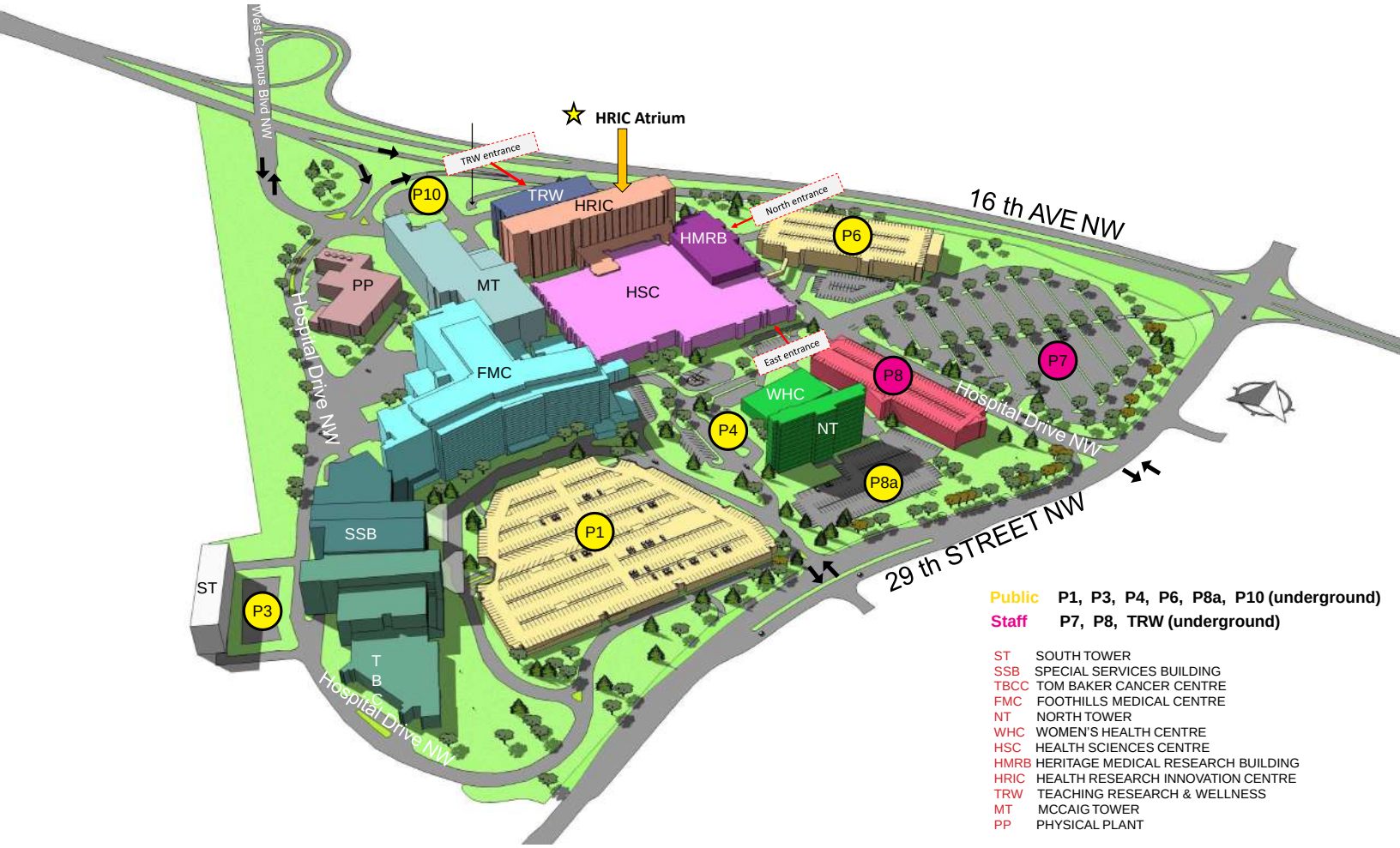
Seventh Canadian Statistics Student Conference
•
Septième Congrès Canadien des Étudiants en Statistique

University of Calgary, Alberta

Saturday • Samedi
May 25 • Mai 2019



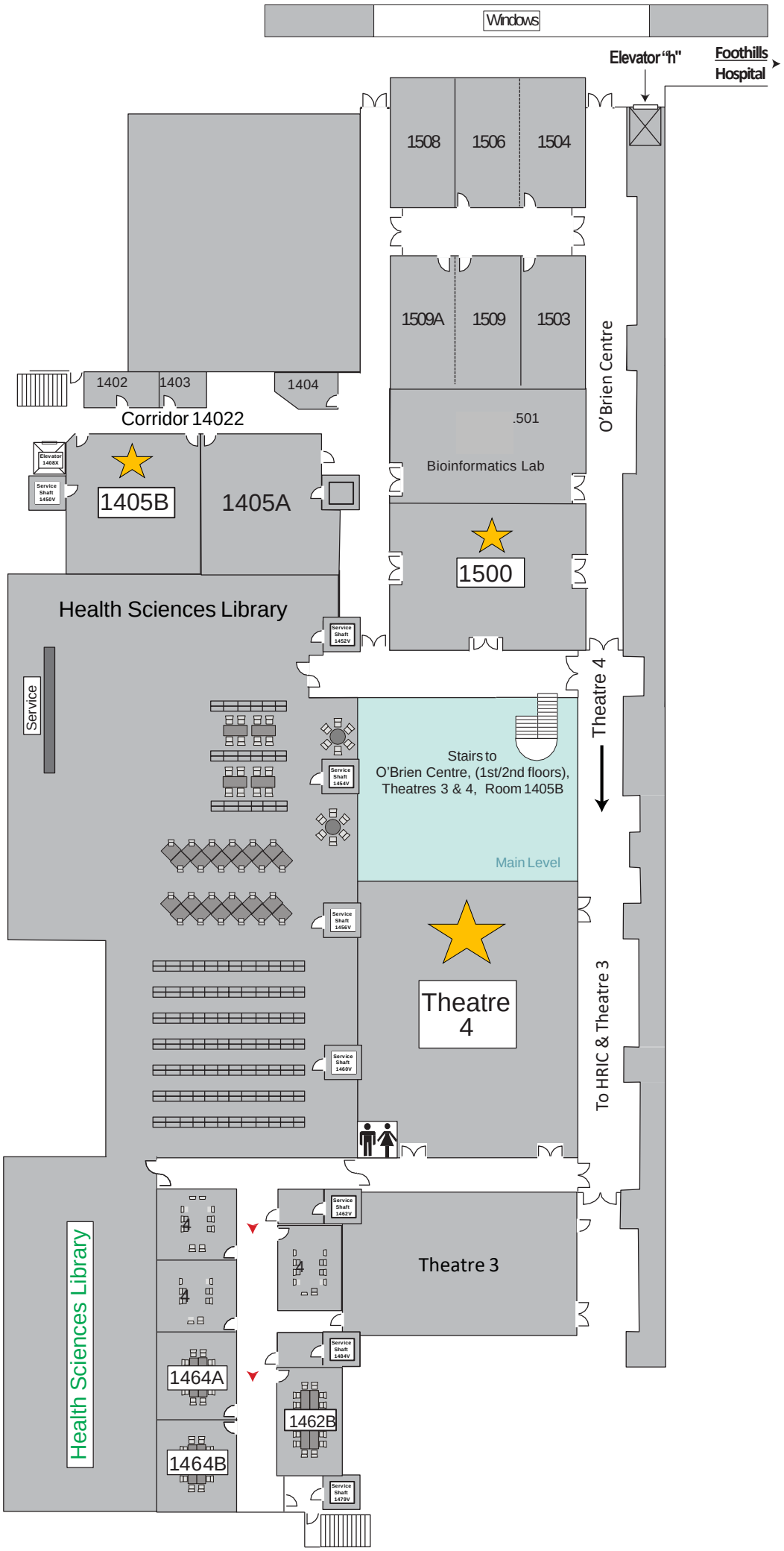
Public **P1, P3, P4, P6, P8a, P10 (underground)**

Staff **P7, P8, TRW (underground)**

ST	SOUTH TOWER
SSB	SPECIAL SERVICES BUILDING
TBCC	TOM BAKER CANCER CENTRE
FMC	FOOTHILLS MEDICAL CENTRE
NT	NORTH TOWER
WHC	WOMEN'S HEALTH CENTRE
HSC	HEALTH SCIENCES CENTRE
HMRB	HERITAGE MEDICAL RESEARCH BUILDING
HRIC	HEALTH RESEARCH INNOVATION CENTRE
TRW	TEACHING RESEARCH & WELLNESS
MT	MCCAIG TOWER
PP	PHYSICAL PLANT

Alberta Health Services

FOOTHILLS CAMPUS MAP


Windows
Elevator "h"
Foothills Hospital
1508
1506
1504
1509A
1509
1503
O'Brien Centre
1402
1403
1404
Corridor 14022
.501
Bioinformatics Lab
1405B
1405A
1500
Health Sciences Library
Service
Service Shaft 1452V
Service Shaft 1454V
Service Shaft 1466V
Service Shaft 1460V
Stairs to O'Brien Centre, (1st/2nd floors), Theatres 3 & 4, Room 1405B
Main Level
Theatre 4
Theatre 3
To HRIC & Theatre 3
Service Shaft 1462V
Service Shaft 1484V
Service Shaft 1479V
1464A
1462B
1464B
Health Sciences Library

1st Floor

Contents • Table des matières

Welcome • Bienvenue

We are pleased to have you at the Canadian Statistics Student Conference!

Our main goal is to offer a space for students and recent graduates to network, learn, and participate in multiple activities that can help them consolidate the direction of their professional path. The CSSC provides a low-stress environment for sharing ideas, developing skills and holding discussions with others about research, while learning about career opportunities. What is best, a wide range of statistical interests are represented, such as biostatistics, industrial statistics; big data analysis; probability theory, Bayesian approaches, and more.

This year's event will put the spotlight on Machine Learning and Big Data, with a talk about machine learning applications in R and an interactive workshop opposing machine learning and logistic regression for big data. Attendees are encouraged to bring their laptops to the workshop. The program also includes a session with representatives from different areas of the statistics workforce, aiming to provide career advice to graduates and recent graduates; as well as multiple sessions dedicated to showcase students' research work through talk and poster presentations.

Finally, the keynote speaker, **Dr. Charmaine Dean** from the University of Waterloo, will talk about Interdisciplinary work and being successful as a leader in the field of statistics.

Nous sommes heureux de vous compter parmi nous au Congrès canadien des étudiants en statistique (CCÉS)! Nous visons principalement à offrir aux étudiants et aux nouveaux diplômés un espace leur permettant de réseauter, d'apprendre, et de participer à de multiples activités susceptibles de les aider à consolider l'orientation de leur parcours professionnel. Le CCÉS offre un environnement sans stress pour partager des idées, développer des compétences et discuter de recherche avec des pairs, tout en en apprenant davantage sur les possibilités de carrière. Qui mieux est, un large éventail d'intérêts statistiques sont représentés, tels que la biostatistique, la statistique industrielle, l'analyse de données volumineuses, la théorie des probabilités, les approches bayésiennes et plus encore. L'événement de cette année mettra en lumière l'apprentissage machine et les données volumineuses, grâce à une présentation sur les applications d'apprentissage machine en R et à un atelier interactif opposant l'apprentissage automatique et la régression logistique pour les données volumineuses. Nous encourageons les participants à apporter leur ordinateur portable à l'atelier. Le programme inclut également une session avec des représentants de différents secteurs de la statistique, dans le but de fournir des conseils de carrière aux diplômés et aux nouveaux diplômés. Il comporte aussi plusieurs sessions qui mettront en valeur le travail de recherche des étudiants, par l'intermédiaire de présentations orales et par affiche.

Sponsors • Commanditaires

Special thanks to all our sponsors who have provided generous support for the various activities of the Canadian Statistics Student Conference. These contributions have made this event possible.

Nous tenons à remercier chacun de nos commanditaires pour leur généreuse contribution au Congrès Canadien des Étudiants en Statistique. C'est grâce à eux que la tenue de ce congrès est possible.

Gold Sponsors • Commanditaires Or

UNIVERSITY OF CALGARY

https://math.ucalgary.ca

NACTRC
Northern Alberta Clinical Trials + Research Centre

Gold Sponsors • Commanditaires Or

CANSSI
INCASS

Gold Sponsors • Commanditaires Or

UNIVERSITY OF WATERLOO
FACULTY OF MATHEMATICS
Department of Statistics
and Actuarial Science

GoldSpot
DISCOVERIES INC.

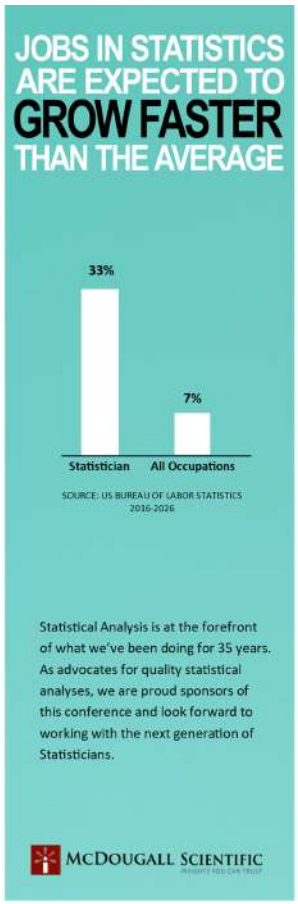
Silver Sponsors • Commanditaires Argent

UNIVERSITY OF CALGARY
CUMMING SCHOOL OF MEDICINE
Department of Community Health Sciences

Silver Sponsors • Commanditaires Argent

RStudio®

SFU STATISTICS AND ACTUARIAL SCIENCE



Silver Sponsors • Commanditaires Argent

The Student Committee of the Canadian Math Society

Apply for Conference Funding!

Read our student publication!

Le comité étudiant.e de la Société mathématique du Canada

Notes from the Margin

Veuillez envoyer votre article à Asmita Sodhi au student-editor@cms.math.ca.

@studccms studc.cms.math.ca issuu.com/cms-studc @studccms

Bronze Sponsors • Commanditaires Bronze

Organizers and volunteers • Organisateurs et bénévoles

Organizing committee • Comité organisateur

Co-chairs / Co-présidentes:
Ms. Anita Brobbey (University of Calagary)
Ms. Myrtha Reyna (University of Toronto)

Local arrangements / organisation locale:
Ms. Fahmida Yeasmin (University of Calgary)
Mr. Charles Sam (University of Calgary)

Fundraising / Collecte de fonds:
Mr. Thai-Son Tang (University of Toronto)
Ms. Melissa Van Bussel (Trent University)
Ms. Lin Ling (University of Toronto)

Translation / Traduction:
Mr. Luc Villandre (McGill University)
Mr. Steve Ferreira (McGill University)
Ms. Marie-Christine Robitaille Grou (Université de Montréal)

Skills session / Séance sur les compétences techniques:
Mr. Olawale Fatai Ayilara (University of Manitoba)
Mr. Jacob Prosser (University of New Brunswick)

Carrer session / Séance sur les carrières:
Ms. Afaf Alzahrani (Dalhousie University)
Mr. Sudipta Saha (University of Toronto)

Scientific Program / Programme scientifique:
Ms. Michela Panarella (Univeristy of Toronto)
Ms. Victoire Michal (Université de Montréal)

Support and thanks • Support et remerciements

SSC President / Président de la SSC: Robert Platt

SSC Administrative assistant / Adjoint Administrative de la SSC: Miaclaire Woodland

SSC Executive assistant / Assistant exécutif de la SSC:: Michelle Benoit

SSC Treasurer / Trésorier de la SSC: Edward Chen

SSC Local organizers / Organisateurs locaux: Karen Kopciuk, Alexander de Leon

SSC meetings coordinator / Coordonnateur des congrès: Changbao Wu

Photographer / Photographe: Peter Macdonald

Volunteers / Bénévoles: Shakiru Alaka, Sarath Kumar Jayaraman, Mohammed Mujaab Kamso, Mili Roy, Oluwaseyi Adetutu Lawal, and Ayoola Ademola.

Judges / Juges: Special thanks to the judges involved in the assessment of abstracts, talks and poster presentations.

Program Overview

Date and times: Saturday, May 25th, from 07:30 to 20:30.
Location: Health Science Centre. Cumming School of Medicine, University of Calgary.

Time	Session	Room	Page
07:30-08:30	Registration Breakfast	HRIC atrium	
08:30-08:45	Presidential address	Theatre 4	
08:50-09:35	**Student research talks I**		
	Applications	Theatre 4	22
	Robust estimators	O1500	25
	Causal inference	1405B	28
09:40-10:25	**Student research talks II**		
	Biostatistics	Theatre 4	31
	Bayesian statistics	O1500	34
	Model selection	1405B	37
10:25-10:45	Coffee break	HRIC atrium	
10:45-11:55	**Skills session** *Machine Learning applications in R*	Theatre 4	16
11:55-12:10	Sponsor Talk	Theatre 4	
12:10-13:20	Lunch Poster session (starts at 12:40)	HRIC atrium	 40
13:20-14:45	**Workshop** *Exploring Machine Learning Classification Methods Using R*	Theatre 4	15
14:45-15:00	Coffee break	HRIC atrium	
15:00-16:00	**Career panel**	Theatre 4	17
16:00-17:15	**Keynote speech** *Interdisciplinary Work and being successful as a leader in this arena*	Theatre 4	14
17:15-18:00	Closing and awards	Theatre 4	
18:00-21:00	Social Evening		

Social evening

A limited quantity of food and beverages will be available at no additional cost.
Address & time • Adresse & heure:
The Den 18:00 hrs. MacEwan Student Centre, 2500 University Drive NW, Calgary, AB. (http://den.su.ucalgary.ca)

Aperçu du programme

Date et plage horaire: Samedi 25 mai de 07h30 à 20h30.
Lieu: Health Science Centre. Cumming School of Medicine, Université de Calgary.

Heure	Séance	Salle	Page
07:30-08:30	Inscription Petit-déjeuner	HRIC atrium	
08:30-08:45	Adresse présidentielle	Theatre 4	
08:50-09:35	**Présentations orales étudiantes I**		
	Applications	Theatre 4	22
	Estimateurs robustes	O1500	25
	Inférence causale	1405B	28
09:40-10:25	**Présentations orales étudiantes II**		
	Sélection de modèle	Theatre 4	31
	Statistique bayesienne	O1500	34
	Biostatistique	1405B	37
10:25-10:45	Pause-café	HRIC atrium	
10:45-11:55	**Session de formation** *Applications en R de l'apprentissage machine*	Theatre 4	16
11:55-12:10	Présentation du sponsor	Theatre 4	
12:10-13:20	Lunch Séance d'affiches (débute à 12:40)	HRIC atrium	40
13:20-14:45	**Atelier statistique** *Exploration de méthodes de classifications en apprentissage machine sur R*	Theatre 4	15
14:45-15:00	Pause-café	HRIC atrium	
15:00-16:00	**Table ronde des carrières**	Theatre 4	17
16:00-17:15	**Présentation d'honneur** *Travail interdisciplinaire et réussir en tant que leader dans cette aréna*	Theatre 4	14
17:15-18:00	Clôture et remise des prix	Theatre 4	
18:00-21:00	Soirée		

Soirée

Une quantité limitée de nourriture et de boissons sera disponible sans frais supplémentaires.
Adresse & heure:
The Den 18:00 hrs. MacEwan Student Centre, 2500 University Drive NW, Calgary, AB. (http://den.su.ucalgary.ca)

Keynote address • Discours d'honneur

Charmaine Dean is Vice-President, Research and International at the University of Waterloo. Her focus is on building upon foundational strengths to heighten the emphasis on collaborations, and link related external portfolios in a systematic approach to industrial partners and entrepreneurship. Dr. Dean's work in space-time analytics for health and forestry has been recognized widely: in 2003, she was awarded the CRM-SSC prize; in 2007 named Fellow of the American Statistical Association and awarded the University of Waterloo Mathematics Alumni Achievement Medal; in 2010 named Fellow of the American Association for the Advancement of Science; in 2012 awarded a Trinidad & Tobago Canadian High Commission Award; and in 2016 elected to the International Statistical Institute.

Abstract

Solving many complex societal problems facing the world, from sustainable development to climate change to understanding natural disasters and global health problems, often involves an interdisciplinary approach. Statisticians have an important role to play in contributing to solutions because of our leadership in developing tools for evidence-based decision-making. Additionally, we are well poised to pull together interdisciplinary teams because generally our work is interdisciplinary in nature. For example, developing new statistical tools to solve a scientific problem requires that we understand the science surrounding the problem. Interdisciplinary research is often of high societal impact and gives researchers valuable exposure to a diversity of research concepts, tools and methodologies beyond their own discipline. What are the key ingredients for creating a successful, innovative and productive interdisciplinary environment? What qualities make for success in this environment and what challenges may hinder success? What are the skills that a leader of a multi-disciplinary team needs? What makes interdisciplinary work fun and exciting? How can you be a confident contributor around an interdisciplinary table of experts? This talk considers these questions, providing examples to illustrate best practices in collaborative training and research environments.

Charmaine Dean est vice-présidente - Recherche et International à l'Université de Waterloo. Dans le cadre de ses fonctions, Dre Dean vise à encourager les collaborations, et à agir comme liaison entre portfolios externes, partenaires industriels et entrepreneurs. Les travaux de Charmaine Dean dans le domaine de l'analyse spatiotemporelle appliquée à la santé et à la foresterie lui ont valu plusieurs prix. En 2003, Dre Dean s'est vue décerner le prix CRM-SSC en statistique, en reconnaissance d'une contribution substantielle à la discipline au cours des quinze années suivant l'obtention du doctorat. En 2007, elle a remporté le "Mathematics Alumni Achievement Medal" de l'Université de Waterloo. En 2010, elle a été nommée Fellow de l'Association américaine pour l'avancement des sciences. En 2012, elle a été lauréate du "Trinidad & Tobago High Commission Award". Enfin, en 2016, elle a été élue membre du International Statistical Institute.

Résumé scientifique

Résoudre les problèmes sociétaux auxquels nous sommes confrontés, du développement durable aux changements climatiques, de la compréhension des catastrophes naturelles à celle des problèmes de santé mondiale, requiert souvent une approche interdisciplinaire. Les statisticiens jouent un rôle crucial dans l'élaboration de solutions, en raison de leur prééminence dans le développement d'outils pour la prise de décisions basée sur des données probantes. Le développement de nouveaux outils statistiques pour résoudre un problème scientifique nécessite notamment une compréhension de la science au coeur du problème. La recherche interdisciplinaire a souvent des répercussions sociétales considérables. Elle est également très bénéfique aux chercheurs eux-mêmes, qui se voient exposés à une variété de concepts, d'outils et de méthodologies allant bien au-delà de leur propre discipline. Quels sont les ingrédients indispensables pour créer un environnement interdisciplinaire novateur, productif et propice au succès? Quelles qualités assurent la réussite dans cet environnement et quels défis peuvent mener à l'échec? Quelles sont les aptitudes dont a besoin le dirigeant d'une équipe multidisciplinaire? Qu'est-ce qui rend le travail interdisciplinaire amusant et excitant? Comment peut-on bien contribuer aux travaux d'un comité interdisciplinaire d'experts? Cette présentation aborde ces questions, et fournit des exemples illustrant les meilleures pratiques dans des environnements de formation coopérative et de recherche.

Statistical Computing Workshop • Atelier en Calculs Statistiques

Brendan Cord Lethebe is an experienced researcher in University of Calgary. He is also the methods/analytics lead of the clinical research unit at the Cumming School of Medicine at University of Calgary. He holds a Bachelor in Actuarial Science (University of Calgary) and a Masters in Biostatistics (University of Calgary). He is mostly focused in Biostatistics and is skilled in Mathematical Modelling, Biostatistics, R, SQL, Python, STATA and Clinical Research.

Abstract

Being able to properly fit machine learning models is becoming an important skill for those in the statistical community. Using a publicly available dataset we will focus on supervised classification models, and techniques for optimal parameter selection. We will explore the LASSO logistic regression, various decision tree algorithms, random forest, and neural net using popular R packages. This will allow us to clearly see the advantages and disadvantages of using interpretable models vs "black-box" algorithms.

Brendan Cord Lethebe est un chercheur aguerri à l'Université de Calgary. Il est également le responsable des méthodes et des analyses au sein de l'unité de recherche clinique au Cumming School of Medicine de l'Université de Calgary. Il détient un baccalauréat en science actuarielle ainsi qu'une maîtrise en biostatistique. Il se concentre surtout sur la biostatistique, mais il est également doué en modélisation mathématique, R, SQL, Python et STATA.

Résumé scientifique

La capacité d'ajuster adéquatement des modèles d'apprentissage machine est de plus en plus importante pour les statisticiens. À l'aide d'un jeu de données public, nous nous concentrerons sur un modèle de classification supervisée et sur des techniques optimales pour la sélection de paramètres. Par l'intermédiaire de librairies R communes, nous aborderons la régression logistique LASSO, les forêts aléatoires, les réseaux de neurones, ainsi que de multiples algorithmes pour les arbres de décision. Ceci nous permettra de bien distinguer les avantages et les inconvénients d'utiliser un modèle interprétable plutôt que des algorithmes "boîte noire".

Machine Learning applications in R
Applications de l'apprentissage machine en R

Tom Loughin is a professor and chair of the Department of Statistics and Actuarial Science at Simon Fraser University in Burnaby, British Columbia, Canada. He got his PhD in Statistics from Iowa State University and spent 13 years at Kansas State University before moving to SFU in 2006. Tom has broad research interests in many areas of statistical application and method development, including statistical learning, modeling categorical data, design and analysis of experiments, and statistics in sports. He has published a book, "Analysis of Categorical Data with R," co-authored with Chris Bilder and available from CRC Press. Tom has extensive experience as a statistical consultant and has PSTAT accreditation from both the ASA and the Statistical Society of Canada. He is also a Fellow of the ASA.

Prof. Tom Loughlin est directeur du département de statistique et de science actuarielle à l'Université Simon Fraser (SFU) à Burnaby, Colombie-Britannique. Il a obtenu son doctorat en statistique de la Iowa State University et a passé treize ans à Kansas State University avant de rejoindre SFU en 2006. Ses intérêts de recherche recoupent plusieurs champs d'applications statistiques et de développement méthodologique. Il s'est penché notamment sur l'apprentissage statistique, la modélisation de données catégoriques, la conception et l'analyse d'expériences, ainsi que sur les statistiques sportives. Il a publié un livre intitulé "Analysis of Categorical Data with R", co-écrit par Chris Bilder et disponible chez CRC Press. Tom a une longue expérience en tant que consultant statistique et possède l'accréditation PSTAT de la American Statistical Association (ASA) et de la Société statistique du Canada (SSC). Il est également Fellow de l'ASA.

Invited career speakers • Conférenciers invités à la séance sur les carrières

Dominique Ibañez

Dominique is Chief, Biostatistics and Risk Modelling Division, Bureau of Food Surveillance and Science Integration, Food Directorate at Health Canada. She graduated with a Master's degree in Biostatistics from the University of Toronto. She joined the Health Canada's Food Directorate four years ago. She arrived with 25 years of experience conducting statistical analysis in a clinical environment – primarily in Rheumatology research. She has over 90 peer-reviewed articles to her credit. She now leads a team of 10 statisticians. Under her leadership, new standards have been introduced to improve performance. Some of these include: mentoring of new staff by more experienced ones, monthly statistical discussion forums and greater focus on statistical research as well as outreach to academic nutrition researchers across Canada. Her team has been involved in providing statistical expertise and analysis in key projects conducted in the Food Directorate such as Sodium Reduction in Canada, pre-market evaluations and Salmonella in Chicken.

Dominique est titulaire d'une maîtrise en biostatistique de l'Université de Toronto. Elle s'est jointe à la Direction des aliments de Santé Canada il y a quatre ans. Elle est arrivée avec 25 ans d'expérience en analyse statistique en milieu clinique, principalement en recherche en rhumatologie. Elle a plus de 90 articles évalués par des pairs à son actif. Elle dirige maintenant une équipe de dix statisticiens. Sous sa direction, de nouvelles normes ont été introduites pour améliorer les performances. Celles-ci incluent: le mentorat du nouveau personnel par des personnes plus expérimentées, des forums de discussion mensuels sur les statistiques, une plus grande attention portée à la recherche statistique ainsi qu'à la création de liens avec les chercheurs universitaires en nutrition à travers le Canada. Son équipe a prodigué une expertise statistique dans le cadre de projets clés menés à la Direction des aliments, tels que la réduction de la teneur en sodium des aliments, les évaluations précédant la mise en marché, et la présence de salmonelle dans le poulet.

Lisa Lix

Dr. Lisa Lix is Professor of Biostatistics and a Tier I Canada Research Chair in Methods for Electronic Health Data Quality in the Department of Community Health Sciences, Max Rady College of Medicine, University of Manitoba. She is also Director of the Data Science Platform in the George & Fay Yee Centre for Healthcare Innovation (CHI), a research unit that is a collaboration between the Winnipeg health region and the University of Manitoba. The CHI aims to strengthen patient-focused research in Manitoba. Her team of 30+ faculty, staff and trainees with expertise in biostatistics, bioinformatics, and clinical research methodology focuses on methodological research, training, and consulting. Dr. Lix's areas of research expertise include methods to address bias and error in electronic health databases, statistical methods for the analysis of patient-reported outcomes, and methods for the analysis of longitudinal data. She is a prolific researcher who has published more than 325 scholarly papers. Dr. Lix is an elected member of the Board of Directors of the Statistical Society of Canada, Program Chair for the Society's 2019 meeting in Calgary, Co-Chair of the Data Quality Working Group for the Canadian Chronic Disease Surveillance System, and Program Chair Elect for the 2020 Joint Statistical Meetings Health Policy Statistics Section.

Dre Lisa Lix est professeure de biostatistique et titulaire de la Chaire de recherche du Canada de niveau 1 sur les méthodes d'assurance de la qualité des données électroniques sur la santé au Collège de médecine Max Rady de l'Université du Manitoba. Elle est aussi directrice de la Plateforme de science des données du Centre George & Fay Yee pour l'innovation dans les soins de santé, une unité de recherche résultant d'une collaboration entre l'Office régional de la santé de Winnipeg et l'Université du Manitoba. Le Centre George & Fay Yee vise à renforcer la recherche ciblée sur le patient au Manitoba. Son équipe, comportant plus de trente professeurs, membres du personnel de soutien et stagiaires possédant une expertise en biostatistique, en bioinformatique et en méthodologie de recherche clinique, est axée sur la recherche méthodologique, la formation et la consultation. Les domaines d'expertise de Dre Lix comprennent les méthodes pour corriger les biais et les erreurs dans les bases de données électroniques sur la santé, les méthodes statistiques pour l'analyse d'issues rapportées par le patient et les méthodes d'analyse pour les données longitudinales. Elle est une chercheuse prolifique ayant publié plus de 325 articles scientifiques. Dr Lix est membre élue du conseil d'administration de la Société statistique du Canada, présidente de la programmation du Congrès annuel de la SSC de 2019, coprésidente du groupe de travail sur la qualité des données du Système canadien de surveillance des maladies chroniques, et responsable désignée de la programmation en statistiques appliquées aux politiques de santé à la conférence Joint Statistical Meetings (JSM) de 2020.

Steven Wu

Steven Wu finished his B. Math Hon. in Statistics at Carleton University and finished his MSc Statistics at Simon Fraser University. He did the co-op programs at both schools, doing his first work term as a manual QA tester at a startup because his programming skills were terrible. It was in that role of rote tasks where he realized how powerful programming is as a tool for productivity. Absorbing tech context at day and developing Python programming skills at night, Steven developed and marketed a web app that allowed Canadian university basketball coaches to inform their game-planning strategies using data and statistical methods. Attempts to sell it failed spectacularly but he learned a ton and it made for a great project to talk about when applying for jobs. He interviewed with a few sports teams, government departments, tech companies, and a hedge fund before accepting a role as a Data Scientist for Shopify's People Analytics team. People Analytics collects and analyzes data about the Shopify team to (a) help leadership make data informed decisions and (b) help make our workplace more efficient and engaging. We regularly use statistical methods to influence decisions around hiring, performance, retention, engagement, learning, culture, andmore.

Steven Wu a obtenu un B.Math Hon. en statistique de l'Université Carleton, puis une M.Sc. en statistique de l'Université Simon Fraser. Il a fait le programme co-op à chaque institution. Pour son premier stage, il a été embauché comme testeur manuel d'assurance qualité (QA) dans une startup, puisque sa connaissance de la programmation était terrible. La succession de tâches monotones lui a permis de réaliser le pouvoir de la programmation comme outil pour accroître la productivité. En s'imprégnant le jour de l'environnement de la haute technologie, puis en raffinant le soir sa maîtrise de Python, Steven a pu développer et mettre en marché une appli permettant aux entraîneurs de basketball dans les universités canadiennes d'améliorer leurs stratégies de jeu à l'aide de données et de méthodes statistiques. Les tentatives de la vendre ont échoué misérablement, mais l'expérience a été malgré tout instructive et lui a de plus donné un bon projet à mentionner dans le cadre de sa recherche d'emploi. Il a postulé auprès de quelques équipes sportives, départements gouvernementaux, et compagnies de haute technologie, ainsi qu'auprès d'un fonds spéculatif avant de dénicher un poste de spécialiste des données dans l'équipe People Analytics de Shopify. L'équipe People Analytics collecte et analyse des données sur le fonctionnement de Shopify afin d'aider les dirigeants à prendre des décisions basées sur les données, et de créer un environnement de travail plus efficace et stimulant. Son travail statistique influence la prise de décision en rapport à l'embauche, la performance, la rétention, l'implication, l'apprentissage, la culture, et bien plus.

Posters • Affiches

No	Title/Titre	Presenter/Presenteur
1	Aint Played Nobody: Building an Optimal Schedule to Secure an NCAA Tournament Berth J'ai pas triché: Construction d'un calendrier optimal pour sécuriser une place au championnat de la NCAA	Kevin Floyd
2	Application of the Distributed Lag Models for Examining Associations Between the Built Environment and Obesity Risk in Children Application de modèles à retards échelonnés pour examiner les associations entre l'environnement bâti et le risque d'obésité chez les enfants	Anna Smyrnova
3	Projected changes of extreme rainfall in the province of Quebec Changements projetés des précipitations extrêmes au Québec	Éloise Nolet-Gravel
4	Computing R-optimal designs for multi-response regression models via interior point method Calcul de schémas R-optimaux pour les modèles de régression à réponses multiples à l'aide d'une méthode du point intérieur	Pengqi Liu
5	Conducting causal inference in the presence of measurement bias using administrative databases Inférence causale en présence de biais de mesure à partir de bases de données administratives	Sumeet Kalia
6	Construction of Block Incomplete Design Under Correlated Error Structure Construction de plans en blocs incomplets avec une structure d'erreurs corrélées	Meixin Liu
7	Covariance-adjusted, sparse, reduced-rank regression with application to imaging-genetics data Régression parcimonieuse, à rang réduit et ajustée pour la covariance, avec applications à des données d'imagerie et de génétique	Haoyao Ruan
8	Grading Gunslingers: A Preliminary Model For Evaluating Pitcher Types in Baseball Classement des tireurs d'élites: Modèle préliminaire pour évaluer les types de lanceurs au baseball - la méthode de réévaluation continue	Alexander Sharp
9	Improving the Hosmer-Lemeshow Goodness-of-Fit Test Amélioration du test d'ajustement de Hosmer-Lemeshow	Nikola Surjanovic
10	Incremental value of AUC, average positive predictive value and Brier Score Valeur ajoutée de l'ASC, de la valeur positive prédictive moyenne et du score de Brier	Zhe Lu
11	Investigating the relationship between temperature and the number of fatalities on Canadian highways using time series analysis. Une étude sur la relation entre la température et le nombre de décès sur les autoroutes canadiennes à l'aide d'une analyse de séries chronologiques.	Alex Mackie
12	Joint Models of Longitudinal and Time-to-event Data: Impact of Data Collection Cycles Modèles conjoints pour données longitudinales et de durée de vie: l'effet des cycles de collectes de données	Yixiu Liu
13	Longitudinal Changes in Colorectal Cancer among Farm and Non-farm Rural Residents Changements longitudinaux dans le cancer colorectal chez les résidents de zones rurales agricoles et non agricoles	Ibrahim Watara Abubakari
14	Parsimonious Gaussian Mixtures via Chimeral Clusters Mélanges gaussiens parcimonieux via des grappes chimères	Jason Hou-Liu
15	Semi-parametric estimation of scoring rates in the English Premier League Estimation semi-paramétrique de la cadence des tirs au but réussis dans la English Premier League	Robyn Ritchie
16	Semi-supervised nonnegative matrix factorization with applications to spectral data Factorisation matricielle non négative semi-supervisée avec applications aux données spectrales	Shreeves Phil
17	Spatial tracking of the current Ebola outbreak in Congo Suivi spatial de l'épidémie actuelle d'Ebola au Congo	Michael Wendlandt
18	Spatio-Temporal Modelling of Ichemic Heart Disease in Manitoba Modélisation spatio-temporelle de la cardiopathie ischémique au Manitoba	Justin Dyck
19	The Effect of Pace on the Performance of the Closers in the Kentucky Derby L'effet de l'allure sur la performance des closers au Kentucky Derby	Miguel Macaraig
20	Understanding Edmonton's Weather: An Analysis of the Mean Temperature and Snowfall Comprendre la météo d'Edmonton : une analyse de la température moyenne et des chutes de neige	David Cao
21	Unsupervised Learning on Functional Data with Application to U.S. Weather Data Apprentissage non supervisé de données fonctionnelles avec application aux données météorologiques américaines	Chuyuan Lin
22	Validation Study on a Screening Tool for Mental Health of Children and Youth in Canada Étude de validation d'un outil de dépistage de la santé mentale des enfants et des jeunes au Canada	Xuejing Jiang

Oral presentations •Présentations orales

Student Research Talks I

No	Title/Titre	Presenter/Presenteur	Category/Catégorie	Page
1	Impact of interest rate portfolio selection techniques L'effet des risques de taux d'intérêt de techniques optimales de sélection de portfolios			22
2	An Application of the Gibbs Sampling to the Battleship Game Une application de l'échantillonnage de Gibbs au jeu Bataille Navale	Dan Richard	Applications	23
3	Examining Age and Sex-related Differential Item Functioning in Seattle Angina Questionnaire Examen du fonctionnement différentiel lié à l'âge et au sexe dans le questionnaire sur l'angine de Seattle d'identification automatique (SIA)	Oluwaseyi A. Lawal	Applications	24
4	Construction of D-optimal Designs in Polynomial Regression Models Formulation de plans D-optimaux pour les modéles de régression polynomiale	Di Wu	Robust estimators	25
5	A new design of the continual reassessment method Un nouveau plan d'expérience pour la méthode de réévaluation continue	Weijia Zhang	Robust estimators	26
6	Time Series Interpolation Algorithms: An Application to Real-World Data Algorithmes d'interpolation pour séries chronologiques : Application à des données réelles	Melissa Van Bussel	Robust estimators	27
7	Sufficient dimension reduction for feasible and robust estimation of average causal effect Réduction suffisante de la dimensionnalité pour une estimation réalisable et robuste de l'effet causal moyen	Trinetri Ghosh	Causal inference	28
8	On Bayesian estimation of causal effect with a latent confounder class Estimation bayésienne d'un effet causal en présence d'une classe de confusion latente	Kuan Liu	Causal inference	29
9	Distance Metrics for Measuring Joint Dependence with Application to Causal Inference Mesures de distance pour mesurer la dépendance conjointe avec application à l'inférence causale	Shubhadeep Chakraborty	Causal inference	30

Student Research Talks II

No	Title/Titre	Presenter/Presenteur	Category/Catégorie	Page
10	Frequentist Model Averaging Estimator of Support Vector Machine Classfiers and Regressors Estimateur par combinaison de modèles fréquentistes pour les classificateurs et régresseurs par machines à vecteurs de support	Kiwon Francis	Model Selection	31
11	Outlier Detection Methods for Quantitative Fatty Acid Signature Analysis Méthodes de détection des valeurs aberrantes pour l'analyse quantitative de la signature en acides gras	Jennifer McNichol	Model Selection	32
12	Predictive Comparison of Vine Copula Models Comparaison prédictive de modèles de copules en vignes	Md Erfanul Hoque	Model Selection	33
13	Bayesian spatial logistic regression model for investigating socio-economic and demographic determinants and geographic variation of pregnancy termination among Bangladeshi women Modèle de régression logistique spatiale bayésienne pour étudier les déterminants socio-économiques et démographiques, et la variation géographique de l'interruption de la grossesse chez les Bangladaises	Rifat Zahan	Bayesian statistics	34
14	Measurement error adjustment in a zero-inflated Poisson model Ajustement pour l'erreur de mesure dans un modèle de Poisson à inflation de zéro	Kangjie Zhang	Bayesian statistics	35
15	Bayesian Approaches to Density Estimation for Use in Functional Linear Regression Approches bayésiennes pour l'estimation de densités pour l'utilisation en régression linéaire fonctionnelle	Shaun McDonald	Bayesian statistics	36
16	Variation Along Continuous Neuroelectric Activity Related to Early Cognitive Impairment Variation le long de l'activité neuroélectrique continue liée à un trouble cognitif précoce	Henry Lu	Biostatistics	37
17	Crossed random effects modelling of binomial data with random cluster sizes La modélisation par effets aléatoires croisés des données binomiales avec des groupages de tailles aléatoires	Jingyu Cui	Biostatistics	38
18	Automated disease detection in dairy cattle using recurrent neural networks Détection automatisée de maladies chez les bovins laitiers à l'aide de réseaux de neurones récurrents	Syed Ali Naqvi	Biostatistics	39

Scientific abstracts: Oral presentations • Résumés scientifiques: présentations orales

Applications

08:50am - 09:35am, Theatre 4, Michela Panarella (Chair • Présidente)
michela.panarella@mail.utoronto.ca

Lin, Wei-Hsiang; Lin, Shih-Kuei; Tsai, Cary Chi-Liang
Impact of interest rate, surrender, and liquidity risks on the surplus of a portfolio of endowment policies using optimal portfolio selection techniques
L'effet des risques de taux d'intérêt, de rachat et de liquidité sur le surplus d'un portfolio de polices de dotation résultant de techniques optimales de sélection de portfolios

A life insurer charges an endowment policyholder high premiums from which the policyholder's cash value is built at an interest rate. The life insurer invests the collected premiums in financial securities to meet or exceed the interest rate, and a policyholder can surrender his policy before maturity and get his cash value back subject to a surrender charge. When lots of policyholders surrender their policies, the life insurer needs to liquidate some securities in a short time, which exposes the insurer to liquidity risk. In this paper, we propose a framework to analyse the impact of interest rate, surrender, and liquidity risks on the surplus of a portfolio of endowment policies. Under the framework, we formulate the fair premium and risk-based reserves calculations. In addition, we adopt optimal portfolio selection methods for maximizing utilities. A series of sensitivity analyses are conducted to illustrate the surplus distributions and corresponding utilities after the adoption.

Un assureur charge au détenteur d'une police de dotation des primes élevées, à partir desquelles se bâtit la valeur en argent de l'assuré, en fonction d'un certain taux d'intérêt. L'assureur investit les primes collectées dans des produits financiers afin d'atteindre ou d'excéder ce taux d'intérêt. L'assuré peut racheter sa police avant maturité et obtenir ainsi sa valeur en argent, moins les frais de rachat. Si un grand nombre d'assurés rachètent leur police, l'assureur doit liquider des actifs financiers en peu de temps, créant un risque de liquidité. Dans cet article, nous proposons une méthode pour analyser l'effet des risques liés aux taux d'intérêt, aux rachats et à la liquidité sur les surplus d'un portfolio de polices de dotation. À partir de cette méthode, nous formulons les calculs pour un montant de prime juste et pour les montants de réserve associés au risque. Nous adoptons également des méthodes optimales de sélection de portfolios afin de maximiser les utilités. Dans ce contexte, nous réalisons un ensemble d'analyses de sensibilité dans le but d'illustrer les distributions des surplus et les utilités correspondantes.

Richard, Dan; Lupul, Nicholas

An Application of the Gibbs Sampling to the Battleship Game
Une application de l'échantillonnage de Gibbs au jeu Bataille Navale

Battleship is a classic two player game where the goal is to sink the opponent's ships. Programming a winning strategy for this game is difficult because the state space representing the possible coordinates for the opponent's ships is huge. To solve this issue, we implemented an algorithm based on the Gibbs sampling to estimate the probability of each coordinate to contain a ship. Simulation results regarding the number of guesses to sink each ship and to complete a game are presented along with strategy Insights.

Bataille Navale est un jeu très populaire, dans lequel deux joueurs visent à couler les bateaux de l'adversaire. Programmer une stratégie gagnante pour ce jeu est difficile, car l'espace d'état comprenant les coordonnées possibles des bateaux de l'adversaire est énorme. Afin de régler ce problème, nous avons implémenté un algorithme basé sur l'échantillonnage de Gibbs (Gibbs sampling) nous permettant d'estimer la probabilité qu'un bateau se trouve à chaque coordonnée. Nous présentons les résultats de simulations en lien au nombre d'essais pour couler chaque bateau et pour compléter le jeu, ainsi que des conseils stratégiques connexes.

Oluwaseyi A. Lawal, Zhiying Liang, Oluwagbohunmi Awosoga, Maria J. Santana, Danielle A Southern, Lisa M. Lix, Colleen Norris, Matthew T. James, Tolulope Sajobi

Examining Age and Sex-related Differential Item Functioning in Seattle Angina Questionnaire

Examen du fonctionnement différentiel lié à l'âge et au sexe dans le questionnaire sur l'angine de Seattle

Background/Aims: Patient-reported outcome measures (PROMs) are increasingly being used in to compare the health status of different population groups. When completing PROMs subgroups of individuals may interpret questions about their health-related quality of life (HRQoL) differently, a phenomenon known as differential item functioning (DIF). This may threaten the overall comparability of PROM scores across population groups and/or over time. This study investigates the presence of DIF in the Seattle Angina Questionnaire (SAQ) administered to individuals with acute coronary syndrome (ACS). Methods: Data are from the Alberta Provincial Project on Outcome Assessment in Coronary Heart Disease registry. PROMs were collected using the 19-item SAQ, a cardiac disease-specific measure of HRQoL. Ordinal logistic regression (OLR) and multi-group confirmatory factor analysis were used to test for DIF at item- and domain level, respectively, across sex and age groups ($<$75 and $\geq$ 75 years). For OLR, DIF in each item was identified using change in Nagelkerke's R2 criterion. For multi-group confirmatory factor analysis (CFA), model fit was assessed using comparative fit index (CFI) and root mean square error of approximation (RMSEA). Changes in CFI $\leq$-0.01 indicates that the null hypothesis of invariance should not be rejected meaning that there is equality of patterns of the configural, weak, strong and strict invariance across sex/age groups; RMSEA $\leq$ 0.05 indicates close model fit. Results: Of the 3864 patients included in this analysis, 3203 (82.89%) were younger than 75 years old while 3006 (77.8%) are male. Several items demonstrated negligible DIF (0 $<$R2 ¡ 0.007). The domain level DIF analysis showed strict invariance across sex groups (CFI = 0.992; RMSEA = 0.035) but partial strong invariance was established across age groups (CFI = 0.982; RMSEA = 0.057). Discussion: Cardiovascular researchers can be confident that the SAQ scores are unbiased and comparable across age and sex groups.

Robust estimators • Estimateurs robustes

08:50am - 09:35am, O1500, Victoire Michal (Chair • Présidente)
victoire.michal@umontreal.ca

Di, Wu
Construction of D-optimal Designs in Polynomial Regression Models
Formulation de plans D-optimaux pour les modèles de régression polynomiale

Whenever we have an appropriate statistical model, it is crucial to have good estimation of the parameters of the model. Optimal design plays a big role on achieving this objective. There are a variety of criteria defining good estimation. Motivated by this fact, we construct D-optimal designs by minimizing the generalized variance of the parameter estimators of some polynomial regression models. In order to construct such designs, we use a class of algorithms, indexed by a function which depends on the derivatives of the criterion function. We also attempt to improve the convergence of the algorithm by using the properties of the directional derivatives of the criterion function.

Même avec un modèle statistique approprié, l'estimation des paramètres, bien que cruciale, peut être difficile. La formulation d'un plan optimal aide à améliorer les estimés obtenus. Il existe une multitude de critères pour quantifier la qualité de l'estimation. Dans cet esprit, nous formulons des plans D-optimaux en minimisant la variance généralisée des estimateurs des paramètres de modèles de régression polynomiale. Nous employons un type d'algorithme indexé par une fonction dépendant de la dérivée de la fonction critère. Nous tentons également d'améliorer la convergence de l'algorithme en exploitant les propriétés des dérivées directionnelles de la fonction critère.

Zhang, Weijia; Yang, Po; Muthukumarana, Saman
A new design of the continual reassessment method
Un nouveau plan d'expérience pour la méthode de réévaluation continue

We propose a new design of the continual reassessment method (CRM) and systematically evaluate its performance on certain operating measures to satisfy the requirements of collective and individual ethics. We consider the cases of a single drug and a combination of two drugs. Simulation results show that our new method works well overall in comparison with currently available designs, on criteria BEARS: Benchmark, Efficacy, Accuracy, Safety. Our new design avoids toxic doses while reliably identifying the maximum tolerated dose.

Nous proposons un nouveau plan d'expérience pour la méthode de réévaluation continue et évaluons systématiquement sa performance sur certaines mesures opérationnelles pour satisfaire aux exigences éthiques collectives et individuelles. Nous considérons le cas d'un seul médicament et celui d'une combinaison de deux médicaments. Les résultats de simulation montrent que notre méthode fonctionne bien globalement en comparaison avec les plans d'expérience actuellement disponibles, selon les critères de Beers: référence, efficacité, précision et sécurité. Notre nouveau plan d'expérience évite les doses toxiques tout en identifiant de manière fiable la dose maximale tolérée.

Van Bussel, Melissa; Castel Sophie; Burr, Wesley

Time Series Interpolation Algorithms: An Application to Real-World Data
Algorithmes d'interpolation pour séries chronologiques : Application à des données réelles

The analysis of complex scientific data observed in the form of time series often uses the power spectrum as an exploratory tool. Robust estimators of this statistic have existed for some time, but typically require that the data set be contiguous, that is, without any missing observations. This presents a problem for many data sets, as observations can be missing for a number of reasons: instrumentation error or fault, data corruption, or observational concerns such as interrupted vision of the observational unit (e.g., satellites losing data coverage due to cloud cover). Interpolators for time series aim to repair the original scientific data by inserting estimated values for the missing quantities. In this talk, we will examine the computational and performance results for a number of modern interpolation algorithms, as applied to various real-world datasets. We conclude with recommendations for interpolator choice based on the structure of the data of interest.

L'analyse de données scientifiques complexes observées sous forme de séries chronologiques fait souvent appel au spectre de puissance comme outil d'exploration. Des estimateurs robustes de cette statistique existent depuis un certain temps, mais ils requièrent habituellement que le jeu de données soit contigu, c'est-à-dire qu'il n'y ait pas de valeurs manquantes. Ceci constitue un problème pour plusieurs bases de données, puisque des observations peuvent être manquantes pour une foule de raisons: erreurs ou défauts dans la prise de mesure, corruption des données, ou problèmes d'observation tels que l'interruption dans l'observation de l'unité, p.ex. un satellite perdant le signal visuel dû au couvert nuageux. Les interpolateurs pour les séries chronologiques visent à arranger les données scientifiques originales en insérant des valeurs estimées pour les quantités manquantes. Dans cette présentation, nous examinerons les résultats calculatoires et la performance d'un certain nombre d'algorithmes d'interpolation modernes, appliqués à divers jeux de données réelles. Nous concluons avec des recommandations pour le choix d'un interpolateur basé sur la structure des données choisies.

Causal inference • Inférence causale

08:50am - 09:35am, 1405B, Thai-Son Tang (Chair • Président)
thaison.tang@mail.utoronto.ca

Ghosh, Trinetri ; Ma, Yanyuan ; Luna, Xavier de
Sufficient dimension reduction for feasible and robust estimation of average causal effect
Réduction suffisante de la dimensionnalité pour une estimation réalisable et robuste de l'effet causal moyen

When estimating the treatment effect in an observational study, we use a semiparametric locally efficient dimension reduction approach to assess both the treatment assignment mechanism and the average responses in both treated and nontreated groups. We then integrate all results through imputation, inverse probability weighting and doubly robust augmentation estimators. Doubly robust estimators are locally efficient while imputation estimators are super-efficient when the response models are correct. To take advantage of both procedures, we introduce a shrinkage estimator to automatically combine the two, which retains the double robustness property while improving on the variance when the response model is correct. We demonstrate the performance of these estimators through simulated experiments and a real dataset concerning the effect of maternal smoking on baby birth weight.

Pour l'estimation de l'effet du traitement dans une étude observationnelle, nous utilisons une approche de réduction de dimensionnalité semi-paramétrique et efficace localement. Elle permet l'évaluation du mécanisme d'assignation du traitement et de la réponse moyenne dans les groupes traités et non-traités. Par la suite, nous intégrons tous les résultats à l'aide d'estimateurs par imputation, par pondération selon la probabilité inverse, et par augmentation doublement robustes. Les estimateurs doublement robustes sont efficaces localement, tandis que les estimateurs d'imputation sont super-efficaces si les modèles pour la réponse sont corrects. Afin de tirer avantage des deux approches, nous présentons un estimateur de rétrécissement (shrinkage estimator) les combinant automatiquement. Il préserve la propriété de double robustesse et améliore la variance quand le modèle pour la réponse est correct. Nous démontrons la performance de l'estimateur à partir d'expériences simulées et via l'analyse d'un véritable jeu de données traitant de l'effet du tabagisme maternel sur le poids des enfants à la naissance.

Kuan Liu; Olli Saarela; Eleanor Pullenayegum

On Bayesian estimation of causal effect with a latent confounder class
Estimation bayésienne d'un effet causal en présence d'une classe de confusion latente

Despite the practicality, observational studies are subjected to selection and confounding bias and often require all confounders to be measured and controlled to infer casual relationship. In practice, it's difficult to ensure and assume all confounders were captured in the data. We consider a causal effect that is confounded by an unobserved latent confounder class. This latent class can be viewed as the unobserved augmented disease-risk/comorbidity profile that functions as a confounder. The observed covariates, instead of being treated directly as confounders, are categorized into two groups: one predicts the latent class (class predictors) and one manifested from the latent class (class indicators). We assume the unobserved latent class 1) captures the true confounding information, 2) can be sufficiently identified (modeled) given the measured covariates and 3) determines both the treatment and outcome process. Furthermore, conditioning on the latent class, treatment assignment is independent of the potential outcomes, which permits a full Bayesian parameterization of the joint distribution of the treatment model, outcome model and the latent class model. Our proposed causal problem is appealing - it features dimension reduction of the measured covariates through modeling the underlying patient augmented confounding in a latent class analysis. The objective of this presentation is to present the proposed causal problem, share existing literature in causal inference with unmeasured (latent) confounder and discuss the planned Bayesian estimation.

Malgré leur aspect pratique, les études observationnelles sont sujettes à du biais de sélection et de confusion. Pour l'inférence d'un lien causal, elles requièrent souvent que tous les facteurs de confusion soient mesurés et contrôlés. En pratique, il est difficile de s'assurer que tous les facteurs de confusion ont été enregistrés. Nous considérons dans cette étude un effet causal confus par une classe de confusion latente non observée. On peut considérer cette classe latente comme le profil non observé et augmenté de comorbidité et de risque de maladie, qui agit en tant que facteur de confusion. Nous ne traitons pas directement les covariables observées comme de simples facteurs de confusion. Nous les subdivisons plutôt en deux groupes: l'une comprend les prédicteurs de classe, et l'autre comprend les effets découlant de la classe (indicateurs de classe). Nous assumons que la classe latente non observée reflète l'information véritable de confusion, qu'elle puisse être suffisamment identifiée (modélisée) à l'aide des covariables mesurées, et enfin, qu'elle détermine autant le processus de traitement que celui de réponse. De plus, en conditionnant sur la classe latente, on obtient que l'assignation du traitement est indépendante des réponses potentielles. Ceci permet une paramétrisation bayésienne complète de la distribution conjointe du modèle de traitement, de réponse et de classe latente. Le problème causal que nous abordons est attrayant: il implique une diminution de la dimensionnalité des covariables mesurées à travers la modélisation, par une analyse de classe latente, de la confusion augmentée sous-jacente pour le patient. Cette présentation vise à expliquer le problème causal proposé, faire connaître la littérature existante en inférence causale en présence d'un facteur de confusion (latent) non mesuré, et mettre en lumière l'estimation bayésienne imaginée.

Chakraborty, Shubhadeep ; Zhang, Xianyang

Distance Metrics for Measuring Joint Dependence with Application to Causal Inference
Mesures de distance pour mesurer la dépendance conjointe avec application à l'inférence causale

Many statistical applications require the quantification of joint dependence among more than two random vectors. In this work, we generalize the notion of distance covariance to quantify joint dependence among $d \geq 2$ random vectors. We introduce the high order distance covariance to measure the so-called Lancaster interaction dependence. The joint distance covariance is then defined as a linear combination of pairwise distance covariances and their higher order counterparts which together completely characterize mutual independence. We further introduce some related concepts including the distance cumulant, distance characteristic function, and rank-based distance covariance. Empirical estimators are constructed based on certain Euclidean distances between sample elements. We study the large sample properties of the estimators and propose a bootstrap procedure to approximate their sampling distributions. The asymptotic validity of the bootstrap procedure is justified under both the null and alternative hypotheses. The new metrics are employed to perform model selection in causal inference, which is based on the joint independence testing of the residuals from the fitted structural equation models. The effectiveness of the method is illustrated via both simulated and real datasets.

Plusieurs applications statistiques nécessitent de quantifier la dépendance conjointe d'un ensemble formé de plus de deux vecteurs aléatoires. Dans nos travaux, nous généralisons la notion de covariance de distance pour quantifier la dépendance conjointe d'un ensemble formé de $d \geq 2$ vecteurs aléatoires. Nous introduisons une distance de covariance d'ordre supérieur pour mesurer la soi-disant dépendance d'interaction de Lancaster. La covariance de distance conjointe est ainsi définie comme une combinaison linéaire des distances de covariance deux-à-deux et de leurs homologues d'ordre supérieur, qui lorsque combinés caractérisent complètement la dépendance mutuelle. Ensuite, nous introduisons certains concepts connexes, comme le cumulant de la distance, la fonction caractéristique de la distance, et la covariance de distance basée sur le rang. Des estimateurs empiriques sont construits à partir de certaines distances euclidiennes entre les éléments de l'échantillon. Nous étudions les propriétés de ces estimateurs pour des échantillons de grande taille, et nous proposons une procédure bootstrap pour approximer leur distribution d'échantillonnage. La validité asymptotique de la procédure bootstrap est justifiée sous les hypothèses nulle et alternative. Les nouvelles mesures sont utilisées pour sélectionner un modèle en inférence causale, celui-ci basé sur un test d'indépendance conjointe des résidus tiré de l'ajustement d'un modèle d'équations structurelles. L'efficacité de la méthode est illustrée à l'aide de données simulées et réelles.

Model selection • Sélection de modèle

09:40am - 10:25am, 1405B, Thai-Son Tang (Chair • Présidente)
thaison.tang@mail.utoronto.ca

Kiwon, Francis; Nolet-Gravel, Eloise; Jalbert, Jonathan
Frequentist Model Averaging Estimator of Support Vector Machine Classfiers and Regressors
Estimateur par combinaison de modèles fréquentistes pour les classificateurs et régresseurs par machines à vecteurs de support

With its capability of reducing model variance while retaining bias, model averaging presents a captivating alternative to model selection for tackling model uncertainty. We propose an application of the Mallows model averaging (MMA) technique suggested by Hansen (2007), which is based on minimizing the Mallows criterion to a set of support vector machines (SVM), for both classification and regression. We compared mean squared error (MSE) of MMA estimator with those of models averaged or selected based on the other information criteria designed for SVM models by Vapnik (1982) and Claeskens (2008). Although there is no single dominant approach over a range of sample sizes and signal-to-noise ratios, from a minimax viewpoint not only is model averaging shown to be more competitive than model selection, but my MMA estimator is competitive among its peer model averaging methods in terms of MSE, especially with smaller sample sizes and larger signal-to-noise ratios. Theoretical underpinnings and an illustrative application are also presented.

Grâce à sa capacité à réduire la variance du modèle tout en limitant le biais, la combinaison de modèles (model averaging) présente une alternative captivante à la sélection de modèles classique. Nous proposons une application de la technique de combinaison de modèles de Mallows (CMM), basée sur la minimisation du critère de Mallows sur un ensemble de machines à vecteurs de support (MVS), à la fois pour la classification et pour la régression. Nous avons comparé l'erreur quadratique moyenne (EQM) de l'estimateur CMM avec celle des combinaisons de modèles ou des sélections de modèles basés sur d'autres critères d'information conçus pour les modèles de MVS. Bien qu'il n'y ait pas une seule approche dominante pour une gamme de tailles échantillonnales et de ratio signal sur bruit, d'un point de vue minimax, la combinaison de modèles s'avère plus performante que la sélection de modèles. En termes d'EQM, elle se compare également avantageusement aux méthodes par combinaison de modèles analogues, en particulier pour des échantillons plus petits et pour des ratios signal sur bruit plus élevés. Des fondements théoriques et une application illustrée sont également présentés.

McNichol, Jennifer

Outlier Detection Methods for Quantitative Fatty Acid Signature Analysis
Méthodes de détection des valeurs aberrantes pour l'analyse quantitative de la signature en acides gras

Quantitative Fatty Acid Signature Analysis (QFASA) is an effective method in the field of ecology used to estimate the proportion of each species in a predator's diet. QFASA uses statistical distances to estimate predator diets using fatty acid signatures obtained from potential prey species and that of the predator. A potential downfall of QFASA is that the mean fatty acid signature of the prey species are used, which may be highly influenced by outliers and lead to inaccurate predator diet estimates. We will explore some outlier detection methods suitable for compositional data, and more specifically for use with QFASA, in order to determine how the removal of outliers in the prey fatty acid signatures affects the predator diet estimates.

L'analyse quantitative de la signature en acides gras (AQSAG) est une méthode dans le domaine de l'écologie utilisée pour estimer efficacement la proportion de chaque espèce dans le régime alimentaire d'un prédateur. L'AQSAG emploie des distances statistiques afin d'estimer le régime alimentaire des prédateurs à l'aide de signatures en acides gras obtenues à partir des proies potentielles et du prédateur. Un problème potentiel de l'AQSAG provient de l'utilisation de la signature moyenne en acides gras des proies, qui peut être fortement influencée par des valeurs aberrantes et ainsi, conduire à des estimés imprécis. Nous explorerons quelques méthodes de détection des valeurs aberrantes adaptées aux données relatives à la composition, et plus particulièrement destinées à l'utilisation avec l'AQSAG. Nous chercherons ainsi à déterminer dans quelle mesure l'élimination des valeurs aberrantes dans les signatures en acides gras des proies affecte les estimés du régime des prédateurs.

Hoque, Md Erfanul; Acar, Elif
Predictive Comparison of Vine Copula Models
Comparaison prédictive de modèles de copules en vignes

Vine copulas are a popular tool for flexible and tractable specification of high dimensional joint densities for representing multivariate data. There are various vine constructions such as D-Vines, C-Vines and more general R-Vines. In recent years, these models have been considered in regression contexts to predict conditional mean and conditional quantiles of a variable of interest given the other variables. In this work, we compare the predictive performance of various vine constructions models and evaluate the predictive utility of vine copula regression over classical regression models through simulation studies and real data applications.

Les copules en vignes sont un outil populaire pour la spécification flexible et accommodante de densités conjointes à haute dimension afin de représenter des données multivariées. Il existe différentes constructions de vignes telles que les vignes-D, vignes-C et plus généralement les vignes-R. Récemment, ces modèles ont été pris en compte dans des contextes de régression pour prédire la moyenne conditionnelle et les quantiles conditionnels d'une variable d'intérêt étant donné d'autres variables. Dans ce travail, nous comparons les performances prédictives de différents modèles de construction de vignes et évaluons l'utilité prédictive de la régression de copule en vignes par rapport aux modèles de régression classiques au moyen d'études de simulation et d'applications à des données réelles.

Bayesian statistics • Statistique Bayesienne

09:40am - 10:25am, O1500, Victoire Michal (Chair • Présidente)
victoire.michal@umontreal.ca

Zahan, Rifat; Feng, Cindy
Bayesian spatial logistic regression model for investigating socio-economic and demographic determinants and geographic variation of pregnancy termination among Bangladeshi women
Modèle de régression logistique spatiale bayésienne pour étudier les déterminants socio-économiques et démographiques, et la variation géographique de l'interruption de la grossesse chez les Bangladaises

Unsafe pregnancy termination is a major public health concern in many developing countries. This study evaluates the role of socio-economic, demographic factors and residual spatial correlation in pregnancy termination among Bangladeshi women. Using Bayesian spatial logistic regression with Integrated Nested Laplace Approximation, we flexibly modeled the non-linear effects of continuous covariates in addition to spatial correlation accounting for the unobserved heterogeneity among individuals living in the same region. Our findings indicate that incorporation of non-linear covariate effects as well as spatial random effects leads to better fitting models and improved statistical inference.

L'interruption de la grossesse non sécuritaire est un problème de santé publique majeur dans plusieurs pays en voie de développement. Cette étude évalue le rôle de facteurs socio-économiques et démographiques, et de la corrélation spatiale résiduelle dans l'interruption de la grossesse chez les femmes bangladaises. En utilisant la régression logistique spatiale bayésienne avec l'approximation imbriquée intégrée de Laplace (INLA), nous avons modélisé de manière flexible les effets non linéaires de covariables continues, en plus de la corrélation spatiale, rendant compte de l'hétérogénéité non observée parmi les individus vivant dans la même région. Nos résultats indiquent que l'incorporation d'effets non linéaires de covariables ainsi que d'effets aléatoires spatiaux mène à de meilleurs modèles d'ajustements et à l'amélioration de l'inférence statistique.

Zhang, Kangjie; Liu, Juxin; Liu, Yang; Zhang, Peng
Measurement error adjustment in a zero-inflated Poisson model
Ajustement pour l'erreur de mesure dans un modèle de Poisson à inflation de zéro

Car crashes are the leading cause of death among teenagers. Our study is motivated by a traffic study on the effect of Graduated Driver Licensing program in Michigan analyzed in Chen et al (2014). Because teenager driver population at county level is not available, Chen et al (2014) used total teenager population in the state level to be a proxy for teenager driver population. We propose including a measurement error model to account for the difference between the teenager population and teenager driver population. To accommodate for the excess amount of zeros, we fit a zero-inflated Poisson model with spatially dependent random effects. To check whether the proposed method is working well, we conduct simulation studies. Both data simulation and the Bayesian MCMC sampling are implemented in rstan.

Les accidents de la route représentent la première cause de décès parmi les adolescents. Notre étude est motivée par une étude de circulation sur les effets du programme de permis de conduire progressif implanté au Michigan, analysés par Chen et al (2014). Puisque la population des conducteurs adolescents par comté n'était pas disponible, Chen et al (2014) ont utilisé la population d'adolescents de l'état en tant que substitut de la population des conducteurs adolescents. Nous proposons d'inclure un modèle d'erreur de mesure pour tenir compte de la différence entre la population d'adolescents et celle des conducteurs adolescents. Pour tenir compte de la quantité excessive de zéros, nous ajustons un modèle de Poisson à inflation de zéro avec effets aléatoires spatialement dépendants. Pour vérifier si la méthode proposée fonctionne bien, nous menons des études de simulation. La simulation des données et l'échantillonnage bayésien de type MCMC sont implémentés en RStan.

McDonald, Shaun; Campbell, David; Leblanc, Alexandre; Muthukumarana, Saman
Bayesian Approaches to Density Estimation for Use in Functional Linear Regression
Approches bayésiennes pour l'estimation de densités pour l'utilisation en régression linéaire fonctionnelle

The theory behind functional linear regression (FLR) in the Bayesian framework is fairly well-established and straightforward: given discrete data, we estimate predictor functions as linear combinations of some basis functions with a smoothness-imposing prior on the coefficients. Scalar outcomes can then be regressed with inner products between these predictors and an unknown regression function. Of interest is the case where our functional covariates are probability densities. However, many theoretical results and desirable properties for Bayesian FLR assume unconstrained predictor functions. Conversely, much of the literature focuses on estimation of densities as the end goal, representing them in forms that would make subsequent inner product calculation difficult in MCMC runs. Here we try to bridge the gap, seeking Bayesian methods to estimate densities that are amenable to being folded into FLR. We compare multiple approaches, using Hamiltonian Monte Carlo as implemented in Stan.

Dans le cadre bayésien, la théorie concernant la régression linéaire fonctionnelle (RLF) est assez bien établie et simple: étant donné des données discrètes, nous estimons les fonctions des prédicteurs sous forme de combinaisons linéaires de certaines fonctions de base avec une densité à priori pour les coefficients imposant un certain lissage. Les résultats scalaires peuvent ensuite être régressés avec des produits internes entre ces prédicteurs et une fonction de régression inconnue. Le cas où les covariables fonctionnelles sont des densités est digne d'intérêt. Cependant, de nombreux résultats théoriques et propriétés souhaitables pour la RLF bayésienne supposent des fonctions de prédiction sans contraintes. À l'inverse, une grande partie de la littérature se concentre sur l'estimation de densités comme objectif final, en les représentant sous des formes qui rendent difficile le calcul ultérieur du produit intérieur dans les exécutions de l'algorithme MCMC. Nous essayons ici de combler le fossé, en cherchant, pour l'estimation de densités, des méthodes bayésiennes susceptibles d'être intégrées à la RLF. Nous comparons plusieurs approches, en utilisant l'algorithme Monte Carlo Hamiltonien, tel qu'implémenté en Stan.

Biostatistics • Biostatistique

09:40am - 10:25am, Theatre 4, Michela Panarella (Chair • Président)
michela.panarella@mail.utoronto.ca

Lu, Hua; Binns, Malcolm; Gardner, Sandra
Variation Along Continuous Neuroelectric Activity Related to Early Cognitive Impairment
Variation le long de l'activité neuroélectrique continue liée à un trouble cognitif précoce

Patients with Mild Cognitive Impairment (MCI) can experience memory loss, concentration problems and judgement difficulties. The disease is hard to be recognized because it falls somewhere between the usual cognitive decline of normal aging and the more serious signs of dementia and Alzheimer's disease. To be diagnosed correctly, patients may need to go through process such as questionnaire followed by neurological exam, and brain imaging, and however, they can be invasive and not available for the population. The goal of this study is to use Functional Data Analysis (FDA) to capture the features in patients' brain activities through Electroencephalogram (EEG); with the characteristics obtained by functional form, MCI could be diagnosed in a easier way.

Les patients avec un trouble cognitif léger (TCL) peuvent subir une perte de mémoire, des problèmes de concentration et des difficultés de jugement. La maladie est difficile à reconnaître parce qu'elle se situe entre le déclin cognitif habituel dû au vieillissement normal et les signes plus sérieux de démence et d'Alzheimer. Pour être diagnostiqués correctement, les patients doivent parfois passer par un processus tel qu'un questionnaire suivi d'un examen neurologique et d'une imagerie cérébrale. Cependant, ces tests peuvent être invasifs ou non disponibles pour la population. L'objectif de cette étude est d'utiliser l'analyse de données fonctionnelles (Functional Data Analysis) pour saisir les caractéristiques des activités cérébrales des patients par l'électroencéphalogramme. À l'aide des caractéristiques obtenues par la forme fonctionnelle, le TCL pourrait être diagnostiqué plus facilement.

Cui,Jingyu; Ma, Renjun; Hasan, Tariq

Crossed random effects modelling of binomial data with random cluster sizes

La modélisation par effets aléatoires croisés des données binomiales avec des groupages de tailles aléatoires

Modelling of binary data with partially crossed random effects was introduced for a salamander mating experiment by McCullagh and Nelder (1989) as it 'is extremely important with a broad range of applications; however, to the best of our knowledge, crossed random effects modelling has not been developed in the literature when binary data are clustered with random cluster sizes. Our research is motivated by an animal fertilization study. Its objective is to evaluate the effect of food supplementation and the time of its administration on the number of oocytes and the number of viable oocytes of donor cows after fertilization with semen from different bulls. In this talk, we introduce a dual Poisson modelling approach to these binomial data with partially crossed donor random effects and bull random effects. We propose an orthodox best linear unbiased predictors approach to this model. Our analysis of animal fertilization data facilitates biological interpretations.

La modélisation de données binaires à l'aide d'effets aléatoires partiellement croisés a été tout d'abord proposée pour une expérience d'accouplement de salamandres par McCullagh et Nelder (1989), qui ont d'ailleurs noté "l'importance et la polyvalence" de la méthode. Toutefois, nous ne sommes pas au fait du développement de modèles à effets aléatoires croisés en présence de données binaires groupées, dont les grappes ont des tailles aléatoires. Notre recherche tire son origine d'une étude sur l'insémination d'animaux. Elle vise à évaluer l'effet de la supplémentation alimentaire et du moment de son administration sur le nombre total d'ovocytes et sur le nombre d'ovocytes viables provenant de vaches donneuses, obtenus après insémination avec le sperme de différents taureaux. Nous proposons une double approche de modélisation de Poisson pour ces données binomiales, comportant des effets aléatoires partiellement croisés pour les vaches donneuses et d'autres effets aléatoires pour les taureaux. Nous proposons également une approche orthodoxe qui permet d'obtenir pour ce modèle le meilleur prédicteur linéaire sans biais. Notre analyse des données d'insémination facilite la formulation d'une interprétation biologique.

Naqvi, S. Ali; King, Meagan T.M.; Champigny, Marc; DeVries, Trevor; Deardon, Rob; Barkema, Herman W.
Automated disease detection in dairy cattle using recurrent neural networks
Détection automatisée de maladies chez les bovins laitiers à l'aide de réseaux de neurones récurrents

Background: As automated milking systems (AMS) continue become more widely adopted in the dairy industry, the need for accurate, early automated disease detection becomes greater. Chemical sensors integrated into AMS are used to detect milk characteristics and identify possible cases of mastitis. Previous studies have developed predictive models for mastitis with varying degrees of success using subsets of these characteristics, while ignoring behavioral and other animal characteristics. Objective(s): To integrate all measurements from AMS to develop accurate mastitis detection models using recurrent neural networks; to determine the relative importance of variables and their effect on model performance; to identify previously undescribed relationships between characteristics or behaviours and health changes preceding diagnosis of clinical mastitis. Materials and methods: Detailed disease recording for individual animals were collected from 13 commercial AMS dairy herds in Ontario for the first 50 days of lactation. Animal and herd-level features that are not directly measured were generated using AMS measurements: incorporating refusals to determine the number of attempted visits, latency to exit milker, daily temperature, and number of cows per robot on the given day. Animal-level features may indicate behavioural changes that are difficult to detect manually, while dynamic or user-supplied herd-level features may affect behavior independently of health. Recurrent neural networks with varying numbers of long short-term memory (LSTM) cells were trained using animals with different lengths of disease episodes and evaluated using accuracy, sensitivity, and specificity. Results and conclusions: Models with 3 LSTM cells showed the best performance of those tested, with accuracy, sensitivity and specificity of 85%. These recurrent neural networks have improved performance compared to previous studies that use feed-forward neural networks for detection of mastitis. This suggests that LSTMs are able to capture temporal trends and patterns too complex to be represented by rolling averages and daily variances. Despite high performance based on those metrics, disease episodes were often identified up to two weeks prior to actual diagnosis. Although the daily performance of the model was good, cases where animals are identified as sick earlier are still contributing to the false-positive rate, while the reason for this misclassification may just be that the model is detecting disease early. Sensitivity and specificity may not be the best metrics for evaluation of these models' performance, and different metrics may be required to develop models which detect disease more accurately and earlier.

Préambule: À mesure que les systèmes de traite automatisés (STA) continuent à être de plus en plus adoptés dans l'industrie laitière, le besoin d'une détection précise et automatisée des maladies devient de plus en plus important. Les capteurs chimiques intégrés aux STA sont utilisés pour détecter les caractéristiques du lait et identifier les cas possibles de mammite. Des études antérieures ont développé des modèles prédictifs de la mammite avec divers degrés de succès en utilisant des sous-ensembles de ces caractéristiques, tout en ignorant les caractéristiques comportementales et d'autres caractéristiques animales. Objectifs: Intégrer toutes les mesures des STA pour développer des modèles de détection de la mammite précis en utilisant des réseaux de neurones récurrents; déterminer l'importance relative des variables et leur effet sur la performance des modèles ; identifier des relations non décrites auparavant entre des caractéristiques ou comportements et les changements de santé précédant le diagnostic clinique de la mammite. Matériels et méthodes: Des enregistrements détaillés de la maladie ont été collectés dans 13 troupeaux de vaches laitières soumises aux STA en Ontario pendant les 50 premiers jours de lactation. Les caractéristiques au niveau de l'animal et au niveau du troupeau qui ne sont pas directement mesurées ont été générées à l'aide de mesures de STA: le temps de latence de sortie du trayeur, la température quotidienne et le nombre de vaches par robot pour un jour donné. Les caractéristiques au niveau de l'animal peuvent indiquer des changements de comportement qui sont difficiles à détecter manuellement, tandis que les caractéristiques au niveau du troupeau, dynamiques ou fournies par l'utilisateur, peuvent affecter le comportement indépendamment de la santé. Des réseaux de neurones récurrents avec un nombre variable de cellules à mémoire court-terme persistante (long short-term memory, LSTM) ont été entraînés en utilisant des animaux présentant des épisodes de la maladie de différentes durées et ont été évalués en termes de précision, de sensibilité et de spécificité. Résultats et conclusions : Les modèles avec trois cellules LSTM ont présenté les meilleures performances parmi ceux testés, avec une précision, une sensibilité et une spécificité de 85%. Ces réseaux de neurones récurrents ont amélioré les performances par rapport aux études précédentes ayant utilisé des réseaux de neurones en aval pour la détection de la mammite. Ceci suggère que les LSTM sont capables de capturer des tendances temporelles et des schémas trop complexes pour être représentés par des moyennes mobiles et des variances quotidiennes. Malgré une bonne performance d'après ces métriques, les épisodes de la maladie ont souvent été identifiés seulement jusqu'à deux semaines avant le diagnostic. Bien que les performances quotidiennes du modèle aient été bonnes, les cas où les animaux sont identifiés comme malades le plus tôt contribuent tout de même au taux de faux positifs, la raison de cette mauvaise classification étant peut-être que le modèle détecte la maladie à un stade précoce. La sensibilité et la spécificité ne sont peut-être pas les meilleures métriques pour l'évaluation de ces modèles, et différentes métriques peuvent être nécessaires pour développer des modèles qui détectent la maladie plus précisément et plus tôt.

Scientific abstracts: Posters • Résumés scientifiques: Posters

Sumeet, Kalia

Conducting causal inference in the presence of measurement bias using administrative databases
Inférence causale en présence de biais de mesure à partir de bases de données administratives

Measurement errors often arise in administrative databases due to rounding or faulty instrument. For example, the end-digit preference of systolic blood pressure in administrative databases requires appropriate methods to account for the underlying heaping behavior. The extent to which observational study design coupled with measurement errors associated with the primary outcome affect statistical inferences are not well understood. The goal of this research is to develop a novel statistical method and to evaluate its small-sample properties using a Monte-Carlo simulation study. The propensity scores are used to account for the systematic differences in baseline characteristics between the intervention and the control arm. This causal model assumes (i) no interference between potential outcomes and treatment; (ii) consistency of outcomes; (iii) no unmeasured confounding, (iv) positivity restrictions. Inverse-probability weights are used to account for the measurement errors in covariates under a marginal structural model. The application of this methodology is illustrated using a Canadian primary care database where the causal effect of beta-blockers on lowering systolic blood pressure is assessed among hypertensive patients.

Les erreurs de mesure sont fréquentes dans les bases de données administratives dues à l'arrondissement ou à des instruments fautifs. Par exemple, la préférence marquée pour certains caractères finaux dans les mesures de pression sanguine requiert l'utilisation de méthodes appropriées pour tenir compte du phénomène d'amoncellement qui en découle. L'effet sur l'inférence statistique des erreurs de mesure sur la variable de résultat primaire dans le contexte d'études d'observation n'est pas bien compris. Notre recherche vise à développer une nouvelle méthode statistique et d'évaluer sa performance pour des échantillons de petite taille à partir d'une étude de simulation Monte-Carlo. Les scores de propension sont utilisés pour tenir compte des différences systématiques entre les groupes témoin et intervention. Ce modèle causal suppose que (i) il n'y a aucune interférence entre les résultats potentiels et la variable traitement; (ii) les variables de résultat sont cohérentes; (iii) il n'y a aucun facteur de confusion non mesuré; (iv) les restrictions de positivité sont vérifiées. Des poids de probabilité inverse sont utilisés pour tenir compte des erreurs de mesure sur les covariables sous l'hypothèse d'un modèle d'équations structurelles. L'application de cette méthodologie est illustrée grâce à une base de données canadienne de soins de santé primaires, pour laquelle l'effet causal des bêtabloquants sur la baisse de la pression sanguine systolique est mesuré chez les patients souffrant d'hypertension.

Dyck, Justin; Torabi, Mahmoud

Spatio-Temporal Modelling of Ischemic Heart Disease in Manitoba
Modélisation spatio-temporelle de la cardiopathie ischémique au Manitoba

The proposed research is a population-based study regarding Ischemic Heart Disease (IHD) prevalence in Manitoba. This study's main objectives are to identify spatial patterns of the disease and also assess temporal trends in IHD prevalence. Prior research conducted by the Manitoba Centre for Health Policy (MCHP) has shown a marginal decrease in IHD prevalence from the 2002/2003-2006/2007 to the 2007/2008-2011/2012 time periods. Therefore, identifying whether this has been a consistent trend before and after these time periods, and consistent throughout all regions in Manitoba is the main aim of this study. To satisfy these objectives, a spatio-temporal model is in development to fit data obtained from MCHP for the years of 1995 to 2018. The algorithm for IHD detection is a data linkage from hospital abstracts, physician claims, and prescription drugs over a three-year period. These counts are then aggregated by year and into 96 small geographic areas. The model being developed is a Poisson generalized linear mixed model that has both spatial and temporal random effects terms to assess time and space patterns of IHD prevalence. A Bayesian approach to parameter estimation will be utilized with a Markov Chain Monte Carlo algorithm. Demographic indicators such as median household income, ethnicity, household education, and net-immigration will be incorporated into the model as covariates. To date there has been limited statistical modelling of IHD for Manitoba in the spatio-temporal context, so this study will add to the understanding of this prevalent chronic disease. Preliminary analysis of the data has shown a link between demographic indicators and IHD, so these are natural predictors of interest for the proposed model. Some spatial disease patterns have also been identified using a preliminary spatial model for the most recent years of the data, which confirms the need of a spatial consideration when modelling this data.

La recherche proposée est une étude populationnelle concernant la prévalence de la cardiopathie ischémique (CPI) au Manitoba. Les objectifs principaux de cette étude sont d'identifier les schémas spatiaux de la maladie et d'évaluer également les tendances temporelles de la prévalence de la CPI. Des études antérieures menées par le Centre d'élaboration de la politique des soins de santé du Manitoba (MCHP) ont montré une baisse marginale de la prévalence de la CPI entre les périodes de 2002/2003-2006/2007 et 2007/2008-2011/2012. Par conséquent, l'objectif principal de cette étude est de déterminer si la tendance avant et après ces périodes est constante et cohérente à travers les régions du Manitoba. Pour répondre à ces objectifs, un modèle spatio-temporel est en cours d'élaboration afin de modéliser les données obtenues du MCHP pour les années 1995 à 2018. L'algorithme de détection de la CPI est un couplage de données provenant de congés d'hôpitaux, de réclamations de médecins et de médicaments de prescription sur une période de trois ans. Ces dénombrements sont ensuite agrégés par année et en 96 petites régions géographiques. Le modèle en cours de développement est un modèle de Poisson mixte linéaire généralisé qui comporte des termes d'effets aléatoires spatiaux et temporels permettant d'évaluer les tendances temporelles et spatiales de la prévalence de la CPI. Une approche bayésienne pour l'estimation des paramètres sera utilisée avec un algorithme de Monte Carlo par chaînes de Markov. Des indicateurs démographiques tels que le revenu médian des ménages, l'appartenance ethnique, le niveau d'éducation au sein du ménage et l'immigration nette seront intégrés au modèle comme covariables. À ce jour, la modélisation statistique de la CPI au Manitoba est limitée au niveau spatio-temporel. Ainsi, cette étude permettra de mieux comprendre cette maladie chronique prévalente. L'analyse préliminaire des données a montré un lien entre les indicateurs démographiques et la CPI, qui sont donc des prédicteurs naturels d'intérêt pour le modèle proposé. Certains schémas spatiaux de la maladie ont également été identifiés à l'aide d'un modèle spatial préliminaire pour les années les plus récentes, venant confirmer le besoin de considérer la composante spatiale lors de la modélisation de ces données.

Jiang, Xuejing; Jiang, Depeng

Validation Study on a Screening Tool for Mental Health of Children and Youth in Canada
Étude de validation d'un outil de dépistage de la santé mentale des enfants et des jeunes au Canada

The Strength and Difficulties Questionnaire (SDQ) is a widely used screening tool for the emotion and behavioural difficulties among children and youth. The normative SDQ scores (i.e., cut-offs for normal, borderline and abnormal) have been well studies in other cultures such as the British and Danish, but not in Canada. In this study we will compare several different statistical methods (logistic regression, ROC curve, discriminant analysis) and traditional methods (Mean pm 2SD or percentiles) to identify the cut-off values. We will use Manitoba Grade 5 Health Survey data to illustrate how a correctly identified statistical technique for the cut-off values can improve the screening tool. The implication of these culture specific normative SDQ scores on the mental health program evaluation will be explored through the PAX Good Behavior Game (PAX) data.

Le Questionnaire sur les points forts et les points faibles (QPFPF) est un outil de dépistage largement utilisé pour les difficultés émotionnelles et comportementales des enfants et des jeunes. Les scores QPFPF normatifs (c'est-à-dire les seuils normaux, limites et anormaux) ont fait l'objet d'études approfondies dans d'autres cultures telles que la Colombie et le Danemark, mais pas au Canada. Dans cette étude, nous comparerons différentes méthodes statistiques (régression logistique, courbe ROC, analyse discriminante) et traditionnelles (moyenne pm 2SD ou percentiles) afin d'identifier les valeurs de seuil. Nous utiliserons les données du Sondage sur la santé mentale des élèves de 5e année du Manitoba afin d'illustrer comment une technique statistique correctement identifiée pour les valeurs seuils peut améliorer l'outil de dépistage. L'implication des scores QPFPF normatifs spécifiques à cette culture sur l'évaluation du programme de santé mentale sera explorée à travers les données de PAX Good Behavior Game (PAX).

Lin, Chuyuan; Yu, Ying; Wu, Yifan; Cao, Jiguo

Unsupervised Learning on Functional Data with Application to U.S. Weather Data
Apprentissage non-supervisé de données fonctionnelles avec application aux données météorologiques américaines

Unsupervised learning is the statistical technique that infers patterns in data without a target variable. One subclass of unsupervised learning is the problem of clustering, which has been developed to functional data analysis over decades. In this paper, we aimed to group U.S. states with similarity in weather forecast performance based on the functional form of data by applying clustering approaches. Besides reviewing the developed functional data clustering algorithms, we extended unsupervised random forest clustering method to functional data and detected its strengths and shortages compared with other clustering methods in simulation studies. Both developed and proposed clustering approaches were applied to U.S. weather data from 2014 to 2017. Through clustering, cluster-specific patterns were visually detected, and the cluster-to-cluster differences were quantified in order to identify the most and least predictable U.S. states.

L'apprentissage non supervisé est une technique statistique pour l'inférence de motifs parmi les données sans l'utilisation de variable cible. Le regroupement (clustering) est un sous-ensemble de l'apprentissage non supervisé. Il a été adapté aux données fonctionnelles au courant des dernières décennies. Dans cet article, à l'aide de méthodes de regroupement, nous tentons de subdiviser les états américains en fonction de la performance au niveau des prévisions météorologiques, en se basant sur la forme fonctionnelle des données. En plus de réviser les algorithmes de regroupement développés pour les données fonctionnelles, nous présentons une extension aux données fonctionnelles de la méthode de regroupement par forêt aléatoire non supervisée. Nous comparons ses forces et ses faiblesses à celles d'autres méthodes de regroupement à l'aide de données simulées. Les méthodes de regroupement proposées et développées ont été appliquées aux données météorologiques américaines pour une période allant de 2014 à 2017. Par l'entremise du regroupement, des motifs propres à chaque grappe ont été détectés visuellement, et les différences entre grappes ont été quantifiées afin d'identifier les états américains les plus et les moins prévisibles.

Ruan, Haoyao

Covariance-adjusted, sparse, reduced-rank regression with application to imaging-genetics data

Régression parcimonieuse, à rang réduit et ajustée pour la covariance, avec applications à des données d'imagerie et de génétique

Alzheimer's disease (AD) is one of the most challenging diseases in the world. There is neither cure for AD, nor a single treatment to prevent it. A various factors may alter the risk of developing AD, and among them genes are considered to play an important role. Therefore, it is crucial for researchers to explore the relationship between AD and genes. We use data from 179 cognitively normal (CN) individuals contain magnetic resonance imaging (MRI) measures of volume or thickness in 56 brain regions of interest (ROIs) and 510 single nucleotide polymorphisms (SNPs) obtained from 33 candidate genes, provided by the AD Neuroimaging Initiative (ADNI). We intend to prioritize the significant SNPs to be further investigated in the future. However, utilizing standard linear regression models that assume independence of observations is inappropriate in this research question, because they cannot account for the spatial correlation between brain regions and lack of enough number of subjects in the imaging-genetics data. Thus, we study the Cov-sRRR method that simultaneously perform variable selection and covariance estimation for high-dimensional data, and evaluate its feasibility to the imaging-genetics data.

La maladie d'Alzheimer (MA) représente un défi considérable: il n'existe aucun remède, et on ne connaît pas de traitement unique pour la prévenir. Plusieurs facteurs affectent le risque de développer la MA. Parmi eux, les gènes auraient une influence prépondérante. Nous utilisons des données tirées de 179 individus normaux sur le plan cognitif. Elles contiennent des mesures d'imagerie par résonance magnétique du volume et de l'épaisseur de 56 régions du cerveau, et de l'information sur 510 polymorphismes mononucléotidiques (SNP) obtenue à partir de 33 gènes candidats. Ces données ont été fournies par le AD Neuroimaging Initiative. Nous comptons prioriser les SNP significatifs pour des analyses subséquentes. Les modèles standards de régression linéaire assument l'indépendance des observations, et sont donc inadéquats pour l'analyse de ces données. En effet, ils sont incapables de refléter la corrélation spatiale entre les régions du cerveau et de compenser le nombre réduit de sujets dans les données d'imagerie et de génétique. Nous proposons donc la méthode Cov-sRRR, qui réalise simultanément la sélection de variables et l'estimation de la covariance pour des données à haute dimensionnalité. Nous évaluons par la suite son applicabilité aux données d'imagerie et de génétique.

Macaraig, Miguel; Dryden, Kaitlyn

The Effect of Pace on the Performance of the Closers in the Kentucky Derby
L'effet du rythme sur la performance des closers au Kentucky Derby

The Kentucky Derby is the longest held sporting event in America, and is often referred to as äóìThe Most Exciting Two Minutes in Sports." In horse racing, the closers or the horses that are coming from the back of the pack excite most viewers. This study attempts to establish the effect of early speed to the performance of closers in the Kentucky Derby based from Kentucky Derby races from 1933 to 2017. The study is designed such that effect of pace is isolated from other factors of the race and is analyzed through different linear models. The results of the study are also used to emphasize the greatness of the late Secretariat, who is considered as one of the greatest horse that ever lived.

Le Kentucky Derby est l'événement sportif le plus ancien en Amérique. On l'appelle communément "les deux minutes les plus excitantes dans le monde du sport". Dans le domaine de la course de chevaux, les closers, les chevaux qui proviennent de la queue du peloton, sont les plus captivants pour l'auditoire. Cette étude tente d'établir l'effet de la vitesse en début de course sur la performance des closers au Kentucky Derby. Les données ont trait aux courses du Kentucky Derby entre 1933 et 2017. Nous avons conçu l'étude pour isoler l'effet de l'allure (pace) de l'effet des autres facteurs de la course. Nous modélisons cet effet à l'aide de modèles linéaires. Les résultats de l'étude mettent l'accent sur l'excellence du défunt Secretariat, considéré le plus grand cheval de course de l'histoire.

Liu, Yixiu; Jiang, Depeng; St John, Philip D; Tate, Robert B

Joint Models of Longitudinal and Time-to-event Data: Impact of Data Collection Cycles
Modèles joints pour données longitudinales et de durée de vie: l'effet des cycles de collectes de données

There are many trade-offs to consider when choosing the frequency of data collection in a longitudinal study (e.g., biennial vs. triennial follow-up strategies). The objective of this study is to examine the impact of follow-up strategies on a study of the quality of life trajectories in older men. Data from the Manitoba Follow-up Study (MFUS) serve as an illustration. The MFUS is among the longest-running studies of health and aging in the world. By the year 2004, there remained 870 study members at a mean age of 83 years old, living across Canada with a geographic distribution similar to that of the national older adult population. This study focuses on the quality of life data collected from 2004 to 2015. Though the annual data for most live members available, we will also build the joint statistical models of the quality of life trajectories and survival outcomes just on the base of biennial data or triennial data. The trajectories of quality of life and its association with mortality derived from those different data collection strategies will be compared. The implication for researchers who are developing study protocol will be discussed.

Dans une étude longitudinale, plusieurs compromis s'imposent dans le choix de la fréquence de collecte des données, p.ex. un suivi biennal ou triennal. Le projet actuel vise à déterminer l'effet de la stratégie de suivi sur les conclusions d'une étude des trajectoires de qualité de vie chez des hommes âgés, basée sur les données du Manitoba Follow-up Study (MFUS). Le MFUS est l'une des études les plus anciennes traitant de la santé et du vieillissement. En 2004, on retrouvait dans l'échantillon 870 individus, dont l'âge moyen était de 83 ans, et dont la distribution géographique reflétait fidèlement celle de la population âgée du Canada. Notre projet se concentre sur les données collectées entre 2004 et 2015. À l'aide des données annuelles pour la plupart des membres vivants de l'échantillon, nous obtenons un modèle statistique conjoint pour les trajectoires de qualité de vie et les taux de survie. Nous comparons également l'association entre trajectoires de qualité de vie et mortalité, inférée sous différentes stratégies de collecte de données. Finalement, nous expliquons comment nos résultats pourraient affecter le développement de protocoles de recherche.

Ritchie, Robyn

Semi-parametric estimation of scoring rates in the English Premier League
Estimation semi-paramétrique de la cadence des tirs au but réussis dans le English Premier League

We study the estimation of scoring rates for teams in the English Premier League (EPL) under the assumption that goal times follow a weighted non-homogeneous Poisson process model with two components: a parametric component related to overall game scoring rates, and a nonparametric component trying to capture the variation in scoring patterns within the course of a complete game (including stoppage time). Many aspects of a soccer game and a teams' performance can be analysed by using scoring patterns, which are estimated using Bernstein polynomials, and used to improve different areas of the game (and possibly related to in-game strategies of teams). We use data collected over four soccer seasons of the EPL, including 1520 games and over 4000 goals, to look at a few specific questions involving team performances. In particular, we consider the problems of comparing the performances of different teams playing at home and on the road, in the first and second halves, and over different seasons.

Nous étudions l'estimation de la cadence des tirs au but réussis (scoring rates) des équipes de la English Premier League (EPL), sous l'hypothèse que le temps entre les buts peut être décrit adéquatement par un processus de Poisson non homogène pondéré avec deux composantes. La première composante est paramétrique et est en lien avec la cadence générale des tirs au but réussis. La seconde est non paramétrique et reflète la variation dans la cadence des tirs au but réussis tout au cours d'une partie, incluant le temps additionnel. Plusieurs aspects d'une partie de soccer et de la performance des équipes se prêtent à l'analyse à partir de la cadence des tirs au but réussis, estimée à l'aide de polynômes de Bernstein. Les estimés peuvent ensuite servir à améliorer certains aspects du jeu, possiblement en lien aux stratégies de jeu des équipes. Nous utilisons des données collectées pendant quatre saisons de la EPL, comprenant 1520 parties et plus de 4000 buts, pour répondre à quelques questions ayant trait à la performance des équipes. Plus particulièrement, nous comparons la performance des équipes entre les parties à domicile et à l'étranger, entre la première et la deuxième période, et entre les différentes saisons.

Wendlandt, Michael; McKinnon, Murdoch; Krishnamurthy, Ashok; Cobb, Loren
Spatial tracking of the current Ebola outbreak in Congo
Suivi spatial de l'épidémie d'Ebola au Congo

We present a spatial Susceptible-Exposed-Infectious-Recovered-Dead (S-E-I-R-D) compartmental model of epidemiology to capture the transmission dynamics and the spatial spread of the ongoing Ebola outbreak in the eastern region of Kivu in Congo. For the current outbreak in Congo we use registered data (province-wide weekly counts of total Ebola cases and confirmed dead) up to Jan 18, 2019 from the World Health Organization (WHO) situation reports. Data Assimilation is a general class of techniques for tracking a state vector in time, using Bayesian updates applied to a dynamic model. Our results for the 2013-16 West African Ebola outbreak suggest that forecasting incidence using data assimilation can be produced in the domain of quantitative tracking of an epidemic across space and time (Krishnamurthy and Cobb, GEOMED, 2015). We observed that the prediction improves as data is assimilated over time. The data assimilation layer receives sparse and error-prone epidemiological data from the field and uses this data to perform corrections to the current state vector of the epidemic. In other words, it enhances the operation of the spatial SEIHRD model by periodically executing a Bayesian correction to the state vector(s), in a way that is, at least arguably, robust and statistically optimal. The projected number of newly infected and death cases up to August 31, 2019 are estimated and presented. We provide a discussion and interpretation of our results. The data assimilation method presented herein can be applied to a large class of compartmental or even agent-based models.

Nous présentons un modèle spatial compartimental Susceptible - Exposé - Infectieux - Remis - Mort (SEIRM), qui vise à refléter la dynamique de transmission et la propagation spatiale d'Ebola dans la région du Kivu, au Congo. Nous utilisons des données officielles, soit le décompte hebdomadaire par province du nombre total de cas d'Ebola et de morts confirmées, dressé par l'Organisation mondiale de la santé (OMS) jusqu'au 18 janvier 2019. L'assimilation des données est une famille générale de techniques permettant de suivre un vecteur d'états dans le temps, basées sur des mises à jour bayésiennes appliquées à un modèle dynamique. Les résultats obtenus pour l'épidémie d'Ebola en Afrique de l'Ouest de 2013-2016 nous portent à croire qu'il est possible de prévoir l'incidence de telles épidémies suivies dans l'espace et dans le temps grâce à l'assimilation de données. Nous observons une amélioration de la prédiction à mesure que les données sont assimilées avec le temps. L'étape d'assimilation de données intègre des données épidémiologiques clairsemées et imprécises collectées sur le terrain. À partir de cette information, elle apporte des corrections au vecteur d'états de l'épidémie. Autrement dit, elle améliore l'application du modèle spatial SEIRM en réalisant périodiquement une correction bayésienne du vecteur d'états, de manière robuste et optimale sur le plan statistique. Nous présentons des estimés du nombre de nouveaux cas et de décès jusqu'au 31 août 2019. La méthode d'assimilation des données présentée ici est applicable à une famille large de modèles compartimentaux et même à des modèles basés sur les agents.

Liu, Pengqi; Zhou, Julie

Computing R-optimal designs for multi-response regression models via interior point method
Calcul de schémas R-optimaux pour les modèles de régression à réponses multiples à l'aide d'une méthode du point intérieur

We study R-optimal designs for multi-response regression models. The R-optimality criterion aims to minimize the product of diagonal elements of the covariance matrix of the estimator of regression parameters. Several theoretical properties of R-optimal designs are derived, including the equivalence theorem, symmetry and scale invariance. We propose a numerical algorithm to construct R-optimal designs on discrete design spaces based on an interior point method, which is powerful to solve convex optimization problems with inequality constraints. The algorithm is flexible and can be applied to any linear/nonlinear multi-response regression model. Applications of R-optimal designs are presented.

Nous analysons des schémas R-optimaux pour les modèles de régression à réponses multiples. Le critère de R-optimalité vise à minimiser le produit des éléments sur la diagonale de la matrice de covariance pour l'estimateur des paramètres de régression. Nous dérivons plusieurs propriétés théoriques des schémas R-optimaux: le théorème d'équivalence, la symétrie et l'invariance au rééchelonnage. Nous proposons un algorithme numérique pour la construction de schémas R-optimaux sur des espaces schématiques (design spaces) discrets. L'algorithme en question, basé sur une méthode du point intérieur, est suffisamment puissant pour résoudre des problèmes d'optimisation convexe avec des contraintes d'inégalité. Il est également flexible et applicable à n'importe quel modèle de régression linéaire ou non linéaire à réponses multiples. Nous présentons finalement des applications des schémas R-optimaux.

Mackie, Alex; Andrews, Adelle

Investigating the relationship between temperature and the number of fatalities on Canadian highways using time series analysis.

Une étude sur la relation entre la température et le nombre de décès sur les autoroutes canadiennes à l'aide d'une analyse de séries chronologiques.

In 2014, transportation Canada released a report stating that 2014 not only saw fewer vehicle fatalities than 2013, but also saw the fewest fatalities in a year since they started collecting data. While it appears that we are seeing fewer fatalities, we are also seeing more extreme weather conditions due to global climate change. We wanted to investigate the relationship between vehicle fatalities and temperature in Canada. Using data obtained from the National Collision Database as well as the Government of Canada website, we conducted time series analysis using bicoherence to test the relationship between temperature and vehicle fatalities in Canada from 1999-2014. We then used our model to see if it had any predictive ability by utilizing the fatality and weather data from 2015 to 2016 and seeing if our model was able to capture the true values.

En 2014, Transports Canada a publié un rapport indiquant qu'en 2014, non seulement y avait-il eu moins de victimes de la route qu'en 2013, mais aussi que cela constituait le plus faible nombre de morts par année depuis le début de la collecte de données. Bien qu'il semble que nous observions moins de décès, nous assistons également à davantage de conditions météorologiques extrêmes dues au changements climatiques mondiaux. Nous cherchions à étudier la relation entre les accidents de la route et la température au Canada. À l'aide de données provenant de la Base nationale de données sur les collisions et du site Web du gouvernement du Canada, nous avons procédé à une analyse de séries chronologiques en utilisant la bicohérence afin de tester la relation entre la température et les accidents de la route au Canada de 1999 à 2014. Nous avons ensuite utilisé notre modèle afin de déterminer s'il possédait une capacité prédictive en nous servant des données de mortalité et de météo de 2015 à 2016 et en vérifiant si notre modèle était capable de capturer les vraies valeurs.

Hou-Liu, Jason; Browne, Ryan P

Parsimonious Gaussian Mixtures via Chimeral Clusters
Mélanges gaussiens parcimonieux via des grappes chimères

We present an unsupervised method of fitting a parsimonious Gaussian partial membership model with intercluster structure. As opposed to a finite mixture model where membership weights are assigned at the observation level, we propose a mixing mechanism at the parameter level. In this model, some clusters act as pure clusters with fully varying parameters while the remaining chimeral clusters assume a convex combination of pure cluster parameters. We observe and mitigate the presence of local minima during parameter estimation by applying stochastic expectation-maximization. Finally, we demonstrate the efficacy of the method on simulated and real-world datasets.

Nous présentons une méthode non supervisée d'ajustement d'un modèle gaussien parcimonieux partiel d'appartenance avec une structure inter-grappes. Contrairement à un modèle de mélange fini, où les poids des membres sont attribués au niveau de l'observation, nous proposons un mécanisme de mélange au niveau des paramètres. Dans ce modèle, certaines grappes agissent comme des grappes pures avec des paramètres complètement variables, tandis que les grappes chimères restantes supposent une combinaison convexe de paramètres des grappe pures. Nous observons et atténuons la présence de minimums locaux lors de l'estimation de paramètres en appliquant un algorithme espérance-maximisation stochastique. Finalement, nous démontrons l'efficacité de la méthode sur des jeux de données simulés et réels.

Smyrnova, Anna; Barnett, Tracie A.; Henderson, Mélanie; Mathieu, Marie-Eve; Kakinami, Lisa
Application of the Distributed Lag Models for Examining Associations Between the Built Environment and Obesity Risk in Children
Application de modèles à retards échelonnés pour examiner les associations entre l'environnement bâti et le risque d'obésité chez les enfants

Objective: The literature on the built environment (BE) associations with health is mixed. Incorrect geographic scale selection can bias estimates. The objective was to compare the Distributed Lag Models (DLM) to a linear regression model. Methods: Data were from the 1st (Mage=11.6) and 2nd (Mage=16.8 years) follow-ups of the QUALITY cohort (n=281). BE features included number of: (1) fast-food restaurants, (2) convenience stores, and (3) fitness facilities. DLMs with 100m ring-shaped areas up to 5km centered on the residential locations were created. Outcomes included age- and sex- adjusted BMI z-scores and minutes of physical activity (PA). Linear regression models used buffers of 500m, 750m and 1000m. Results: The number of fast-food restaurants was associated with an increase in BMI z-score up to 600m at F1 and up to 900m at F2. The number of convenience stores was associated at F2 only. No significant association between fitness facilities and PA were detected. The DLM-estimated associations and standard errors were smaller than with the linear regression models. Conclusion: Different distances of association for F1 and F2 were detected. As DLMs do not require pre-specified fixed buffer sizes, their use can help identify the appropriate spatial scale for a given population.

Objectif: La littérature traitant des associations entre les environnements bâtis (EB) et la santé contient des contradictions. La sélection d'une échelle géographique incorrecte peut biaiser les estimés. Cette étude vise à comparer les modèles à retards échelonnés (MRE) au modèle conventionnel de régression linéaire. Méthodes: Les données proviennent du premier (Mage= 11.6 ans) et du deuxième (Mage= 16.8 ans) suivi de la cohorte QUALITY (n = 281). Les caractéristiques de l'EB incluent le nombre d'établissements de restauration rapide, de dépanneurs et de centres sportifs. Des MRE avec des zones circulaires de 100m jusqu'à 5 km, centrées sur les emplacements résidentiels ont été créés. Les variables réponses comprennent les scores-z d'IMC ajustés pour l'âge et le sexe, ainsi que le nombre de minutes dédiées à l'activité physique. Les modèles de régression linéaire utilisent des zones tampons de 500 mètres, de 750 mètres et de 1000 mètres. Résultats: Le nombre d'établissements de restauration rapide était associé à une augmentation du score-z d'IMC jusqu'à 600 mètres selon les données du premier suivi, et jusqu'à 900 mètres selon celles du deuxième suivi. Nous avons observé une association avec le nombre de dépanneurs selon les données du deuxième suivi seulement. Nous n'avons détecté aucune association significative entre le nombre de centres sportifs et le degré d'activité physique. Les MRE ont produit des associations et des erreurs-type plus basses que celles estimées à partir des modèles de régression linéaire. Conclusion: Nous avons détecté une distance d'association différente entre le premier et le deuxième suivi. Puisque les MRE ne nécessitent pas de spécification a priori de la taille des zones tampons, ils peuvent aider à identifier l'échelle spatiale appropriée pour une population donnée.

Nolet-Gravel, Éloise; Jalbert, Jonathan

Projected changes of extreme rainfall in the province of Quebec
Changements projetés des précipitations extrêmes au Québec

According to the latest report of the Intergovernmental Panel on Climate Change (IPCC), the province of Quebec is part of the regions in the world where extreme rainfall events are most likely to increase substantially if the global warming goes from 1.5ºC to 2ºC (IPCC, 2018, p.33). Intensity-Duration-Frequency (IDF) curves of precipitation are the principal tool for dimensioning infrastructures exposed to climate hazards and for the definition of flood plains (Maillot, Duchesne, Caya, and Talbot, 2007, p.197). With climate changes, IDF curves are likely to evolve with time. That is not currently taken into account in the IDF curves provided by Environment and Climate Change Canada (ECCC). Therefore, the risk of failure of an infrastructure dimensioned with the actual curves is subject to evolve according to climate change. It becomes primordial to include their effects in the estimation of IDF curves to ensure infrastructures a certain security level throughout their useful life. The proposed project consists of including the effects of climate change in the IDF rainfall curves. To update these curves, a spatial and non-stationary model has been developed for the extreme precipitations simulated by a climate model. The developed methodology will take into account the spatial and non-stationary aspects of intense precipitations to develop IDF curves in future climate.

Selon le dernier rapport du Groupe d'experts intergouvernemental sur l'évolution du climat (GIEC), la province de Québec fait partie des régions mondiales où les précipitations extrêmes sont les plus susceptibles d'augmenter considérablement si le réchauffement de la planète passe de 1,5 ºC à 2 ºC (GIEC , 2018, p.33). Les courbes Intensité-Durée-Fréquence (IDF) des précipitations constituent le principal outil pour le dimensionnement des infrastructures exposées aux risques climatiques et pour la définition des plaines inondables (Maillot, Duchesne, Caya et Talbot, 2007, p.197). Avec les changements climatiques, les courbes IDF évolueront probablement avec le temps. Cela n'est présentement pas pris en compte dans les courbes IDF fournies par Environnement et Changement climatique Canada (ECCC). Par conséquent, le risque de défaillance d'une infrastructure dimensionnée avec les courbes réelles est susceptible d'évoluer en fonction du changement climatique. Il devient primordial d'inclure leurs effets dans l'estimation des courbes IDF pour assurer aux infrastructures un certain niveau de sécurité tout au long de leur vie utile. Le projet proposé consiste à inclure les effets du changement climatique dans les courbes IDF des précipitations. Pour mettre à jour ces courbes, un modèle spatial et non stationnaire a été développé pour les précipitations extrêmes simulées par un modèle climatique. La méthodologie développée tiendra compte des aspects spatiaux et non stationnaires des précipitations intenses afin de développer les courbes IDF pour les condition climatiques futures.

Liu, Meixin; Zhou, Julie; Dukes, Peter

Construction of Block Incomplete Design Under Correlated Error Structure
Construction de plans en blocs incomplets avec une structure d'erreurs corrélées

Consider an experimental situation where the experimenter wants to compare v treatments. Usually we use balanced incomplete block design to analyze the treatment effect using a fixed effect model, where we consider error independent. However, for some experiments, errors are correlated under specific error structure. So we investigate the construction of a binary block design under correlated error structure where comparison needed to be selected in a balance manner, that is the covariance matrix of least squared estimator to be completely symmetric. To help analysis of the criteria, we introduce graph as representation of block. We show that some particular graph can be used as block to obtain the covariance structure we want.

Dans le cadre d'une expérience où l'on vise à comparer des traitements, on utilise généralement un plan en blocs incomplet pour analyser l'effet de traitement à partir d'un modèle à effets fixes, dans lequel on considère l'erreur indépendante. Or, pour certaines expériences, l'erreur est corrélée selon une structure spécifique. Nous investiguons donc la construction d'un plan en blocs binaire avec une structure de l'erreur corrélée, où la comparaison doit être faite d'une manière balancée. En d'autres mots, la matrice de covariance de l'estimateur par moindres carrés doit être complètement symétrique. Pour aider à analyser ce critère, nous introduisons des graphes comme représentation des blocs. Nous démontrons que certains graphes peuvent être utilisés comme blocs pour obtenir la structure de covariance souhaitée.

Sharp, Alex; Browne, Ryan

Grading Gunslingers: A Preliminary Model For Evaluating Pitcher Types in Baseball - the continual reassessment method

Classement des tireurs d'élites: Modèle préliminaire pour évaluer les types de lanceurs au baseball - la méthode de réévaluation continue

We explore what it means to be an effective pitcher in baseball. We survey the currently used metrics used to evaluate a pitcher's success. We evaluate the significance of a pitcher's arsenal in relation to these metrics. We introduce a metric that includes the importance of the pitches with the quality of each pitch in that arsenal. We then augment this evaluation by exploring the significance of factors that are not directly related to the quality of the arsenal. Such factors include pitch sequence selection, performance under pressure, and pitch masking. These analyses are then combined into a general model which takes all of the relevant parameters into consideration and predicts the success metric for a given pitcher. Finally, we use this model to evaluate rookie pitchers, and attempt to predict their future effectiveness and team value.

Nous explorons ce que signifie être un lanceur efficace au baseball. Nous survolons les métriques présentement utilisées pour évaluer le succès d'un lanceur. Nous évaluons l'importance de l'arsenal d'un lanceur en termes de ces métriques. Nous introduisons une métrique qui incorpore l'importance des lancers avec la qualité de chaque lancer dans cet arsenal. Nous améliorons ensuite cette évaluation en explorant l'importance de facteurs n'étant pas directement reliés à la qualité de l'arsenal. Ces facteurs incluent la sélection de la séquence de lancers, la performance sous pression et le masquage du lancer. Ces analyses sont ensuite combinées dans un modèle général qui prend en considération tous les paramètres pertinents, et qui prédit la mesure du succès d'un lanceur donné. Finalement, nous utilisons ce modèle pour évaluer les lanceurs recrues, et nous tentons de prédire leur efficacité future et la valeur qu'ils apporteront à l'équipe.

Surjanovic, Nikola; Loughin, Thomas
Improving the Hosmer-Lemeshow Goodness-of-Fit Test
Amélioration du test d'ajustement de Hosmer-Lemeshow

Goodness-of-fit (GOF) tests help to identify when a model is a poor fit for a given dataset. One such test for logistic regression models is the Hosmer-Lemeshow test, which is used extensively in applications in many disciplines. Despite its common use, there is evidence that the test functions poorly under certain circumstances. We explore the origins of some of these properties and suggest a modification to the test based on these findings. Also, we discuss the performance of a new and parallel test for Poisson regression models, as well as other generalized linear models (GLMs).

Les tests d'ajustement (goodness-of-fit) aident à déterminer si un modèle s'ajuste mal à un certain jeu de données. Pour les modèles de régression logistique, le test de Hosmer-Lemeshow est communément utilisé, et ce, dans de multiples disciplines. Malgré son usage fréquent, on suspecte que le test fonctionne mal dans certaines circonstances. Nous explorons les origines de certaines de ses propriétés et suggérons une modification au test basée sur nos découvertes. Nous abordons également la performance d'un nouveau test analogue pour les modèles de régression de Poisson, de même que pour les autres modèles linéaires généralisés.

Cao, David; Velupillai, Nirudika

Understanding Edmonton's Weather: an Analysis of the Mean Temperature and Snowfall
Comprendre la météo d'Edmonton : une analyse de la température moyenne et des chutes de neige

In this work, we aim to study the average monthly temperature and snowfall in Edmonton, Alberta. Some of the methods used in arriving to a conclusion include fitting seasonal ARIMA models to both datasets, analyzing weather patterns and lastly, coming up with forecasts. Additionally, we aim to study the sample spectrum and wish to investigate the relationship between temperature and snowfall in Edmonton using measures like cross-correlation and bi-coherence.

Dans ce travail, nous visons à étudier la température mensuelle moyenne et les chutes de neige à Edmonton, Alberta. Les méthodes utilisées pour arriver à une conclusion incluent l'ajustement de modèles ARIMA saisonniers aux deux bases de données, l'analyse des conditions météorologiques et, finalement, la formulation de prévisions. De plus, nous cherchons à étudier le spectre de l'échantillon, ainsi que la relation entre la température et les chutes de neige à partir des mesures comme la corrélation croisée et la bi-cohérence.

Floyd, Kevin; Loughin, Thomas

Ain't Played Nobody: Building an Optimal Schedule to Secure an NCAA Tournament Berth
J'ai pas triché: Construction d'un calendrier optimal pour sécuriser une place au championnat de la NCAA

Each spring, American men's college basketball teams compete for the National Collegiate Athletics Association (NCAA) national championship, determined through a 68-team, single-elimination tournament known as "March Madness". Participants in this tournament either qualify automatically, through their conferences' individual year-end tournaments, or are chosen by a selection committee based on various measures of regular season success. This committee aims to select teams that prove themselves to be the strongest and most deserving both in the conference games that are scheduled in advance and nonconference games scheduled by teams themselves. Using historical data, we find the criteria that the committee values most when selecting teams for this tournament. Along with these criteria, we use prior seasons' success and projected returning players to forecast every team's strength for the upcoming season. Using the selection criteria and these projections, we develop a tool to help college basketball teams build the optimal nonconference schedule to increase the probability of being invited to the NCAA tournament.

Chaque printemps, les équipes masculines américaines de basketball universitaire se disputent le championnat national du National Collegiate Athletics Association (NCAA), déterminé par un tournoi à élimination simple à 68 équipes dénommé le "March Madness". Les participants de ce tournoi se qualifient soit automatiquement à travers le championnat de fin d'année de leur conférence, soit à travers un comité de sélection basé sur diverses mesures de succès de la saison régulière. Ce comité vise à sélectionner les équipes qui se sont montrées les plus fortes et les plus méritantes à la fois lors des parties de conférence planifiées à l'avance et lors des parties hors conférence planifiées par les équipes elles-mêmes. À l'aide de données historiques, nous trouvons les critères auxquels le comité attache le plus d'importance dans la sélection d'équipes pour ce tournoi. Parallèlement à ces critères, nous utilisons le succès des années précédentes et les retours projetés des joueurs pour prévoir la force de chaque équipe pour la saison à venir. En utilisant les critères de sélection et ces projections, nous développons un outil pour aider les équipes de basketball universitaire à bâtir un calendrier hors conférence optimal pour augmenter la probabilité d'être invité au tournoi NCAA.

Abubakari, Ibrahim Watara; Pahwa Punam; Khan, Shahed; Ahmed, Shahid, Karunanayake, Chandima; James, Dosman

Longitudinal Changes in Colorectal Cancer among Farm and Non-farm Rural Residents
Changements longitudinaux dans le cancer colorectal chez les résidents de zones rurales agricoles et non agricoles

Objective: To determine longitudinal changes in colorectal (CRC) prevalence and associated risk factors in farm and non-farm rural residents in Saskatchewan, Canada. Methods: Data from the Saskatchewan Rural Health Study (SRHS) were collected from 8,261 individuals nested within 4,624 households at baseline and 4,867 individuals within 2,797 households at follow-up. The study sample consists of 5,599 individuals at baseline and 3,933 at follow-up (¿= 50 years). Clustering effects of individuals within a household were accounted for using the generalized estimating equations (GEE) and repeated measurements using jackknife robust variance estimation. Multilevel marginal logistic regression models based on GEE were formulated to obtain the odds ratio (OR) for determining risk factors. Results: Prevalence of CRC decreased over time among farm (baseline: 3.1%; follow-up: 1.3%, p¡0.05), however increased among non-farm (baseline: 1.4%; follow-up: 2.0%, p¿0.05) residents. Individuals who spent their first year of life on farm had higher risks of developing CRC than those who did not (OR = 1.64, p¡0.05). Exposure to grain dust and radiation were significant (p¡0.05) determinants of longitudinal changes in CRC prevalence. Conclusion: Longitudinal changes in the CRC prevalence among farming and non-farming rural residents appear to depend on a complex combination of individual and contextual factors.

Objectif: Déterminer les changements longitudinaux dans la prévalence du cancer colorectal (CCR) et les facteurs de risque connexes chez les résidents de zones rurales agricoles et non agricoles en Saskatchewan, Canada. Méthodes: Les données de l'étude, portant sur la santé rurale en Saskatchewan, ont été collectées auprès de 8 261 individus imbriqués dans 4 624 ménages au début de l'étude, puis auprès de 4 867 individus dans 2 797 ménages au moment du suivi. Nous avons retenu 5 599 individus au début de l'étude et 3 933 au suivi (50 ans et plus). Les effets de grappes dus aux individus à l'intérieur d'un même ménage ont été pris en compte par l'intermédiaire des équations d'estimation généralisées (EEG) et les mesures répétées, par l'intermédiaire de l'estimateur robuste de la variance de type jackknife. Des modèles de régression logistique marginale à niveaux multiples basés sur les EEG ont été formulés pour obtenir les rapports de cotes (RC) permettant de déterminer les facteurs de risque. Résultats : La prévalence du CCR a diminué au fil du temps chez les résidents des fermes (début: 3,1%; au suivi: 1,3%, p ¡ 0,05), mais a augmenté chez les résidents des zones non agricoles (début: 1,4%; au suivi: 2,0%, p ¿ 0,05). Les individus ayant passé leur première année de vie sur une ferme couraient un risque plus élevé de développer un CCR que les autres (RC = 1,64, p ¡ 0,05). L'exposition à la poussière de grains et à la radiation étaient des déterminants significatifs (p ¡ 0,05) des changements longitudinaux de la prévalence du CCR. Conclusion : Les changements longitudinaux dans la prévalence du CCR parmi les résidents des zones rurales agricoles et non agricoles semblent dépendre d'une combinaison complexe de facteurs individuels et contextuels.

Shreeves, Phillip; Andrews, Jeffrey L.; Jirasek, Andrew

Semi-supervised nonnegative matrix factorization with applications to spectral data
Factorisation matricielle non négative semi-supervisée avec applications aux données spectrales

A commonly known technique in the world of unsupervised learning is nonnegative matrix factorization, a method used to decompose a data matrix X into two lower rank nonnegative matrices W and H. Herein, we develop a semi-supervised form of the method by further decomposing the H matrix into matrices A and S, where A is an auxiliary matrix and S is a matrix of bases that can be specified. We demonstrate the technique via application to medical physics data — specifically, Raman spectroscopy of irradiated cells. We specify matrix S to contain the pure spectra of basic chemicals commonly found in these cell samples, and investigate how the cells behave with respect to these chemicals following radiation.

Une techniques bien connue dans le monde de l'apprentissage non supervisé est la factorisation matricielle non négative, une méthode utilisée pour décomposer une matrice de données X en deux matrices non négatives W et H de rang inférieur. Ici, nous développons une forme semi-supervisée de la méthode en décomposant la matrice H en matrices A et S, où A est une matrice auxiliaire et S est une matrice de bases pouvant être spécifiées. Nous démontrons la technique via une application à des données de physique médicale – en particulier, à la spectroscopie de Raman de cellules irradiées. Nous spécifions la matrice S pour qu'elle contienne le spectre pur des produits chimiques de base que l'on retrouve couramment dans ces échantillons de cellules, et nous étudions comment les cellules se comportent vis-à-vis de ces produits chimiques suite à la radiation.

Zhe, Lu

Incremental value of AUC, average positive predictive value and Brier Score
Valeur ajoutée de l'ASC, de la valeur positive prédictive moyenne et du score de Brier

Decisions about whether to include a new biomarker in prognosis depends on how the new marker improves the model prediction performance. AUC, average positive predictive value (AP) and Brier score are all threshold-free proper scoring measures for model performance assessment. As they focus on different aspects of a model's performance, they are not always consistent with each other. We conduct simulation studies to investigate and compare the statistical properties of these three measures. Data were generated from a logistic model that includes two markers X1 (old marker) and X2 (new marker) as well as their interaction. Two logistic models were fit: (i) with X1 alone, and (ii) with only the main effects of X1 and X2. We quantify the incremental value of X2 in the form of ratios of AUC, AP, and Brier score from these two models. We found that the correlation of the incremental values in AP and Brier score is much higher, compared to the correlation of the incremental values in AUC and Brier score. Interestingly, under certain simulation settings, adding X2 makes the prediction worse in terms of AP and Brier, but AUC improves.

La décision d'inclure un nouveau biomarqueur dans un pronostic dépend de l'amélioration qu'il apporte à la performance du modèle de prédiction. L'ASC (aire sous la courbe), la valeur positive prédictive moyenne (VP) et le score de Brier ne dépendent pas d'une coupure numérique, et sont tous des mesures de pointage propres pour l'évaluation de la performance d'un modèle. Puisqu'ils se concentrent sur différents aspects de la performance d'un modèle, ils peuvent toutefois se contredire. Nous réalisons une étude de simulation afin de déterminer et de comparer les propriétés statistiques de ces trois mesures. Nous avons généré des données à partir d'un modèle logistique incluant deux marqueurs, X1 (l'ancien marqueur) et X2 (le nouveau), puis un terme d'interaction. Nous avons ajusté deux modèles logistiques, l'un avec X1 seulement, l'autre avec seulement les effets principaux de X1 et X2. Nous quantifions la valeur ajoutée de X2 à l'aide des ratios d'ASC, de VP et de score de Brier obtenus à partir des deux modèles. Nous avons remarqué que la corrélation pour les valeurs ajoutées de la VP et du score de Brier est beaucoup plus élevée, en comparaison avec celle obtenue pour l'ASC et le score de Brier. D'ailleurs, sous certains scénarios de simulation, l'ajout de X2 nuit à la prédiction en termes de la VP et du score de Brier, mais améliore tout de même l'ASC.