Translations of Abstracts

PRESIDENTIAL ADDRESS

Nous évoquons ici la mise en place de la maîtrise de la connaissance statistique, principalement dans les pays en voie de développement. Nous soulignons quelques uns des propos des anciens présidents de l'IBS, et les actions menées, qui montrent le rôle de l'IBS dans l'éducation, avec un accent particulier au sujet des pays en développement. Quelques exemples montrent que les échanges internationaux sont une méthode bien connue maintenant pour améliorer la qualité de l'enseignement, de l'apprentissage, et de la recherche, et nous soulignons le fait que si les initiatives individuelles sont très importantes (et profitables) pour l'éducation, des actions plus générales doivent être encouragées par nous.

Dans mon allocution présidentielle, au Congrès International de Biométrie de Florianopolis en 2010, j'avais présenté la structure modifiée de gouvernance pour la Société Internationale de Biométrie (IBS). Cette structure avait été ensuite aménagée, et la version finale approuvée par les adhérents de la Société en juin 2012. Les adhérents avaient aussi approuvé la fusion de la constitution avec nos règlements, aboutissant à l'abrogation de l'ancienne constitution (qui n'était d'ailleurs plus cohérente avec la pratique). A partir du 1er janvier 2013, la responsabilité de la gouvernance et de la direction de l'IBS repose sur un Comité Exécutif de 15 membres, qui sera assisté par un Conseil Représentatif élargi dont les membres sont sélectionnés par et à partir de chacune des sociétés régionales. Le Conseil représentatif assure la supervision de la formation du Comité Exécutif, donne des avis sur les questions stratégiques et politiques, et contribue au fonctionnement de l'IBS. Ce Conseil sera un relais efficace entre les régions et le Comité Exécutif, aidé par la Présidence qui est présente à toutes les réunions du Comité.

BIOMETRIC METHODOLOGY

En recherche vaccinale, les marqueurs immunologiques pouvant prédire de façon fiable l'effet de la vaccination sur la réponse clinique (i.e., marqueurs de substitution) sont des outils importants pour guider le développement d'un vaccin. Ce papier s'intéresse à

l'optimisation du plan d'échantillonnage en deux phases d'études visant à évaluer les marqueurs de substitution, motivée par la conception d'un futur essai vaccinal contre le VIH. Pour traiter le problème de manque de résultats potentiels dans un essai standard, de nouveaux essais utilisant des prédicteurs à l'inclusion des réponses de biomarqueurs immunitaires et/ou augmentant l'essai en vaccinant des sujets non infectés recevant un placebo à la fin de l'essai et en mesurant leurs biomarqueurs immunitaires ont été proposés. Toutefois, l'utilisation inefficace de l'information augmentée peut conduire à des résultats contre-intuitifs sur la précision de l'estimation. Pour remédier à ce problème, nous proposons un estimateur de type pseudo-score approprié pour un schéma augmenté et caractérisons ses propriétés asymptotiques. Cet estimateur a des performances supérieures par rapport aux estimateurs existants et permet de calculer les variances analytiques très utiles pour la conception d'études. Sur la base du nouvel estimateur, nous étudions en détail le problème d'optimisation du plan d'échantillonnage d'un biomarqueur dans un essai d'efficacité vaccinale pour estimer efficacement son effet de substitution, telle qu'elle est caractérisée par la courbe d'efficacité vaccinale (une courbe de prédictibilité de l'effet causal) et par l'efficacité globale prédite du vaccin en utilisant le biomarqueur.

L'effet naturel direct (e.g. intrinsèque), ou l'effet d'une exposition sur une réponse si une variable intermédiaire avait été réglé à son niveau en l'absence de l'exposition, est souvent intéressant pour les chercheurs. En général, le paramètre statistique associée à l'effet intrinsèque est difficile à estimer non-paramétriquement, notamment lorsque la variable intermédiaire est continue ou de grande dimension.
Dans cet article, nous introduisons un nouveau paramètre causal appelé effet naturel direct parmi les non traités, les hypothèses d'identifiabilité sont discutées, une analyse de sensibilité pour certaines des hypothèses est proposée et l'on établit que ce nouveau paramètre est équivalente à l'effet intrinsèque dans un essai randomisé contrôlé. Nous présentons également un estimateur par minimisation de perte ciblée (TMLE), un estimateur doublement robuste et localement efficace pour le paramètre statistique associée à ce paramètre de causalité. De plus, nous introduisons trois quantités causale voisine l'effet intrinsèque chez les traités, l'effet indirect chez les non-traités et l'effet indirect chez les traités

L'évaluation des traitements médicaux se voit fréquemment parasitée par la présence d'importants effets placebo, en particulier quand les critères sont relativement subjectifs ; la solution classique à ce problème est d'effectuer des essais cliniques randomisés en double-aveugle, contrôlés par un bras placebo. Cependant, le masquage du traitement ne garantit pas que tous les patients aient les mêmes convictions ou mentalisations quant à la nature du traitement qu'ils ont reçu (ce qu'on pourrait appeler, par jeu de mots, la « traitementalité »). Cela complique quelque peu l'interprétation de la

traditionnelle évaluation « en intention de traitement » comme un effet purement induit par le traitement. Nous discutons les relations causales entre le traitement, sa mentalisation (la « traitementalité ») et la réponse clinique, et nous proposons un modèle causal pour l'évaluation conjointe de l'effet du placebo et de l'effet spécifique du traitement. Ce modèle souligne à quel point il est important de mesurer et de prendre en compte cette « traitementalité » ; pour bien faire, il faudrait considérer les groupes de traitement comme des études observationnelles séparées au sein desquelles on userait de la « traitementalité » comme d'un facteur d'exposition non contrôlé. Cette perspective nous permet d'adapter les méthodes existantes de traitement des facteurs confondants à ce problème de l'évaluation conjointe de l'effet du placebo et de l'effet spécifique du traitement, en utilisant les mesures de « traitementalité » (on demande aux patients quel traitement ils pensent avoir reçu, recueillant ainsi ce que l'on appelle les « données d'évaluation de l'aveugle »). Nous employons successivement une évaluation catégorielle courante de cette « traitementalité » – illustrée par un exemple dans l'asthme –, puis une évaluation continue – notamment par le biais d'une probabilité subjective –, pour laquelle nous décrivons les méthodes analytiques associées.

Dans les essais cliniques contre placebo randomisés mesurant l'efficacité d'un vaccin préventif contre le VIH, un objectif est d'évaluer la relation entre l'efficacité du vaccin à prévenir l'infection et les écarts génétiques entre les souches d'exposition aux VIH et les multiples séquences du VIH incluses dans la construction vaccinale. L'ensemble des écarts génétiques est considéré comme la « marque » observée multivariée continue seulement chez les sujets infectés. Ce travail développe un modèle multivarié à risques proportionnels marque-spécifique dans le cadre des risques compétitifs en analyse de survie pour l'évaluation de l'efficacité d'un vaccin marque-spécifique. Cela permet d'améliorer l'efficacité de l'estimation en utilisant la méthode semi-paramétrique du maximum de « vraisemblance profil » dans le modèle à rapport de densités vaccin contre placebo à marque. Le modèle permet également l'utilisation d'une méthode d'estimation plus efficace pour le logarithme du risque relatif total dans le modèle de Cox. De plus, nous proposons de tester des procédures pour évaluer deux hypothèses pertinentes relatives à l'efficacité des vaccins marque-spécifiques. Les propriétés asymptotiques ainsi que la performance des procédures d'inférences sur de petits échantillons ont été étudiés. Finalement, nous appliquons les méthodes proposées aux données recueillies dans l'essai thaïlandais RV144 mesurant l'efficacité d'un vaccin anti-VIH.

Nous proposons des méthodes semi-paramétriques pour estimer l'effet d'une covariable dépendante du temps sur la survie sans traitement. La structure des données d'intérêt consiste en une séquence longitudinale de mesures et un délai de survie potentiellement censuré. Le facteur d'intérêt dépend du temps. La prise d'un traitement est considérée comme une censure dépendante pour la survie sans traitement. Les patients peuvent être

exclus selon des considérations liées au traitement, de manière temporaire ou définitive. Les méthodes proposées combinent une analyse « landmark » et une régression à risque partiellement conditionnel. Une série d'intervalles de temps est spécifiée, et la survie (depuis la date du début de l'intervalle) est modélisée à l'aide d'une régression de Cox pondérée. Le modèle théorique pour le décès est marginal dans le sens où les covariables dépendantes du temps sont prises comme fixes à chaque repère, avec la fonction de risque de mortalité implicitement moyennée à travers les futures trajectoires des covariables. La censure dépendante est surmontée par une variante de la méthode de probabilité inverse de censure pondérée (IPCW). Les estimateurs proposés sont convergents et asymptotiquement normaux, avec des estimateurs de la covariance convergents. Des études de simulation montrent que les procédures d'estimation proposées sont appropriées à une utilisation pratique. Nous appliquons les méthodes proposées à la mortalité pré-transplantation parmi des patients atteints de maladies du foie en phase terminale (ESLD).

Pour explorer les interactions non-linéaires entre une variable indicatrice et des covariables, des modèles de hasards proportionnels partiellement linéaires ont été proposés pour des données de survie censurées. Cependant la spécification de la structure partiellement linéaire était usuellement assurée de manière ad-hoc en faisant tout d'abord l'ajustement d'un modèle complet à coefficients variables, puis en examinant visuellement les résultats de l'ajustement pour identifier la partie linéaire. Dans cet article, nous examinons le problème de l'estimation de coefficients et de l'identification de coefficients constants en nous basant sur une approche de double contraction. La sélection de variables est aussi considérée dans un cadre cohérent d'estimation, aboutissant à une procédure de double pénalisation. Sous des hypothèses modérées, nous établissons des propriétés asymptotiques de cette procédure, telles que la consistence, la stabilité au travers de la « sparesistency » (la probabilité que soient nuls tous les paramètres identifiés comme égaux à zéro tend vers 1), la « constansistency » (la probabilité que les paramètres constants identifiés comme non nuls soient constants non-nuls tend vers 1), et la normalité asymptotique. Nous évaluons la performance de la méthode proposée par des simulations numériques, et nous illustrons son application sur un ensemble de données du cancer du sein.

Le modèle de vie accélérée classique a été largement étudié en analyse de survie du fait de l'interprétation directe des effets des covariables sur le temps de survie moyen. Cependant, ce modèle de vie accélérée classique et les méthodologies associées sont construits sur l'hypothèse fondamentale d'homoscédasticité des données. En conséquence, quand l'hypothèse d'homoscédasticité n'est pas vérifiée comme souvent dans les applications réelles, les estimateurs perdent en efficacité et l'inférence associée n'est pas fiable. De plus, aucune des méthodes existantes ne fournit un estimateur consistant de la constante. Pour remédier à ses inconvénients, nous proposons dans ce

papier une approche semiparamétrique pour des données homoscédastiques comme hétéroscédastiques. Cette approche utilise une équation des moindres carrés pondérés avec des observations de synthèse pondérées par la racine carrée de leur variance où les variances sont estimées par régression polynomiale locale. Nous établissons les distributions limites des estimateurs des coefficients qui en résultent et montrons qu'à la fois les paramètres de pente et la constante ont des estimateurs consistants. Nous évaluons la performance de l'approche proposée en échantillon fini à l'aide d'études de simulations et démontrons sa supériorité sur les méthodes existantes quand les données sont hétéroscédastiques par le biais d'un exemple réel sur son efficacité et sa fiabilité.

Dans les études cliniques et enquêtes longitudinales, l'événement d'intérêt peut souvent être récurrent pour le même individu. S'il existe aussi un événement terminal, par exemple le décès, qui met fin au processus d'événements récurrents, la situation est plus complexe. L'événement terminal est alors souvent fortement corrélé avec le processus des événements récurrents. Pour analyser ce type de données, nous modélisons la dépendance entre événements récurrents et terminal par une fragilité partagée de type Gamma, incluse simultanément dans les fonctions de risque des événements récurrents et de l'événement terminal. Conditionnellement à cette fragilité, un modèle est spécifié uniquement pour le processus marginal d'événements récurrents, ce qui nous libère de la condition forte de processus de Poisson. L'analyse est basée sur des fonctions d'estimation, qui permettent d'estimer les effets de covariables sur les événements récurrents et l'événement terminal. Il est aussi possible d('estimer le degré d'association entre les deux processus. Des estimateurs asymptotiques de la variance sont proposés sous forme analytique. Nous évaluons par des simulations l'applicabilité des résultats asymptotiques à des échantillons finis, ainsi que la sensibilité de la méthode par rapport aux hypothèses sous-jacentes. Les méthodes présentées peuvent être directement généralisées au cas d'événements terminaux ou récurrents multiples. Enfin, nous illustrons nos méthodes sur des données d'hospitalisation de patients sous dialyse dans une étude internationale multi-centres.

L'évaluation des effets de covariables sur le temps écoulé entre des événements récurrents consécutifs est d'un intérêt majeur dans de nombreuses études médicales et de santé publique. Alors que les méthodes actuelles pour l'analyse d'événements récurrents s'intéressent à la modélisation de la fonction de risque instantané, une interprétation directe des effets des covariables sur les temps de récidive entre évènements consécutifs n'est pas disponible par ces méthodes. Dans cet article, nous considérons une régression par quantile qui peut fournir une évaluation directe des effets des covariables sur la valeur des quantiles de la distribution des temps de récidive. Dans l'esprit de la méthode des ensembles de risques ('risk-set') pondérés proposée par Luo et Huang (2011, Statistics in Medicine 30, 301-311), nous étendons la méthode d'estimation par équation basé sur des martingales considérée par Peng et Huang (2008, Journal of the American Statistical Association 103, 637-649) pour des données de survie uni variées permettant l'analyse des temps entre des évènements récurrents consécutifs. La procédure d'estimation

proposée peut être facilement implémentée dans les logiciels existants pour des régressions par quantile censurée uni variées. La consistance uniforme et la convergence faible des estimateurs proposés sont établies. Des études de Monte-Carlo démontrent l'efficacité de la méthode proposée. Une application aux données du Registre Psychiatrique Central Danois est présentée pour illustrer les méthodes développées dans cet article.

Dans les études observationnelles, l'estimation se fait souvent en utilisant des données longitudinales ainsi l'intérêt réside en particulier dans l'estimation de la relation entre les variables explicatives et les variables dépendantes au niveau de la population. Les données longitudinales comportent souvent des biais dus à l'échantillonnage et au caractère non aléatoire des sorties d'étude. Toutefois, l'inclusion d'information au niveau de la population peut améliorer l'efficacité des estimateurs. Dans cet article, nous considérons un modèle partiel linéaire généralisé pour données longitudinales incomplètes en présence d'information au niveau populationnel. Nous présentons une méthode basée sur la vraisemblance pseudo-empirique pour intégrer des informations au niveau de la population. Nous corrigeons les biais de sortie d'étude non aléatoire en utilisant une méthode d'estimation par équations généralisées pondérées. Une procédure d'estimation en trois étapes est proposée ce qui rend les calculs plus faciles. Plusieurs méthodes, souvent utilisés dans la pratique, sont comparées dans des études de simulation. Nous démontrons que notre méthode permet de corriger les biais liés au caractère non aléatoire des sorties d'études et d'augmenter l'efficacité d'estimation, en particulier pour les échantillons de petite taille ou lorsque la proportion de perdus de vue est élevée. Nous appliquons cette méthode à l'étude de la maladie d'Alzheimer.

La reconstruction des gradients de données fonctionnelles observées de manière clairsemée est un problème inverse difficile. A partir d'observations de courbes lisse (e.g. des courbes de croissance) à des instants isolés, le but est de proposer des estimateurs des gradients sous-jacents (ou vitesses de croissance). Pour ce faire, nous développons une approche d'inversion bayésienne qui modélise les trous entre les observations par un mouvement brownien contraint, conditionnellement aux points observés. La moyenne et le noyau de covariance a posteriori des vitesses de croissance sont alors explicites et représentables sous formes de splines quadratiques qui peuvent être évalués. Les hyperparamètres de la loi apriori sont spécifiés par une approche bayésienne non paramétrique empirique avec pour la matrice de précision aux dates d'observations une minimisation contrainte l_1. La variance infinitésimale a priori du mouvement brownien est choisie par validation croisée. Cette approche est illustrée à la fois sur des données simulées et des données réelles.

Dans un travail motivé par les études d'association de polymorphisme nucléotidique simple (SNP) avec l'incertitude sur le génotype, nous proposons une généralisation u test de Kruskall-Wallis qui incorpore une incertitude sur le groupe dans la comparaison de k échantillons. La statistique de test étendue est basée sur une somme de rangs pondérés par des probabilités, et qui suit asymptotiquement, sous l'hypothèse nulle, une distribution de khi-deux à $k - 1$ degrés de liberté. Des études de simulation confirment la validité et la robustesse du test proposé sur des échantillons de taille finie. Une application à une étude d'association génomique des complications du diabète de type 1, démontre ensuite l'utilité de ce test de Kruskal-Wallis généralisé pour des études avec incertitude de groupe. Le modèle a été implémenté en un programme R, ressource d'accès libre, GKW.

La sélection de variables caractéristiques est devenue un problème important dans la littérature statistique récente. Dans les applications, quelquefois, plusieurs variables sont introduites pour réduire des possibles biais de modélisation, mais le nombre de variables qu'un modèle peut incorporer est souvent limité par la quantité de donnée disponible. En d'autres mots, le nombre de variables considéré dépend de la taille d'échantillon, qui reflète la possibilité d'estimation du modèle paramétrique. Dans cet article, nous envisageons le problème de la sélection de variables caractéristiques dans un mélange fini de modèles de régression lorsque le nombre de paramètres du modèle peut augmenter avec la taille d'échantillon. Nous proposons une approche par vraisemblance pénalisée pour sélectionner des variables caractéristiques dans ces modèles. Sous certaines conditions de régularité, notre approche conduit à une sélection consistante de variables. Nous avons mené de nombreuses études par simulation pour évaluer la performance de l'approche proposée pour divers schémas prédéfinis. Nous appliquons aussi cette méthode à deux ensembles de données réelles. Tout d'abord sur le télémonitoring de la maladie de Parkinson (PD), lorsque le problème est de savoir si certaines caractéristiques de dysphonie, extraites d'enregistrements de signaux vocaux des patients faits à domicile, peuvent être utilisés comme indicateurs de la sévérité et de la progression de la maladie de Parkinson. Dans le second cas, il s'agit du pronostic du cancer du sein, pour lequel nous sommes intéressés à connaître quelles caractéristiques des noyaux cellulaires peuvent donner des valeurs de pronostic à long terme sur la survie des patients ayant un cancer du sein. Notre analyse, dans chacune de ces applications, montre une structure de mélange dans la population étudiée et qui ne traduit pas une relation unique entre les caractéristiques et la variable de réponse dans chacune des composantes du mélange.

Nous décrivons dans cet article une technique de statistique bayésienne afin de (a) réaliser une sélection conjointe parcimonieuse de variables explicatives et de coefficients de la matrice inverse de covariance dans le cadre d'un système de régressions apparemment non liées dans un espace linéaire gaussien de haute dimension et (b)

réaliser une analyse d'association entre l'ensemble des variables explicatives et à expliquer dans un tel cadre. Afin de parcourir l'espace de haute dimension du modèle, dans lequel autant le nombre de variables explicatives que celui des variables à expliquer peut être plus grand que la taille de l'échantillon, nous montrons qu'un échantillonneur de Gibbs basé sur la marginalisation, associé à des a priori de type « *spike and slab* » (« pic sur pavé »), est une solution raisonnable en terme de temps de calcul et efficace. A titre d'exemple, nous appliquons notre méthode à l'analyse de *loci* associés à des traits quantitatif d'expression (eQTL) sur des données publiques humaines de polymorphismes d'un seul nucléotide (SNP) et d'expression de gènes qui ont pour but premier de trouver des associations significatives entre des ensembles de SNPs et des transcrits qui sont probablement corrélés entre eux. Notre méthode permet également l'inférence d'un réseau d'interactions entre transcrits (les variables à expliquer) après avoir pris en compte l'effet des SNPs (les variables explicatives). Nous exploitons pour cela des propriétés issues de la théorie des modèles graphiques gaussiens, propriétés qui nous servent à poser des hypothèses sur les indépendances conditionnelles entre les variables à expliquer. Notre méthode se mesure favorablement à d'autres méthodes bayésiennes développées dans des buts similaires aux nôtres.

Nous nous intéressons à l'inférence bayésienne pour les modèles statistiques dépendant de l'évaluation d'un programme ou d'une simulation coûteux en calcul. Dans de telles situations, le nombre d'évaluations de la fonction de vraisemblance, et par conséquent de la fonction non normalisée de densité a posteriori, est déterminé par la ressource de calcul disponible et peut être limité de ce fait. Nous présentons un nouvel exemple d'un tel simulateur décrivant les propriétés de cellules souche humaines embryonnaires utilisant des données d'expériences par piégeage optique. Cette application est utilisée pour motiver une stratégie nouvelle d'inférence bayésienne exploitant une approximation par processus gaussien du simulateur, et permet une inférence MCMC efficiente du point de vue des calculs. Les avantages de cette stratégie sur les précédentes méthodologies sont d'une part une moindre dépendance à la détermination des paramètres de réglage, et d'autre part qu'elle permet l'application de procédures de diagnostic de modèles ne nécessitant pas d'évaluations additionnelles du simulateur. Nous montrons les avantages de notre méthode sur des exemples synthétiques, et présentons son application sur des expériences sur des cellules souche.

Dans cet article, nous considérons un échantillonnage en deux phases dans lequel toutes les covariables sont catégorielles. Les plans en deux phases sont attirants d'un point de vue de l'efficacité, car s'ils sont mis en place soigneusement, ils permettent de se concentrer sur les cellules informatives. Un certain nombre de méthodes basées sur la vraisemblance ont été développées pour l'analyse des données en deux phases, mais nous décrivons ici une approche bayésienne qui était précédemment non disponible. Les méthodes sont d'abord comparées avec des approches existantes à travers une étude de

simulation puis appliquées à des données collectées sur la tumeur de Wilms. Les bénéfices d'une approche bayésienne incluent de relâcher la confiance dans l'inférence asymptotique, particulièrement pour des données dispersées et la capacité à modéliser des données ayant des dépendances complexes, par exemple en introduisant des effets aléatoires. La situation de données dispersées est illustrée via un exemple simulé.

Davide Altomare, Guido Consonni, and Luca La Rocca **478**
Objective Bayesian Search of Gaussian Directed Acyclic Graphical Models for Ordered Variables with Non-Local Priors

Les modèles décrits par des graphes acycliques orientés (DAG) sont de plus en plus utilisés dans l'étude des systèmes physiques ou biologiques; ils permettent de modéliser des influences directes entre variables. Pouvoir identifier le graphe à partir des données constitue un challenge qui peut être raisonnablement abordé si les variables sont supposées satisfaire un ordre donné; dans ce cas, il faut simplement estimer la présence/absence de chaque liaison potentielle. Travaillant sous cette hypothèse, nous proposons pour la recherche de l'espace des modèles décrits par un DAG gaussien une approche bayésienne objective qui fournit un output riche pour un input minimal. Nous faisons reposer notre analyse sur des a priori non-locaux qui conviennent tout particulièrement à des graphes clairsemés; ils permettent en effet d'avoir une vitesse d'apprentissage supérieure à celle observée avec des paramètres locaux ordinaires, et ce quand la (vraie) distribution d'échantillonnage appartient à un modèle simple. Nous mettons en oeuvre un algorithme stochastique efficace de recherche qui permet de traiter de façon effective des jeux de données ayant une taille inférieure à celle du nombre de variables, et appliquons notre méthode à une grande variété de jeux de données réelles et simulées. Si l'on compare notre approche à l'approche fréquentiste standard actuelle (qui repose sur une hypothèse de variables ordonnées), celle-ci se comporte bien si la comparaison est faite en termes de courbe ROC du taux de vraies liaisons contre le taux de fausses alertes. Sous cette hypothèse, celle-ci présente un avantage sur l'algorithme PC, qui peut-être considéré comme une référence de l'approche fréquentiste pour des variables ordonnées. Notre approche est également avantageuse s'agissant de l'apprentrissage du squelette du DAG, quand l'ordre des variables est modérément mal-spécifié. Notre méthode pourra à l'avenir être couplée avec une stratégie d'apprentissage de l'ordre des variables, en renonçant ainsi à l'hypothèse d'un ordre connu.

J. Besag and D. Mondal **488**
Exact Goodness-of-Fit Tests for Markov Chains

Les tests d'ajustement sont utiles pour mesurer si un modèle statistique est consistant avec les données disponibles. Cependant, les propriétés asymptotiques usuelles du chi-deux ne sont pas vérifiées, soit par manque de données soit parce qu'un test statistique standard est intéressant. Dans cet article nous décrivons des tests d'ajustement exacts pour des chaînes de Markov d'ordre un ou d'ordre plus élevé, avec une attention particulière pour les chaînes réversibles. Les tests sont obtenus en conditionnant sur les statistiques exhaustives pour les probabilités de transitions et sont implémentés par échantillonnage de Monte Carlo ou par des méthodes MCMC. Ils s'appliquent à la fois à

des séquences simples et des séquences multiples et permettent un libre choix de la statistique de test. Trois exemples sont donnés. Le premier concerne des séquences multiples de jours secs ou pluvieux en janvier pour les années de 1948 à 1983 à Snoqualmie Falls, état de Washington, et suggère que l'analyse standard peut être trompeuse. Le second exemple est pour une séquence ADN à quatre états et sert de support à la conclusion originale qu'une chaîne de Markov de second ordre produit un ajustement adéquat aux données. Le dernier exemple concerne des données atomiques à six états dans une simulation de la dynamique moléculaire de la variabilité conformationnelle de dipeptides et souligne l'existence de preuves solides contre une chaîne de Markov réversible d'ordre un avec des pas de temps de 6 picosecondes.

Pour tester la stationnarité d'une distribution spatiale de points, Guan (2008) a proposé un modèle statistique libre, fondé sur les écarts entre les effectifs observés et attendus de points dans l'expansion des régions à l'intérieur de la fenêtre d'échantillonnage. Cet article étend sa méthode à une classe générale de statistiques en intégrant également cette information quand les points sont projetés sur les axes et par différents moyens permettant de construire les régions dans lesquelles les écarts sont considérés. Les distributions limites des nouvelles statistiques peuvent être exprimées en termes d'intégrales d'une feuille brownienne et donc les valeurs critiques asymptotiques peuvent être estimées. Une étude de simulation montre que les nouveaux tests sont toujours plus puissants que ceux de Guan. Lorsqu'on les applique aux données de pin des marais où le test de Guan n'a pas donné de réponse concluante, les nouveaux tests indiquent un rejet clair de l'hypothèse de stationnarité.

La régression logistique est un outil populaire pour l'analyse de risque tant dans le domaine des sciences médicales que dans celui des sciences de la santé. Avec des données dont la réponse est continue, il est habituel de créer une variable dichotomique pour spécifier un seuil pour le caractère positif. En supposant un modèle d'échantillonnage logistique des données de l'ajustement d'une régression linéaire à une réponse non dichotomique on a montré de manière empirique qu'on obtient des estimations plus efficaces des rapports de chance que par
une régression logistique habituelle des points extrêmes dichotomiques. Nous illustrons que l'inférence sur le risque n'est pas robuste pour des écarts à la distribution logistique paramétrique. De plus, la supposition de proportionnalité des rapports des chances n'est pas satisfaite si la condition de distribution logistique est violée sur les données, conduisant à une inférence biaisée à partir de l'analyse paramétrique de la logistique. Nous développons une nouvelle méthodologie bayésienne semi-paramétrique pour faire le test de bon ajustement de la régression logistique paramétrique avec des données continues. Les procédures de test sont conservées pour tout seuil limite et notre approche

fournit simultanément la possibilité de réaliser une analyse semi-paramétrique du risque. Les facteurs de Bayes sont calculés avec le rapport de Savage-Dickey pour faire le test de l'hypothèse nulle de la régression logistique contre une généralisation semi-paramétrique. Nous proposons une approche entièrement bayésienne et une approche empirique efficace pour les calculs pour le test, et nous proposons des méthodes d'estimation semi-paramétrique des risques, des risques relatifs et des rapports des chances quand la régression logistique paramétrique ne marche pas. Des résultats théoriques établissent la consistance du test de Bayes empirique. Des résultats obtenus à partir de données simulées montrent que l'approche proposée fournit une inférence précise indépendamment du fait que les suppositions sur le caractère paramétrique soient ou non satisfaites. Une évaluation des facteurs de risque d'obésité montre que différentes inférences sont obtenues à partir d'une analyse d'un ensemble de données réelles quand des écarts à une distribution logistique sont permis dans un cadre semi-paramétrique flexible.

La méthode des séries de cas est utilisée pour étudier la relation entre des temps d'expositions, par exemple à des infections, et des événements aigus observés sur des individus au cours de périodes définies. Elle fournit des estimations de l'incidence relative des événements en périodes à risque (par exemple, 30-jours après infection) par rapport aux périodes basales. Lorsque les temps de début d'exposition ne sont pas connus avec précision, l'application d'un modèle de séries de cas en ignorant l'erreur liée à la mesure de début d'exposition conduit à des estimations biaisées. La correction du biais est nécessaire pour comprendre la direction et la taille des effets vrais associés à la période d'exposition, bien que les estimateurs non corrigés aient une variance plus faible. Ainsi, l'inférence via des tests d'hypothèses basés sur des statistiques de test non corrigées, si elle est valide, a potentiellement plus de puissance. En outre, les tests peuvent être mis en œuvre à l'aide d'un logiciel standard et ne nécessitent pas d'autres données auxiliaires. Dans ce travail, nous examinons la validité et la puissance des tests d'hypothèses naïfs en utilisant des analyses de séries de cas sur des données imprécises sans correction de l'erreur. Sur la base d'études de simulation et de calculs théoriques, on détermine la validité et la puissance relative des tests d'hypothèses d'intérêt dans l'analyse des séries de cas. En particulier, on montre que les tests de l'hypothèse nulle globale et de l'ensemble des hypothèses nulles associées à toutes les périodes à risque ou à tous les effets de l'âge sont valides. Cependant, les tests sur les paramètres des périodes à risque individuelles ne sont généralement pas valides. Un guide pratique est fourni et illustré par des données provenant de données de patients sous dialyse.

Nous reprenons la question de la détermination de la taille d'échantillon (SSD) dans le cas où l'on teste le logarithme de l'odds-ratio (OR) vis-à-vis de la valeur nulle pour deux

binomiales indépendantes. Quatre approches classiques sont considérées : une formule de taille d'échantillon sous forme fermée basée sur le test de Wald (nW), des formules fermées par le test du score et un test exact (nS et nE), et une approche numérique de calcul de la taille d'échantillon basée sur les tests du rapport des vraisemblance (nL). Plusieurs remarques utiles en pratique sont présentées. D'abord, nW est une fonction strictement convexe de l'OR pour un OR > 1 et pour un OR < 1 respectivement, ce qui implique que la taille d'échantillon calculée par la méthode nW ne décroît pas nécessairement lorsque l'OR s'éloigne de 1. Cependant une taille d'échantillon (SS) minimale est souvent obtenue pour des valeurs de l'OR considérées comme des valeurs relativement extrêmes et rarement observées en réalité. Les quantités nS, nE et nL sont monotones décroissantes lorsque l'OR s'écarte de 1. Deuxièmement le ratio d'échantillonnage optimal (OSR) entre deux binomiales indépendantes qui donne la puissance maximale pour une taille d'échantillon (SS) totale fixée n'est pas toujours égal à 1 : 1 mais dépend des chances de l'événement dans chaque bras. La méthode nW est celle qui bénéficie le plus de l'application de l'OSR, dans la mesure où la taille d'échantillon totale (SS) peut être significativement réduite en comparant au rapport 1 : 1 classique. D'un point de vue pratique, les avantages obtenus par OSR pour cette taille totale par les méthodes nS, nL et nE sont peu appréciables. Enfin, nous développons des études par simulation pour étudier la puissance (du oint de vue de sa fidélité) de chaque approche, et nous envisageons la vraisemblance pénalisée comme remède à ce défaut.

Le modèle avec retards distribués (DLM), utilisé principalement dans les études sur la pollution de l'air, trouve son utilisation partout où l'effet d'une covariable est décalée et distribuée au cours du temps.
Nous spécifions des formulations modifiées des DLM pour fournir un calcul avantageux, des coefficients des modèles souples qui sont applicables dans tous les contexte où les covariables décalées sont régressées sur une réponse dépendante du temps. Nous étudions l'utilisation de ces modèles à la pluviométrie et au débit fluvial et en particulier leur rôle dans la compréhension de l'impact des variables cachées dans les systèmes fluviaux. Nous appliquons deux modèles aux données d'une rivière montagnarde Ecossaise et nous utilisons des données simulées afin de vérifier l'efficacité de notre approche. Dans des conditions de fortes précipitations, les variations de leur influence sur les flux proviennent d'une interaction complexe entre l'humidité antérieure du sol et un délai temporel des précipitations. Les modèles identifient des changements subtils dans la réactivité aux précipitations, en particulier dans la position du pic d'influence dans la structure des délais.

Les observations spatiales binaires sont courantes en épidémiologie et en santé publique. La régression logistique est un modèle naturel pour ces données en raison de l'élégante interprétation de ses coefficients en termes de log-odds. Afin de prendre en compte l'existence de variables de confusion non observées qui pourraient présenter une structure

spatiale (comme des caractéristiques socioéconomiques, biologiques ou environnementales), il est d'usage d'inclure des effets aléatoires spatiaux à distribution gaussienne. Conditionnellement à ces effets aléatoires, les coefficients représentent des log d'odds ratio. Cependant, les coefficients marginaux par rapport aux effets aléatoires ne conservent pas cette interprétation et leurs estimations sont difficiles à interpréter ou à généraliser à d'autres régions. Pour résoudre ce problème, nous proposons une nouvelle distribution des effets aléatoires spatiaux, dans le cadre des copules, qui garantit que l'interprétation en termes de log-odds reste valide conditionnellement et marginalement par rapport aux effets aléatoires spatiaux. Nous évaluons la robustesse de notre approche vis-à-vis de diverses distributions des effets aléatoires à l'aide de simulations ; nous l'appliquons à un intéressant jeu de données sur la santé parodontale d'afro-américains parlant le créole Gullah. La souplesse de cette méthodologie permet de traiter des données territoriales ou géo-statistiques ainsi que des modèles hiérarchiques avec plusieurs niveaux d'ordonnées à l'origine aléatoires.