Translations of Abstracts

Biometric Methodology

L'évaluation diététique des aliments consommés épisodiquement conduit souvent à des données non négatives avec un excès de valeurs nulles et d'erreurs de mesures. Tooze et al (2006, *Journal of the American Dietetic Association* ; **106**, 1575-1587) décrivent une approche statistique générale (méthode de NCI) pour modéliser de telles ingestions de nourriture et démontrer son usage pour estimer la distribution de la prise usuelle de nourriture dans la population. Dans cet article nous proposons une extension de cette méthode pour prédire la prise usuelle individuelle de tels aliments et pour évaluer les relations entre ces prises individuelles et les résultats sur la santé. En suivant l'approche « régression calibration » pour la correction d'erreur de mesure, la prise usuelle individuelle est généralement prédite par la moyenne conditionnelle d'une prise donnée rapportée sur 24h et les autres covariables du modèle de santé. Une caractéristique de la méthode proposée est que les covariables potentielles en liaison avec la prise usuelle peuvent être utilisées pour améliorer la précision des estimateurs de la prise usuelle et les résultats d'association du régime de santé.
En appliquant la méthode sur des données d'alimentation nous quantifions l'accroissement de la précision obtenue en incluant dans le modèle de calibration comme covariable la fréquence de prise présente dans un questionnaire approprié (FFQ). Nous démontrons ensuite la méthode en évaluant la relation linéaire entre le logarithme du niveau de mercure dans le sang et la prise de poisson chez des femmes en utilisant les données d'un sondage provenant de la « National Health and Nutrition Examination » et nous montrons l'amélioration lorsque sont prises en compte les informations du questionnaire FFQ. Enfin nous présentons des résultats de simulation évaluant la performance de la méthode dans ce contexte.

Cet article présente un modèle latent et une méthode d'inférence basée sur la vraisemblance afin de détecter des regroupements temporels d'événements. Le modèle simule des processus susceptibles de générer typiquement les données observées. Nous appliquons des techniques de choix de modèle pour déterminer le nombre de groupes et développons une inférence de vraisemblance et un algorithme EM de Monte-Carlo pour estimer les paramètres du modèle, détecter les regroupements et les localiser. Notre méthode diffère de la statistique de balayage classique en ce qu'elle permet de détecter simultanément des regroupements multiples de taille variable. Nous illustrons la méthodologie sur deux applications de données réelles et évaluons son efficacité par des études de simulation. Pour les processus habituels de génération de données, notre méthode est plus efficace qu'une procédure concurrente qui s'appuie sur les moindres carrés.

L'analyse de données de grande dimension telles que les données de microarray ont apporté de nouveaux problèmes statistiques. Par exemple, l'utilisation d'un grand nombre de gènes pour classer des échantillons à partir d'un petit nombre de microarray reste toujours une tâche difficile. L'analyse discriminante diagonale, les machines à support vectoriel et la méthode des k-plus proches voisins sont considérées comme faisant partie des méthodes les plus appropriées pour analyser des petits échantillons mais aucune n'est supérieure aux autres. Dans cet article, nous proposons une amélioration de l'analyse discriminante diagonale au

travers l'utilisation de méthodes de lissage et de régularisation des variances. Les performances de notre nouvelle méthode ainsi que celles des méthodes existantes sont étudiées au travers de simulations et d'applications à des données réelles. Ces travaux montrent que notre approche par analyse discriminante diagonale basée sur des méthodes de lissage et de régularisation a des erreurs de mis-classification plus faibles que les méthodes existantes dans beaucoup de situations.

W. Chen, D. Ghosh, T. E. Raghunathan, and D. J. Sargent **1030**
Bayesian Variable Selection with Joint Modeling of Categorical and Survival Outcomes: An Application to Individualizing Chemotherapy Treatment in Advanced Colorectal Cancer

Le cancer colorectal est la seconde plus importante cause de décès par cancer aux Etats-Unis, avec plus de 130000 nouveaux cas de cancers colorectaux diagnostiqués chaque année. Des études cliniques ont montré que certaines altérations génétiques conduisent à différentes réponses au même traitement, malgré des tumeurs morphologiquement similaires. Un test moléculaire avant le traitement pourrait permettre de déterminer pour un patient le traitement optimal en termes de toxicité et d'efficacité. Cet article introduit une méthode statistique adaptée à la prédiction et à la comparaison de critères multiples étant donnés plusieurs options de traitements et profils moléculaires d'un individu. Nous considérons ici un modèle de régression multivariée basé sur des variables latentes avec une matrice structurée de variance-covariance. Les variables latentes tiennent compte de la nature corrélée des critères multiples et intègre le fait que certains critères cliniques sont des variables catégorielles et d'autres des variables censurées. La structure hiérarchique de mélanges normaux admet une règle naturelle de sélection de variables. L'inférence a été réalisée en échantillonnant la distribution a posteriori par la méthode de Monte Carlo par chaînes de Markov. Nous analysons les propriétés de la méthode proposée sur des échantillons finis par des simulations. L'application à l'étude du cancer colorectal avancé révèle des associations entre les critères multiples et des biomarqueurs particuliers, ce qui démontre le potentiel de la méthode à individualiser le traitement sur la base de profils génétiques.

L. Xu, T. D. Johnson, T. E. Nichols, and D. E. Nee **1041**
Modeling Inter-Subject Variability in fMRI Activation Location: A Bayesian Hierarchical Spatial Model

Le but de ce travail est de développer un modèle spatial pour des données d'IRM fonctionnelle multi-sujets. De nombreux travaux traitent de la modélisation univariée de chaque voxel pour des études à sujet unique ou multi-sujets, ou encore de modélisation spatiale pour des études à sujet unique, et récemment pour des études multi-sujets. Il n'y a cependant eu aucun travail sur des modèles spatiaux prenant explicitement en compte la variabilité des lieux d'activation entre les sujets. Dans cet article, nous utilisons la notion de centre d'activation et modélisons directement la variabilité des lieux d'activation entre les sujets. La spécification bayésienne hiérarchique que nous avons retenue permet d'inférer aussi bien au niveau de la population, de l'individu qu'au niveau du voxel. Nous utilisons une mélange de gaussiennes pour la probabilité qu'un individu ait une activation spécifique, ce qui permet de répondre à une question non élucidée par les approches usuelles : quelle proportion des sujets ont une activation significative dans une zone donnée ? Notre approche considère le nombre de composantes du mélange comme un paramètre du modèle dont la loi a posteriori est estimée par une méthode MCMC à sauts réversibles. Nous illustrons notre méthode sur une étude IRM fonctionnelle portant sur l'interférence proactive et montrons l'amélioration drastique de la localisation en comparaison aux méthodes univariées usuelles. Bien que la motivation initiale concerne les données d'IRM fonctionnelle, ce modèle peut être adapté à d'autres formes d'images.

J. Cao, C. Z. He, K. M. S. Wells, J. J. Millspaugh, and M. R. Ryan **1052**
Modeling Age and Nest-Specific Survival Using a Hierarchical Bayesian Approach

Des études récentes ont montré que les oiseaux des prés déclinent plus rapidement que les autres groupes d'oiseaux terrestres. Les méthodes actuelles d'estimation des taux de survie des nids spécifiques à l'âge requièrent la connaissance des âges des nids, en supposant les nids homogènes en terme de taux de survie ou en traitant la fonction de risque comme une fonction constante par morceaux. Dans ce papier, nous proposons un modèle bayésien hiérarchique avec des covariables spécifiques aux nids pour estimer des probabilités de survie journalières spécifiques à l'âge sans les contraintes ci-dessus. Le modèle fournit un

estimateur lisse de la courbe de survie des nids et identifie les facteurs associés à la survie des nids. Le modèle peut traiter des calendriers de visite irrégulier et il comporte les hypothèses les moins restrictives parmi toutes les méthodes existantes. Sans hypothèse de risques proportionnels, nous utilisons un modèle logistique multinomial semiparamétrique pour spécifier une relation directe entre la probabilité d'échec des nids spécifique à l'âge et les covariables spécifiques aux nids. Un *a priori* auto-régressif intrinsèque est employé pour l'effet âge du nid. Cet *a priori* non paramétrique fournit une alternative plus souple aux hypothèses paramétriques. Le calcul bayésien est efficace car les distributions postérieures conditionnelles complètes ont une forme analytique ou sont log-concaves. Nous utilisons cette méthode pour analyser le jeu de données Missouri dickcissel (passereaux) et nous trouvons que (1) la survie des nids n'est pas homogène pendant la période de nidification et elle atteint son niveau le plus bas à la transition entre incubation et oisillon et (2) la survie du nid est lié à la couverture herbeuse et à la hauteur de la végétation dans la zone étudiée.

C. X. Mao and J. Li **1063**
Comparing Species Assemblages Via Species Accumulation Curves

La comparaison de rassemblements d'espèces connaissant les données d'incidence est importante dans les études écologiques, souvent faite par une inspection visuelle des courbes d'accumulation des espèces estimées ou par une utilisation ad hoc de bandes de confiance à 95% de ces courbes. Il est montré que comparer des rassemblements d'espèces est un problème stimulant. Un test du χ^2 est proposé. Un ajustement utilisant une décomposition en valeurs propres est proposé pour surmonter les difficultés calculatoires. La méthode bootstrap est aussi suggérée pour approximer la distribution de la statistique de test proposée.
Le test du χ^2 ajusté sur les valeurs propres (Eva) et le test Eva-bootstrap sont évalués par une étude de simulation. Les deux tests (Eva-χ^2 et Eva-bootstrap) sont appliqués sur une étude concernant les rassemblements de deux espèces de plantes ligneuses.

J. S. Yap, J. Fan, and R. Wu **1068**
Nonparametric Modeling of Longitudinal Covariance Structure in Functional Mapping of Quantitative Trait Loci

L'estimation de la structure de covariance des processus longitudinaux est un préalable fondamental pour appliquer les méthodes d'association fonctionnelle, qui permettent d'étudier les régulations génétiques et les variations des loci à effets quantitatif (QTL) complexes et dynamiques. Nous présentons ici une approche non paramétrique pour l'estimation de la structure de covariance d'un QTL mesuré de façon répétée à une suite d'instants. Plus précisément, nous adoptons l'approche de Huang et al. (2006a), qui se base sur la décomposition de Cholesky modifiée. On est alors ramené à la modélisation d'une suite de régression sur des réponses. On obtient un estimateur de covariance régularisé en utilisant une vraisemblance normale pénalisée avec une pénalité L2. Cette approche, incluse dans un cadre de vraisemblance de mélange, conduit à accroitre la justesse, la précision et la flexibilité de l'association fonctionnelle tout en préservant sa pertinence biologique. Les propriétés statistiques et les avantages de la méthode proposée sont présentés à l'aide de simulations. Un exemple tiré d'un projet sur le génome de la sourie est analysé pour illustrer l'utilisation de notre méthodologie. Cette nouvelle méthode fournira un outil utile pour rechercher à l'échelle du génome l'existence et la distribution de QTL liés à un caractère dynamique important en agriculture, en biologie ou en médecine.

S. W. Thurston, D. Ruppert, and P. W. Davidson **1078**
Bayesian Models for Multiple Outcomes Nested in Domains

Nous considérons le problème d'estimer l'effet d'une exposition sur de multiples variables réponses continues, lorsque ces variables sont mesurées avec différentes échelles et sont emboitées dans plusieurs classes ou "domaines". Notre modèle bayésien étend l'approche par modèles mixtes linéaires pour permettre à l'effet de l'exposition d'être différent selon les domaines et selon les variables réponse dans les domaines. Notre modèle peut être paramétré pour permettre la contraction des effets dans les différents niveaux d'emboitement ou pour permettre des effets fixes spécifiques aux domaines sans contraction. Notre modèle permet aussi que les effets des covariables diffèrent selon les variables à expliquer et les domaines.

Notre méthodologie est appliquée à des données d'exposition prénatale au méthylmercure et de multiples variables réponses dans quatre domaines, mesurées à l'âge de 9 ans auprès d'enfants inclus dans l'étude sur le développement de l'enfant aux Seychelles. Nous utilisons trois a priori différents et montrons que nos principales conclusions ne sont pas sensibles aux choix des a priori. Des études de simulation examinent la performance du modèle selon différents scenarii. Nos résultats démontrent qu'une augmentation mesurable de la puissance est possible.

J. A. L. Gelfond, M. Gupta, and J. G. Ibrahim **1087**
A Bayesian Hidden Markov Model for Motif Discovery Through Joint Modeling of Genomic Sequence and ChIP-Chip Data

Nous proposons un cadre d'analyse des puces « ChIP-chip » à immunoprécipitation (IP) de chromatine (Ch) pour la détection de sites de fixation des facteurs de transcription (TFBS). Dans ces expériences, des fragments d'ADN contenant des TFBS sont tout d'abord isolés puis déposés sur une puce munie de sondes recouvrant le génome. Les méthodes d'analyse actuelles procèdent en 2 étapes : (i) analyse de la puce pour estimer les pics d'intensité IP, puis (ii) analyse des séquences d'ADN correspondantes indépendamment de l'information sur l'intensité des pics. Le modèle proposé permet la détection du pic et l'identification des motifs simultanément grâce à une approche combinant analyse bayesienne et modèle de Markov caché (HMM) qui permet de prendre en compte l'incertitude qui existe dans les deux mesures. Nous proposons un algorithme de Monte Carlo par chaîne de Markov pour l'estimation des paramètres, en adaptant les techniques récursives utilisées pour les HMM. Sur des simulations et des données de levure RAP1, nous montrons que la méthode proposée a de bonnes performances pour trouver les TFBS comparée aux méthodes actuelles procédant en deux étapes, autant en terme de sensitivité que de spécificité.

Y. Li and B. I. Graubard **1096**
Testing Hardy–Weinberg Equilibrium and Homogeneity of Hardy–Weinberg Disequilibrium using Complex Survey Data

En génétique des populations, l'utilisation d'échantillons aléatoires représentatifs de la population étudiée permet d'éviter des biais d'échantillonnage. Des données génétiques sur le polymorphisme d'une centaine de gènes ont été recueillies dans un échantillon représentatif de la population américaine, l'étude NHANES III (Third National Health and Nutrition Examination Survey). Une étude comme NHANES III est basée sur un plan d'expérience complexe incluant une stratification hiérarchique par sous-groupes où l'utilisation de pondérations dans le mode d'échantillonnage peut conduire à une surestimation des variances et à modifier la distribution attendue des tests statistiques. Les tests d'équilibre d'Hardy-Weinberg (*HWE* en anglais) et d'homogénéité du déséquilibre d'Hardy-Weinberg (*HHWD* en anglais) utilisés habituellement pour des échantillons aléatoires ne peuvent donc pas s'appliquer à de tels plans d'expérience. Nous proposons des versions modifiées des tests de Wald pour le *HWE* et du score généralisé pour le *HHWD* applicables à ces plans d'expérience complexes. Des études de simulation par Monte Carlo sont utilisées pour vérifier les propriétés de ces tests à distance finie. L'application des corrections de Rao-Scott a permis d'améliorer les performances de ces tests en terme d'erreur de type I. Ces méthodes sont appliquées à des données génétiques de l'étude NHANES III concernant trois loci impliqués dans le métabolisme.

L. Chen, L. Hsu, and K. Malone **1105**
A Frailty-Model-Based Approach to Estimating the Age-Dependent Penetrance Function of Candidate Genes Using Population-Based Case-Control Study Designs: An Application to Data on the BRCA1 Gene

L'étude basée sur une population cas-témoins est peut-être l'une sinon la plus communément utilisée pour rechercher les contributions génétiques et environnementales au risque de maladie dans les études épidémiologiques. L'âge au commencement de la maladie et le statut par rapport à la maladie des membres de la famille sont systématiquement collectés parmi les participants à l'essai. En considérant l'âge du début de la maladie chez les parents comme résultat, cet article se focalise sur l'utilisation de l'information de l'histoire familiale pour obtenir la fonction de risque c'est-à-dire la fonction de pénétrance âge-dépendante de gènes candidats dans les études cas-témoins. Une approche basée sur un modèle de fragilité est proposée pour intégrer le risque partagé par les membres de la famille qui n'est pas pris en compte par les facteurs de risques observés. Cette approche est prolongée pour associer les génotypes manquants chez les membres de

la famille et un plan d'échantillonnage cas-témoins en deux phases. Des résultats de simulation montrent que la méthode marche bien dans des contextes réalistes. Pour finir la méthode est illustrée par une étude du gène BRCA1 dans le cancer du sein dans une étude cas-témoin en deux phases .

J. Joo, M. Kwak, K. Ahn, and G. Zheng **1115**
A Robust Genome-Wide Scan Statistic of the Wellcome Trust Case–Control Consortium

Dans les études d'association pangénomiques, les statistiques de test qui sont efficaces et robustes pour l'ensemble des modèles génétiques sont bien sur préférables, en particulier lorsque l'on étudie plusieurs maladies comme dans le projet mené par le Wellcome Trust Case-Control Consortium (WTCCC,2007). Une nouvelle statistique de test, la p-value minimale entre le test de tendance et le test de Pearson, a été proposée par le WTCCC. Elle sera dénotée MIN2. Parce la plus petite p-value de deux p-values n'est plus une p-value valide elle-même, le WTCCC l'a seulement utilisée pour classer les SNPs mais n'a pas rapporté les p-values des SNPs associés a/aux maladie(s) étudiée(s) lorsque MIN2 a été utilisé pour le classement. Compte tenu de son importance pratique, nous dérivons ici la distribution asymptotique de MIN2 sous l'hypothèse nulle, nous étudions certaines de ses propriétés analytiques dans le contexte des études d'association pangénomiques et nous le comparons aux méthodes existantes les plus communément utilisées (test de tendance, test de Pearson, MAX3 et test du rapport de vraisemblance restreint CLRT) au travers d'études de simulation sous différents modèles génétiques (i.e. récessif, additif, multiplicatif, dominant et surdominant). Les résultats montrent que MAX3 et CLRT sont plus robustes et une plus efficaces que les autres tests lorsque les modèles récessif, additif/multiplicatif et dominant sont possibles tandis que MIN2 et le test de Pearson sont plus performants lorsque les modèles génétiques incluent le modèle de surdominance. Nous concluons que les tests robustes (MIN2, MAX3, CLRT et Pearson) sont préférables à un simple test de tendance pour l'analyse initiale des études d'association pangénomiques. Finalement, les quatre tests robustes sont appliqués à plus de 100 SNPs associés avec onze maladies identifiées dans les 2 études d'association pangénomiques du WTCCC.

D. A. Costain **1123**
Bayesian Partitioning for Modeling and Mapping Spatial Case–Control Data

La modélisation et la représentation des variations spatiales d'un risque de maladie constituent toujours un sujet de recherche actif. Outils d'exploration des hétérogénéités régionales en santé, les analyses spatiales peuvent fournir des indications sur l'étiologie des maladies et la gestion du système de santé ou générer des hypothèses de recherche. Cet article présente une approche bayésienne du partitionnement en analyse de données individuelles géo-référencées. Le modèle exige peu d'hypothèses sur la forme de la surface de risque, il est adaptatif et il permet de prendre en compte les déterminants connus de la maladie. Cette approche est utilisée pour modéliser les variations spatiales de la mortalité néonatale à Porto Alegre, Brésil.

Y. Huang and M. S. Pepe **1133**
A Parametric ROC Model-Based Approach for Evaluating the Predictiveness of Continuous Markers in Case–Control Studies

La courbe de prédiction montre la distribution dans la population du risque attaché à un marqueur ou à un modèle de prédiction du risque. Elle fournit un moyen d'évaluer la capacité du modèle à stratifier une population en fonction du risque. Des méthodes d'inférence sur la courbe de prédiction ont été mises au point pour des données transversales ou de cohortes. Ici nous considérons une inférence basée sur des études épidémiologiques beaucoup plus couramment utilisées, par comparaison de cas observés à des témoins. Nous étudions la relation entre la courbe ROC (qui mesure la caractéristique d'opération du récepteur dans la détection des signaux) et la courbe de prédiction. L'approfondissement de cette relation offre un éventail d'interprétations ROC pour les courbes de prédiction et pour un index synthétique proposé précédemment pour les résumer. En outre la relation plaide en faveur des méthodes ROC pour estimer la courbe de prédiction. Un avantage important de ces méthodes sur celles précédemment proposées est leur invariance relativement au risque. Elles offrent aussi un moyen de combiner l'information couvrant un ensemble de populations ayant des ROC similaires mais des prévalences variables du phénomène étudié. Nous appliquons la méthode au marqueur PSA de prédiction du risque de cancer de la prostate.

La modélisation traditionnelle par classe latente a été largement utilisée pour évaluer l'exactitude de tests de diagnostic dichotomiques. Cependant ces modèles supposent que les résultats sont indépendants conditionnellement au statut réel de la maladie, cela est rarement le cas en pratique. D'autres modèles, utilisant les probits, ont été proposés qui tiennent compte d'une dépendance entre les tests mais en n'utilisant que quelques structures de corrélations. Dans ce papier nous proposons un modèle de classe latente basé sur les probits qui permet d'utiliser une structure de corrélation quelconque. Quand on le combine avec quelques diagnostics utiles, ce modèle fonctionne dans un cadre plus souple qui permet l'évaluation de la structure de corrélation et ainsi que de l'ajustement du modèle. Notre modèle incorpore plusieurs autres modèles de classe latente avec probits mais utilise un algorithme PX-EM (parameter expanded Monte Carlo EM) pour obtenir les estimations par maximum de vraisemblance. Cet algorithme est conçu pour accélérer la convergence de l'algorithme EM en développant le modèle pour les données complètes pour y inclure un ensemble plus large de paramètres et assurer une solution simple à l'ajustement du modèle de classe latente en probits. On démontre les performances de notre estimation et sa méthode de sélection du modèle en utilisant une simulation et deux études médicales publiées.

Le boosting est une puissante approche pour ajuster des modèles de régression. Cet article décrit un algorithme de boosting pour une estimation basée sur la vraisemblance avec des données incomplètes. L'algorithme combine le boosting avec une variante de l'approximation stochastique qui utilise la chaine de Monte Carlo pour prendre en compte les données manquantes. Des applications pour ajuster des modèles linéaires généralisés et des modèles additifs avec des covariables manquantes sont données. La méthode est appliquée aux données de diabète des indiens Pima où près de la moitié des observations contiennent des données manquantes.

Dans l'analyse de données manquantes, on utilise habituellement des analyses de sensibilité pour vérifier la sensibilité des paramètres d'intérêt au regard du mécanisme de données manquantes et à d'autres hypothèses portant sur le modèle ou la distribution. Dans cet article, nous développons formellement une méthode générale d'influence locale pour mener les analyses de sensibilité à de petites perturbations dans les modèles linéaires généralisés en présence de covariables manquantes. Nous examinons deux types de schémas de perturbation (le cas simple et les schémas de pertubation globale) des diverses hypothèses dans ce contexte. Nous montrons que le tenseur métrique d'une variété de perturbations fournit une information utile pour sélectionner une perturbation appropriée. Nous développons aussi plusieurs mesures d'influence locale pour identifier les points influents et tester la mauvaise spécification du modèle. Des simulations sont conduites pour évaluer nos méthodes, et des données réelles sont analysées pour illustrer l'utilisation de nos mesures d'influence locale.

Dans les études basées sur des données longitudinales ou de panel, les réponses manquantes dépendent souvent des valeurs des réponses pour un effet-sujet aléatoire non observé. Au delà de l'approche par vraisemblance basée sur des modèles paramétriques existe une méthode semiparamétrique, le modèle conditionnel approché (ACM) , qui repose sur la disponibilité d'une statistique de résumé et une approximation linéaire ou polynomiale des effets aléatoires. Cependant, deux éléments importants doivent être bien pris en considération pour appliquer ACM. Tout d'abord comment trouver une statistique de résumé, et ensuite comment estimer les paramètres du modèle original à l'aide des paramètres estimés par ACM. Notre travail concerne ces deux aspects. Pour le premier, nous obtenons des statistiques de résumé

dans des contextes variés. Pour le second point, nous proposons une méthode de regroupement, au lieu de l'approximation linéaire ou polynomiale pour les effets aléatoires. La méthode de regroupement étant une approche basée sur les moments, les conditions que nous supposons pour l'obtention des statistiques de résumé sont plus faibles que celles de la littérature. Lorsque la statistique de résumé obtenue est continue, nous proposons d'utiliser une méthode par arbre de classification pour obtenir une statistique de résumé approchée pour le regroupement. Nous présentons des résultats de simulation pour l'étude des performances de la méthode proposée avec des échantillons finis. Une application est présentée en utilisant des données d'une étude sur la modification de régime dans l'insuffisance rénale.

Y. Liu, H. Zhou, and J. Cai **1184**
Estimated Pseudopartial-Likelihood Method for Correlated Failure Time Data with Auxiliary Covariates

Comme les études biologiques deviennent de plus en plus couteuses à conduire, des méthodes statistiques, qui considèrent l'information auxiliaire existante à propos d'une variable d'exposition couteuse, sont souhaitables en pratique. De telles méthodes devraient améliorer l'efficacité de l'étude et augmenter la puissance statistique pour un nombre de manipulations donné. Dans cet article, nous considérons une procédure inférentielle pour l'analyse multivariée avec l'information auxiliaire sur les covariables. Nous proposons un estimateur de la vraisemblance pseudo-partielle dans le cadre d'un modèle à risque marginal et nous développons les propriétés asymptotiques de l'estimateur proposé. Nous réalisons une étude de simulations pour évaluer la performance de la méthode proposée dans des situations pratiques. La méthode est appliquée à des données de l'étude sur l'insuffisance ventriculaire gauche (SOLVD, 1991).

A. G. DiRienzo **1194**
Flexible Regression Model Selection for Survival Probabilities: With Application to AIDS

Les cliniciens s'intéressent souvent à l'effet de covariables sur les probabilités de survie à des délais pré-définis. Comme différents facteurs peuvent être associés avec le risque d'échec à court et à long terme, une stratégie de modélisation flexible est poursuivie. Etant donné un ensemble de multiples modèles candidats possibles, une méthodologie objective est proposée. Son but est de construire des estimateurs convergents et asymptotiquement normaux des coefficients de régression et de l'erreur de prédiction moyenne pour chaque modèle, qui sont affranchis de la censure, variable de nuisance. Cela nécessite la modélisation de la distribution de la censure conditionnelle aux covariables. La stratégie de sélection de modèle utilise des procédures de tests d'hypothèses multiples ascendantes ou descendantes qui contrôlent soit la proportion de faux positifs soit le taux d'erreur lié à la famille généralisé quand on compare des modèles basées sur des estimateurs de l'erreur de prédiction moyenne. Le contexte peut en effet être envisagé comme un problème de données manquantes, où les estimateurs des coefficients de régression dérivés de cas complets augmentés par pondération de probabilité inverse (AIPWCC) et les erreurs de prédiction peuvent être utilisés (Tsiatis, 2006). Une étude de simulation et une analyse intéressante d'un essai récent dans le SIDA sont présentées.

S.-H. Jung, J.-H. Jeong, and H. Bandos **1203**
Regression on Quantile Residual Life

Une méthode de régression log-linéaire avec un effet temps sur le quantile de durée de vie résiduelle est proposée. Sous le modèle proposé avec censure à droite, n'importe quel quantile de la distribution du délai jusqu'à l'événement parmi les survivants au-delà d'un certain temps est associée à des covariables sélectionnées. La consistance et la normalité asymptotique de l'estimateur de la régression sont établies. Un test statistique asymptotique est proposé pour évaluer les effets des covariables sur le quantile de durées de vie résiduelles à un temps donné. L'évaluation de la statistique du test ne nécessite pas l'estimation d'une matrice de variance-covariance des coefficients de régression, qui implique la fonction de densité de la distribution de survie avec censure. Des études de simulations sont réalisées pour évaluer sur des échantillons finis les propriétés de l'estimateur des paramètres de la régression et de la statistique de test. La nouvelle méthode de régression est appliquée sur une base de données de cancer du sein avec un suivi à long terme pour estimer les durées de vie résiduelles *médianes*, en ajustant sur les facteurs pronostiques importants.

Une procédure de test de comparaison multiple contrôle versus traitement est développée pour des échantillons de sous-ensembles triés. Le test, construit sur les *K*-indépendants intervalles de confiance exacts de la médiane correspondant aux *K* populations, rejette l'hypothèse nulle de l'égalité des médianes si les intervalles de confiance de la médiane du groupe contrôle et n'importe lequel des *K*-1 autres groupes traitement sont disjoints. L'utilisation du test proposé est illustrée avec des données d'une expérimentation agricole.

Biometric Practice

L'évaluation des vaccins candidats contre le VIH chez les primates non-humains (NHPs) constitue un pas critique vers le développement d'un vaccin réussi pour contrôler la pandémie de VIH. Historiquement, les régimes de vaccins contre le VIH ont été testés sur les NHPs en administrant une unique dose élevée du virus en compétition. Plus récemment, l'évaluation des vaccins candidats contre le VIH a entraîné des infections d'épreuve répétées à faible dose, qui mimaient plus précisément l'exposition dans le dispositif de transmission naturelle. Dans ce papier, nous considérons l'évaluation du type et de l'ampleur de l'efficacité du vaccin à partir de telles expériences. Basée dans un cadre de stratification principale, nous discutons aussi l'évaluation de potentiels marqueurs immunologiques pour l'infection.

La restriction à la croissance fœtale est une cause majeure de la morbidité et de la mortalité périnatale qui pourrait être réduite si les enfants à haut risque étaient identifiés précocement durant la grossesse. Nous proposons un modèle bayésien pour agréger dix huit mesures par ultrasons de taille fœtale et de flux sanguin en trois facteurs latents continus sous-jacents. Notre procédure est plus flexible que les méthodes habituelles à variables latentes en ce que nous relaxons la supposition de normalité en permettant aux facteurs latents de suivre des distributions de mélange fini. Utiliser des distributions de mélange nous permet aussi de grouper des individus avec des caractéristiques observées similaires et d'identifier des classes latentes de sujets qui ont plus tendance à se développer ou à avoir une restriction de flux sanguin durant la grossesse. Nous utilisons aussi notre modèle de mélange pour identifier une classe latente à signification clinique de sujets de faible poids à la naissance et d'âge gestationnel précoce. Nous examinons alors l'association des classes latentes de croissance intra-utérine restreinte avec les classes latentes d'issues de la naissance et avec des covariables maternelles observées incluant le sexe de l'enfant, la race, la parité, l'indice de masse corporelle (IMC) et la taille. Nos méthodes ont identifié une classe latente de sujets avec une restriction de flux sanguin croissante et une taille intra-utérine sous la moyenne qui avaient plus tendance à avoir une croissance restreinte à la naissance qu'une classe avec taille et flux sanguin typique.

Dans beaucoup d'applications concernant des données indexées géographiquement, l'intérêt se concentre sur l'identification de régions avec des changements rapides dans la surface spatiale, ou bien sur le problème vu sous un autre angle de la construction ou de la reconnaissance de frontières séparant les régions avec des valeurs observées sensiblement différentes. Ce processus est souvent référencé dans la littérature comme l'analyse de frontières par la méthode de Womble (wombling). De récents développements dans les modèles hiérarchiques pour des données prises en des points prédéterminés (géostatistique) et dans des grilles (lattice), ont conduit à des méthodes de wombling, mais celles-ci n'apparaissent dans aucune littérature sur le sujet dans le cas de processus ponctuels, où les positions elles-mêmes sont supposées aléatoires, et l'évaluation de la vraisemblance est notoirement difficile. Nous

étendons les méthodes de wombling développées pour les points prédéterminés et dans des grilles à ce cas, en obtenant une inférence a posteriori complète pour les effets aléatoires multivariés qui, lorsqu'on les représente sur une carte, peuvent suggérer des covariables spatiales qui manquent encore dans le modèle. Dans le cas des grilles nous pouvons aussi construire des cartes selon la méthode de Womble, montrant des frontières significatives sur la surface d'intensité ajustée, tandis que la formulation avec des points prédéterminés permet de tester la significativité des frontières postulées. En ce qui concerne les besoins calculatoires dans le cas des points prédéterminés, notre algorithme combine des approches Monte Carlo de la vraisemblance avec une étape prédictive du processus pour réduire la dimension du problème à une taille convenable. Nous appliquons ces techniques à l'analyse de données de cancers colorectaux et de la prostate recueillies dans le nord du Minnesota, où un élément clé possible provient de similitudes possibles dans leur répartition spatiale, et si elles sont affectées ou non par l'éloignement de chaque patient des installations capables d'effectuer des examens diagnostiques du cancer.

C. Czado, T. Gneiting, and L. Held **1254**
Predictive Model Assessment for Count Data

Nous envisageons des outils pour l'évaluation de prédictions probabilistes et la critique de modèles statistiques appliqués aux données de comptages. Nous proposons une transformation non randomisée de la fonction de répartition, des diagrammes de calibration marginaux, et des règles de quantification appropriées, comme la déviance prédictive. Dans les études de cas, nous discutons des modèles de régressions appliqués à des nombres de brevets et nous évaluons la performance prédictive de modèles âge-période-cohorte bayésiens concernant l'incidence des cancers du larynx en Allemagne. Nos outils sont utilisables en statistique bayésienne ou classique, dans un cadre paramétrique ou non paramétrique et ils s'appliquent à n'importe quel type de données discrètes ordonnées.

T. Maiti and V. Pradhan **1262**
Bias Reduction and a Solution for Separation of Logistic Regression with Missing Covariates

La régression logistique est une procédure statistique importante utilisée dans de nombreuses disciplines. La plupart des logiciels d'analyse de données comprennent une fonction permettant d'obtenir les estimateurs du maximum de vraisemblance de ce modèle de manière itérative. Il est toutefois bien connu que les estimateurs obtenus pour des échantillons de taille petite ou moyenne sont biaisés. De plus, un phénomène dit de ``séparation'' se produit souvent, phénomène selon lequel l'estimateur du maximum de vraisemblance de l'un au moins des paramètres diverge vers l'infini. Les approches usuelles d'estimation ne prennent pas en compte ces problèmes, ni celui des valeurs manquantes dans les variables explicatives du modèle. Dans cet article, nous traitons ces trois problèmes fréquents en pratique - le biais, la séparation et les covariables manquantes- par de simples ajustements de la procédure d'estimation. La technique proposée est appliquée à des données réelles et à des données simulées. Dans tous les cas, les estimateurs convergent et sont moins biaisés que selon l'approche usuelle. Une macro SAS implémentant la méthode décrite ici est disponible auprès des auteurs.

P. Chen, J. M. Tebbs, and C. R. Bilder **1270**
Group Testing Regression Models with Fixed and Random Effects

Pour le dépistage des maladies infectieuses, on recourt, depuis très longtemps et avec succès, à des tests statistiques réalisés sur données groupées (plutôt que sur données individuelles). Dans cet article, nous développons des modèles de régression pour données groupées, intégrant les effets de certaines covariables lorsque ceux-ci sont plutôt de nature aléatoire. Nous présentons différentes approches permettant d'estimer des modèles mixtes par le maximum de vraisemblance, d'examiner le comportement du rapport de vraisemblance et du test du score pour les composantes de la variance, et d'évaluer par simulation la qualité des résultats obtenus avec de petits échantillons. Nous illustrons nos méthodes sur des données collectées par l'état du Nebraska, dans le cadre d'un Projet de Prévention de l'Infertilité, concernant les blennoragies et les infections à chlamydia.

W. Liu, F. Bretz, A. J. Hayter, and H. P. Wynn **1279**

Assessing Nonsuperiority, Noninferiority, or Equivalence When Comparing Two Regression Models Over a Restricted Covariate Region

Dans beaucoup de problèmes scientifiques, le but de la comparaison de deux modèles de régression décrivant, pour deux groupes différents, la relation entre une même variable réponse et les mêmes variables explicatives, est de démontrer qu'aucun des deux modèles n'est, de façon autre que négligeable, meilleur que l'autre, ou de démontrer que les différences entre lesdits modèles sont si faibles que l'on peut considérer qu'ils décrivent pratiquement la même relation entre la variable réponse et les covariables. Dans cet article, nous proposons des méthodes basées sur des bandes de confiance unilatérales afin d'évaluer la non-supériorité d'un modèle sur l'autre et l'équivalence de deux modèles de régression. Nous illustrons ces méthodes à l'aide d'exemples issus d'une étude QT/QTC et d'une étude de stabilité d'un médicament.

M. Kwak, J. Joo, and G. Zheng **1288**
A Robust Test for Two-Stage Design in Genome-Wide Association Studies

Un plan d'analyse en deux étapes des études d'association pangénomiques (GWAS) testant des centaines de milliers de polymorphismes mononucléotidiques (SNPs) est une stratégie efficace en termes de cout. Dans ce type de plan d'analyse, chaque SNP est génotypé à l'étape 1 en utilisant une partie seulement des cas et des témoins. Les SNPs les plus prometteurs sont sélectionnés et génotypés à l'étape 2 en utilisant les échantillons supplémentaires. Une analyse conjointe combinant les statistiques des deux étapes est utilisée lors de l'étape 2. Les études de suivi incluant des échantillons de validation indépendants peuvent être considérées comme un plan d'analyse en deux étapes. Lorsque des SNPs d'intérêt ont été identifiés, ils sont génotypés dans des échantillons indépendants supplémentaires et sont analysés séparément ou conjointement avec les données initiales pour confirmer les résultats de la première analyse. Lorsque le modèle génétique sous-jacent est connu, un test de tendance asymptotiquement optimal peut être utilisé pour chaque analyse. En pratique, cependant, les modèles génétiques des SNPs réellement associés sont habituellement inconnus. Dans ce cas, les méthodes existantes pour l'analyse des plans d'analyse en 2 étapes et des études de suivi ne sont pas robustes au travers des différents modèles génétiques. Nous proposons une procédure simple et robuste avec sélection du modèle génétique dans les GWAS en deux étapes. Nos résultats montrent que si le test de tendance optimal à une puissance de 80% lorsque le modèle génétique est connu, les méthodes d'analyse existantes pour l'analyse des études en deux étapes ont des puissances minimales autour de 20% lorsque l'on considère les quatre modèles génétiques les plus communs (le modèle vrai étant inconnu) tandis que notre approche à des puissances minimales de l'ordre de 70%. Ces résultats sont aussi applicables aux études de suivi et aux études de réplication avec analyse conjointe.

S. M. DeSantis, E. A. Houseman, B. A. Coull, D. N. Louis, G. Mohapatra, and R. A. Betensky **1296**
A Latent Class Model with Hidden Markov Dependence for Array CGH Data

La puce à ADN est une technique à haut débit permettant de détecter des altérations génomiques liées au développement et à la progression d'un cancer. La technique conduit à des ratios de fluorescence qui caractérisent le nombre de copies d'ADN dans la tumeur versus dans les cellules saines. La classification des tumeurs, basée sur les profils de CGH est intéressante scientifiquement mais l'analyse de ces données est rendue complexe par le grand nombre de mesures hautement corrélées. Dans ce papier, nous développons une approche de classification latente bayésienne, supervisée, qui repose sur un modèle de Markov caché afin de prendre en compte la dépendance des ratios d'intensité. La supervision signifie que la classification est guidée par un événement clinique. Les inférences a posteriori sont faites sur des classes spécifiques de gains et pertes du nombre de copies. Nous démontrons notre technique sur une étude de tumeurs du cerveau, pour laquelle notre approche est capable d'identifier des sous-ensembles de tumeurs ayant des profils génomiques différents et de différencier des classes par survie bien mieux que par des méthodes non supervisées.