## Translations of Abstracts

### Biometric Methodology



Nous proposons une nouvelle méthodologie de classification en présence de données éparses. Les données éparses sont définies comme dissimilaires à toute autre, conduisant ainsi les approches traditionnelles à des conclusions erronées en les forçant dans des groupes. L'approche que nous proposons est un schéma qui, sous hypothèses de classes sphériques, construit itérativement des noyaux autour de leurs centres, regroupe les points de chaque noyau et identifie les points extérieurs comme épars. En l'absence de données éparses, l'algorithme se réduit à celui des k-means. Nous proposons également une méthodologie pour initialiser l'algorithme et pour estimer le nombre de classes dans les données. Nos résultats pour des situations expérimentales montrent d'excellentes performances, particulièrement pour des classes à symétrie elliptique. La méthodologie est appliquée à l'analyse des rapports sur les relargages industriels de mercure en 2000 établis par l'Agence Américaine de Protection de l'Environnement (EPA) pour l'inventaire des relargages toxiques (TRI).



La classification spatiale est communément modélisée par une méthode bayésienne dans le cadre des modèles linéaires généralisés à effets mixtes (GLMM). Les classes spatiales sont communément détectées par une méthode fréquentiste reposant sur des tests d'hypothèses. Dans cet article, nous proposons une méthode fréquentiste pour établir les propriétés spatiales des GLMM. Nous suggérons une stratégie de détection des classes spatiales par l'estimation des paramètres des associations spatiales, et qui établit les qualités spatiales au travers des résidus itérés. Des simulations et une étude de cas montrent que la méthode proposée est capable de détecter les positions et les dimensions des classes spatiales, avec consistance et efficacité.

Les modèles linéaires mixtes généralisés (GLMMs) sont largement utilisés dans l'analyse de données classifiées. Cependant la validité de l'inférence basée sur la vraisemblance dans de telles analyses peut être très dépendante du modèle supposé pour les effets aléatoires. Nous proposons une méthode de diagnostic pour l'erreur de classification par modèle à effets aléatoires dans les GLMM en cas de réponse binaire. Nous fournissons une justification théorique de la méthode proposée et nous étudions par simulation ses performances sur échantillon de taille fini. La méthode proposée est appliquée à des données d'une étude longitudinale d'infection respiratoire.



Décider quels effets prédictifs peuvent varier entre sujets est un problème difficile. Les procédures de test et les critères de sélection de modèles classiques sont souvent inappropriés pour comparer des modèles avec des nombres différents d'effets aléatoires en raison de contraintes sur l'espace des paramètres des composantes de la variance. Tester sur la frontière de l'espace des paramètres modifie la distribution asymptotique de quelques tests classiques et crée des problèmes pour l'approximation des facteurs de Bayes. Nous proposons une approche simple pour tester des effets aléatoires dans le modèle linéaire mixte en utilisant les facteurs de Bayes. Nous normalisons chaque effet aléatoire par la variance résiduelle et introduisons un paramètre qui contrôle la contribution de chaque effet aléatoire indépendamment de l'échelle de mesure des données. Nous intégrons les effets aléatoires et les composantes de la variance en utilisant des formules explicites. Les intégrales résultantes nécessaires pour calculer le facteur de Bayes sont des intégrales de dimension réduite sans les composantes de la variance et peuvent être efficacement approchées avec la méthode de Laplace. Nous proposons une distribution a priori par défaut sur le paramètre contrôlant la contribution de chaque effet aléatoire et effectuons des simulations pour montrer que notre méthode a de bonnes propriétés pour la sélection de modèles. Enfin, nous illustrons nos méthodes sur les données d'un essai clinique dans le trouble bipolaire et sur les données d'une étude épidémiologique du lien entre produits de désinfection de l'eau et fertilité masculine.



Dans les analyses longitudinales il est souvent supposé que les temps d'observations sont prédéterminés et les mêmes pour tous les sujets. Une telle supposition néanmoins est souvent violée en pratique. En conséquence, les temps d'observation peuvent être hautement irréguliers. Il est bien connu que si le schéma d'échantillonnage est corrélé avec les résultats, l'analyse statistique usuelle peut conduire à un biais. Dans ce papier, nous proposons une modélisation jointe et une analyse des données longitudinales avec des temps d'observation possiblement informatifs via des variables latentes. Une procédure d'estimation en deux étapes est développée pour l'estimation des paramètres. Nous montrons que les estimateurs résultants sont consistants et asymptotiquement normaux et que la variance asymptotique peut être estimée de façon consistante en

utilisant la méthode du bootstrap. Des études de simulation et une analyse de données réelles démontrent que notre méthode marche bien avec des tailles d'échantillon réalistes et est appropriée pour un usage pratique

Y. Cheng, J. Fine, and M. R. Kosorok 385
*Nonparametric Association Analysis of Exchangeable Clustered Competing Risks Data*

Regularized estimation for the accelerated failure time model Lu Tian, J. Huang and T. CaiCe travail a été motivé par la Cache County Study of Aging (étude du vieillissement du Cache County), une étude de population en Utah, dans laquelle on étudie les associations au sein des fratries du début de la démence. Les difficultés tiennent au fait que seule une fraction de la population sera atteinte de démence, la majorité décédant sans démence. L'application des analyses standard de dépendance pour des données censurées indépendamment à droite peut ne pas être appropriée avec de telles données de risques compétitifs multivariées, où le décès peut ne pas satisfaire l'hypothèse de censure indépendante. Des estimateurs non paramétriques de la fonction de risque cumulée bivariée et la fonction d'incidence cumulée bivariée sont adaptés du contexte bivarié simple non échangeable aux données regroupées échangeables, comme cela est nécessaire avec les grandes fratries de la Cache County Study. Des mesures d'association dépendant du temps sont évaluées avec ces estimateurs. Des inférences pour grands échantillons sont étudiées de façon rigoureuse en utilisant des techniques de processus empirique. L'utilité pratique de la méthodologie est démontrée à l'aide d'échantillons réalistes à la fois par des simulations et par une application à la Cache County Study, où le regroupement des débuts de la démence au sein des fratries varie fortement avec l'âge.

T. Cai, J. Huang, and L. Tian 394
*Regularized Estimation for the Accelerated Failure Time Model*

En présence de prédicteurs de grandes dimensions, développer des modèles de régression fiables pouvant être utilisés pour prédire des résultats futurs est une tâche difficile. Des complications supplémentaires surgissent quand le résultat d’intérêt est un temps d’événement qui n’est souvent pas complètement observé du fait de la censure. Dans cet article nous développons des modèles de prédiction robustes pour les temps d’événements en régularisant l’estimateur de Gehan pour le modèle à temps accélérés (Tsiatis, 1990) au moyen d’une pénalité LASSO. Contrairement aux méthodes existantes basées sur la pondération par probabilité inverse et l’estimateur de Bukley et James (Buckley et James, 1979), l’approche proposée ne nécessite pas d’hypothèse supplémentaire sur la censure, et fournit toujours une solution convergente. De plus l’estimateur proposé conduit à un modèle de régression stable pour de la prédiction même lorsque le modèle à temps accéléré est en défaut. Afin de faciliter la sélection adaptative du paramètre de réglage, nous détaillons un algorithme numérique efficace pour obtenir le chemin de régularisation complet. Les procédures proposées sont appliquées à des données de cancer du sein afin de déduire un modèle de régression fiable pour prédire la survie des patients à partir d’un ensemble de facteurs pronostiques et de signatures génétiques. Les performances des procédures sur des échantillons finis sont évaluées par une étude par simulation.

Une étude dentaire rétrospective a été réalisée pour évaluer l'influence de l'envahissement pulpaire sur la survie de la dent. A cause du regroupement des dents, les temps de survie au sein de chaque sujet peuvent être corrélés et donc la méthode conventionnelle pour les études cas-témoins ne peut pas être utilisée directement. Dans cet article, nous proposons une approche de modèle marginal pour ce type de données corrélées d'études cas-témoins au sein d'une cohorte. Des équations d'estimations pondérées sont proposées pour l'estimation des paramètres de régression. Différents types de poids sont aussi considérés pour améliorer l'efficacité. Les propriétés asymptotiques des estimateurs proposés sont examinées et leurs propriétés pour des échantillons finis sont établies par des études de simulation. La méthode proposée est appliquée à l'étude dentaire ci-dessus.



Les études longitudinales rassemblent des informations sur l’arrivée d’événements cliniques clefs et sur des caractéristiques spécifiques qui décrivent ces événements. Les variables aléatoires qui mesurent les aspects qualitatifs et quantitatifs associés à un événement sont appelées des marques. Un processus de points marqués récurrents consiste en l’arrivée d’événements possiblement récurrents (avec une possible exposition à un risque) avec une marque mesurée seulement si un événement se produit. Les choix d’analyse dépendent des aspects des données qui sont considérés comme scientifiquement prioritaires. En premier lieu, les facteurs qui influent sur l’arrivée de l’événement dans le temps peuvent être caractérisés par des méthodes d’analyse d’événement récurrent. Ensuite, s’il y a plus d’un événement par sujet, l’association entre l’exposition au facteur de risque et la marque peut être quantifiée par des méthodes de régression de mesures répétées. Nous détaillons les hypothèses requises dans tout processus d’exposition dépendant du temps et d’arrivée des événements dans le temps pour assurer l’obtention d’estimées valides dans des modèles linéaires mixtes simples ou généralisés avec des équations d’estimation généralisée. Nous donnons des preuves théoriques et empiriques que si ces conditions ne sont pas réalisées, alors une équation d’estimation indépendante doit être utilisée pour obtenir une estimation cohérente de l’association. Nous recommandons en conclusion que les analystes examinent soigneusement les processus d’exposition au risque et d’arrivée des événements dans le temps avant d’appliquer une analyse de mesures répétées à des données résultant d’un processus de points marqués récurrents.



Nous proposons une méthode pour tester des différences significatives dans les niveaux d'agrégats entre deux processus ponctuels spatiaux (cas et témoins) prenant en compte les

différences dans leurs intensités de premier ordre. L'avancée clef des méthodes précédentes était que les témoins n'étaient pas supposés être un processus de Poisson. L'inférence et les diagnostics sont basés sur une K-fonction inhomogène avec une enveloppe de confiance obtenue à partir soit des événements ré-échantillonnés par une approche de bootstrap non-paramétrique, soit en simulant de nouveaux événements comme dans un bootstrap paramétrique. Les méthodes développées sont expliquées en utilisant des emplacements d'arbres jeunes et adultes dans une forêt tropicale. Une étude de simulation examine brièvement l'exactitude et la puissance des procédures inférentielles.



Dans le cadre de l’analyse des données de temps jusqu’à événement, on accorde de plus en plus d’attention aux modèles marginaux des processus temporels pour leurs hypothèses plus faibles en comparaison avec les modèles d’intensité traditionnels. Les travaux récents relatifs aux modèles pleinement fonctionnels de régression des processus temporels offrent une grande flexibilité : tous les coefficients de régression peuvent y varier avec le temps, de façon non paramétrique. Néanmoins, la procédure existante d’estimation ne permet pas de réaliser des tests successifs d’adéquation aux données pour juger des coefficients des covariables par comparaison de modèles emboîtés. Cet article propose donc un modèle partiellement fonctionnel de régression pour les processus temporels, dans la droite ligne des modèles marginaux. Les effets de certaines covariables peuvent y être indépendants du temps tandis que la relation au temps d’autres covariables peut n’y être absolument pas spécifiée. Cette classe de modèles est très riche, englobant notamment le modèle pleinement fonctionnel et le modèle semi-paramétrique. Pour en estimer les paramètres, nous proposons des équations d’estimation de profil semi-paramétrique, résolues par algorithme itératif, avec pour valeurs de départ les estimateurs consistants issus du modèle pleinement fonctionnel correspondant. Aucun lissage n’est nécessaire, contrairement au cas d’autres méthodes à coefficients variables. La convergence faible des estimateurs vient de la théorie des processus empiriques. On peut alors procéder à des tests successifs des effets variant avec le temps et à de la sélection de modèle pas-à-pas, par élimination de covariables. L’intérêt pratique d’une telle méthodologie est démontré sur une étude de simulation ainsi que sur un exemple réel portant sur la récurrence des rechutes chez des patients atteints de fibrose cystique.



Cet article approfondit un intérêt récent pour les modèles bayésiens hiérarchiques en génétique quantitative par le développement de modèles de processus spatiaux pour l’inférence sur la variance génétique dominante ou additive dans le contexte des grands essais avec référencement spatiale des données. L’application directe de ces modèles à des grands jeux de données spatiales est cependant numériquement infaisable car

l’estimation implique des algorithmes de matrice cubique. La situation est encore pire dans le contexte des chaînes de Markov Monte Carlo (MCMC) où ces calculs sont effectués pour plusieurs itérations. Ici, nous discutons des approches qui permettent d'éviter ces obstacles sans sacrifier la richesse dans la modélisation. Pour les effets génétiques, nous montrons comment une décomposition spectrale initiale des matrices de relations annule les inversions de matrices coûteuses nécessaires dans les méthodes MCMC proposées précédemment. Pour les effets spatiaux, nous présentons deux méthodes pour contourner les décompositions de matrices de coût prohibitif: la première utilise les résultats analytiques des processus Ornstein-Uhlenbeck qui conduisent à des structures tri-diagonales facilement calculables tandis que la seconde dérive un modèle de processus prédictif modifié à partir du modèle original en projetant ses réalisations sur un sous-espace de moindre dimension, réduisant ainsi la charge calculatoire.
Nous illustrons les méthodes proposées en utilisant un jeu de données synthétique avec des effets génétiques additifs ou dominants et des résidus spatiaux anisotropes, et un grand jeu de données d'une étude de descendance de pin sylvestre (Pinus sylvestris L.) menée dans le nord de la Suède. Nos approches nous permettent de fournir une analyse complète de ce grand essai qui démontre amplement que, en plus de violer les hypothèses de base du modèle linéaire, ignorer les effets spatiaux peut conduire à une sous-estimation des mesures d’héritabilité.



Dans les évaluations de toxicité on a recours à de nombreux moyens de jugement, pour évaluer de façon complète la présence ou la gravité des effets des traitements, on fait appel à échelles de jugement multiples et ordinales sur la sécurité, la douleur ou la progression de la maladie. Les tables de contingences sous-jacentes sont souvent creuses et déséquilibrées, ce qui rend les résultats asymptotiques discutables et les ajustements de modèles d'une complexité prohibitive, sauf à faire des hypothèses exagérément simplificatrices sur les lois marginales.ou conjointe. Plutôt qu'une modélisation, nous considérons un ordre stochastique et une inhomogénéité marginale comme l'expression de l'effet du traitement avec des hypothèses bien plus faibles. On peut souvent regrouper les critères de jugement par ce qu'ils décrivent, domaine physiologique ou groupe de fonctions biologiques. Nous en déduisons des tests basés sur ces sous-groupes qui peuvent compléter ou remplacer les critères individuels grâce à une meilleure puissance. La distribution bootstrap ou de permutation est utilisée intensivement pour obtenir les niveaux de signification soit globalement, soit par sous-groupe ou pour chaque critère individuellement, en incorporant naturellement leur corrélation. Un théorème est fourni qui établit le lien entre l'homogénéité marginale et l'hypothèse plus forte d'échangeabilité de l'approche par permutation.. Les ajustements pour multiplicité des tests des critères individuels sont obtenus par une procédure de désescalade, alors que les niveaux de signification des sous groupe sont ajustés par la procédure de test close. La méthodologie proposé est illustrée en l'appliquant à l'évaluation de la toxicité d'un composé chimique mesurée par un ensemble de 25 critères de jugement corrélés, groupés en 6 domaines.

*Estimating a Multivariate Familial Correlation Using Joint Models for Canonical Correlations: Application to Memory Score Analysis from Familial Hispanic Alzheimer's Disease Study*

L’analyse de caractères multiples peut fournir des informations supplémentaires à l’analyse d’un caractère unique, permettant de mieux comprendre les mécanismes génétiques qui participent à la survenue d’une maladie multifactorielle. Afin de prendre en compte des caractères multiples dans une analyse de corrélation familiale ajustée sur des facteurs de confusion, nous avons développé un modèle de régression pour corrélations canoniques et proposons une modélisation jointe des paramètres de centralité et d’échelle. La méthode proposée est plus puissante que le modèle de régression étant donné qu’il prend en compte l’agrégation familiale des caractères multiples par les corrélations canoniques.



La cartographie de caractère quantitatif dans les organismes expérimentaux est d’une grande importance scientifique et économique. Il y a eu un avancement rapide dans les méthodes statistiques pour la cartographie de caractère quantitatif. Des méthodes variées pour les caractères distribués normalement ont été bien établies. Certaines d’entre elles ont aussi été adaptées pour d’autres types de caractères tels que les caractères binaires, de comptage ou catégoriels. Dans cet article, nous considérons un modèle linéaire généralisé de mélange unifié pour la cartographie d’intervalles multiples dans les croisements expérimentaux. L’approche de cartographie d’intervalles multiples a été proposée par Kao et al. (1999) pour les caractères normalement distribués. Cependant son application à des caractères non normalement distribués a été largement entravée par le manque d’algorithme de calcul efficace et une procédure de cartographie appropriée. Dans cet article, un algorithme EM efficace pour le calcul du modèle linéaire généralisé de mélange et une procédure de cartographie d’intervalles multiples ajustée d’un effet épistasie sont développés. Un jeu de données réelles, les données Radiata Pine, est analysé et la structure des données est utilisée dans les études de simulation pour démontrer les caractéristiques désirables de la méthode développée.



Les modèles de sélection et les modèles par mélanges sont souvent utilisés pour traiter les sorties non ignorables dans les études de longitudinales. Ces deux classes de modèles sont basées sur des factorisations différentes de la distribution jointe du processus étudié et du processus de sortie d’étude. Nous considérons une nouvelle classe de modèles, appelée modèles hybrides à effets mixtes, où la distribution jointe du processus étudié et du processus de sortie d’étude est factorisée au sein de la distribution marginale des effets aléatoires, le processus de sortie d’étude conditionnellement aux effets aléatoires, et le processus étudié conditionnellement aux processus des sorties et aux effets aléatoires. Les

modèles hybrides à effets mixtes combinent les caractéristiques des models de sélections et des modèles par mélanges : ils modélisent directement le processus manquant comme dans les modèles de sélection, et bénéficient de la même facilité de calcul que les modèles par mélanges. Le modèle hybride à effets mixtes fournit une généralisation des modèles à variables partagées en se libérant de l'hypothèse d'indépendance conditionnelle entre le processus de mesure et le processus de sortie sachant les effets aléatoires. Comme les modèles à paramètres partagés sont emboîtés dans les modèles hybrides à effets mixtes, les tests du rapport de vraisemblance peuvent être construit pour évaluer l'hypothèse d'indépendance conditionnelle des modèles à variables partagées. Nous utilisons des données issues d'un essai clinique en SIDA pédiatrique pour illustrer les modèles.



Envisageons une méta-analyse d'études avec des proportions variables de données manquantes pour les niveaux des patients, et supposons que pour chaque étude primaire on ait fait certains ajustements pour des données manquantes afin de rendre valables les estimations utilisées de l'ampleur de l'effet du traitement et de la variance. Ces estimations des effets des traitements peuvent être combinées dans l'ensemble des études par les méthodes standard de la méta-analyse, qui emploient un modèle à effet aléatoire pour tenir compte de l'hétérogénéité entre les études. Toutefois, nous notons qu'une méta-analyse basée un modèle standard d'effets aléatoires va conduire à des estimations biaisées quand les taux d'attrition d'études primaires dépendent de l'ampleur de l'effet traitement de l'étude sous-jacente. Alors qu'on peut vraisemblablement l'ignorer à l'intérieur de chaque étude, ce type de données manquantes ne l'est plus dans une méta-analyse. Nous proposons trois méthodes pour corriger le biais résultant de telles données manquantes dans une méta-analyse: la repondération de l'estimation de DerSimonian-Laird par le taux d'achèvement; l'incorporation du taux d'achèvement dans un modèle à effet aléatoire bayesien ; et une inférence bayesienne fondée sur l'inclusion du taux d'achèvement. Nous illustrons ces méthodes par le biais d'une méta-analyse de 16 essais randomisés publiés qui examinent le traitement combiné de pharmacothérapie et de psychologie de la dépression.



Nous nous intéressons à l'estimation de l'effet causal d'un traitement sur le statut fonctionnel d'individus, à un temps donné *t*, après qu'ils aient vécu un événement catastrophique. Ce travail est réalisé à partir de données observationnelles dont les caractéristiques sont les suivantes : (1) le traitement est prescrit très peu de temps après l'événement et n'est pas randomisé, (2) il est prévu d'interroger les personnes en vie au temps *t*, (3) il y a des non répondeurs parmi les individus interviewés, (4) les facteurs de confusion concernant la période avant l'événement sont en données manquantes pour les individus décédés avant *t*, (5) les informations concernant les facteurs de confusion,

relevés après l'événement et avant le traitement, sont recueillies pour tous les patients dans les dossiers médicaux. En raison du biais lié à la sélection des vivants, nous essayons d'estimer l'effet causal moyen chez les survivants (SACE), l'effet du traitement sur le statut fonctionnel dans une cohorte d'individus qui devrait survivre au temps *t*, qu'ils soient ou non sous traitement. Pour estimer cet effet à partir de données observationnelles, nous devons imposer des hypothèses que l'on ne peut pas vérifier, qui dépendent de l'ensemble des facteurs de confusion. Comme les informations antérieures à l'événement sont manquantes pour les individus décédés avant *t*, il est peu probable que ces données soient manquantes de façon aléatoire (MAR). Nous introduisons une méthode d'analyse de sensibilité pour évaluer la stabilité des inférences SACE face à des écarts à l'hypothèse MAR. Nous appliquons notre méthode à l'évaluation de l'effet d'un centre de soins pour traumatisés sur des critères de vitalité, à partir des données de l'Etude Nationale sur les Coûts et Devenirs des Prises en Charge des Traumatismes.

J. Y. Lin, T. R. Ten Have, and M. R. Elliott 505
*Nested Markov Compliance Class Model in the Presence of Time-Varying Noncompliance*

Dans le contexte du modèle de compliance de Imbens-Rubin (1997), nous nous intéressons à une structure markovienne pour des classes de compliance dépendant du temps et partiellement non-observées. Le contexte est celui d'une étude d'intervention longitudinale randomisée, où les sujets sont randomisés une fois pour le point de base, les valeurs suivies et l'adhésion des patients sont mesurés lors de suivis multiples, l'adhésion des patients pouvant varier avec le temps. Nous proposons un modèle emboîté, à classe latente de compliance, dans lequel nous utilisons des strates principales de compliance invariantes dans le temps et sujet-spécifiques pour résumer les tendances longitudinales de compliance sujet-spécifique et dépendant du temps. Les strates principales sont constituées à l'aide de modèles de Markov reliant le comportement de compliance courant à l'historique de compliance. Les effets traitement sont estimés dans un contexte intention-de-traiter dans la strate principale de compliance.

A. Sjölander, K. Humphreys, S. Vansteelandt, R. Bellocco, and J. Palmgren 514
*Sensitivity Analysis for Principal Stratum Direct Effects, with an Application to a Study of Physical Activity and Coronary Heart Disease*

De nombreuses études cherchent à mettre en évidence l'effet direct de l'exposition à un facteur, c'est-a-dire l'effet non relayé par une variable intermédiaire. Par exemple, dans les études sur les pathologies de la circulation sanguine il peut être intéressant d'évaluer si un niveau adapté d'activité physique peut réduire la morbidité, même si cela n'évite pas l'obésité. Il est connu que la stratification sur la variable intermédiaire peut induire un « biais de sélection post-traitement ». Pour tenir compte de ce problème, nous nous plaçons dans le contexte de la *stratification principale* (Frangakis and Rubin 2002) et nous définissons un estimateur pertinent du lien de causalité, le PSDE (effet direct de la strate principale). Dans le cadre expérimental que nous étudions, le PSDE n'est pas identifié. Nous proposons une méthode d'analyse de sensibilité qui fournit une plage de valeurs plausibles pour le paramètre de causalité. Nous comparons notre approche à des

méthodes similaires proposées dans la littérature pour gérer le problème ressemblant de « censure par décès ».



L'identification de biomarqueurs et autres mesures biologiques pouvant être utilisés comme critères de substitutions dans les essais cliniques a fait récemment l'objet d'une attention particulière. Nous nous plaçons dans la situation d'un seul essai clinique. Dans ce papier, nous considérons un cadre de travail dans lequel nous établissons la substituabilité à partir de la région où le critère de substitution doit être observé avant le véritable critère de jugement. Cela amène à considérer le critère de substitution et le critère final de jugement dans un cadre de risques semi-compétitifs ; cette approche est nouvelle dans la littérature consacrée à la substituabilité et conduit à un traitement asymétrique des deux critères. Cependant, un tel cadre complique conceptuellement beaucoup des mesures précédentes de substituabilité. Nous proposons de nouvelles procédures d'estimation et d'inférence pour l'effet relatif et l'association ajustée proposés par Buyse et Molenberghs (*1998, Biometrics, 1014-1029*). La méthodologie proposée est illustrée sur des données simulées ainsi que sur les données d'un essai dans la leucémie.



A ce jour existent quatre grands cadres d'analyse des critères de substitution dans les essais cliniques. L'un s'appuie sur l'indépendance conditionnelle des variables observables, un autre sur les effets directs et indirects, un troisième s'inscrit dans le champ des méta-analyses, et un quatrième est basé sur la stratification principale. Les deux premiers cadres cités ci-dessus appartiennent au paradigme des effets causaux : l'effet d'un traitement sur un bon critère de substitution combiné à l'effet de ce critère sur le critère clinique doit permettre de prédire l'effet du traitement sur le critère clinique. Les deux autres cadres relèvent du paradigme dit de l'association causale : l'effet du traitement sur le critère de substitution est simplement associé à l'effet du traitement sur le critère clinique. Dans cet article, nous nous penchons d'abord sur le premier paradigme, en examinant sous quelles hypothèses l'identification des paramètres est possible et en étudiant quelques procédures simples d'estimation ; puis nous examinons le second paradigme, celui de l'association causale. Nous étudions les liens entre ces différentes approches ainsi qu'entre les estimateurs qui en sont issus. Une petite étude de simulation nous permet d'illustrer les propriétés des estimateurs selon différents scénarios. Enfin nous concluons par une discussion quant aux cas où chaque paradigme est applicable.

Nous produisons une méthode qui fournit des intervalles de confiance, des estimations ponctuelles et des valeurs de *p* pour le paramètre principal de taille d'effet à l'issue d'un essai clinique séquentiel groupé à 2 bras pour lequel des modifications adaptatives sont intervenues en cours de réalisation. La méthode est basée sur l'application de la procédure de test d'hypothèse adaptatif de Müller et Schäffer (2001) à une suite de tests duaux dérivés de l'intervalle de confiance ajusté par étape de Tsiatis, Rosner et Mehta (1984). Dans un cadre non adaptatif cet intervalle de confiance est connu pour fournir une couverture exacte. Dans un cadre adaptatif une couverture exacte est garantie pourvu que l'adaptation ait lieu à l'avant-dernière étape. D'une façon générale, cependant, tout ce qu'on peut affirmer théoriquement est que la couverture est conservatrice. Néanmoins l'expérimentation par simulation extensive, confortée par la caractérisation empirique de la fonction d'erreur conditionnelle, montre de manière convaincante que dans la pratique la couverture est exacte et l'estimation ponctuelle est non biaisée en médiane. Il n'y a pas actuellement de procédure disponible pour produire, dans un cadre de groupes séquentiels adaptatifs, des intervalles de confiance et des estimations ponctuelles possédant ces propriétés souhaitables. La méthodologie est illustrée par application à un essai clinique sur la stimulation profonde du cerveau dans la maladie de Parkinson.



Un modèle de mélange est un choix naturel pour prendre en compte l'hétérogénéité individuelle dans les études de capture-recapture. Pledger (2000, 2005) a proposé l'utilisation du modèle de mélange à deux points. Dorazio et Royle (2003, 2005) ont suggéré que le modèle beta-binomial avait des avantages. La controverse est liée à la non identifiabilité des la taille de la population (Link 2003) et certains problèmes de borne. Le biais total est décomposé en un biais intrinsèque, une approximation du biais et un biais d'estimation. Nous proposons d'évaluer le biais d'approximation, le biais d'estimation et la variance, avec le biais intrinsèque exclus quand les différents estimateurs sont comparés. Les problèmes de borne dans les deux modèles, et leurs impacts sont étudiés. Des exemples épidémiologiques et écologiques réels sont analysés.



En modélisation écologique de l'habitat d'une espèce, le coût de détermination de l'absence d'une espèce peut être prohibitif. Les données de présence consistent en un échantillon de positions avec des présences observées et un groupe séparé de positions échantillonnées à partir d'un paysage complet, avec des présences inconnues. Nous proposons un algorithme EM pour estimer le modèle logistique sous-jacent de présence-absence pour les données de présence. L'algorithme peut être utilisé avec tout modèle logistique prêt à l'usage. Pour les modèles avec des procédures d'ajustement pas à pas, tels que les arbres « dopés » (boostés), le processus d'ajustement peut être accéléré en intercalant les étapes d'espérance à l'intérieur de la procédure. Des analyses préliminaires basées sur un échantillonnage à partir d'enregistrements de présence absence de poissons dans les rivières de Nouvelle Zélande montrent que cette nouvelle procédure peut réduire

à la fois les estimations de la déviance et du rétrécissement de l'effet marginal qui surviennent dans le modèle naïf souvent utilisé en pratique. Finalement, nous montrons que la prévalence dans la population d'une espèce est seulement identifiable quand il y a des contraintes irréalistes sur la structure du modèle logistique. En pratique, il est fortement recommandé qu'une estimation de la prévalence dans la population soit fournie.



Les méthodes de sélection de variables de type rétrécissement ont récemment fait l'objet de nombreuses applications dans la recherche biomédicale. Cependant, leur performance peut être diminuée par les valeurs aberrantes dans la variable de réponse ou dans les variables explicatives. Ce papier propose une méthode de type Wilcoxon pondérée, basée sur la déviation absolue raccourcie de manière lisse (WW-SCAD), qui traite simultanément le problème de la sélection de variables et de l'estimation de façon robuste. La nouvelle procédure peut être facilement implémentée à l'aide du logiciel statistique R. Nous montrons que la méthode WW-SCAD identifie correctement les coefficients nuls avec une probabilité approchant 1 et estime les coefficients non-nuls au taux $n^{1/2}$. De plus, avec des coefficients de pondération correctement choisis, la méthode WW-SCAD est robuste face aux valeurs aberrantes aussi bien dans la variable de réponse que dans les variables explicatives. Le cas particulier important de coefficients de pondération constants donne un estimateur de type oracle qui se caractérise par une grande efficacité en présence d'erreurs aléatoires à queue lourde. La robustesse de la méthode WW-SCAD est en partie justifiée par sa performance asymptotique en cas de contamination rétrécissante locale. Nous proposons une méthode de sélection des paramètres basée sur le critère BIC. La performance de la méthode WW-SCAD est démontrée à l'aide d'une simulation et d'une application à une étude sur l'effet des caractéristiques individuelles et du régime alimentaire sur le niveau de beta-carotène dans le plasma.

**Biometric Practice**



Nous comparons deux procédures de Monte Carlo, l'échantillonnage séquentiel par importance (SIS) et la procédure de Monte Carlo par chaînes de Markov (MCMC), pour réaliser des inférences bayésiennes concernant des états inconnus et des paramètres de modèles espace-états pour des populations animales. Les procédures sont appliquées aussi bien sur des données simulées et réelles de comptage de jeunes phoques à partir de la méta population britannique des phoques gris, que sur des données simulées de population de saumons quinnats. L'implémentation MCMC est basée sur des distributions fait-main combinées à une intégration analytique de certains états et paramètres. SIS a été implémentée d'une manière plus générique. Pour le même temps de calcul, MCMC tend à

produire des distributions a posteriori avec moins de variation entre les différentes séquences de l'algorithme, comparé à l'implémentation SIS, à l'exception, dans le modèle des phoques, de quelques états et d'un des paramètres qui se mélangent vraiment lentement. L'efficacité de l'échantillonneur SIS s'accroit beaucoup en intégrant analytiquement des paramètres inconnus dans le modèle d'observation. Nous considérons qu'une implémentation attentive de MCMC dans les cas où les données sont informatives relativement aux lois a priori, demeure le gold standard, bien que les échantillonneurs SIS soient une alternative viable qui peuvent être programmés plus rapidement. Notre implémentation SIS est particulièrement compétitive dans des situations où les données sont relativement non informatives; dans d'autres cas, SIS peut nécessiter une puissance de calcul substantiellement plus grande qu'une implémentation efficace de MCMC pour atteindre la même erreur de Monte Carlo.



Les études de jumeaux sont essentielles pour déterminer la composante génétique d’une maladie multifactorielle. Les données générées dans ce cadre sont traditionnellement analysées à l’aide de programmes spécialisés. Pour de nombreux chercheurs, surtout ceux qui travaillent depuis peu dans ce domaine, la compréhension et l’utilisation de ces logiciels représentent un problème. SAS étant l’un des logiciels les plus couramment utilisés, nous proposons son utilisation pour l’analyse d’études de jumeaux. Nous montrons que l’utilisation de SAS permet d’obtenir des résultats similaires à ceux obtenus avec les logiciels spécialisés. Cette validation a une utilité pratique, car elle répond aux questions qui se posent quant à l’utilisation de logiciels « généralistes » pour des schémas expérimentaux particuliers tels que les études de jumeaux et sa capacité à tester une hypothèse particulière. Une étude de simulation nous permet de conclure que les procédures SAS peuvent être utilisées facilement et représentent une alternative à l’utilisation de logiciels plus spécialisés.



Quand on dispose d’une variable auxiliaire *x* suffisamment corrélée à la variable d’intérêt *y*, les estimateurs de la somme des *y* dans une population finie obtenus à partir de ratios peuvent être beaucoup plus efficaces que les estimateurs qui ignorent la variable auxiliaire. Les propriétés bien connues des estimateurs par ratios sont affectées par des erreurs de mesure sur la variable auxiliaire. Nous étudions le cas d’une erreur systématique comme celui d’erreurs distribuées selon une loi spécifiée. Outre les expressions du biais et de la variance de ces estimateurs contaminés par les erreurs de mesure, nous présentons des résultats numériques concernant une population particulière. La présence d’une erreur systématique introduit un biais qui est une fonction non symétrique par rapport à *0* de cette erreur et la précision peut-être augmentée ou diminuée selon l’amplitude de l’erreur. En présence d’erreurs aléatoires, le biais de l’estimateur usuel par le ratio-des-moyennes augmente légèrement avec la variance des erreurs, mais beaucoup moins que le biais de l’estimateur usuel par la moyenne-des-ratios. De même,

la précision de l’estimateur par la moyenne-des-ratios, qui diminue quand la variance des erreurs augmente, est davantage affectée que celle des autres estimateurs étudiés. Globalement, l’estimation par le rapport-des-moyennes apparaît remarquablement robuste en présence d’erreurs de mesure sur la covariable.



Des méthodes pour l'analyse d'études cas-témoin avec une information sur la localisation de la tumeur et la localisation de la dose sont décrites. Celles-ci incluent des méthodes de vraisemblance qui utilisent des cas avec une information précise sur la localisation de la tumeur et d'autres avec une information imprécise. La théorie établit que chacune de ces méthodes basées sur la vraisemblance estime les mêmes paramètres de rapport de radiation dans le contexte du modèle approprié pour prendre en compte les effets des covariables de localisation et dépendantes du sujet. Les suppositions sous-jacentes sont caractérisées et la force potentielle et les limites de chaque méthode sont décrites. Les méthodes sont illustrées et comparées sur l'étude WECARE sur les radiations et le cancer du sein asynchrone controlatéral.



Les courbes de fluorescence sont utilisées pour suivre les changements de l’activité de photosynthèse. Diverses mesures ont été utilisées pour quantifier les différences entre des courbes de fluorescence correspondant à différents traitements, mais ces approches peuvent conduire à des pertes d’information utile. Puisque chaque courbe individuelle de fluorescence est une observation fonctionnelle, il est naturel de choisir un modèle de régression fonctionnelle. Le modèle proposé comprend une composante non paramétrique qui capture la forme générale des courbes et une composante semi-paramétrique qui décrit les différences entre traitements et permet les comparaisons entre traitements. Plusieurs vérifications graphiques de la validité du modèle sont introduites. A la fois des intervalles de confiance asymptotiques approchés et des intervalles de confiance basés sur des simulations sont disponibles. L’analyse d’une expérience de production végétale utilisant le modèle proposé montre que les traits saillants des courbes de fluorescence sont adéquatement capturés. Le modèle de régression fonctionnelle proposé est utile pour l’analyse des données de fluorescence à haut débit obtenues ans le suivi de la croissance des plantes.



Récemment les modèles médians de régression ont suscité une attention croissante. Lorsque les réponses continues suivent une distribution éloignée de la loi normale, les

modèles de régression usuels peuvent faillir pour produire des estimateurs efficaces. En revanche les modèles médians se comportent de façon satisfaisante.
Dans cet article nous discutons de l'utilisation de ce type de modèle pour traiter des données longitudinales avec rupture. Des équations d'estimation pondérées sont proposées pour estimer les paramètres médians de régression pour des données longitudinales incomplètes où les poids sont déterminés par modélisation du processus de rupture. La consistance et la distribution asymptotique des estimateurs résultants sont établies. La méthode proposée est utilisée pour l'analyse d'un jeu de données longitudinales issues d'un essai contrôlé sur le virus HIV (Volberding et al., 1990). Des études de simulation sont conduites pour évaluer la performance de la méthode proposée dans différentes situations. Une extension à l'estimation des paramètres d'association est décrite.



Le choix du modèle et la sélection de variables sont des questions importantes dans l'analyse pratique de la régression. Elle apparaît dans de nombreuses applications biométriques telles les analyses de convenance de l'habitat où le but est d'identifier l'influence de potentiellement beaucoup de conditions environnementales sur certaines espèces. Nous décrivons des modèles de régression pour l'élevage de communauté d'oiseaux qui facilitent à la fois le choix du modèle et la sélection de variables, par un algorithme performant qui fonctionne dans une classe de modèles de régression géo additive comprenant des effets spatiaux, des effets non paramétriques de covariables continues, des surfaces d'interaction, et des coefficients variables.
Tous les termes du modèle de lissage sont représentés comme une somme d'une composante paramétrique et d'une composante de lissage avec un degré de liberté pour obtenir une comparaison équitable entre les termes du modèle. Une représentation générique du modèle géo additif permet d'imaginer un algorithme général performant qui effectue automatiquement le choix du modèle et la sélection des variables.



Nous développons des formules de calcul d'effectif pour des études visant à tester la différence moyenne entre un groupe traitement et un groupe contrôle, dans le cas où l'on s'attend à des biais de sélection, et à ce que certains patients ne se conforment pas au protocole (cette propriété étant mesurée en « tout ou rien »). Sous l'hypothèse d'une absence de biais de sélection, un article récent de Fay, Halloran et Follmann (*Biometrics* 2007 **63**, 465-474) s'est intéressé au problème de l'augmentation de la variance au sein des groupes de traitement (définis selon le traitement attribué), augmentation due à des défauts d'observance. Dans cet article, nous élargissons l'approche de ces auteurs en supposant la possibilité de biais de sélection qui prendraient la forme de différences systématiques de moyennes et de variances entre des sous-groupes latents définis par le traitement assigné et le fait de le prendre ou non. Nous illustrons cette approche en

comparant des calculs d'effectif, selon que les études concernées sont précédées ou non d'études pilotes sur l'observance des patients. Des formules de calcul d'effectif, ainsi que des tests applicables à des variables gaussiennes, sont également développés dans une Annexe disponible sur Internet. Ces calculs tiennent compte de l'incertitude des estimations issues de données pilotes, que celles-ci proviennent d'études externes ou internes.



Nous considérons l'analyse d'essais cliniques qui comprennent une randomisation entre un traitement actif ($T$=1) et un traitement contrôle ($T$=0), quand le traitement actif est sujet à une compliance de type tout ou rien. Nous comparons trois approches pour estimer l'efficacité du traitement dans cette situation : analyse comme effectivement traité, analyse conformément au protocole, et estimation d'une variable instrumentale (IV), où l'effet traitement est estimé en utilisant l'indicatrice de randomisation comme une variable instrumentale. Des estimateurs de IV à la fois basés sur un modèle ou sur la méthode des moments sont étudiés. Les hypothèses sous-tendant ces estimateurs sont évaluées, les écart-types et les carrés moyens des écarts des estimations sont comparées, et les implications au niveau du plan des trois méthodes sont examinés. Des extensions des méthodes pour inclure les covariables observées sont ensuite discutées, mettant en avant le rôle des méthodes de propension à la compliance et le rôle contrastant des covariables dans ces extensions. Les méthodes sont illustrées par des données de l'étude « Women Take Pride », une évaluation des traitements comportementaux pour les femmes avec une maladie cardiaque.



De nombreuses hormones sont sécrétées par bouffées (pulses). Les relations pulsatiles entre hormones assurent la régulation de plusieurs processus biologiques. Pour comprendre le système de régulation endocrine, des séries chronologiques de concentrations hormonales sont recueillies. L'objectif est de caractériser les formes pulsatiles et les associations entre hormones. Habituellement on étudie et on ajuste chaque série hormonale pour chaque sujet, dans un cadre univarié. Ceci conduit à des estimations du nombre de pulses et à des estimations de la quantité d'hormone, la détection des pulses sécrétée ; cependant lorsque le rapport signal-sur-bruit est faible la détection des pulses et l'estimation des paramètres est difficile avec les approches existantes. Dans cet article, nous présentons un modèle bivarié de déconvolution pour de données de pulsations hormonales en nous attachant à la prise en compte des associations de pulses. Par la simulation nous montrons que l'utilisation de l'association entre pulses pour deux hormones améliore l'estimation du nombre de pulses et des autres paramètres relatifs à chaque hormone. Nous montrons comment les modèles de naissance et mort par MCMC peuvent être utilisés pour l'estimation. Nous illustrons par une étude par simulation et sur la relation entre FSH et LH.