Translations of Abstracts

DISCUSSION PAPER

Chi Wang, Giovanni Parmigiani, and Francesca Dominici
Bayesian Effect Estimation Accounting for Adjustment Uncertainty 661

Le modèle sur lequel repose l'estimation de l'effet d'une exposition sur un résultat est généralement sensible au choix des facteurs de confusion inclus dans le modèle. Nous proposons une nouvelle approche que nous appelons *Bayesian Adjustment for Confounding* (en abrégé BAC) pour estimer l'effet d'une exposition d'intérêt sur le résultat tout en prenant en compte l'aléa d'ajustement. Notre approche repose sur deux modèles. Dans le premier (appelé modèle résultat), le résultat est fonction de l'exposition et des sources de confusion potentielles. Dans le second (appelé modèle exposition), l'exposition est fonction des sources de confusion potentielles. Nous mettons en oeuvre une procédure bayésienne de sélection de variables pour les deux modèles, et relions ces deux modèles à l'aide un paramètre (noté ω) qui représente les rapports de cote (a priori) d'inclusion d' un prédicteur dans le modèle résultat, sachant que celui-ci est présent dans le modèle exposition. En l'absence de dépendance ($\omega = 1$), le BAC se ramène au traditionnel BMA (*Bayesian Model Averaging*). Dans des études par simulation nous montrons qu'un BAC avec un $\omega > 1$ estime l'effet d'exposition avec un biais plus petit que le traditionnel BMA, et améliore la couverture. Nous comparons ensuite le BAC avec l'approche récente de Crainiceanu et al. (2008), puis avec le traditionnel BMA, et ce, sur des données de séries temporelles d'admission à l'hopital, de niveaux de pollution de l'air, et de variables météorologiques de Nassau (New York) entre 1999 et 2005. Nous avons utilisé chacune de ces approches et estimé l'effet à court terme du PM2,5 sur les admissions en urgence des patients atteints de maladies cardio-vasculaires en prenant en compte la confusion. Cet exemple illustre les pièges d'une utilisation erronée des méthodes de sélection de variables en présence d'un aléa d'ajustement.

BIOMETRIC METHODOLOGY

Ying Huang, Peter B. Gilbert, and Holly Janes 687
Assessing Treatment-Selection Markers using a Potential Outcomes Framework

Les marqueurs sélectifs de traitement sont des molécules biologiques ou des caractéristiques des patients, associées à la réponse de chacun à un traitement. Ils peuvent être utilisés pour prédire les effets d'un traitement pour des sujets individualisés et par conséquent aider à prescrire un traitement à ceux qui en auront un réel bénéfice. Des instruments statistiques sont nécessaires pour évaluer la capacité d'un tel marqueur à permettre une sélection de traitement. Le critère communément adopté pour un bon marqueur sélectif de traitement a été l'interaction entre le marqueur et le traitement. Bien

qu'une forte interaction soit importante, elle n'est cependant pas suffisante pour signer une bonne qualité du marqueur. Dans cet article, nous proposons de nouvelles mesures pour évaluer un marqueur sélectif de traitement, continu, basées sur un ensemble de résultats potentiels. Sous un ensemble d'hypothèses, nous obtenons la règle optimale de décision pour un marqueur, afin de classer les individus selon leur bénéfice par le traitement, et nous caractérisons les performances du marqueur en utilisant la précision de la classification correspondante ainsi que la distribution globale du classificateur. Nous développons une méthode par maximum de vraisemblance contraint pour l'estimation et le test dans un dispositif d'essai randomisé. Nous avons effectué des études par simulation pour montrer la performance de nos méthodes. Enfin, nous illustrons ces méthodes au travers d'un essai vaccinal contre le VIH où nous explorons la valeur de l'immunité préalable vis-à-vis de l'Adenovirus sérotype 5 pour prédire une augmentation induite par le vaccin du risque de contamination par VIH.

Zhiwei Zhang, Zhen Chen, James F. Troendle, and Jun Zhang 697
Causal Inference on Quantiles with an Obstetric Application

La littérature actuelle sur les statistiques d'inférence causale est principalement liée à des moyennes de populations sur des résultats potentiels, alors que la pratique de la statistique actuelle implique aussi d'autres quantités significatives telles que les quantiles. Motivé par le consortium sur la sécurité au travail (CSL), avec une grande étude observationnelle de la progression du travail obstétrique nous proposons et nous comparons des méthodes d'estimation des quantiles marginaux de résultats potentiels.
En adaptant les méthodes et techniques existantes nous obtenons des estimateurs basés sur les résultats de la régression , la pondération inverse des probabilités, la stratification ainsi que sur l'estimateur doublement robuste. En incorporant la stratification dans l'estimateur doublement robuste, nous développons un estimateur hybride avec une stabilité numérique améliorée obtenue au détriment d'un léger biais et un manque de spécification des résultats du modèle de régression. Les méthodes proposées sont illustrées avec les données du CSL et évaluées par des expérimentations simulées qui reproduisent le CSL.

Yu-Jen Cheng and Mei-Cheng Wang 707
Estimating Propensity Scores and Causal Survival Functions Using Prevalent Survival Data

Cet article développe des approches semi-paramétriques pour l'estimation des scores de propension et des fonctions de survie causale à partir de données de survie prévalentes. Le problème analytique apparaît lorsque les temps d'échecs sont obtenus par échantillonnage prévalent et, en conséquence, les covariables sont incomplètement observées à cause de leur association avec les temps d'échecs. La procédure proposée pour estimer les scores de propension partage d'intéressantes caractéristiques avec la formulation par vraisemblance dans les études cas-contrôle, mais dans notre situation elle nécessite des considérations supplémentaires pour le terme constant. Le résultat montre que les scores de propension corrigés dans le contexte de la régression logistique peuvent être obtenus au moyen d'une procédure d'estimation standard avec des ajustements spécifiques sur le terme constant. Pour l'estimation causale, nous rencontrons deux types distincts de sources manquantes : l'une peut être expliquée par le cadre de résultats

potentiels ; l'autre est causée par le schéma d'échantillonnage prévalent. Une analyse statistique sans ajustement pour les biais des deux sources d'incomplétude conduira à des résultats biaisés dans l'inférence causale. Les méthodes proposées ont été partiellement motivées par, et appliquées à la Surveillance, l'Epidémiologie et les Résultats Finaux (SEER) de données liées à Medicare pour les femmes diagnostiquées atteinte de cancer du sein.

Hongwei Zhao, Chen Zuo, Shuai Chen, and Heejung Bang 717
Nonparametric Inference for Median Costs with Censored Data

De plus en plus, des estimations de coûts des soins de santé sont utilisés pour évaluer des traitements concurrents ou pour apprécier des dépenses attendues associées à certaines maladies. Dans la politique et l'économie de la santé, le centre d'intérêt principal de ces estimations est le coût moyen, parce que le coût total peut être déduit du coût moyen, et que l'information sur l'ensemble des ressources utilisées est hautement pertinent pour les décideurs. Cependant le coût médian pourrait être également important, à la fois comme mesure intuitive de tendance centrale de distribution des coûts et comme sujet d'intérêt pour les payeurs et les consommateurs. Dans de nombreuses études prospectives la collecte des données de coûts est incomplète chez certains sujets du fait de la censure à droite, typiquement causée par la perte de vue ou la durée limitée de l'étude. La censure pose un problème particulier pour l'analyse des données de coûts à cause de la censure informative induite en ceci que les méthodes traditionnelles adaptées aux données de survie se trouvent généralement invalides pour l'estimation de coûts censurés. Dans cet article, nous proposons des méthodes pour estimer le coût médian et son intervalle de confiance quand les données sont sujettes à la censure à droite. Nous considérons également l'estimation du rapport ou de la différence de deux coûts médians ainsi que leurs intervalles de confiance. Ces méthodes peuvent être étendues à l'estimation d'autres quantiles et d'autres données informativement censurées. Nous réalisons des simulations et analyses de données réelles afin d'examiner les performances des méthodes proposées.

Lu Wang, Pang Du, and Hua Liang 726
Two-Component Mixture Cure Rate Model with Spline Estimated Nonparametric Components

Dans certaines analyses de survie pour des études médicales, il y a souvent des patients à survie longue qui peuvent être considérés comme guéri de manière permanente. Le but de ces études est d'estimer la probabilité de non guérison de la population entière et le taux de hasard de la population sensible. Quand il y a des covariables, comme c'est souvent le cas en pratique, comprendre les effets des covariables sur la probabilité de non guérison ou sur le taux de hasard est d'importance équivalente. Les méthodes existantes sont limitées à des modèles paramétriques ou semi paramétriques. Nous proposons un modèle de mélange à deux composantes pour le taux de guérison avec des formes non paramétriques pour la probabilité de guérison et la fonction de hasard. L'identifiabilité du modèle est garantie par une hypothèse d'additivité qui interdit les interactions entre covariables temporelles dans le logarithme du taux de hasard. L'estimation est conduite par le biais d'un algorithme EM maximisant la vraisemblance pénalisée. Pour des raisons inférentielles, nous appliquons la formule de Louis pour obtenir des intervalles de confiance ponctuels de la probabilité de non guérison et du taux de hasard. Les vitesses

de convergences de nos estimateurs fonctionnels sont établies. Nous évaluons ensuite la méthode proposée par simulation. Nous analysons les données de survie d'une étude sur le mélanome et trouvons des éléments intéressants pour cette étude.

Birgit Schrödle, Leonhard Held, and Havard Rue 736
Assessing the Impact of a Movement Network on the Spatiotemporal Spread of Infectious Diseases

Un défi majeur en épidémiologie des maladies infectieuses est l'étude du lien entre l'information sur les réseaux de déplacements et les données spatio-temporelles pour l'incidence des maladies. Dans cet article, nous proposons et nous comparons deux environnements statistiques pour cet objectif, des modèles dirigés par les paramètres et ceux dirigés par les observations. On procède à une inférence bayésienne dans les modèles guidés par les paramètres en utilisant des approximations de Laplace intégrées emboîtées, tandis que les modèles guidés par les observations peuvent être aisément ajustés à l'aide de programmes existants utilisant le maximum de vraisemblance. La performance prédictive de chacune de ces formulations est évaluée en utilisant des règles de notation correctes. Nous étudions l'exemple de l'impact du commerce de bétail sur la diffusions spatio-temporelle de la Coxiellose (ou fièvre Q) dans le cheptel suisse entre 2004 et 2009.

Judith J. Lok and Victor DeGruttola 745
Impact of Time to Start Treatment Following Infection with Application to Initiating HAART in HIV-Positive Patients

Nous estimons comment l'effet du traitement antiviral dépend du temps écoulé jusqu'à l'initiation du traitement depuis l'infection par le VIH, à partir de données observationnelles. Un défi majeur dans la réalisation d'inférences à partir de telles données observationnelles vient des biais associés à l'affectation non randomisée des traitements, par exemple le biais induit par le lien entre l'état de la maladie et le moment où le traitement a été initié. Pour tenir compte de ces éléments, nous développons une nouvelle classe de Modèles Structurels à Moyennes Emboîtées (SNNMs) pour estimer l'impact du temps écoulé entre l'infection et l'initiation du traitement sur une mesure faite un temps fixé après l'initiation du traitement, comparée à l'effet lorsque le traitement n'est pas initié. Ceci conduit à un SNNM qui modélise l'effet pour plusieurs dosages de traitement sur une variable dépendant du temps, contrastant avec la plupart des SNNMs existant, qui s'intéressent à l'effet d'un seul dosage sur une variable mesurée à la fin de l'étude. Notre hypothèse est qu'il n'existe pas de confusion avec des éléments non mesurés. Nous illustrons notre méthode avec la base de données observationnelles AIEDRP (Infection Aiguë et Programme de Recherche sur le Début de la Maladie) Core01 pour l'infection à VIH. La prise en charge standard des patients infectés par le HIV est définie par le HAART (Traitement Anti Rétrovirus Hautement Actif) ; cependant le temps optimal de début du HAART n'a pas encore été identifié. La nouvelle classe de SNNMs permet d'estimer le lien entre l'effet d'une année de HAART et le temps écoulé entre la date estimée de l'infection et le début du traitement, ainsi qu'avec les caractéristiques du patient. Les résultats de l'ajustement de ce modèle montrent qu'un début précoce du HAART améliore substantiellement la reconstitution immunitaire dans la phase initiale et dans la phase aiguë de l'infection par le VIH.

Chris Groendyke, David Welch, and David R. Hunter 755
A Network-based Analysis of the 1861 Hagelloch Measles Data

Dans cet article, nous décrivons une méthode statistique pour ajuster les paramètres d'un réseau sophistiqué et d'un modèle épidémique à des données d'une maladie. La structure des contacts entre les hôtes est décrite par une classe de modèles graphiques aléatoires de la famille exponentielle (ERGMs) alors que le processus de transmission qui s'achemine dans le réseau est modélisé par une épidémie stochastique susceptible-latent-infectieux-immun (SEIR). Nous ajustons ces modèles à des données très fines d'une épidémie de rougeole apparue en 1861 à Hagelloch en Allemagne. Les modèles de réseau inclus les paramètres pour toutes les covariables collectées des hôtes incluant l'âge, le sexe, l'appartenance à un foyer ou une classe et la localisation du foyer alors que les temps de transmissions dans le modèle épidémique SEIR sont distribués exponentiellement avec des périodes de latence et infectieuse distribuées selon une loi Gamma. Cette approche nous permet d'obtenir des enseignements riches concernant la structure de la population – distincte du processus de transmission – ainsi que de fournir des estimations de nombreuses quantités biologiques d'intérêt, telles que le taux de reproduction R. En utilisant des méthodes de Monte Carlo par chaînes de Markov à sauts, nous produisons des échantillons à partir de la distribution jointe a posteriori de tous les paramètres de ce modèle – le réseau, l'arbre de transmission, les paramètres du réseau et les paramètres SEIR- et nous sélectionnons le modèle bayésien en déterminant le modèle de réseau le mieux ajusté. Nous comparons nos résultats avec les résultats d'analyses précédentes et montrons que le modèle de réseau ERGM est s'ajuste mieux aux données que le modèle de réseau Bernoulli utilisé précédemment. Nous fournissons également un logiciel, écrit en R, qui réalise ce type d'analyse.

Holger Schwender, Margaret A. Taub, Terri H. Beaty, Mary L. Marazita, and Ingo Ruczinski 766
Rapid Testing of SNPs and Gene–Environment Interactions in Case–Parent Trio Data Based on Exact Analytic Parameter Estimation

Les études de triplets cas-parents portent sur des enfants atteints d'une maladie et leurs parents. Elles visent à identifier des polymorphismes nucléotidiques simples (SNP) révélant une transmission préférentielle de certains allèles parentaux à l'enfant atteint. L'un des tests statistiques les plus fréquemment utilisés dans ce schéma d'étude est le test de transmission/déséquilibre génotypique (gTDT). Il repose sur un modèle de régression logistique conditionnelle ajusté aux observations par une procédure itérative. Dans cet article nous présentons les expressions analytiques des estimateurs des paramètres du modèle logistique conditionnel utilisé pour tester l'effet additif, dominant ou récessif d'un SNP, puis nous montrons qu'une expression analytique de l'estimateur existe aussi quand on s'intéresse aux interactions gène-environnement qui impliquent des variables environnementales binaires. Dans la mesure où, le plus souvent, on ignore quel modèle génétique sous-tend l'association entre un SNP et une maladie, on peut avoir intérêt à utiliser comme statistique de test le maximum des statistiques gTDT calculées sous les différents modèles. Nous proposons donc une procédure permettant le calcul rapide de la statistique de test et du p associé à ce MAX gTDT. Ces méthodes sont appliquées aux criblages du génome entier réalisés sur les triplets cas-parents réunis par le Consortium

International pour l'étude des fentes palatines. Ces applications révèlent une réduction spectaculaire du temps de calcul par rapport aux méthodes itératives conventionnelles : pour plusieurs centaines de milliers de SNP, on passe ainsi de plusieurs heures à quelques minutes.

Pei Fen Kuan and Derek Y. Chiang 774
Integrating Prior Knowledge in Multiple Testing under Dependence with Applications to Detecting Differential DNA Methylation

La méthylation de l'ADN est apparue comme la marque essentielle de l'épigénétique. On dispose de nombreuses plates-formes mettant en œuvre des puces à ADN de type « tiling» et du séquençage de nouvelle génération, ainsi que de protocoles expérimentaux pour évaluer des profils de méthylation de l'ADN. Comme d'autres données de puces à ADN de ce type, les données de méthylation de l'ADN présentent une structure inhérente de corrélation entre des sondes voisines les unes des autres. Mais, à la différence des données d'expression des gènes ou de liaison ADN-protéine, la densité variable de CpG génère des iles de CpG et permet de définir par rapport à ces « iles » des « rivages » et des « écueils » qui constituent une information exogène utilisable dans la détection de méthylation différentielle. Cet article propose un test robuste et une procédure de classement des sondes basés sur un modèle de Markov caché non homogène qui incorpore les propriétés décrites ci-dessus pour la détection de méthylation différentielle. Nous réexaminons l'article majeur de Sun et Cai (2009, *Journal of the Royal Statistical Society* ; *Series B* (*Statistical Methodology*) **71**, 393424) et nous proposons de modéliser l'hypothèse non-nulle à l'aide d'une distribution symétrique non paramétrique dans un test d'hypothèse bilatéral. Nous montrons par des simulations à grande échelle que notre modèle améliore le classement des sondes et qu'il est robuste vis-à-vis de modèles mal spécifiés. Nous illustrons par ailleurs sur des données réelles de méthylation d'ADN les bonnes caractéristiques opérationnelles de l'analyse que nous proposons par rapport aux méthodes communément utilisées pour détecter les sites de méthylation différentielle.

Mehdi Maadooliat, Jianhua Z. Huang, and Jianhua Hu 784
Analyzing Multiple-Probe Microarray: Estimation and Application of Gene Expression Indexes

L'estimation de la valeur d'expression de chaque gène (index d'expression) est une étape essentielle dans l'analyse des données de biopuces d'expression. Différentes méthodes ont été proposées dans ce domaine. Parmi celles-ci, la méthode de Li et Wong (2001) basée sur un modèle multiplicatif est très populaire. Elle est similaire à la méthode d'Irizarry et al. (2003a) reposant sur une modèle additif en échelle logarithmique. Et dans la même veine, Hu et al (2006) ont proposé de transformer les données pour améliorer l'estimation de l'index d'expression par une méthode ad hoc (critères d'entropie, recherche exhaustive naïve sur grille). Dans ce travail, nous avons ré-examiné ce problème en proposant une nouvelle approche d'estimation utilisant une transformation basée sur le profil de vraisemblance. Cette approche est statistiquement plus élégante et computationnellement plus efficace. Nous démontrons l'applicabilité de la méthode proposée en utilisant les données de référence Affymetrix U95A. Par ailleurs, nous avons introduit un nouvel index d'expression multivarié et utilisé une étude empirique pour montrer son potentiel comparé à l'indice univarié classique. Dans une autre partie

importante de ce travail, nous discutons deux problèmes couramment rencontrés en pratique : la normalisation et la statistique utilisées pour détecter les différences d'expression. Notre étude empirique montre des résultats quelque peu différents de ceux du projet MAQC (MAQC, 2006).

Christopher S. McMahan, Joshua M. Tebbs, and Christopher R. Bilder 793
Two-Dimensional Informative Array Testing

Les algorithmes de tests groupés basés sur des tableaux pour l'identification de cas sont très souvent utilisés dans les tests de maladies infectieuses, la découverte de médicaments et la génétique. Dans ce papier, nous généralisons un travail statistique antérieur sur les tests de tableaux pour prendre en compte l'hétérogénéité entre les individus testés. Nous dérivons d'abord des expressions formelles pour le nombre attendu de tests (efficacité) et les probabilités de mauvais classement (sensibilité, spécificité, valeurs prédictives) pour des tests de tableaux de dimension 2 dans une population hétérogène. Nous proposons ensuite deux techniques de construction de tableaux « informatifs » qui exploitent l'hétérogénéité de la population de façons qui peuvent considérablement améliorer l'efficacité du test par comparaison aux approches classiques qui considèrent la population comme homogène. De plus, un corollaire utile de notre méthodologie est que les probabilités de mauvais classement peuvent être estimées par individu. Nous illustrons nos procédures nouvelles par des données de tests de chlamydia et gonorrhée collectées au Nebraska dans le cadre du projet de prévention de l'infertilité.

Nicoleta Serban and Huijing Jiang 805
Multilevel Functional Clustering Analysis

Dans ce papier, nous étudions les méthodes de classification pour des données fonctionnelles à niveaux multiples, qui consistent en des fonctions aléatoires répétées qui sont observées pour un grand nombre d'unités (par exemple des gènes) sur des sous unités multiples (par exemple des types de bactéries). Afin de décrire la variabilité interne et externe induite par cette structure hiérarchique, nous considérons une approche par composantes principales fonctionnelles à niveaux multiples (MFPCA). Nous développons et comparons une méthode dure de classification appliquée aux scores dérivés de la MFPCA et une méthode souple utilisant une décomposition MFPCA. En étudiant des données simulées, nous évaluons l'estimation de la précision des classes prédites ainsi que la structure des groupes en considérant la série de paramètres suivante : petit *vs* grand nombre de points de mesure temporels, différents niveaux de bruit et différents nombres d'unités par sous unité. Nous montrons l'applicabilité de cette méthode de classification sur un jeu de données réel consistant en des profils d'expression de gènes activés dans des cellules du système immunitaire. Des schémas de réponse immunitaire bien connus ont été correctement identifiés en agglomérant les profils d'expression avec notre méthode.

Christian H. Weiß and Philip K. Pollett 815
Chain Binomial Models and Binomial Autoregressive Processes

Ce travail décrit la relation entre une classe de modèle binomiaux par chaîne couramment utilisé en écologie et épidémiologie et les processus binomiaux autorégressifs. Pour ces

derniers, de nouveaux développements sont proposés notamment en termes d'expression des distributions conditionnelles au décalage h et de ces paramètres dérivés. Ce travail se concentre sur deux classes de modèles binomiaux par chaîne, les modèles extinction-colonisation et colonisation-extinction, deux approches d'estimation sont décrites pour ces derniers et une application à des données réelles est réalisée. Ce travail rapproche les modèles autorégressifs standards des modèles binomiaux par chaîne avec des implications en termes d'estimation de leurs paramètres.

Giovanni Motta and Hernando Ombao 825
Evolutionary Factor Analysis of Replicated Time Series

Dans cet article nous développons une nouvelle méthode expliquant la structure dynamique des électroencéphalogrammes (EEG) multi-canaux enregistrés en plusieurs essais dans une expérimentation à tâches motrices et visuelles. Les analyses préliminaires de nos données évoquent deux problématiques statistiques. Tout d'abord, la variance de chaque canal et la covariance croisée entre les paires de canaux évolue dans le temps. De plus les profils de covariance croisée montrent une structure commune au sein de toutes les paires, et cette particularité apparaît de façon consistante entre tous les essais. A la lumière de ces éléments, nous développons un nouveau modèle factoriel évolutif (EFM) pour les données des EEG multi-canaux qui intègre systématiquement l'information entre essais répétés et permet l'usage de coordonnées factorielles à variation temporelle lisse. Les séries EEG individuelles ont en commun certains aspects entre essais, suggérant ainsi le regroupement de l'information entre essais, ce qui justifie l'utilisation de l'EFM pour des séries temporelles répliquées. Nous expliquons la dynamique commune des signaux EEG par l'existence d'un petit nombre de facteurs communs. Ces facteurs latents sont essentiellement responsables de l'organisation de la tâche motrice et visuelle qui, par les composantes, définit le comportement des signaux observés sur différents canaux. L'estimation des composantes dépendant du temps est basée sur la décomposition spectrale de la matrice de covariance estimée, dépendant elle aussi du temps.

Veronica J. Berrocal, Alan E. Gelfand, and David M. Holland 837
Space-Time Data fusion Under Error in Computer Model Output: An Application to Modeling Air Quality

Nous fournissons une méthode qui peut être utilisée pour obtenir une évaluation plus valide d'exposition environnementale. En particulier, nous proposons deux approches de modélisation pour combiner les données mesurées en un point avec les prédictions d'un modèle numérique au niveau de la cellule d'une grille, ce qui engendre une amélioration de la prédiction de l'exposition ambiante au niveau du point. En étendant notre précédent modèle (*Berrocal et al*., 2010b), ces nouveaux modèles sont destinés à prendre en compte deux préoccupations avec les prédictions du modèle. La première reconnait qu'il peut y avoir des informations utiles dans les prédictions pour les cellules d'une grille qui sont au voisinage de celle qui a été localisée. La seconde tient compte d'un potentiel non alignement spatial entre une station et la cellule de la grille associée.
Le premier modèle est un réducteur d'échelle lissé basé sur des champs aléatoires Markovien et Gaussien qui lie les données enregistrées en station et les prédictions du modèle grâce à l'introduction d'un champ aléatoire Markovien Gaussien latent qui lie les deux sources de données. Le second modèle est un réducteur d'échelle lissé avec des

pondérations variant aléatoirement dans l'espace définies à travers un processus Gaussien latent et une fonction exponentielle à noyaux, qui génère, à chaque site, une nouvelle variable sur laquelle les données de la station de monitorage sont régressées avec un modèle linéaire spatial. Nous avons appliqué les deux méthodes à des données de concentration quotidienne d'ozone à l'Est des Etats-Unis au cours des mois d'été de Juin, Juillet et Aout 2001, obtenant, respectivement, 5% et 15% de gain de prédiction sur l'erreur de moindre carrés par rapport à notre précédent modèle de réduction d'échelle (*Berrocal et al.,* 2010b). Plus encore, le gain de prédiction était plus important dans les sites éloignés des sites de monitorage.

Lauren Hund, Jarvis T. Chen, Nancy Krieger, and Brent A. Coull 849
A Geostatistical Approach to Large-Scale Disease Mapping with Temporal Misalignment

Le non alignement des bornes temporelles se produit lorsque les bornes se déplacent au cours du temps (par exemple, les bornes des secteurs du recensement changent à chaque année de recensement), ce qui complique la modélisation des tendances temporelles sur l'espace. De grands échantillons de zones géographiques présentant au cours du temps des limites non alignées sont de plus en plus souvent rencontrés en pratique. La plupart des approches géographiques pour tenir compte du non alignement au sein de données géographiques aborde le non alignement à travers des échelles multiples de l'espace/géographie (Gotway and Young, 2002). Les quelques approches existantes pour des données temporellement non alignées ne tiennent pas compte de la corrélation spatiale des effets aléatoires au cours du temps. Pour surmonter les problèmes associés à un non alignement temporel, nous construisons un modèle géostatistique pour des données de comptage agrégées en supposant qu'il existe un risque sous-jacent continu de la surface lié à la corrélation spatiales entre les zones. Nous mettons en œuvre le modèle dans le contexte des modèles linéaires généralisés (GLMM) en utilisant des splines à base radiale. En utilisant cette approche, le non alignement des bornes n'est plus un problème. De plus, ce contexte de cartographie des maladies est rapide et l'ajustement du modèle, en utilisant une approximation PQL pour l'estimation par maximum de vraisemblance, est simple. Nous prévoyons que la méthode sera aussi utile pour les grands ensembles de données de cartographie des maladies pour lesquels les approches fully-bayésiennes ne sont pas réalisables. Nous appliquons notre méthode pour évaluer les tendances économiques de l'incidence des cancers du sein à Los Angeles entre les périodes 1988-1992 et 1998-2002.

Luis E. Nieto-Barajas, Peter Müller, Yuan Ji, Yiling Lu, and Gordon B. Mills 859
A Time-Series DDP for Functional Proteomics Profiles

Par le biais d'une nouvelle technologie de puce biologique, la puce à protéines en phase inverse (RPPA), nous mesurons la concentration de protéines au cours du temps correspondant à un ensemble de marqueurs biologiques connus pour jouer un rôle biologique en interagissant dans une voie métabolique. Afin de prendre en compte la nature complexe et interactive du jeu de données, c'est-à-dire la corrélation temporelle couplée à la structure d'interdépendances décrite par la voie métabolique des marqueurs protéiques, nous proposons un modèle à effets mixtes avec des composantes spécifiques à la dépendance temporelle et aux protéines. Nous développons une séquence de mesures de probabilités aléatoires (RPM) dans le but de prendre en compte le temps dans la

mesure des concentrations en protéines. Marginalement, nous supposons pour chaque RPM un modèle de processus de Dirichlet (DP). La dépendance est introduite en définissant des distributions *beta* multivariées pour modéliser les poids non normalisés de la représentation « *stick-breaking* ». Nous prenons également en compte la structure de dépendance décrite par la voie métabolique à travers un modèle conditionnellement autorégressif (CAR). En appliquant notre modèle aux données RPPA, nous révélons un profil fonctionnel dépendant de la voie métabolique pour l'ensemble des protéines ainsi que des profils d'expression marginale en fonction du temps pour chaque marqueur considéré individuellement.

Jinbo Chen, Dongyu Lin, and Hagit Hochner 869
Semiparametric Maximum Likelihood Methods for Analyzing Genetic and Environmental Effects with Case-Control Mother–Child Pair Data

Les études cas-témoins de paires mère-enfant représentent un avantage unique pour disséquer
la susceptibilité génétique des maladies multifactorielles, car il permet l'évaluation de la composante génétique chez la mère et l'enfant. Cette approche a été largement adoptée dans les études de complications obstétricales et néonatales. Dans ce travail, nous avons développé une méthode statistique efficace pour évaluer conjointement les effets génétiques et environnementaux sur un phénotype binaire. En utilisant un modèle de régression logistique pour décrire la relation entre le phénotype et les facteurs de risque génétiques et environnementaux chez la mère et l'enfant, nous avons développé une méthode semi-paramétrique par maximum de vraisemblance pour l'estimation des paramètres d'association (odds ratio). Notre méthode est originale car elle exploite deux caractéristiques uniques des données pour estimer les paramètres. Premièrement, la corrélation entre les génotypes de la mère et de l'enfant peut être spécifiée sous les hypothèses de panmixie, d'équilibre de Hardy-Weinberg, et d'hérédité mendélienne. Deuxièmement, les expositions environnementales ne sont généralement pas modifiées par les génotypes de l'enfant conditionnellement aux génotypes maternels. Notre méthode donne des estimations plus efficaces, comparée à la méthode prospective standard de régression logistique sur données cas-témoins. Nous démontrons les performances de notre méthode sur des études de simulation détaillées et sur des données réelles de l'Étude Périnatale de Jérusalem.

Souparno Ghosh, Alan E. Gelfand, Kai Zhu, and James S. Clark 878
The k-ZIG: Flexible Modeling for Zero-Inflated Counts

De nombreuses applications utilisent des données de comptage issu d'un processus qui engendre un grand nombre de zéros. Les modèles pour les comptages avec excès de zéros, en particulier, les modèles de Poisson (ZIP) et à binomiale négative (ZINB) sont communément utilisés pour gérer ce problème. Cependant, ces modèles gèrent difficilement les incidences de zéros extrêmes (disons plus de 80%), en particulier pour trouver des covariables importantes. En fait, le ZIP peut même être en difficulté quand les proportions ne sont pas extrêmes. Pour gérer ce problème, nous proposons une classe de modèles k-ZIG. Ces modèles permettent une plus grande flexibilité de modélisation des grand nombre de zéros ainsi que des comptages non nuls, permettant une relation entre ces deux composantes. Nous développons les propriétés de cette nouvelle classe de

modèles, incluant une reparamétrisation avec une fonction de lien naturelle. Les modèles sont estimés sans difficulté avec une approche Bayésienne. La méthodologie est illustrée avec des exemples de données simulées ainsi qu'un jeu de données de plantation forestière obtenu à partir du programme du USDA Forest Service's Forest Inventory and Analysis (FIA).

Stefan Englert and Meinhard Kieser 886
Improving the Flexibility and Efficiency of Phase II Designs for Oncology Trials

Les essais de phase II en oncologie sont généralement menés avec des schémas à simple bras, en deux étapes avec des terminaisons binaires. Les schémas adaptatifs couramment disponibles sont configurés pour des études comparatives avec des statistiques de test continues. Le transfert direct de ces méthodes à des statistiques de test discrètes conduit à des procédures conservatives et donc à une perte de puissance. Nous proposons une méthode basée sur un principe de fonction d'erreur conditionnelle prenant directement en compte la nature discrète de l'issue. Nous montrons comment l'application de cette méthode peut être utilisée pour construire de nouveaux schémas de phase II, plus efficients que ceux appliqués de manière courante, et qui permettent des modifications flexibles de schéma à mi-parcours. La méthode proposée est illustrée avec plusieurs schémas de phase II fréquemment utilisés.

Björn Bornkamp 893
Functional Uniform Priors for Nonlinear Modeling

Cet article s'intéresse à la recherche des distributions a priori lorsqu'un composant essentiel d'un modèle statistique dépend d'une fonction non linéaire. En utilisant des résultats sur la façon de construire des distributions uniformes dans des espaces métriques généraux, nous proposons une distribution a priori uniforme dans l'espace des formes fonctionnelles de la fonction non linéaire sous-jacente, et d'appliquer la transformation inverse pour obtenir une distribution a priori pour les paramètres originaux du modèle. Le premier contexte envisagé dans cet article est celui de la régression non linéaire, mais l'idée peut être appliquée bien au delà. Pour la régression non linéaire, les a priori ainsi construits sont invariants par paramétrisation et ils n'enfreignent pas le principe de vraisemblance, contrairement aux distributions uniformes sur les paramètres ou aux a priori de Jeffrey respectivement. L'utilité des a priori proposés est démontrée dans le contexte de la modélisation par régression non linéaire pour des essais cliniques de recherche de dose, à travers un exemple sur des données réelles et par simulation.

Mélanie Prague, Daniel Commenges, Julia Drylewicz, and Rodolphe Thiébaut 902
Treatment Monitoring of HIV-Infected Patients based on Mechanistic Models

Chez la plupart des patients infectés par le VIH, le virus ne peut pas être éradiqué mais la charge virale peut être rendue indétectable par des traitements antirétroviraux hautement actifs. Les traitements doivent alors être pris à vie; l'enjeux majeur est donc la réduction de leurs effets secondaires. Nous proposons de contrôler la dose de traitement afin de trouver la dose minimale qui permet l'indétectabilité de la charge virale. Cette approche se base sur une modélisation mécaniste de l'interaction entre le virus et le système immunitaire avec le modèle à « cellules T activées » qui présente de bonnes qualités

prédictives quant à l'effet de changement de doses. L'existence d'un équilibre non-trivial dans les modèles dynamiques, avec une charge virale non nulle, est caractérisée par un nombre de reproduction de base R_0 plus grand que 1. Pour réduire les effets secondaires on peut donc donner la dose de traitement juste au dessus de la dose critique pour laquelle R_0=1.
Nous utilisons une distribution a priori pour l'ensemble des paramètres du modèle, définie comme l'a posteriori issu de l'analyse d'essais cliniques précédents. Dans une optique bayésienne, les observations de la charge virale et du nombre de CD4 d'un patient donné permettent de mettre à jour la dose de manière à ce qu'il y ait une forte probabilité que le nombre de reproduction de base soit en dessous de 1. L'avantage de cette approche est qu'elle ne repose sur aucune fonction de coût arbitraire pondérant les effets du traitement et son efficacité. Nous montrons qu'il est possible d'approcher la dose critique si le modèle est correctement spécifié. Une analyse de sensibilité a aussi permis de confirmer la robustesse de notre approche.

Anne Chao and Chih-Wei Lin 912
Nonparametric Lower Bounds for Species Richness and Shared Species Richness under Sampling without Replacement

Divers estimateurs de la richesse en espèces ont été développés lorsque les individus (ou les unités d'échantillonnage) sont échantillonnés avec remise. Cependant si l'échantillonnage est sans remise, si bien qu'aucune unité échantillonnée ne peut être observée de façon répétée, les estimateurs usuels pour l'échantillonnage avec remise tendent à surestimer la richesse pour des forts taux d'échantillonnage (rapport de la taille de l'échantillon par rapport au nombre total d'unités cibles de l'échantillonnage) et ne converge pas vers la vraie richesse spécifique quand la fraction échantillonnée se rapproche de 1. Nous proposons une borne inférieure non paramétrique de la richesse d'une seule communauté ou de la richesse partagée (nombre d'espèces communes à plusieurs communautés) pour des données d'abondance ou des données de présence/absence répétées. Les bornes proposées sont obtenues sous des modèles d'échantillonnage très généraux. Elles sont valides pour tous types de distribution d'abondance et toutes probabilités de détection des espèces. Pour des données d'abondance, la détectabilité des individus peut être hétérogène entre espèces. Pour des données de présence-absence répétées, les unités d'échantillonnage sélectionnées (par exemple des quadrats) peuvent présenter une agrégativité spatiale. Toutes les bornes convergent vers les vraies valeurs des paramètres quand la fraction échantillonnée tend vers 1. Des jeux de données réels sont utilisés pour illustration. Nous éprouvons également les bornes inférieures proposées, et en comparons les performances par rapport à celles de certains estimateurs existants, en utilisant des sous-échantillons générés à partir d'enquêtes ou de recensements de grande ampleur.

BIOMETRIC PRACTICE

Corwin M. Zigler and Thomas R. Belin 922
A Bayesian Approach to Improved Estimation of Causal Effect Predictiveness for a Principal Surrogate Endpoint

La littérature adossée au concept de « résultats potentiels » *(potential outcomes)* a montré

que les méthodes traditionnelles de caractérisation des critères de substitution dans les essais cliniques, lorsqu'elles ne se basent que sur les quantités observées, peuvent ne pas détecter les relations de causalité entre traitements, critères de substitution et résultats cliniques. En formulant le *critère de substitution principal* dans le cadre mathématique associé à ce domaine des « résultats potentiels », nous présentons une méthode bayésienne permettant d'estimer la *surface de Prédictivité de l'Effet Causal (CEP)* et de quantifier la capacité d'un critère de substitution à prédire de façon fiable des résultats cliniques. Cette méthode, au regard de la distribution jointe de toutes les quantités potentiellement observables, a des caractéristiques appréciables. Premièrement, elle permet de repérer des hypothèses qui étaient implicites dans d'autres stratégies d'estimation peu performantes. Deuxièmement, elle rend explicites et scientifiquement interprétables des hypothèses portant sur des associations entre variables, associations sur lesquelles les données observées n'apportent pas d'information. Notre approche bayésienne, au vu de simulations réalisées sur l'essai d'un vaccin anti-VIH, montre qu'elle aboutir à de meilleures estimations de la surface CEP que d'autres méthodes. Elle permet enfin d'estimer l'effet du critère principal de substitution dans un cadre plus large que celui considéré dans notre exemple de vaccin, où le critère-candidat de substitution est constant dans l'un des bras de l'étude ; nous l'appliquons en effet dans le contexte d'un essai sur le sida, où, cette fois-ci, le critère-candidat de substitution varie dans les deux bras traités.

Nanhua Zhang and Roderick J. Little 933
A Pseudo-Bayesian Shrinkage Approach to Regression with Missing Covariates

Nous considérons la régression linéaire de Y sur les régresseurs W et Z avec des valeurs de W manquantes, lorsque l'intérêt principal est l'effet de Z sur Y, et que W est une variable de contrôle. Les trois approches habituelles qui existent lorsque des covariables sont manquantes sont : (a) : l'analyse des cas complets (CC) par éliminations des cas incomplets, (b) les méthodes de vraisemblance tronquées avec une inférence sur les données observées en supposant que les données manquantes le sont par hasard (Rubin 1976), et (c) une modélisation qui propose une distribution conjointe des variables et des indicatrices des données manquantes. Une autre approche simple et pratique qui n'a pas beaucoup reçu d'attention sur le plan théorique est d'ignorer dans la modélisation de la régression les variables régresseurs contenant des valeurs manquantes (DV, pour variables à ignorer). DV ne conduit pas à un biais quand soit (a) le coefficient de régression de W est nul soit (b) les variables W et Z ne sont pas corrélées. Nous proposons une approche pseudo-bayésienne pour la régression avec des covariables manquantes qui est un compromis entre les estimateurs CC et DV en exploitant les informations des cas incomplets lorsque les données présentent les hypothèses DV. Nous présentons les bonnes propriétés de la méthode grâce à des simulations, puis nous appliquons la méthode proposée sur une étude du cancer du foie. Un prolongement de la méthode est également envisagé lorsque plus d'une covariable est manquante.

Yangxin Huang and Getachew Dagne 943
Bayesian Semiparametric Nonlinear Mixed-Effects Joint Models for Data with Skewness, Missing Responses, and Measurement Errors in Covariates

Classiquement, dans l'analyse de données longitudinales complexes par des modèles semi-paramétriques non linéaires à effets mixtes (SPNLEM) une distribution normale est utilisée. Cette hypothèse de normalité des erreurs au sein du modèle peut cependant masquer des aspects importants de la variabilité. Par ailleurs, une erreur de mesure substantielle est souvent attachée à certaines des covariables qui sont inclues dans le modèle pour expliquer en partie la variabilité inter et intra-sujets et enfin la réponse peut être manquante avec un processus informatif sous-jacent cette absence. Les procédés d'inférence sont compliqués dans une mesure dramatique en présence de données asymétriques, manquantes et entachées d'erreur de mesure. Si la littérature est riche d'études sur la gestion de l'asymétrie, des données manquantes ou des erreurs systématiques en modélisation, peu ont considérés les trois phénomènes simultanément. L'objectif de ce travail est justement de traiter l'impact simultané de l'asymétrie, des données manquantes et des erreurs systématiques par la modélisation jointe de la réponse et de covariables « processus » avec une version Bayésienne et flexible d'un modèle SPNLEM. Cette méthode est illustrée sur un jeu de donnée réel dans le SIDA, où différents modèles, scénarios et spécifications de distribution sont comparés.

Hongtu Zhu, Joseph G. Ibrahim, Yueh-Yun Chi, and Niansheng Tang 954
Bayesian Influence Measures for Joint Models for Longitudinal and Survival Data

Cet article développe dans un cadre bayésien différentes mesures d'influence afin de réaliser une analyse de perturbation (analyse de sensibilité) dans des modèles de données longitudinales et de survie. Un modèle de perturbation est introduit qui permet de caractériser des perturbations de type individuel ou global; cette modélisation concerne, tout à la fois, la loi a priori, la loi a posteriori, et la loi des observations. Des mesures d'influence locale sont aussi proposées afin de quantifier le degré de perturbation. Les méthodes proposées permettent la détection des points aberrants et des points influents. Elles permettent également d'évaluer la sensibilité de l'inférence vis à vis d'hypothèses non vérifiables. Des études par simulation, ainsi qu'un jeu de données réel, sont utilisés pour illustrer le large spectre des applications des mesures bayésiennes d'influence que nous proposons dans cet article.

Megan D. Higgs and Jay M. Ver Hoef 965
Discretized and Aggregated: Modeling Dive Depth of Harbor Seals from Ordered Categorical Data with Temporal Autocorrelation

Les données catégorielles ordonnées sont fréquemment rencontrées dans les données environnementales et écologiques, et souvent résultent de contraintes qui nécessitent de discrétiser une variable continue en catégories ordonnées. Une grande quantité de données ont été collectées dans une étude portant sur les comportements des mammifères marins à l'aide de satellites enregistreurs des profondeurs (SDRs), qui souvent discrétisent une variable continue comme la profondeur. En outre, les contraintes de stockage ou de transmission des données peuvent nécessiter l'agrégation des données au cours du temps, à l'aide d'intervalles de longueur spécifiée. La catégorisation et l'agrégation conduit à des séries temporelles correspondant à des comptages des catégories multiples ordonnées pour chaque animal, ce qui représente un défit en termes de modélisation statistique et d'interprétation. Nous décrivons une stratégie intuitive pour la modélisation de telles données agrégées ou catégorielles ordonnées, permettant une inférence en regardant les probabilités des catégories et la tendance de la mesure centrale

sur l'échelle originale des données (par exemple, mètres), en introduisant une corrélation temporelle et de la sur-dispersion. La stratégie étend les modèles de censure à covariables spécifiques pour des données ordinales. Nous montrons la méthode en analysant les données issues des SDR enregistreurs des profondeurs des phoques en Alaska. L'objectif principal de l'analyse est d'évaluer la relations entre les covariables, comme le temps de la journée, avec le nombre de plongées et le nombre maximum de plongées. Nous prédisons aussi les données manquantes et introduisons une nouvelle approche graphique de résumés des données et des résultats.

David Todem, Wei-Wen Hsu, and KyungMann Kim 975
On the Efficiency of Score Tests for Homogeneity in Two-Component Parametric Models for Discrete Data

Dans les modèles de mélange à deux composantes sur données discrètes – par exemple les modèles où l'occurrence de la valeur zéro est atypique –, on se pose couramment des problèmes d'inférence à propos des poids qui régissent la répartition des probabilités associées à chacune des deux composantes. Pour réaliser cette inférence, on utilise notamment des tests du score élaborés à partir du modèle marginal sur lequel les poids négatifs sont autorisés. Cependant, les procédures de test existantes, souvent, sont construites à partir d'hypothèses restrictives telles que l'invariabilité des poids, et ignorent la plupart du temps les contraintes structurelles du modèle marginal. Nous développons, dans cet article, un test d'homogénéité (un test du score, de nouveau) qui repousse les limites des procédures existantes. Sa technique repose sur une décomposition des poids en des termes qui ont une interprétation statistique évidente ; c'est cette décomposition qui sert de socle à la construction du test. Par des simulations, nous montrons que notre statistique de test, où les poids peuvent être ajustés sur des covariables, s'avèrent souvent plus efficaces que des tests où les poids sont constants. Nous illustrons notre méthodologie sur un exemple réel concernant les caries dentaires.

Shaowu Tang and Jong-Hyeon Jeong 983
Median Tests for Censored Survival Data; a Contingency Table Approach

Le temps de survie médian est souvent utilisé pour résumer les données de survie car son interprétation est plus simple en pratique pour les investigateurs que la populaire fonction de risque. Cependant, les méthodes existantes pour comparer des temps de survie médians pour des données de survie censurées soit requièrent l'estimation de la fonction de densité de probabilité soit impliquent des formules compliquées pour calculer la variance des estimateurs. Dans ce papier, nous modifions le test des médianes entre K échantillons pour données de survie censurées (*Brookmeyer & Crowley*, 1982) par une approche de table de contingence simple où chaque cellule compte le nombre d'observations dans chaque échantillon qui sont supérieures à la médiane sur l'ensemble des échantillons ou réciproquement. En présence de censure, cette approche générerait des entrées non entières pour les cellules de la table de contingence. Nous proposons de construire une statistique de test pondérée asymptotique qui agrège les statistiques du khi-2 dépendantes constituées des points arrondis à l'entier le plus proche à partir des entrées non entières. Nous montrons que cette statistique suit approximativement une distribution de khi-2 à k–1 degrés de liberté. Pour le cas d'un échantillon de petite taille, nous proposons une statistique de test fondée sur les p-valeurs combinées des tests exacts de

Fisher et qui suit une distribution de khi-2 à deux degrés de liberté. Des études de simulation sont réalisées pour monter que la méthode proposée produit des probabilités d'erreur de type I et des puissances raisonnables. La méthode proposée est illustrée avec deux jeux de données réelles issues d'essais cliniques de phase III pour le cancer du sein.