de gènes au sein de catégories ordonnées est fréquemment réalisée pour mieux comprendre les fonctions de ces gènes et les mécanismes biologiques sous-jacents. C'est le cas des études où un tissu ou une lignée cellulaire est exposé à une substance chimique avec des doses et/ou une durée variables. Le but de telles études est d'identifier des motifs/profils d'expression au sein des catégories ordonnées. Ce type de problème peut se voir comme un problème de tests multiples où l'on teste pour chaque gène l'hypothèse nulle d'absence de différence entre les moyennes consécutives de son expression et où, si cette hypothèse est rejetée, des décisions sont prises sur les directions à suivre. La plupart des procédures traitant du problème des tests multiples ont été conçus pour contrôler le taux de faux positifs (FDR) et non le FDR mixte directionnel qui est la proportion attendue d'erreur de type I et de direction parmi l'ensemble des hypothèses rejetées. Benjamini et Yekutieli (2005) ont montré qu'une augmentation de la procédure classique de Benjamini-Hochberg (BH) controlait le FDR mixte directionnel lorsque l'on testait des hypothèses nulles simples versus des hypothèses alternatives bilatérales en fonction d'un seul paramètre. Dans cet article, nous abordons le problème du contrôle du FDR mixte directionnel en présence de paramètre multidimensionnel. Pour traiter ce problème, nous développons une procédure basée sur un test de Bonferroni appliqué à chaque gène et qui étend celle de Benjamini et Yekutieli. Nous prouvons que cette procédure contrôle le FDR mixte directionnel quand les tests statistiques sous-jacents sont indépendants entre gènes. Une étude de simulation est réalisée pour évaluer les performances de cette méthode par rapport à d'autres procédures, à la fois sous l'hypothèse d'indépendance mais également lorsque les tests statistiques ne sont plus indépendants entre gènes. Enfin, cette méthodologie est appliquée aux données obtenues par Lobenhofer et al. (2002). Nous identifions ainsi plusieurs gènes importants du cycle cellulaire, tel que le gène MCM4 et celui de la sous-unité C2 du facteur de réplication impliqués dans la réplication et la réparation de l'ADN, gènes qui n'avaient pas été identifiés par les précédentes analyses de ce même jeu de données par Lobenhofer et al. (2002) et Peddada et al. (2003).
Bien que certains de nos résultats coïncident avec des résultats précédents, nous identifions plusieurs autres gènes qui complètent les résultats de Lobenhofer et al. (2002)

L. Wang and D. B. Dunson **493**
Semiparametric Bayes Multiple Testing: Applications to Tumor Data

Les investigateurs des études du National Toxicology Program (NTP) cherchent à évaluer, en général ou pour un certain type de tumeur, si un agent est carcinogène, tout en appréciant les profils des dose-réponse. A cause d'une corrélation potentielle entre les tumeurs, on préfère une analyse conjointe à des analyse séparées par type de tumeur. Pour cela nous proposons un modèle logistique à effet aléatoire avec une matrice de coefficients représentant les logarithmes des OR pour des doses adjacentes et des types de tumeur à différents sites. On propose des a-priori non paramétriques adaptées pour ces coefficients de manière à caractériser les corrélations et à permettre des transferts d'information entre les groupes de doses et les types de tumeur. Les hypothèses globales ou locales sont aisément évaluées à partir d'une seule chaîne de résultats MCMC. Deux procédures, basées sur les a posteriori d'alternatives locales, sont appliquées pour les tests multiples des hypothèses locales. Des simulations et l'analyse d'un jeu de données du

NTP illustrent l'approche proposée.

En recherche cancérologique, il est important d'évaluer la performance d'un marqueur biologique (par exemple, moléculaires, génétiques, ou de l'imagerie), qui corrèle l'évolution des patients ou la prédiction de leur réponse à un traitement dans une grande étude prospective. En raison de la contrainte budgétaire globale et les coûts élevés associés aux essais biologiques, les enquêteurs ont souvent à choisir un sous-ensemble de patients pour évaluer les bio-marqueurs. Pour détecter une éventuelle association entre ces derniers et le résultat, les chercheurs ont besoin de décider comment sélectionner le sous-ensemble de taille fixe de telle sorte que l' efficacité de l'étude puisse être renforcée. Nous montrons que, plutôt que de tirer un échantillon aléatoire simple de l'étude de la cohorte, une plus grande efficacité peut être atteinte en permettant à la probabilité de sélection de dépendre des résultats et d'une variable auxiliaire ; nous nous référons à un tel schéma d'échantillonnage comme un *sous-échantillonnage dépendant des résultats et d'une variable auxiliaire* (OADS). Cette étude est motivée par la nécessité d'analyser les données d'une étude des bio-marqueurs du cancer du poumon en adoptant cette planification OADS pour évaluer les mutations EGFR comme bio-marqueur prédictif pour savoir si un sujet répond mieux à des médicaments inhibiteurs de l'EGFR. Nous proposons une méthode d'estimation du maximum de vraisemblance qui tient compte de la planification OADS et qui utilise toutes les informations observées, en particulier celles contenues dans le score du risque des mutations EGFR (une variable auxiliaire de mutations EGFR), qui est disponible pour tous les patients. Nous obtenons les propriétés asymptotiques de l'estimateur proposé et nous évaluons ses propriétés d'échantillonnage fini par simulation. Nous illustrons la méthode proposée sur un exemple.

Dans beaucoup de maladies pour lesquelles on dispose de différentes options de traitement, il n'y a souvent pas de consensus sur le meilleur traitement à donner à des patients en particulier. Dans de tels cas, il peut être nécessaire de définir une stratégie de prescription ; c'est-à-dire un algorithme qui indique le traitement qu'un individu devrait recevoir fondé sur les mesures de ses caractéristiques. Une telle stratégie ou algorithme est aussi référencé comme un régime de traitement. Le régime de traitement optimal est la stratégie qui devrait fournir le plus grand bénéfice de santé publique en minimisant autant que possible les mauvais résultats. En utilisant une mesure qui est une généralisation du risque attribuable et des notions de résultats potentiels, nous donnons un estimateur de la proportion d'évènements que le régime optimal de traitement aurait pu prévenir s'il avait été mis en œuvre. Alors que les traditionnelles études de risques attribuables considèrent le risque supplémentaire qui peut être attribué à l'exposition à un certain contaminant, ici nous étudions le bénéfice que l'on peut attribuer en utilisant la stratégie de traitement optimale.

Nous montrons comment des modèles de régression peuvent être utilisés pour estimer la stratégie optimale de traitement et le bénéfice que l'on peut lui attribuer. Nous donnons aussi les propriétés asymptotiques de cet estimateur. Comme exemple motivant nous appliquons nos méthodes à l'étude de l'observation de 3856 patients traités au Duke University Medical Center, avec d'abord une chirurgie de pontage d'artère coronaire suivie de problèmes cardiaques reliés, nécessitant l'usage d'un cathéter. Les patients peuvent être traités, soit par une thérapie médicale seule, soit par une combinaison de thérapie médicale et d'intervention percutanée de la coronaire, sans qu'il y ait un consensus général sur le traitement qui serait le meilleur pour des patients en particulier.

Y. Li, J. M. G. Taylor, and M. R. Elliott **523**
A Bayesian Approach to Surrogacy Assessment Using Principal Stratification in Clinical Trials

Un marqueur supplétif de substitution (*S*) est une variable qui peut être mesurée plus précocement et souvent plus facilement que le véritable critère de jugement (*T*). De nombreuses études passées ont été ‘consacrées’ au développement de mesures supplétives afin d’évaluer de quelle façon *S* peut remplacer *T* ou bien pour examiner l’utilisation de *S* dans la prédiction de l’effet d’un traitement (*Z*). Cette étude nécessite cependant que l’on définisse des modèles de la distribution de *T* conditionnellement à *S* et *Z*. Il est bien connu que de tels modèles n’ont pas d’interprétation causale parce qu’ils conditionnent sur une variable *S* postérieure à la randomisation. Dans cet article nous modélisons directement la relation entre *T*, *S* et *Z* en nous servant d’un cadre de réponses potentielles introduit par Frangakis et Rubin (2002). Nous proposons une méthode d’estimation bayésienne pour évaluer les probabilités causales associées à la validation croisée des réponses potentielles de *S* et *T* lorsque *S* et *T* sont toutes les deux des variables binaires. Nous utilisons un modèle log-linéaire pour modéliser directement l’association entre les réponses potentielles de *S* et *T* à l’aide du odds-ratio. Les quantités calculées par cette démarche ont toujours une interprétation causale. Cependant ce modèle causal n’est identifiable à partir des données qu’au prix d’hypothèses supplémentaires. Pour réduire le problème de non-identifiabilité et accroître la précision des inférences statistiques, nous devons supposer la monotonicité et introduire des connaissances a priori plausibles dans le contexte de la variable supplétive par l’intermédiaire de lois a priori. Nous explorons aussi la relation entre les mesures supplétives basées sur les modèles traditionnels et celles du modèle proposé. Cette méthode est appliquée à des données d’une étude sur le traitement du glaucome.

N. Houede, P. F. Thall, H. Nguyen, X. Paoletti, and A. Kramar **532**
Utility-Based Optimization of Combination Therapy Using Ordinal Toxicity and Efficacy in Phase I/II Trials

Un plan expérimental fondé sur une méthode bayesienne adaptive est proposée pour choisir un couple de doses optimal d’une association thérapeutique entre un agent cytotoxique et un agent biologique dans un essai clinique de phase I/II. Le résultat est modélisé par un vecteur défini par deux variables ordinales : l’efficacité et la toxicité. Les probabilités marginales des relations dose-toxicité et dose-efficacité de chaque couple de

doses sont estimées par un modèle généralisé flexible d'Aranda-Ordaz (1983, *Biometrika* **68**, 357-363). Une copula Gaussienne est utilisée pour estimer la distribution jointe des deux variables d'évaluation. Des scores numériques d'utilité sont donnés à chaque combinaison de résultats, avec la possibilité de juger un traitement non évaluable pour efficacité en cas de toxicité sévère. Ces scores sont obtenus par une technique permettant de recueillir un consensus au sein d'un groupe de médecins experts. Pour chaque cohorte de patients inclus, un couple de doses est choisi afin de maximiser la moyenne du score d'utilité a posteriori. Cette méthode est illustrée par un essai clinique dans le cancer de la vessie. Des simulations sont présentées pour tester la sensibilité de la méthode en fonction des paramètres estimés a priori, des scores d'utilité, de la correlation entre les résultats, de la taille de l'échantillon, de la taille des cohortes de patients et du couple de doses initial.

B. N. Bekele, Y. Li, and Y. Ji **541**
Risk-Group-Specific Dose Finding Based on an Average Toxicity Score

Nous proposons un dispositif bayésien de détermination de dose qui tienne compte de deux importants facteurs : la gravité de la toxicité et l'hétérogénéité des sensibilités des patients au toxique. On définit une gradation appropriée pour les différents niveaux de gravité des manifestations toxiques, puis on utilise une vraisemblance multinomiale avec une loi a priori de Dirichlet pour les probabilités de ces grades de toxicité à chaque dose. La toxicité globale est caractérisée par un score moyen. Pour traiter la question de l'hétérogénéité des sujets, on les range dans des groupes de risque en fonction de leur sensibilité. Une transformation isotonique bayésienne est appliqué pour produire une inférence a posteriori respectant une contrainte d'ordre sur les scores de toxicité moyens. On montre les performances de notre méthode de recherche de dose par une simulation basée sur un essai clinique concernant le myélome multiple.

O. Davidov, K. Fokianos, and G. Iliopoulos **549**
Order-Restricted Semiparametric Inference for the Power Bias Model
Le modèle de puissance biaisé (power bias model), qui est une généralisation de l'échantillonnage avec biais de longueur, est introduit et étudié en détail. En particulier, notre attention se concentre sur une inférence ordonnée restreinte. Nous montrons que le modèle de puissance biaisé est un exemple du modèle du rapport des densités, ou en d'autres termes, c'est un modèle semi-paramétrique qui est spécifié en supposant que le rapport de plusieurs densités de probabilité inconnues a une forme paramétrique. Des procédures d'estimations et de tests sous différentes contraintes sont développées en détail. Nous montrons que le modèle de puissance biaisé peut être utilisé pour tester le rapport de vraisemblance ordonné, parmi de multiples populations, sans recourir à des hypothèses paramétriques. Des exemples et l'analyse de données réelles démontrent l'utilité de cette approche.

D. Todem, J. Fine, and L. Peng **558**
A Global Sensitivity Test for Evaluating Statistical Hypotheses with Nonidentifiable Models

Nous nous intéressons au problème de l'évaluation d'une hypothèse statistique lorsque certaines caractéristiques du modèle sont non-identifiables à partir des données observées. Un tel scénario est commun pour la détermination du biais de publication en méta-analyse, ou pour l'évaluation de l'effet d'une covariable dans les études longitudinales lorsque les perdus de vue sont certainement à ne pas ignorer. Une approche possible à ce problème est de fixer un ensemble minimal de paramètres de sensibilité conditionnellement auxquels les paramètres de l'hypothèse sont identifiables. Ici, nous étendons cette idée et nous montrons comment évaluer l'hypothèse étudiée en utilisant une statistique infimum sur le support complet du paramétrage de sensibilité. Nous caractérisons la distribution limite de la statistique comme un processus du paramètre de sensibilité, qui implique une étude théorique fine de son comportement en cas de mauvaise spécification du modèle. En pratique nous suggérons une procédure de rééchantillonnage (bootstrap) nonparamétrique pour implémenter ce test infimum et pour construire des bornes de confiance pour des tests ponctuels simultanés au travers de toutes les valeurs du paramètre sensibilité, avec ajustement pour tests multiples. L'utilité pratique de cette méthodologie est illustrée par l'analyse d'une étude longitudinale en psychiatrie.

Cet article propose des approximations de point selle à l'espérance et à la fonction de variance covariance de processus de branchement multitypes dépendant de l'âge. Nous trouvons que les approximations proposées sont précises, faciles à mettre en œuvre et beaucoup plus rapides à calculer que par simulation du processus. Nous présentons de multiples applications, incluant les analyses de données clonales sur la génération d'oligodendrocytes provenant des cellules qui les engendrent immédiatement, et sur la prolifération de cellules Hela. Nous construisons de nouveaux estimateurs pour analyser les données clonales. Nous utilisons les méthodes proposées pour approcher la distribution de la génération, ce qui a récemment trouvé de nombreuses applications en biologie cellulaire.

Un modèle de régression fonctionnelle pour prédire la concentration de pollen de cyprès dans un intervalle de temps donné, en prenant comme régresseur la température de l'air dans un intervalle préalable, est obtenu au moyen d'une procédure en deux étapes. Cette estimation est obtenue par une analyse en composantes principales et le bruit résiduel est également modélisé par une régression fonctionnelle sur composantes principales prenant la concentration en pollen dans l'intervalle préalable comme processus explicatif. La performance de la prédiction est ensuite évaluée sur des séries de données concernant ce pollen enregistrées à Grenade (Espagne) sur une période de dix ans.

Biometric Practice

Les études actuelles génome large association sont une puissante approche pour découvrir des variantes génétiques fréquentes causant des maladies fréquentes et complexes. La découverte de ces variantes génétiques est un atout majeur dans le diagnostic précoce de maladie et donc dans la prévention, et dans l'individualisation des traitements. Nous décrivons ici une méthode pour combiner des variantes génétiques basées sur l'optimalité de la théorie du maximum de vraisemblance. Une telle théorie montre simplement que la courbe de la caractéristique de fonctionnement de récepteur (ROC) basée sur le rapport de vraisemblance (LR) présente un maximum de performance sur chaque point de rupture et que l'aire (AUC) en dessous de la courbe ROC ainsi obtenue est la plus importante parmi toutes les approches . A l'aide de données simulées et de données réelles nous comparons la méthode avec les approches classiques de régression logistique et d'arbre de classification. Les trois approches présentent des performances similaires si le modèle sous jacent de la maladie est connu. Cependant pour la plupart de ces maladies nous disposons de peu de connaissance à priori sur le modèle de la maladie et dans ce cas la nouvelle méthode présente un avantage sur les approches de régression logistique et d'arbre de classification. Nous appliquons la nouvelle méthode pour le génome large association du diabète de Type 1 sur des données provenant de Wellcome Trust Case Control Consortium. Basé sur cinq polymorphismes simples de nucléotide (SNPs) l'essai atteint l'exactitude de niveau moyen de classification. Avec plus de résultats découverts dans le futur sur les gènes nous pensons qu'un essai génétique prédictif sur le diabète de type 1 peut être construit avec succès et par la suite mis en application pour l'usage clinique.

La disponibilité croissante de registres de données pour diverses maladies a donné aux épidémiologistes de précieuses opportunités de compréhension de ces maladies. Elle a également ouvert des défis dans les méthodes traditionnelles d'analyse en raison des schémas compliqués de censure et troncature, et des dynamiques temporelles d'effets de covariables. Dans le cas d'une étude-type portant sur les données du registre des patients de la Fondation sur la mucoviscidose, nous proposons d'analyser la progression des symptômes en utilisant des processus temporels régressifs au lieu des modèles de hasards proportionnels habituellement employés. Deux objectifs sont envisagés, la prévalence permanente et momentanée d'infection pulmonaire à Pseudomona Aeruginosa (PA), qui reflète différents aspects du processus pathologique. L'analyse de l'infection permanente à PA par un modèle à coefficient dépendant du temps montre un défaut d'ajustement et une perte d'information potentielle de l'analyse usuelle par les hasards proportionnels.

L'analyse de l'infection momentanée à PA conduit à des résultats cliniquement significatifs et jusqu'à présent non reportés dans la littérature sur la mucoviscidose. Nos analyses démontrent que la détection prénatale/néonatale diminue la prévalence de l'infection à PA, en comparaison aux méthodes traditionnelles de diagnostic sur signes et symptômes, mais que ce bénéfice diminue avec l'âge. Le risque d'infection à PA est également affecté par l'année calendaire de diagnostic ; les patients diagnostiqués dans des cohortes récentes témoignent d'une prévalence plus élevée d'infection permanente à PA, mais aussi d'une prévalence abaissée d'une infection momentanée à PA.

L'exposition à l'amiante est un facteur de risque bien connu de cancer et autres maladies des poumons. Lorsque des ouvriers développent de telles maladies, il est nécessaire d'établir l'éventuel lien avec une exposition professionnelle qui donnerait droit au versement d'indemnités de compensation. En l'absence d'une exposition professionnelle établie, des prélèvements peuvent être effectués dans les poumons pour mesurer les fibres d'amiante longues et courtes ainsi que procéder au comptage des corps asbestosiques. On dispose donc de données provenant d'un ou plusieurs prélèvements sur les poumons pour estimer l'exposition à l'amiante, en comparant souvent avec des prélèvements de référence faits sur une population non exposée. Comme il n'existe aucune méthodologie standardisée pour le traitement de ce type de données à la fois discrètes et continues et présentant également la particularité que plusieurs prélèvements peuvent être effectués sur le même sujet et ainsi être corrélés, nous proposons plusieurs modèles de classes latentes prenant en compte ces spécificités. Ces méthodes peuvent utiles aux comités d'indemnisation en leur fournissant une estimation de la probabilité individuelle d'exposition professionnelle fondée sur les données disponibles, aux chercheurs qui étudient les propriétés des tests développés dans ce domaine et plus généralement dans le cas de tests fondés sur des données de structure similaire.

L'identification des régions ayant les taux de mortalité les plus haut et les plus bas, et la cartographie avec codage par couleurs correspondante peuvent aider les épidémiologistes à identifier des voies prometteuses pour des études étiologiques analytiques. En nous basant sur un modèle Poisson-Gamma à deux niveaux, avec covariables, nous utilisons l'information sur les facteurs de risque connus, tels que la prévalence du tabagisme, pour ajuster les taux de mortalité et montrer la variation résiduelle sur les risques relatifs qui pourraient refléter des associations étiologiques cachées auparavant. En plus de l'ajustement par les covariables, nous étudions des classements basés sur les taux de mortalité standardisés, sur les estimateurs empiriques bayésiens, et sur une méthode de classement par percentile a posteriori, et nous indiquons les conditions nécessaires aux procédures plus complexes pour obtenir des probabilités élevées de classement des régions situées dans les 100□% plus basses et dans les 100□% plus hautes en terme de

risque relatif, pour □ = 0.05, 0.1 et 0.2. Nous donnons également des approximations analytiques pour les probabilités de bonne classification des régions situées dans les 100□% les plus hautes en termes de risque relatif pour ces trois méthodes de classement. En utilisant des données de mortalité par pathologie cardiaque, nous montrons que l'ajustement par la prévalence du tabagisme a un impact important pour le classement des régions en risque élevé et risque faible. Les trois méthodes ont des performances similaires pour une pathologie aussi commune. Cependant, pour des maladies moins fréquentes telles que certains cancers, et de larges fluctuations de présence entre régions, les méthodes empirique bayésienne et par percentile a posteriori surpassent la méthode basée sur les taux de mortalité standardisés.

Dans une expérience base sur micropuces, un plan expérimental est utilisé pour obtenir des mesures d'expression des gènes. Une méthode populaire d'analyse, implique d'ajuster le même modèle linéaire mixte pour chaque gène, obtenant des valeurs p spécifique aux gènes pour les tests d'intérêt impliquant des effets fixes et donc de choisir un seuil de significativité destiné à contrôler le taux de fausses découvertes (FDR) à un niveau désiré. Lorsqu'un ou plusieurs facteurs aléatoires ont des composantes de variance nulle pour certains gènes, la pratique standard d'ajuster le même modèle linéaire saturé à tous les gènes peut conduire à un échec dans le contrôle du FDR. Nous proposons une nouvelle méthode qui combine les résultats de l'ajustement du modèle saturé et des modèles linéaires mixtes sélectionnés pour identifier l'expression différentielle des gènes et fournit un contrôle du FDR à des niveaux cibles lorsque la vraie structure sous-jacente des effets aléatoires varie selon les gènes.

Les puces de type '*tiling arrays*' de haute densité sont conçues pour couvrir avec une grande résolution la totalité d'une région du génome avec des oligonucléotides qui se chevauchent ('*tiling*'), et sont utilisées dans de nombreuses applications biologiques. Les expériences sont généralement réalisées en plusieurs étapes au cours desquelles des variations techniques non contrôlées sont introduites. Les '*tiling arrays*' devenant de plus en plus populaires et étant utilisées par de nombreux laboratoires de recherche, il est urgent de développer des outils de contrôle de qualité comme ce qui a été fait pour les puces d'expression. Nous proposons ici un ensemble de critères statistiques pour vérifier la qualité des données issues de '*tiling arrays*' similaires à ceux utilisés pour les puces d'expression. Nous développons également une méthode pour estimer le seuil de significativité d'une mesure de qualité à partir de tests de randomisation. Ces méthodes, appliquées à plusieurs jeux de données réelles, incluant trois expériences indépendantes d'immunoprécipitation de la chromatine (ChIP-chip) et une étude transcriptomique, se sont montrées très efficaces pour identifier les puces de bonne qualité et les valeurs

extrêmes de chaque étude.

Les tests de permutation bases sur les distances entre distributions multivariées ont trouvé de nombreuses applications dans les sciences biologiques. Deux contextes importants pour ceci sont les procédures de permutation multiréponse, et les tests pseudo-F provenant de l'extension à des distances de l'analyse de variance multivariée. Dans cet article, nous donnons les conditions sous lesquelles ces deux contextes sont équivalents. Les méthodes et le résultat d'équivalence sont illustrés en réanalysant un ensemble de données écologiques et avec une nouvelle application à des données d'imagerie par résonance magnétique fonctionnelle.

Les méthodes MCMC à saut réversible sont utilisés pour ajuster des modèles bayesiens de capture-recapture avec hétérogénéité des individus et des échantillons. L'hétérogénéité dans les probabilités de capture provient de mélanges en nombre fini et/ou d'effets fixes des échantillons avec d'éventuelles interactions. L'estimation par la méthode MCMC à saut réversible permet une sélection automatique des modèles et/ou une estimation moyenne. Les distributions a priori des paramètres stabilisent les estimateurs et fournissent des intervalles de crédibilité réalistes pour la taille de la population dans le cas de modèles sur-paramétrés contrairement aux méthodes fondées sur la vraisemblance. Pour illustrer cette approche nous analysons le classique fichier de données sur le lièvre raquettes et le lapin à queue blanche.