Fifth Annual Canadian Statistics Student Conference

•

La cinquième édition du congrès canadien des étudiants en statistique

University of Manitoba

Winnipeg, Manitoba

Saturday • Samedi
10 June • juin 2017

Contents

Welcome • Bienvenue

The Canadian Statistics Student Conference's main goal is simple: to bring together students and recent graduates from departments of statistics across Canada to network, learn about career and academic opportunities, and share their ideas and research with each other. A wide range of statistical interests are represented, including actuarial science, biostatistics, business and industrial statistics, probability, statistical education, and survey methods. In addition to providing the opportunity for students to present their work with their peers, the Student Conference features an impressive slate of invited speakers sharing their experiences in both industry and academia.

This year, several sessions also focus on the importance of collaboration in statistics, with the goal of equipping students with the tools required to successfully navigate this increasingly important facet of many careers in statistics. This year's program also features a statistical computing workshop led by **Dr. Chel Hee Lee**, University of Calgary and a keynote address by **Dr. Jiahua Chen**, 2014 SSC Gold Medalist.

The CSSC Student Conference organizing committee have been working hard this past year to prepare for this conference. On behalf of the organizing committee, we would like to extend a warm welcome to all attendees! Your annual attendance is greatly valued and we hope you thoroughly enjoy what this conference has to offer!

L'objectif du congrès étudiant de la CSSC est simple: amener les étudiants et les diplômés récents provenant de différents départements à travers le Canada à se rencontrer, à apprendre à propos des possibilités d'emplois et les opportunités académiques et à partager leurs idées et leur recherche avec d'autres. Une variété d'intérêts reliés ù la statistique seront représentés, incluant les sciences actuarielles, la biostatistique, les statistiques industrielles et de gestion, la probabilité, l'éducation en statistique et les méthodes d'enquête. En plus d'offrir aux étudiants et diplômés la possibilité de présenter leur travail, le congrès étudiant se verra présenter plusieurs conférenciers invités qui partageront leur expérience, tant dans le domaine industriel qu'académique.

Cette année, plusieurs sessions se concentreront aussi sur l'importance de la collaboration en statistique et ces sessions auront pour but de donner aux étudiants les outils nécessaires pour traiter de cette facette, de plus en plus importante dans plusieurs carrières en statistique. Le programme de cette année contient aussi un atelier sur le calcul statistique présenté par **Dr Chel Hee Lee**, de l'Université de Calgary, et un discours d'honneur par le **Dr Jiahua Chen**, récipiendaire de la médaille d'or de la SSC de 2014.

Le comité organisateur du congrès étudiant de la CSSC a travaillé fort dans la dernière année pour préparer cette conférence. De la part du comité organisateur, nous aimerions souhaiter la bienvenue à tous les participants! Votre participation annuelle est grandement appréciée et nous espérons que vous profiterez de tout ce que cette conférence a à offrir.

Campus Shuttle Information
umanitoba.ca/campus/parking/shuttle

Parking Services Game Day Parking Information
umanitoba.ca/campus/parking

University of Manitoba Events Calendar
events.umanitoba.ca

Your Source for U of M News
umtoday.ca

Sponsors • Commanditaires

We would like to thank all of our sponsors for their generous support of the Canadian Statistics Student Conference; these contributions have made this conference 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.

Société statistique du Canada
Statistical Society of Canada

University of Manitoba

Gold Sponsors • Commanditaires Or

CANSSI INCASS

FIELDS

Statistics Canada
Statistique Canada

pims

Roche

biostats
Biostatistics Section | Groupe de biostatistique

Statistical Society of Canada
Société statistique du Canada

Silver Sponsors • Commanditaires Argent

Canadian Tire Financial Services

CMS SMC

Bronze Sponsors • Commanditaires Bronze

sas

Organizers and Volunteers • Organisateurs et bénévoles

Organizing Committee • Comité organisateur:

Co-Chairs / Présidentes:

Osvaldo Espin-Garcia (University of Toronto)

Kuan Liu (University of Toronto)

Local Arrangements / Arrangements locaux:

Lin Xue (University of Manitoba)

Session organizers and committee members / Organisateurs des sessions et membres du comité:

Sheri Albers (Canada Revenue Agency),

Janie Coulombe (McGill University),

Katherine Daignault (University of Toronto),

Saima Khan Khosa (University of Saskatchewan),

Marie-Christine Robitaille-Grou (Université de Montréal),

Dinara Salaeva (Centre for Addiction and Mental Health),

Yidan Shi (University of Waterloo)

Support and thanks / Support et remerciements:

Past CSSC organizers / Les derniers organisateurs du congrès étudiant de la CSSC:
Nathalie Moon, Caitlin Daly

SSC liaison / Liaison avec la SSC:
Léo Belzile (SARGC), Miaclaire Woodland (Administrator), Rachel Boutet (Accountant)
Alex Leblanc (2017 SSC organizer), Erica E.M. Moodie (2017 SSC organizer),
Changbao Wu (2017 SSC organizer), Edward Chen (Treasurer), Jill Weldon (Liaison), Larry Weldon (Liaison)

Photographer / Photographe:
Peter Macdonald

Translations / Traduction:
Janie Coulombe, Marie-Christine Robitaille-Grou, Gabrielle Simoneau, Sangook Kim

Sponsors / Commanditaires:
Paul Mahoney (Roche), Yvette Roberts (CANSSI), Denise Feighan (CANSSI&PIMs), Esther Berzunza (Fields), Jack Gambino (Statistics Canada), Tracy Chen (Canadian Tire Financial Services),
Lindsay Hart (SAS Canada), Mark Morreale (SAS Canada), Matt Malczewski (SAS Canada)

Program Overview

Time	Session	Room	Page
8:45-9:30	Registration and Refreshments	EITC atrium	
9:30-9:40	Welcome Session with SSC President Address	E3-270	
9:40-9:55	Sponsor Talk I: CANSSI	E3-270	
10:00-10:45	Scientific Presentations I		
	Biostatistics applications and Causal Inference	E2-125	16
	Methodological advancements in Measurement Error	E2-130	18
	Longitudinal Data and Multistate Models	E2-150	20
	Time Series Analysis	E2-155	22
10:50-11:20	Scientific Presentations II		
	Statistical learning and classification I	E2-125	24
	Statistical learning and classification II: Methods in Multi-Label Classification	E2-130	25
	Stochastic Processes and Probability Theory	E2-150	26
	Sports Analytics	E2-155	27
11:25-12:15	Invited Career Speakers Session	E3-270	13
12:15-1:15	Lunch (Provided) Networking Activity	EITC atrium	15
1:15-2:35	Workshop	E3-270	10
2:40-3:30	Skills session	E3-270	11
3:30-4:30	Poster Session and Refreshments	EITC atrium	28
4:30-5:15	Keynote Address	E3-270	8
5:15-5:25	Sponsor Talk II: SAS	E3-270	
5:25-5:40	Awards and Closing Remarks	E3-270	

Note: All activities will take place at the Engineering and Information Technology Complex (EITC)

Aperçu du Programme

Heure	Session	Local	Page
8:45-9:30	Inscription et rafraîchissements	EITC atrium	
9:30-9:40	Séance de bienvenue et mot du Président	E3-270	
9:40-9:55	Mot du commanditaire I: INCASS	E3-270	
10:00-10:45	Présentations scientifiques I		
	Applications de biostatistique et inférence causale	E2-125	16
	Avancements méthodologiques en erreur de mesure	E2-130	18
	Données longitudinales et modèles multi-états	E2-150	20
	Analyse de séries chronologiques	E2-155	22
10:50-11:20	Présentations scientifiques II		
	Apprentissage statistique et classification I	E2-125	24
	Apprentissage statistique et classification II: Méthodes en classification multi-label	E2-130	25
	Processus stochastiques et théorie de probabilité	E2-150	26
	Analytique du sport	E2-155	27
11:25-12:15	Séance sur invitation sur les carrières	E3-270	13
12:15-1:15	Dîner (Fourni) Séance de résautage	EITC atrium	15
1:15-2:35	Atelier	E3-270	10
2:40-3:30	Séance de compétence	E3-270	11
3:30-4:30	Séance d'affiches et Rafraîchissements	EITC atrium	28
4:30-5:15	Discours d'honneur	E3-270	8
5:15-5:25	Mot du commanditaire II: SAS	E3-270	
5:25-5:40	Prix et séance de clôture	E3-270	

Keynote Address • Discours d'honneur

Dr. Jiahua Chen

Research ideas, motivations and rewards •
Les idées de recherche, les motivations et les récompenses

All researchers have an ambition from a tender age: making ground shaking scientific discoveries. Against all odds, new researchers are finally at the door of discovery. Yet many may find a research career not as glorious and get completely lost at the very beginning. It is extremely hard to come up with novel research ideas and insights. Remaining enthusiastic on research itself may become tough and one may settle at working merely for promotion and career advancement. Yet from time to time, statistical research revives itself with industrial statistics, biostatistics, bioinformatics, MCMC, statistical learning, big data and whatever is coming. It is up to us to stay relevant and motivated. In this talk, I wish to share my research experiences with graduate students, to reveal my struggles and to explain some of my research discoveries. I hope to convince you that the best remedy for a successful research career is to stay faithful to the spirit of science. There are few shortcuts to discoveries. You should first prepare yourself with technical strength and sharp observing eyes. Opportunities are more abundant than you may believe. You will ultimately be rewarded with your own discoveries.

•

Depuis la tendre enfance, chaque chercheur partage une certaine ambition: celle de faire une découverte scientifique qui créera une onde de choc. Contre toute attente, les nouveaux chercheurs sont finalements arrivés à la porte de la découverte. Il n'empêche que plusieurs chercheurs peuvent se retrouver pris dans une carrière qui n'est pas aussi glorieuse et se sentir complètement perdus à leurs débuts. Il est extrêmement difficile de penser à des idées de recherche novatrices et d'avoir des connaissances approfondies. Il peut devenir difficile de rester enthousiaste par rapport à la recherche et certains peuvent préférer s'en remettre à travailler dans le domaine de la promotion ou de l'avancement de carrière. De temps en temps, il arrive tout de même que la statistique puisse se raviver à travers des statistiques industrielles, de la biostatistique, de la bioinformatique, des MCMC, de l'apprentissage statistique, des grands jeux de données ou de tout autre sujet chaud à venir. Il est de notre devoir de demeurer motivé et pertinent. Lors de cet exposé, je souhaite partager mes expériences de recherche avec les étudiants gradués, révéler les problèmes rencontrés durant mon parcours, et expliquer certaines de mes découvertes de recherche. J'espère vous convaincre que le meilleur remède pour une carrière de recherche empreinte de succès est de rester fidèle à l'esprit scientifique. Il existe quelques raccourcis pour faire des découvertes. Vous devriez d'abord vous assurer d'avoir une grande connaissance du point de vue technique et une vision aiguisée. Les opportunités sont plus abondantes que vous ne le croyez. Vous serez ultimement récompensés par vos propres découvertes.

Dr. Jiahua Chen

Professor Jiahua Chen holds a Tier 1 Canada Research Chair in the Department of Statistics at the University of British Columbia. He earned an undergraduate degree in mathematics at the University of Science and Technology of China in 1982 followed by a Masters degree in statistics from Academia Sinica in Beijing in 1985. He completed his PhD under Professor C. F. Jeff Wu at the University of Wisconsin in 1990 and was a postdoctoral fellow under Professor Jack Kalbfleisch. In 1991, he joined the University of Waterloo as Assistant Professor and was promoted to Associate Professor and Professor in 1996 and 2001, respectively. In 2007 he took up a professorial appointment and his Research Chair at the University of British Columbia. Jiahua has made outstanding research contributions in many areas of statistics including experimental design, sampling theory, empirical likelihood, mixture models, variable selection and applications in genetics. In addition, he has served the Canadian and international statistical communities exceptionally well, for example, as Editor of The Canadian Journal of Statistics and President of the International Chinese Statistical Association. He has a tremendous record as educator, having supervised many PhD and Master's students; many of his PhD advisees are now active researchers and teachers at Canadian and other universities. Jiahua's work in survey sampling that deals with nearest neighbor imputation for non-response, as well as assessing the accuracy of estimates has been widely recognized by survey sampling methodologists. His 1993 Biometrika paper with Jing Qin and others on empirical likelihoods for finite populations has formed the basis of extensive additional work. Jiahua has also contributed much to the area of mixture models, by producing innovative new results as well as refining and extending existing methods. He and his co-authors have published numerous papers in many extremely well-respected journals. Recently he has worked on developing methods for testing the order of a mixture, which is particularly relevant in statistical genetics.

•

Professeur Jiahua Chen détient une chaire de recherche du Canada (I) au département de statistique de l'Université de la Colombie-Britannique. Il a obtenu un diplôme de premier cycle en mathématiques à l'Université des sciences et technologies de Chine en 1982 suivi d'un diplôme de maîtrise en statistique de l'Academia Sinica à Beijing en 1985. Il a complété ses études doctorales sous la supervision de professeur C. F. Jeff Wu à l'Université du Wisconsin en 1990 et a reçu une bourse postdoctorale avec laquelle il a travaillé avec le professeur Jack Kalbfleisch. En 1991, il a rejoint l'Université de Waterloo en tant qu'assistant professeur et fut promeut comme professeur associé et professeur en 1996 et 2001 respectivement. En 2007, il obtenait la chaire de recherche à l'Université de la Colombie-Britannique en plus de son titre de professeur. Les contributions de Jiahua en recherche sont immenses, et ce dans plusieurs domaines liés à la statistique, incluant les plans d'expérience, la théorie sur l'échantillonnage, la vraisemblance empirique, les modèles mixtes, la sélection de variables et les applications en génétique. De plus, il a servi de façon exceptionnelle la communauté statistique canadienne et internationale, en étant par exemple l'éditeur de la revue canadienne de statistique et en tant que Président de l'association statistique chinoise internationale. Il présente un dossier d'enseignant remarquable, ayant supervisé plusieurs étudiants doctoraux et à la maîtrise; plusieurs des étudiants qu'il a supervisés au doctorat sont maintenant des chercheurs actifs et des professeurs dans les universités canadiennes et ailleurs. Le travail de Jiahua en méthodes de sondage, entre autres sur l'imputation par le plus proche voisin en cas de non-réponse, ainsi que l'évaluation de la précision des estimations a été grandement reconnu par les méthodologistes en méthodes de sondage. Son article de 1993 dans Biometrika avec Jing Qin et cie sur la vraisemblance empirique pour des populations finies a déclenché de nombreux autres travaux. Jiahua a aussi contribué à la recherche sur les modèles mixtes, en produisant de nouveaux résultats innovateurs tout en raffinant et en étendant les méthodes existantes. Avec ses co-auteur(s), il a publié un grand nombre d'articles dans plusieurs journaux renommés. Il a récemment travaillé à développer des méthodes pour tester l'ordre des mixtures, ce qui est particulièrement utile en statistiques génétiques.

Statistical Computing Workshop • Atelier en Calcul Statistiques

Dr. Chel Hee Lee received his Ph.D. in Biostatistics from the University of Saskatchewan in 2014 under the supervision of Prof. Mikelis Bickis, and was working as Biostatistician at the College of Medicine, University of Saskatchewan before joining the University of Calgary. His research interests mainly lie in developing and applying statistical methods to advance knowledge in medicine and public health, and currently working on numerical and statistical computing, classification, imprecise probabilities, high-dimensional data analysis, and survival and longitudinal data analysis.

Abstract

This workshop focuses on practical aspects of R programming for optimization and simulation in computational statistics. We consider an estimating equation that does not have a closed-form solution. The solutions are found using different numerical methods and they are compared. Participants will learn the followings from this problem-solving:

- types of numerical errors
- basics of writing user-defined function
- flow control with loop and conditional expressions
- exceptions, messages, and error handling, and
- objective-oriented programming.

If time permits basic debugging tools will be introduced. This workshop will ultimately take you to the next stage of programming in R.

Dr Chel Hee a obtenu son doctorat en biostatistique de l'Université de Saskatchewan en 2014, sous la supervision du professeur Mikelis Bickis, et travaillait comme biostatisticien au collège de médecine, Université de Saskatchewan, avant de rejoindre l'Université de Calgary. Ses intérêts de recherche se concentrent sur le développement et l'application de méthodes statistiques dans le domaine des connaissances avancées de médecine et de la santé publique, et il travaille présentement sur le calcul numérique et statistique, la classification, les probabilités imprécises, l'analyse de grands jeux de données et l'analyse de données longitudinales et de survie.

Résumé

Cet atelier se concentre sur les aspects pratiques de la programmation en R pour l'optimisation et la simulation dans le domaine des statistiques computationnelles. Nous considérons une équation d'estimation qui ne possède pas de solution analytique explicite. Les solutions sont trouvées à partir de diverses méthodes numériques et sont comparées. Les participants pourront apprendre sur:

- les types d'erreurs numériques
- la base pour programmer une fonction définie par l'utilisateur
- le contrôle du débit avec des boucles et des expressions conditionnelles
- les exceptions, les messages et le traitement des erreurs, et
- la programmation orientée sur l'objet.

Si le temps le permet, des outils pour le débogage seront présentés. Ultimement, cet atelier vous amènera à un autre niveau en termes de programmation en R.

Skills Sessions • Sessions sur les compétences professionnelles

Mohammad Jafari Josani

Mohammad Jafari Jozani is currently an Associate Professor with the Department of Statistics at the University of Manitoba. Recently, he has been working on statistical learning problems with high dimensional aspects in Engineering and Sustainable Energy; small area estimation as well as statistical inference with complex sampling designs using order statistics and rank information. He has applied his research in areas such as breast cancer study, BMD analysis, pollution, age determination of fish spices, and recently in the calibration problems to design simulators for training purposes in order to make surgeries safer. So far, he has successfully supervised 3 Postdocs, 6 PhD students and 4 MSc students. Currently he has 2 MSc and 3 PhD students. In 2012, Mohammad was the recipient of the prestigious Rh award, for outstanding contributions to scholarship and research in the natural science category. The Rh award is awarded to scholars who show "exceptional innovation, leadership, and promise" early in their careers. He got merit awards for research in 2011 and 2014. Mohammad has collaborations with Manitoba Hydro, Invenia company, NeuroArm, etc. and his research is supported by NSERC Discovery grant, NSERC Engage grant, CANSSI Kick start program, as well as a number of internal funds from the University of Manitoba.

•

Mohammad Jafari Jozani est présentement professeur associé au département de statistique de l'Université du Manitoba. Il a récemment travaillé sur les problèmes d'apprentissage statistique sur des données de grandes dimensions en ingénierie et en énergie renouvelable, ainsi que sur l'estimation des petites zones et sur l'inférence statistique avec des plans d'échantillonnage complexes, à partir des statistiques d'ordre et d'information sur les rangs. Ses recherches se sont appliquées aux domaines de l'étude du cancer du sein, de l'analyse de la densité minérale osseuse, de la pollution, de la détermination de l'âge de différentes espèces de poissons, et récemment aux problèmes de calibration dans la conception de simulateurs pour l'entraînement, afin de rendre les chirurgies plus sécuritaires. À ce jour, il a supervisé avec succès 3 étudiants post-doctoraux, 6 étudiants au doctorat et 4 étudiants à la maîtrise. Il supervise présentement 2 étudiants à la maîtrise et 3 étudiants doctoraux. En 2012, Mohammad fut le récipiendaire du prestigieux prix Rh pour ses contributions remarquables à l'érudition et à la recherche dans la catégorie des sciences naturelles. Le prix Rh est remis aux chercheurs qui démontrent une 'innovation exceptionnelle, du leadership et qui sont très prometteurs' tôt dans leur carrière. Il a reçu deux prix de mérite pour sa recherche en 2011 et en 2014. Mohammad collabore avec Hydro Manitoba, la compagnie Invenia, NeuroArm, etc. et sa recherche est financée par le programme de subventions à la découverte du CRSNG, le programme de subventions à l'implication du CRSNG, le programme de relance du CANSSI, ainsi que plusieurs autres fonds internes de l'Université du Manitoba.

Olli Saarela

Olli Saarela is an assistant professor at the Dalla Lana School of Public Health of University of Toronto. He studied at the University of Helsinki. After obtaining his master's degree in 2004, he worked at the National Institute for Health and Welfare of Finland. Meanwhile he also pursued PhD studies and defended his thesis in 2010. He came to Canada for a postdoctoral fellowship at the Department of Epidemiology, Biostatistics and Occupational Health of McGill University, before joining University of Toronto in Summer 2014. His areas of interest include Bayesian inference, causal inference, epidemiological study designs, and survival and event history analysis.

Olli Saarela est assistant professeur à l'école de santé publique Dalla Lana de l'Université de Toronto. Il a fait ses études à l'Université de Helsinki. Après avoir obtenu son diplôme de maîtrise en 2004, il a travaillé à l'institut national de la santé et du bien-être de Finlande. Pendant ce temps, il a aussi poursuivi ses études doctorales et a défendu sa thèse en 2010. Il est venu au Canada dans le cadre d'un programme de bourse postdoctorale, pour travailler au département d'épidémiologie, de biostatistique et de santé au travail à l'Université McGill, avant de se diriger vers l'Université de Toronto, à l'été 2014. Ses intérêts de recherche incluent l'inférence bayésienne, l'inférence causale, les plans d'études épidémiologiques, et l'analyse de survie et de l'historique d'événements.

Nathalie Moon

Nathalie Moon is a PhD candidate in biostatistics at the University of Waterloo. She holds a MMath degree in biostatistics from the University of Waterloo (2013) and a BScH in statistics from Queen's University (2011). Over the course of her graduate studies, Nathalie has been awarded several competitive graduate scholarships including NSERC CGS-M, NSERC CGS-D, and the Ontario Graduate Scholarship (OGS). Her current research interests involve developing methods to improve efficiency in cohort studies in the presence of dropout, by optimally selecting a subset of individuals for extended follow-up (tracing).

Nathalie Moon est candidate au doctorat en biostatistique à l'Université de Waterloo. Elle détient un diplôme MMath en biostatistique de l'Université de Waterloo (2013) et un BSc avec honneurs en statistique de l'Université Queen's (2011). Pendant ses études graduées, Nathalie a reçu plusieurs prix de compétitions incluant NSERC CGS-M, NSERC CGS-D et la bourse d'études supérieures de l'Ontario (OGS). Ses intérêts de recherche se concentrent présentement sur le développement des méthodes pour améliorer l'efficacité dans les études de cohortes, en présence de décrochage, en sélectionnant de façon optimale un sous-groupe d'individus pour un suivi prolongé (repérage).

Invited Career Speakers • Séance sur les carrières avec des conférenciers invités

Kathryn Mills - Canada Revenue Agency
Opportunities available to statisticians in the federal government • Occasions pour statisticiens en gouvernement fédéral

Kathryn Mills is presently the Manager of Advanced Analytics in the Innovation Lab at the Canada Revenue Agency. There she leads a team of Data Scientists in collaborating on innovative data projects in support of service and compliance. Kathryn's career path has covered designing software and electronic circuitry in products used by top telecommunications providers worldwide, performing statistical analyses on the Register of Electors and the Census, working in the protection of electronic information and communication, graduating from a three-year specialized mathematics intern training program in the US, and managing a section tasked with providing advanced analysis in support of securing our nation's borders. She earned a B. Eng. in Electrical Engineering and a M.Sc. specializing in Applied Statistics, both from Carleton University. The two degrees combine well to address the analytical challenges associated with Big Data. Kathryn's interests are in supporting evidence-based decision making.

Kathryn Mills est présentement directrice de la section d'analyses avancées au laboratoire d'innovation de l'agence du revenu du Canada. Là-bas, elle supervise une équipe de scientifiques de données dans la collaboration sur des projets de données innovateurs, liés au support au service et à la conformité. Le cheminement de carrière de Kathryn a couvert la conception de logiciels et les circuits électroniques pour des produits utilisés par les plus importants fournisseurs de télécommunication au monde, elle a fait des analyses statistiques pour le registre des électeurs et pour le recensement, elle a travaillé dans la protection de l'information électronique et de la communication, elle fut diplômée d'un programme de stage spécialisé de trois ans en mathématiques aux États-Unis et elle a aussi géré une section qui offrait des analyses avancées en soutien à la protection des frontières de notre nation. Elle s'est vue obtenir un Baccalauréat en ingénierie électrique et une maîtrise spécialisée en statistique appliquée, tous deux de l'Université Carleton. Ces deux diplômes se complémentent bien pour résoudre les défis analytiques liés aux problèmes de grands jeux de données. Kathryn s'intéresse au soutien à un processus décisionnel fondé sur la preuve.

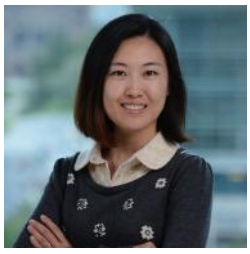
Dr. Karla Fox - Canada Revenue Agency
The pros and cons of a statistics PhD in a non-academic career • (Dés)Advantages d'avoir un doctorat en statistiques pour les carrière non académique

Dr. Karla Fox is presently the Manager of Research and Analytics, GST/HST at the Canada Revenue Agency. She has worked in several government departments designing experiments, surveys, and epidemiological studies. Just prior to working at the CRA she was the Research Manager for Data Analysis as well as chief of Generalized Systems and the Data Analysis and Resource Center at Statistics Canada. While working full time, she completed her doctorate at Queen's University on the meta-analysis of survey data. Her research interests are in tax compliance, study design, the analysis of complex survey data, meta-analysis, record linkage, and microsimulation.

Dr. Karla Fox est présentement directrice du groupe d'analytique et de recherche, GST/HST à l'agence du revenu du Canada. Elle a travaillé dans plusieurs départements gouvernementaux, à planifier des expériences, des sondages et des études épidémiologiques. Juste avant de travailler à la CRA, elle était directrice de la recherche pour la section analyse des données, et chef des systèmes généralisés et du centre pour l'analyse de données et des ressources, chez statistique Canada. Elle a complété son doctorat sur la méta-analyse de données de sondage à l'Université Queen's, tout en travaillant temps plein. Ses intérêts de recherche sont la conformité fiscale, la conception des études, l'analyse de données de sondage complexes, la méta-analyse, le couplage de dossiers et la micro-simulation.

Dr. Zhihui (Amy) Liu - University of Toronto
Biostatistics and medical research • Biostatistiques et la recherche médicale

Dr. Amy Liu is a biostatistician in the Strategic Analytics team at Cancer Care Ontario, where she provides statistical support in an internal consulting unit. She is a status Assistant Professor at the Dalla Lana School of Public Health of University of Toronto. She received her BSc and MSc in Statistics from McMaster University and PhD in Biostatistics from McGill University. She has worked in a number of projects initiated by the Kidney Cancer Research Network of Canada since 2011. She was in one of the winning teams in the 2011 SSC Case Studies in Data Analysis competition, and was a member of the organizing committee in the inaugural SSC Student Conference in 2013.



Dr. Amy Liu est biostatisticienne dans l'équipe d'analyse stratégique chez Action Cancer Ontario, où elle offre un support statistique dans une unité de consultation interne. Elle est professeure titulaire à la Dalla Lana School of Public Health de l'Université de Toronto. Elle a reçu son baccalauréat et sa maîtrise en statistique à l'Université McMaster et son doctorat en biostatistique à l'Université McGill. Elle a travaillé dans plusieurs projets initiés par le réseau de recherche sur le cancer du rein du Canada depuis 2011. Elle faisait partie de l'une des équipes gagnantes d'une étude de cas de la SSC en 2011 et fut membre du comité organisateur de la première conférence étudiante de la SSC, en 2013.

Networking Session • Séance de réseautage

Chel Hee Lee	University of Calgary	Statistical computing
Karla Fox	Canada Revenue Agency	Opportunities for methodologists in government
Kathryn Mills	Canada Revenue Agency	Opportunities for data scientists in government
Amy Liu	University of Toronto	How to do well in a job interview
Osvaldo Espin-Garcia	University of Toronto	Reproducible research
Sheri Albers	Canada Revenue Agency	Getting involved with SARGC
Olli Saarela	University of Toronto	Planning for an academic career
Peter Macdonald	McMaster University	Planning for an academic career
Katherine Daignault	University of Toronto	Work/life balance as a graduate student
Nathalie Moon	University of Waterloo	Scholarship applications
Kuan Liu	University of Toronto	Join the CSSC/CCÉS committee next year
Janie Coulombe	McGill University	

For more networking opportunities, you can join the SSC's Student and Recent Graduate Committee (SARGC) Facebook page

•

Pour d'autres occasions de résautage, vous pouvez rejoindre la page Facebook du comité des étudiants et diplômés récents (CÉDIR) de la SSC

SARGC Facebook page • Page Facebook de CÉDIR

Scientific Abstracts: Oral Presentations • Résumés Scientifiques: Présentations Orales

Biostatistics applications and Causal Inference • Applications de biostatistique et inférence causale

10:00am - 10:45am **E2-125**

Tayler Scory (University of Calgary)

Assessing Effect Modification using Statistical Analyses in Epidemiology
Évaluation de la modification d'effet à partir d'analyses statistiques en épidémiologie

Epidemiology is the study of the distribution of diseases in relation to their causes. One of the main goals of analytical epidemiology is to determine associations between a risk factor and a disease. However, these associations may be distorted by extraneous variables. An example of this phenomenon is effect modification: when the association differs depending on the level of a third variable. In my talk I will describe how statistical tests are used to examine effect modification in analytical epidemiology studies. Specifically, I will describe how to use analytical approaches such as stratified analysis or logistic regression to determine if a variable is an effect modifier, and how to appropriately report test results in the presence of effect modification.

L'épidémiologie est l'étude de la distribution des maladies en relation avec leurs causes. Un des buts premiers de l'épidémiologie analytique est de déterminer les associations entre une maladie et un facteur de risque. Cependant, ces associations peuvent être déformées par des variables externes. Un exemple de ce phénomène est la modification d'effet, où l'association diffère selon le niveau d'une troisième variable. Lors de ma présentation, je décrirai comment les tests statistiques sont utilisés afin d'examiner la modification d'effet dans les études d'épidémiologie analytique. Plus spécifiquement, je décrirai comment l'on peut utiliser des approches analytiques, comme l'analyse stratifiée ou la régression logistique, afin de déterminer si une variable est un modificateur d'effet, et comment rapporter correctement les résultats de tests en présence de modification d'effet.

Pai-Shan Cheng (University of Toronto), Rosalie H Wang (University of Toronto and Toronto Rehabilitation Institute - University Health Network), Marge Coahran (University of Toronto and Toronto Rehabilitation Institute - University Health Network), Jose Zariffa (University of Toronto and Toronto Rehabilitation Institute - University Health Network)

Effectiveness of Robotic System for Stroke Rehabilitation
Efficacité des systèmes robotiques pour la réhabilitation après un AVC

Stroke survivors often suffer from loss of upper limb movement control. To facilitate rehabilitation, a robotic system is used to provide treatment to nine patients and to collect data on each patientÍs performance. Through a two-level modelling approach, it is possible to combine the data collected on each patient in the study, and determine the average treatment effect and whether this treatment effect is statistically significant. Analysis based on the proposed two-level model reveals a statistically significant treatment effect when treating each of several variables measuring a patient's motor performance as the outcome variable.

Les patients ayant survécu à un AVC souffrent souvent d'une perte de contrôle au niveau des mouvements des membres supérieurs. Pour faciliter leur réhabilitation, un système robotique est utilisé afin d'offrir un traitement à neuf patients et de collecter les données sur la performance de ces patients. À partir d'une approche de modélisation à deux niveaux, on peut combiner toutes les données collectées de chaque patient de l'étude et déterminer l'effet moyen de traitement et si l'effet du traitement est statistiquement significatif. L'analyse basée sur l'approche à deux niveaux révèle un effet statistiquement significatif du traitement lorsque l'on traite de chacune des variables mesurant la performance motrice des patients comme d'une variable réponse.

Kuan Liu (University of Toronto), George Tomlinson (University Health Network)

Assessing the Safety and Efficacy of Current Consensus Treatments on Juvenile Dermatomyositis using a Complex Multi-center Pilot Registry

Évaluation de la sécurité et de l'efficacité des traitements faisant présentement consensus pour la dermatomyosite juvénile à partir d'une base de données enregistrées pilote multicentrique complexe

Juvenile dermatomyositis (JDM) is a rare, chronic multisystem disease with estimated incidence about 2-4 per million pediatric population in North America. Children and adolescents with JDM often have long periods of active disease that increase their risk of skin, joint and muscle damage. Knowing which treatment strategies are most effective is critically important. Owing to the limited event size and long disease course, designing well powered clinical trials are deemed unfeasible. In 2012, CARA initiated a multiple-center registry on pediatric arthritis in light to examine and improve current treatment protocols. We analyzed the pilot portion of this registry to study the safety and efficacy of three current consensus treatments on JDM. To overcome data limitations including limited sample size and missing data, as well as the issue with complex causal inference on treatment effect under observational studies, various advanced statistical approaches were adopted, including multiple imputation, generalized boosted models on Propensity Score estimation and inverse probability weighting.

La dermatomyosite juvénile (DMJ) est une maladie chronique multi-système rare qui présente une incidence estimée à 2-4 par million dans la population pédiatrique d'Amérique du Nord. Les enfants et adolescents atteints de la DMJ présentent souvent de longues périodes actives de DMJ qui augmentent leurs risques de dommages à la peau, aux jointures et aux muscles. Il est très important et même critique de connaître les stratégies de traitement qui sont les plus efficaces. On considère qu'il est presque impossible de planifier des essais cliniques avec assez de puissance dans le cas de cette maladie, à cause du faible taux d'événements et de la longue période sur laquelle s'étend la maladie. En 2012, CARA a initié un essai multicentrique pour l'arthrite juvénile afin d'examiner et d'améliorer les protocoles actuels de traitement. Nous avons analysé la portion pilote de cet essai afin d'étudier la sécurité et l'efficacité de trois traitements consensuels de la DMJ. Pour pallier aux limitations dues aux données, incluant la taille limitée de l'échantillon et les données manquantes, ainsi que les problèmes liés à la complexité des études observationnelles et de l'inférence causale sur l'effet de traitement, plusieurs approches statistiques avancées ont été adoptées, incluant l'imputation multiple, les modèles généralisés renforcés sur le score de propension et la pondération selon la probabilité inverse.

Methodological advancements in Measurement Error • Avancements méthodologiques en erreur de mesure
10:00am - 10:45am E2-130

Rajib Dey (Memorial University of Newfoundland), Dr. Noel Cadigan (Fisheries and Marine Institute of Memorial University of Newfoundland), Taraneh Abarin (Memorial University of Newfoundland)

Sensitivity to Model Misspecification of a VonB Growth Model with Measurement Error in Age
Sensibilité à l'erreur de spécification d'un modèle de croissance VonB avec erreur de mesure dans l'âge

The Von Bertalanffy growth function (VBGF) specifies the length of fish as a function of age. However, in practise age is measured with error. We study the structural errors-in-variables (SEV) approach to account for ageing error. Recent research also proposed this approach for fish growth data. It was assumed that the distribution of true unobserved age was a Gamma distribution. It is evident from the result that particular SEV approach provided improved regression parameter inferences. Here we investigate whether SEV VBGF parameter estimators are robust to misspecification of the gamma true age distribution. By robust we mean no large sample bias. Our simulation results demonstrate that the SEV VBGF using a gamma distribution for unobserved age is not robust. For large ME the bias of estimators are large.

La fonction de croissance Von Bertalanffy (VBGF) spécifie la longueur d'un poisson en fonction de son âge. Cependant, en pratique, l'âge est mesuré avec une erreur. Nous étudions l'approche structurelle d'erreur dans les variables (SEV) afin de tenir compte de l'erreur sur l'âge. Les recherches récentes ont aussi suggéré cette approche pour les données de croissance des poissons. On assumait alors que la distribution de l'âge véritable non observé suivait une loi gamma. Suite aux résultats obtenus, il est évident que l'approche SEV a mené à une inférence améliorée quant aux paramètres de régression. Ici, nous vérifions si les estimateurs de paramètre SEV VBGF sont robustes à l'erreur de spécification sur la distribution gamma de l'âge véritable. Par robuste, nous entendons qu'il n'y aurait pas de biais pour de grands échantillons. Les résultats de nos simulations montrent que le SEV VBGF avec distribution gamma pour l'âge n'est pas robuste pour l'âge non observé. Lorsque l'erreur de mesure est importante, le biais des estimateurs est considérable.

Md Erfanul Hoque (University of Manitoba), Mahmoud Torabi (University of Manitoba)

Modeling the Random Effects Covariance Matrix for Longitudinal Data with Covariates Measurement Error
Modéliser la matrice de covariance des effets aléatoires pour des données longitudinales avec erreur de mesure sur les covariables

Generalized linear mixed models (GLMMs) are commonly used to analyze longitudinal data. It is typically assumed that the random effects covariance matrix is constant across the subject (and among subjects) in these models. In many situations, however, this correlation structure may differ among subjects and ignoring this heterogeneity can cause the biased estimate of model parameters. Covariates measured with error also happen frequently in the longitudinal data setup. Ignoring this issue in the data may produce bias in model parameters estimate and lead to wrong conclusions. In this work, we propose an approach to properly model the random effects covariance matrix based on covariates in the class of GLMMs with covariates measurement error using modified Cholesky decomposition. The resulting parameters from the decomposition have a sensible interpretation. Performance of the proposed approach is evaluated through simulation studies and a data application from Manitoba Follow-up study.

Les modèles mixtes linéaires généralisés (MMLG) sont communément utilisés dans l'analyse de données longitudinales. Dans ces modèles, on assume généralement que la matrice de covariance des effets mixtes est constante pour un même sujet (et entre les sujets). Cependant, dans plusieurs situations, cette structure de corrélation peut différer entre les observations; ignorer l'hétérogénéité peut mener à des estimations biaisées des paramètres du modèle. Dans un contexte de données longitudinales, les covariables présentent elles-aussi fréquemment des erreurs de mesure. Ignorer ce problème peut mener à un biais dans les paramètres du modèle et mener à de fausses conclusions. Dans ce travail, nous proposons une approche permettant de modéliser correctement la matrice de covariance des effets mixtes, dans un contexte de MMLG avec erreur de mesure sur les covariables, en utilisant une modification de la décomposition de Cholesky. Les paramètres résultant de cette décomposition ont une interprétation sensée. On évalue la performance de l'approche proposée à partir d'études de simulation et d'une application aux données d'une étude longitudinale du Manitoba.

Longitudinal Data and Multistate Models • Données longitudinales et modèles multi-états
10:00am - 10:45am E2-150

Nathalie Moon (University of Waterloo), Leilei Zeng (University of Waterloo), Richard Cook (University of Waterloo)

Efficiency Gains from Optimal Selection for Tracing in Incomplete Cohorts due to Loss to Follow-up
Gains d'efficacité de la sélection optimale pour le traçage dans des cohortes incomplètes en raison de la perte au suivi

While prospective cohort studies yield data on the course of chronic disease, loss-to-follow-up is a common concern. If dropouts arise from a sequentially missing at random mechanism, less efficient estimates are obtained, while inconsistent estimates are obtained from sequentially missing not and random mechanisms. We propose selecting a subset of individuals who are lost-to-follow-up and tracing them to determine their status, to augment phase-I data. A hybrid information-based criterion is presented, combining observed information collected prior to dropout and the expected contribution of individuals selected for tracing, with the goal of minimizing the expected variance of various parameters subject to cost constraints. Properties of the methodology are explored via simulation and applied to data from the Toronto Psorisis Clinic.

Alors que les études de cohorte prospectives fournissent des données sur l'évolution de maladies chroniques, la perte au suivi est une préoccupation usuelle. Si l'attrition découle d'un mécanisme de données séquentiellement manquantes au hasard, des estimations moins efficaces sont obtenues, tandis que des estimations non convergentes sont obtenues dans le cas de données séquentiellement manquantes non au hasard. Nous proposons de sélectionner un sous-ensemble d'individus qui sont perdus au suivi et de les tracer pour déterminer leur statut, afin d'augmenter les données de la phase I. Un critère hybride basé sur l'information est présenté, combinant l'information observée collectée avant la perte au suivi et la contribution attendue des individus sélectionnés pour le tracé, avec comme but de minimiser la variance attendue de divers paramètres sujets à des contraintes de coûts. Les propriétés de la méthodologie sont explorées par simulation et appliquées aux données de la Toronto Psorisis Clinic.

Xiaoming Lu (Memorial University of Newfoundland), Zhaozhi Fan (Memorial University of Newfoundland)

Generalized Linear Quantile Mixed Regression for Longitudinal Data
Régression quantile linéaire mixte généralisée pour des données longitudinales

This presentation proposes a new generalized linear quantile mixed model for longitudinal data considering both subject-specific and observation-specific random effects. Random effects are estimated by using the best linear unbiased predictors (BLUP) which is based on the Tweedie exponential dispersion model distributions. The proposed method applies a generalized quasi-likelihood approach to account for correlations between repeated measurements. We redefined the estimating functions as smoothed functions which could be differentiated with respect to regression parameters. And the parameter estimation is based on the Newton-Raphson iteration method. Our proposed quantile mixed model gives consistent estimates that have asymptotically normal distributions. An application to epilepsy data and simulation studies are carried out to evaluate the performance of our proposed method

Cette étude propose un nouveau modèle quantile linéaire mixte généralisé pour les données longitudinales tenant compte des effets aléatoires spécifiques aux sujets et spécifiques aux observations. Les effets aléatoires sont estimés en utilisant le meilleur estimateur linéaire non-biaisé, lequel se base sur les distributions de modèles de dispersion exponentiels Tweedie. La méthode proposée utilise l'approche de type quasi-vraisemblance généralisée afin de tenir compte des corrélations entre les mesures répétées. Nous avons redéfini les fonctions d'estimation pour qu'elles soient lisses et puissent être dérivées par rapport aux paramètres de régression. L'estimation de paramètres est faite à partir de la méthode itérative de Newton-Raphson. Le modèle quantile mixte que nous proposons mène à des estimateurs convergents qui ont une distribution asymptotique normale. Une application aux données d'épilepsie et des études de simulations ont été faites afin d'évaluer la performance de la méthode proposée.

Maxime Turgeon (McGill University), Sahir Bhatnagar (McGill University), James Hanley (McGill University), Olli Saarela (University of Toronto)

A Novel Approach To Competing-Risk Analysis Using Case-Base Sampling
Une approche novatrice dans l'analyse de risques concurrents, qui utilise l'échantillonnage de type 'case-base'

In competing risk settings, clinicians are often interested in cause-specific absolute risks, i.e. probability of an event given a covariate profile. Semiparametric methods are used to assess the effect of covariates on failure time; however, they typically treat the baseline hazard as a nuisance parameter, and its non-parametric estimates lead to step-wise estimates of the event-specific cumulative incidence that are difficult to interpret. Using case-base sampling, we explain how a competing-risk analysis that directly models the hazard function parametrically can be performed using multinomial regression. We thus obtain smooth estimates of the cumulative incidence. Using simulations, we compare our approach to other competing-risk methods. We then demonstrate our approach using data from patients who received stem-cell transplant for acute leukemia.

Dans un contexte de risques concurrents, les médecins s'intéressent souvent aux risques absolus spécifiques à la cause, c'est-à-dire la probabilité d'un événement, conditionnelle au profil des covariables. Les méthodes semi-paramétriques sont utilisées afin d'évaluer l'effet de covariables sur le temps de défaillance; cependant, elles traitent généralement le risque de base comme un paramètre de nuisance et leurs estimations non-paramétriques mènent à des estimations par étape de l'incidence cumulative spécifique à l'événement, qui sont difficiles à interpréter. En utilisant l'échantillonnage de type 'case-base', nous expliquons comment une analyse de risques concurrents modélisant directement la fonction de risque de façon paramétrique peut être produite en utilisant la régression multinomiale. Nous obtenons ainsi des estimations lisses de l'incidence cumulative. Nous comparons notre approche à d'autres méthodes de risques concurrents à partir de simulations. Nous démontrons ensuite notre approche à partir de données de patients atteints de leucémie aiguë ayant reçu des transplantations de cellules souches.

Time Series Analysis • Analyse de séries chronologiques
10:00am - 10:45am E2-155

Lenin Arango-Castillo (Queen's University), G. Takahara (Queen's University)

Robust Multitaper Spectral Estimator of the Hurst Parameter for Gaussian Long-range Dependent Processes
Estimateur spectral multitaper robuste du paramètre Hurst pour les processus gaussiens dépendants à long terme

Three new estimators of the Hurst parameter are developed in the context of Gaussian Long-Range Dependent processes, namely, Fractional Gaussian Noise (FGN) and Fractional Autoregressive Moving Average (FARIMA) with Gaussian innovations. Our estimation method is based on the multitaper spectral estimation technique. It is shown that our estimators are robust against process process misspecification. A second key feature of our approach is that it allows us to differentiate between FGN and FARIMA with Gaussian innovations. The use of such information and correct Hurst parameter estimation are important for statistical inference about the process location and scale parameters. Statistical, computational, and numerical comparisons are made against traditional estimators including parametric and semi-parametric methods in frequency and wavelet domains.

Trois nouveaux estimateurs du paramètre Hurst sont développés dans le contexte des processus gaussiens dépendants à long terme, soient le bruit gaussien fractionnaire (BGF) et la moyenne mobile autorégressive fractionnaire (FARIMA) avec innovations gaussiennes. Notre méthode d'estimation se base sur la technique d'estimation spectrale multitaper. Il est démontré que les estimateurs obtenus à partir de cette méthode sont robustes aux erreurs de spécification du processus. Une autre caractéristique importante de notre méthode est qu'elle nous permet de différentier le FGN et le FARIMA avec innovations gaussiennes. L'utilisation d'une telle information et l'estimation appropriée du paramètre de Hurst sont importants pour l'inférence statistique à propos des paramètres de location et d'échelle du processus. Des comparaisons statistiques, computationnelles et numériques sont faites par rapport aux méthodes traditionnelles d'estimation, incluant les méthodes paramétriques et semi-paramétriques dans les domaines des fréquences et des ondelettes.

François Marshall (Queen's University)

Mode Detection in a Non-Stationary Environment
Détection du mode dans un environnement non-stationnaire

Model assumptions about the forms of non-stationary process are often too restricted, so a richer class of processes must be incorporated. For K tapers, the test statistic for detecting non-stationary processes is an average of K multitaper dual-frequency coherences between windowed Fourier transforms. When the dual-frequency spectrum of the process rapidly fluctuates, some of the net coherence is accounted for by dual-frequency, dual-window coherences between smeared transforms. A new coherence estimator has been constructed, which incorporates these component coherences. Its performance is demonstrated in a space-physics dataset.

Les hypothèses de modèles sur la forme des processus non-stationnaires sont souvent trop restrictives, et il est donc nécessaire d'incorporer une classe plus riche de processus. Pour K pointes, la statistique de test utilisée pour détecter des processus non-stationnaires représente la moyenne de K cohérences multi-pointes à fréquences duales entre les fenêtres des transformations de Fourier. Lorsque le spectre à fréquence duale du processus fluctue rapidement, une partie de la cohérence nette est prise en charge par les cohérences à fréquence duale, à double fenêtre, entre des transformations dégagées. Un nouvel estimateur de cohérence a été construit, lequel incorpore ces composantes de cohérences. Sa performance est démontrée à partir d'un jeu de données espace-physique.

Claire Boteler (Queen's University), David J. Thomson (Queen's University), Troy Day (Queen's University)

Seasonality in Influenza Mortality: Moving Beyond the Annual Cycle
Saisonnalité de la mortalité due à l'influenza : voir plus loin que le cycle annuel

Influenza is an infectious disease, and its periodic patterns are commonly modeled focusing on a yearly cycle. However, considering that four pandemics have occurred in the past century, this poses the question of whether there could be any hidden patterns. A times series analysis and Multitaper spectral analysis were performed on the dataset: mortality counts due to influenza. Signals with one-year and 30-year periods were found to be significant. The analysis also revealed significant frequencies of 2, 4 and 5 cycles per year. These frequencies could have a biological connection to influenza, or could be harmonics of the yearly cycle. This analysis suggests there might be a more complex periodic structure to this influenza mortality data.

L'influenza est une maladie infectieuse, et l'on modélise généralement sa tendance périodique en se concentrant sur son cycle annuel. Cependant, considérant que quatre pandémies se sont produites dans le dernier siècle, on peut se demander si d'autres tendances se cachent derrière ce cycle annuel. Une analyse de séries chronologiques et une analyse spectrale de type multitaper ont été faites sur le jeu de données, plus précisément sur le nombre de décès dus à l'influenza. Les signaux sur des périodes de 1 et de 30 ans étaient significatifs. L'analyse a aussi révélé des fréquences significatives de 2, 4 et 5 cycles par année. Ces cycles pourraient avoir une connexion biologique avec le virus influenza ou pourraient être harmoniques au cycle annuel. Cette analyse suggère une structure périodique plus complexe que le cycle annuel pour les données de mortalité due à l'influenza.

Statistical learning and classification I • Apprentissage statistique et classification I
10:50am - 11:20 am E2-125

Charmaine Navis (University of Calgary)

A Brief Review of Twin Support Vector Regression
Une brève revue sur la régression à vecteurs de support jumeaux

Multiple linear regression is one of the most widely used statistical applications. However, in some very high dimensional spaces it may not be the most effective method. Two decades ago Drucker et al. (1997) presented a new method to address this issue based on the concept of support vector machines, an algorithm which finds the optimal hyperplane between two classifiers. Recently this has been expanded further to incorporate the innovative twin support vector machine. In this talk I will present a literature review of twin support vector regression (TSVR), a tool for performing regression in these high-dimensional spaces, beginning with an introduction to twin support vector machines. I will then present an overview of recent research being performed in TSVR.

La régression linéaire multiple est une des applications statistiques les plus largement utilisées. Or, dans certains espaces de très haute dimension, ce n'est peut-être pas la méthode la plus efficace. Il y a deux décennies, Drucker et al. (19997) ont présenté une nouvelle méthode pour aborder ce problème basée sur le concept de machines à vecteurs de support, un algorithme qui trouve l'hyperplan optimal entre deux classificateurs. Cette méthode a été récemment étendue pour incorporer la novatrice machine à vecteurs de support jumeaux. Dans cette conférence, je présenterai une revue de littérature concernant la régression à vecteurs de support jumeaux, un outil pour effectuer une régression dans ces espaces à haute dimension, commençant avec une introduction aux machines à vecteurs de support jumeaux. Je présenterai ensuite un survol de la recherche récente en régression à vecteurs de support jumeaux.

Jordan Gardener (University of British Columbia Okanagan), Jeffrey Andrews (University of British Columbia Okanagan)

Robust Variable Selection for Clustering and Classification
Sélection de variables robuste pour le regroupement par grappes et la classification

With increased data collection, the importance of variable selection techniques has risen greatly. We introduce a feature reduction method that increases the robustness of an established variable selection technique for clustering and classification problems (namely, VSCC). To increase the robustness of the method, outliers are removed from the calculations by first clustering the data using contaminated Gaussian mixture models and tossing out observations that are classified into contaminated components. This new robust technique will be contrasted against the classic version, as well as other comparable variable selection techniques, on both simulated and real data sets.

Avec la collecte accrue de données, l'importance des techniques de sélection de variables a considérablement augmenté. Nous introduisons une méthode de réduction des prédicteurs qui augmente la robustesse d'une technique de sélection de variables établie pour les problèmes de regroupement par grappes et de classification (à savoir, VSCC). Pour augmenter la robustesse de la méthode, les données aberrantes sont supprimées des calculs en regroupant d'abord les données en utilisant des modèles de mélange gaussiens contaminés, puis en enlevant les observations qui sont classées dans des composantes contaminées. Cette nouvelle technique robuste sera comparée à la version classique, de même qu'à d'autres techniques de sélection de variables comparables, à la fois sur des ensembles de données simulés et réels.

Statistical learning and classification II: Methods in Multi-Label Classification • Apprentissage statistique et classification II: Méthodes en classification multi-label

10:50am - 11:20 am E2-130

Zhoushanyue He (University of Waterloo), Matthias Schonlau (University of Waterloo)

Ensemble of Iterative Classifier Chains for Multi-label Classification
Ensemble de chaînes de classification itératives pour la classification multi label

Multi-label classification is a classification problem in which each instance can belong to more than one categories. Ensemble of Classifier Chains (ECC) is widely known as one of the benchmark methods for solving multi-label classification problems, in which classifier chains were applied in an ensemble framework. We extend the ECC approach further by incorporating the probabilistic predictions instead of the 0/1 label relevance predictions and running classifier chains in iteration. The ensemble of iterative classifier chains (EICC) and ECC were evaluated on some multi-label datasets with a variety of evaluation metric. The empirical evaluations suggest EICC outperforms ECC in most of the benchmark datasets in terms of macro F-measure.

La classification multi label est un problème de classification dans lequel chaque observation peut appartenir à plus d'une catégorie. L'ensemble de chaînes de classification (ECC) est l'une des méthodes les plus utilisées pour résoudre les problèmes de classification multi label, avec laquelle les chaînes de classification sont appliquées dans un contexte d'ensemble. Nous amenons l'approche des ECC plus loin, en incorporant les prédictions probabilistes au lieu des prédictions de labels 0/1, et en itérant les chaînes de classification. L'ensemble des chaînes de classification itératives (EICC) et d'ECC ont été évaluées sur des jeux de données multi labels à partir d'une variété de métriques d'évaluation. Les évaluations empiriques suggèrent que les EICC performent mieux que les ECC dans la plupart des jeux de données de référence en termes de mesure-F macro.

Xin Liu (Western University), Wenqing He (Western University)

Kernel Based Data-Adaptive Support Vector Machines in Multi-class Cases
Machines à vecteurs de support adaptatif basé sur le noyau dans un contexte multi-classes

Support Vector Machines can be used for multi-class classification problem. However, it suffers when the data in real application are imbalanced. In this talk, a new way to enhance the performance of an multi-class SVM classifier is proposed. By conformally rescaling the initial kernel functions, the separating boundary among different classes can be adaptively enlarged based on the prior knowledge from the traditional SVM in a robust way, and only limited numbers of parameters are required. Consequently, the new proposed classifier considers the spacial distribution of the support vectors in the feature space, Numerical studies shows Improvement of prediction accuracy with this data-adaptive SVM.

Les machines à vecteurs de support peuvent être utilisées pour les problèmes de classification multi-classes. Cependant, cette méthode souffre lorsque les données d'applications réelles sont déséquilibrées. Dans cette présentation, une nouvelle façon d'améliorer la performance d'un classificateur par machine à vecteurs de support multi-classes est proposée. En rééchelonnant de manière conforme les fonctions de noyau initiales, la frontière entre différentes classes peut être agrandie de façon robuste par un processus adaptatif basé sur la connaissance antérieure provenant de la machine à vecteurs de support traditionnelle, et seul un nombre limité de paramètres est requis. Conséquemment, le nouveau classificateur proposé considère la distribution spatiale des vecteurs de support dans l'espace des prédicteurs. Des études numériques montrent l'amélioration de la précision de prédiction avec cette machine à vecteurs de support adaptative.

Stochastic Processes and Probability Theory • Processus stochastiques et théorie de probabilité
10:50am - 11:20 am E2-150

Creagh Briercliffe (University of British Columbia), Alexandre Bouchard-Côté (University of British Columbia), Paul Gustafson (University of British Columbia)

Poisson Process Infinite Relational Model: a Bayesian nonparametric model for transactional data
Modèle relationnel infini pour un processus de Poisson: Un modèle bayésien non paramétrique pour les données transactionnelles

Transactional data consists of instantaneously occurring observations made on ordered pairs of entities. For example, a set of timestamped emails between coworkers, with one sender and one recipient. Visually, it can be represented as a network, or more specifically, a directed multigraph with edges possessing unique timestamps. In this talk, I explore a Bayesian nonparametric model for discovering latent class-structure in transactional data. By pooling information within clusters of entities, this model can be used to infer the underlying dynamics of the time-series data.

Les données transactionnelles consistent en des observations produites de façon instantanée, faites sur des paires ordonnées d'entités. Par exemple, un ensemble de courriels estampillés entre collègues, avec un expéditeur et un destinataire. Visuellement, on peut représenter ceci par un réseau, ou plus spécifiquement, par un multigraphe direct avec des bords possédant des horodatages uniques. Lors de cette présentation, j'explore un type de modèle bayésien non paramétrique afin de découvrir une structure de classe latente dans les données transactionnelles. En combinant l'information entre les groupes d'entités, ce modèle peut servir à inférer sur les dynamiques sous-jacentes aux données de séries temporelles.

Hongcan Lin (University of Waterloo), David Saunders (University of Waterloo), Chengguo Weng (University of Waterloo)

Optimizing Performance Ratio via Martingale Approach
Ratio des performances d'optimisation sous l'approche des Martingales

We consider the continuous time portfolio selection problem for an investor seeking to maximize a performance ratio. We show that the problem is unbounded for some performance measures popular in practice (the Omega measure in particular), and then analyze a modified problem that is well-posed. In particular, we derive semi-analytical expressions for the optimal strategy in the case where the reward and risk are power functions of the excess and deficit with respect to a fixed benchmark level.

Nous considérons le problème de sélection du portefeuille en temps continu pour un investisseur qui désire maximiser un ratio de performances. Nous montrons que le problème n'est pas délimité pour certaines mesures de performance qui sont populaires en pratique (la mesure Omega, en particulier), et analysons ensuite un problème modifié qui est bien posé. En particulier, nous dérivons des expressions semi-analytiques pour la stratégie optimale, dans le cas où les remises et le risque sont des fonctions de puissance de l'excès et du déficit, par rapport à un certain niveau de référence fixé.

Sports Analytics • Analytique du sport
10:50am - 11:20 am E2-155

Matthew van Bommel (Simon Fraser University), Luke Bornn (Simon Fraser University)

Adjusting for Scorekeeper Bias in NBA Box Scores
Ajustement pour le biais dû au marqueur dans les résultats de la NBA

Box score statistics in the National Basketball Association are used to measure and evaluate player performance. Some of these statistics are subjective in nature and since they are recorded by scorekeepers hired by the home team for each game, there exists potential for inconsistency and bias. Using box score and optical player tracking data from the 2015-2016 season, we estimate a model able to quantify both the bias and the generosity of each scorekeeper in awarding assists. From this model, we present results measuring the impact of the scorekeepers and of the other contextual variables on the probability of a pass being recorded as an assist.

Les statistiques concernant les points et les passes dans l'Association nationale de basketball (NBA) sont utilisées afin de mesurer et d'évaluer la performance des joueurs. Certaines de ces statistiques sont subjectives de par leur nature, et puisqu'elles sont enregistrées par des marqueurs qui ont été engagés par l'équipe locale pour chaque partie, il y a la possibilité de biais et d'inconsistance. En utilisant les données de la saison 2015-2016 contenant les résultats des joueurs et leur suivi des lectures optiques, nous estimons un modèle qui peut quantifier le biais et la générosité de chaque marqueur quant à l'attribution de points d'assistance. À partir de ce modèle, nous présentons les résultats mesurant l'impact des marqueurs et d'autres variables contextuelles sur la probabilité qu'une passe soit enregistrée comme une assistance.

Nathan Sandholtz (Simon Fraser University), Luke Bornn (Simon Fraser University)

Replaying the NBA: Using Markov decision processes to test decision-making from the 2015-2016 regular season
Rejouer la NBA: Utiliser des processus de décision de Markov pour tester la prise de décision durant la saison régulière de 2015-2016

Last year, the Cleveland Cavaliers took 152 contested long range 2-point shots with at least 14 seconds remaining on the shot clock. What couldÍve happened if they had instead reset these possessions with a pass, creating potential for more valuable shots? We attempt to answer these types of questions by modeling possessions from the 2015-2016 NBA regular season as Markov chains realized from team-specific Markov decision processes. Using spatiotemporal data which tracks the movements of all players simultaneously, we use hierarchical Bayesian methods to estimate each teamÍs latent decision policy. Using posterior draws from the estimated policies, we simulate ñalternateî regular seasons for various NBA teams, identify suboptimal transition patterns, and explore potential outcomes under more efficient policies.

L'année dernière, les Cavaliers de Cleveland ont effectué 152 tirs contestés de 2 points avec au moins 14 secondes restantes au chronomètre de tirs. Qu'aurait-il pu arriver s'ils avaient plutôt réinitialisé leur possession du ballon avec une passe, créant ainsi un potentiel pour des tirs de plus grande valeur? Nous essayons de répondre à ce type de questions en modélisant les possessions de ballon de la saison régulière 2015-2016 de la NBA comme des chaînes de Markov réalisées à partir de processus de décision de Markov spécifiques à l'équipe. En utilisant des données spatio-temporelles qui suivent simultanément les mouvements de tous les joueurs, nous utilisons des méthodes bayésiennes hiérarchiques pour estimer la politique de décision latente de chaque équipe. À l'aide des tirages postérieurs des politiques estimées, nous simulons des saisons régulières «alternatives» pour diverses équipes de la NBA, identifions des modèles de transition sous-optimaux et explorons les résultats potentiels sous des politiques plus efficaces.

Scientific Abstracts: Poster Presentations • Résumés Scientifiques: Présentations d'Affiches

Mehdi Rostami (University of Toronto), Dr. Wendy Lou (University of Toronto)

Predictive Models for Prediction of Hospitalization in Canadian Long-Term Care Facilities
Modèles Prédictifs pour l'Hospitalisation dans les Centres Canadiens de Soins de Longue Durée

Prediction of hospitalization of residents of the long-term care facilities (LTCFs) is one of the challenges whose prevention can protect residents from deteriorating their health conditions and reduce health care costs. This research attempts to build predictive models to predict future hospitalization in Canadian LTCFs. A retrospective cohort study was carried out for new residents of the Canadian LTCFs in 2013. The Continuing Care Reporting System and Discharge Abstract Databases hosted by CIHI are main data sources. Three methods, logistic regression, decisions trees, and support vector machines, with and without rank reduction methods are applied and compared. 5-fold validation is used to avoid over-fitting and area under ROC is used to assess the performance of machine learning methods.

La prédiction de l'hospitalisation pour les résidents de centres de soins de longue durée (CSLD) est un défi pour lequel la prévention peut éviter aux patients une détérioration de leur condition de santé et peut réduire les coûts des soins de santé. Avec ce projet de recherche, nous tentons de construire des modèles prédictifs pour prédire l'hospitalisation future dans les CSLD canadiens. Une étude de cohorte rétrospective a été menée pour les nouveaux résidents de CSLD canadiens en 2013. Les bases de données des Congés sur les Patients et du Système d'Information sur les Soins de Longue Durée, hébergées par l'Institut canadien d'information sur la santé (ICIS), sont les principales sources de données. Trois méthodes sont appliquées et comparées: la régression logistique, les arbres de décisions, et les machines à vecteurs de support, avec ou sans l'utilisation de méthodes de réduction des rangs. La validation en 5-pli est utilisée pour éviter l'«overfitting», et la région sous la courbe ROC est utilisée pour évaluer la performance des méthodes de «machine learning».

Yuyan Yang (University of Victoria), Laura L.E. Cowen (University of Victoria), Maycira Costa (University of Victoria), Ziwei Wang (University of Victoria)

Evaluation of Ocean Colour Spectra Acquired by Ferry Passengers in the Salish Sea
Évaluation du spectre de couleurs de l'océan acquis par des passagers de traversier dans la mer des Salish

In situ water reflectance data is of great value due to the requirements for validation of satellite images and development of satellite-based regional models for estimating biogeochemical properties of the ocean. The increasing popularity of mobile devices provides an opportunity for citizen scientists to use the camera in their devices as a sensor to acquire data for scientific research. This research assesses the quality of ocean colour data acquired by ferry passengers using tablets. Paired T-tests and linear regression are used to evaluate the relationship between the traditional instrument data and the citizen application data under different environmental conditions. This data can be used to validate satellite data if there is a fixed relationship between instrument and application data.

Les données de réflectance de l'eau in situ sont d'une grande valeur en raison des exigences de validation des images satellite et du développement de modèle régionaux satellitaires pour l'estimation des propriétés biogéochimiques de l'océan. La popularité croissante des appareils mobiles permet aux scientifiques citoyens d'utiliser la caméra de leur appareil comme senseur pour acquérir des données pour la recherche scientifique. Cette recherche évalue la qualité de données de couleur océanique acquises par les passages d'un traversier en utilisant des tablettes. Les tests-t appariés et la régression linéaire sont utilisés pour évaluer la relation entre les données provenant d'instruments traditionnels et les données d'applications de citoyens sous différentes conditions environnementales. Ces données peuvent être utilisées pour valider les données satellites s'il existe une relation fixe entre les données d'instruments et celles d'applications.

Yue Yin (University of Victoria), Dr. Julie Zhou (University of Victoria), Dr. Weng Kee Wong (University of California)

Using SeDuMi to Compute Various Optimal Designs for Regression Models
Utilisation de SeDuMi pour obtenir différents modèles de régression optimaux

Optimal regression design problems have been studied for various linear and nonlinear models in the literature. For many regression models and optimality criteria, it is hard to derive optimal designs analytically. Several numerical algorithms have been developed to compute the optimal designs. However, there are some issues with these numerical algorithms. In this paper we propose an efficient and effective algorithm based on a powerful optimization tool in MATLAB, SeDuMi (self-dual minimization), for finding optimal designs. We can apply this algorithm to compute various optimal designs including A-, As-, c-, I-, and L-optimal designs. This algorithm is flexible. It can be applied for any linear or nonlinear models. Several examples are presented, and the results are all verified by using the Kiefer-Wolfowitz equivalence theorem.

Dans la littérature, les problèmes de modèles de régression optimaux ont été étudiés pour divers modèles linéaires et non-linéaires. Pour plusieurs modèles de régression et critères d'optimalité, il est difficile de dériver des modèles optimaux de façon analytique. Plusieurs algorithmes numériques ont été développés afin d'atteindre ces modèles optimaux. Cependant, ces algorithmes numériques présentent certains problèmes. Dans cet article, nous proposons un algorithme efficace pour trouver des modèles optimaux, basé sur un outil d'optimisation puissant du logiciel MATLAB, SeDuMi (self-dual minimization). Nous pouvons utiliser cet algorithme afin de trouver plusieurs types de modèles optimaux incluant les modèles A-, As-, c-, I- et L-optimal. L'algorithme proposé est flexible et peut s'appliquer à n'importe quel modèle, linéaire ou non. Plusieurs exemples sont présentés et les résultats sont vérifiés à partir du théorème d'équivalence de Kiefer-Wolfowitz.

Sahar Ahmed (Montclair State University), Andrew McDougall (Montclair State University)

Analysis of Daily New Jersey Precipitation
Analyse des précipitations quotidiennes au New Jersey

Global warming is a contentious topic since modern climate records only exist for the last 100+ years (in contrast to ice-core analysis that establishes ice ages tens of thousands of years ago). Nevertheless, patterns associated with precipitation amounts over the last century can provide a useful indicator of climate change. This project focuses on daily precipitation totals in the state of New Jersey over the last 100 to 150 years from 19 meteorological recording sites. Our aim is to see if these data show an increase in major precipitation events over recent years and if so, then are these events localized or statewide. We present several approaches to the statistical and exploratory analysis of this dataset.

Le réchauffement climatique est un sujet controversé puisque les archives climatiques modernes n'existent que depuis 100+ ans (contrairement à l'analyse des carottes glaciaires qui établit l'ère de glaces à il y a des dizaines de milliers d'années). Néanmoins, des motifs associés aux quantités de précipitations au cours du siècle dernier peut fournir un indicateur utile des changements climatiques. Ce projet se concentre sur les totaux quotidiens de précipitations dans l'état du New Jersey au cours des 100 à 150 dernières années provenant de 19 sites d'enregistrement météorologique. Notre objectif est de voir si ces données montrent une augmentation des précipitations majeures ces dernières années, et si oui, de déterminer si ces événements sont locaux ou à l'échelle de l'état. Nous présentons plusieurs approches à l'analyse statistique et exploratoire de cette base de données.

Miguel Macaraig (MacEwan University)

A Time Series Approach for Forecasting the Weekly Percentages of Influenza in Canada
Une approche par séries temporelles pour prédire les pourcentages hebdomadaires d'influenza au Canada

Influenza is the most common respiratory virus in Canada. This work is an attempt to establish a time series model, based on weekly reports from September 7, 2003 to August 23, 2015 of percentage of flu in Canada (from Canada's FluWatch). As expected the weekly percentage of flu is seasonal, and our analysis show that a SARIMA (2,1,2)x(0,1,1)52 model is the best fit. Using the proposed model we obtain accurate 10 weeks ahead forecasts. The ability to forecast the rate of flu is important for conducting preventive measures which will lower the incidence of flu. We have conducted also a frequency domain analysis and a cross correlation analysis between the data from Canada's FluWatch and Google Flu Trends.

L'influenza est le virus respiratoire le plus commun au Canada. Ce travail est une tentative d'établir un modèle de série temporelle basé sur les rapports hebdomadaires du 7 septembre 2003 au 23 août 2015 du pourcentage de grippe au Canada (provenant de FluWatch Canada). Comme prévu, le pourcentage hebdomadaire de grippe est saisonnier, et notre analyse montre qu'un modèle SARIMA (2,1,2)x(0,1,1)52 constitue le meilleur ajustement. En utilisant le modèle proposé, on obtient des prévisions précises 10 semaines à l'avance. La capacité de prévoir le taux de grippe est importante pour établir des mesures préventives qui réduiront l'incidence de la grippe. Nous avons également effectué une analyse de domaine fréquentiel et une analyse de corrélation croisée entre les données de FluWatch Canada et de Google Flu Trends.

Logan Ewanchuk (MacEwan University)

Time Series Analysis of National League Slugging Percentage
Analyse de séries chronologiques des moyennes de puissance dans la ligue nationale

Major League Baseball records yearly National League slugging percentage values since 1901. For this project 113 observations, from 1901-2013, were examined using a time series model. We started with an exploratory data analysis, followed by model selection. Once a list of possible Autoregressive Integrated Moving Average (ARIMA) models was made, the next step was fitting the models to the series and estimating the coefficients. The observations for the years 1901-2003 were used as the training set, and we predicted the values for the next ten years, from 2004-2013. Based on the accuracy of the predictions, an ARIMA(1,0,0) model was chosen. A bicoherence analysis was also performed, comparing American League slugging percentage over the same time period to the National League.

La ligue majeure de baseball enregistre les moyennes de puissance annuelles de la ligue nationale depuis 1901. Pour ce projet, 113 observations mesurées de 1901-2013 ont été examinées à partir de modèles de séries chronologiques. Nous avons d'abord fait une analyse exploratoire des données, suivie par une sélection de modèle. Une fois que nous avions choisi une liste de modèles potentiels de type ARIMA, la prochaine étape était l'ajustement des modèles aux séries et l'estimation de coefficients. Nous avons utilisé les observations de 1901-2003 comme données d'apprentissage et avons prédit les valeurs pour les dix prochaines années, de 2004-2013. En se basant sur la précision des prédictions, nous avons finalement choisi un modèle ARIMA(1,0,0). Une analyse de bi cohérence a été performée afin de comparer les moyennes de puissance de la ligue américaine aux moyennes de la ligue nationale sur la même période.

Radia Taisir (University of Guelph)

Dealing Non-ignorable Missingness in Longitudinal Data with EM Algorithm
Traiter les données manquantes non ignorable d'études longitudinales avec l'algorithme espérance-maximisation (EM)

It is very natural to occur non-ignorable missingness in longitudinal studies where the same experimental units are observed repeatedly over time. To model the binary response, a Markov model may be used along with a suitable non-response model for the missing portion of the data. Similar model exists for such incomplete longitudinal data where the estimation of the regression parameters is carried out by likelihood method, summing over all possible values of the missing observations. This study introduces an EM algorithm technique for the estimation purpose, which is computationally simple and produces similar efficient estimates as the existing complex estimation method. The Health and Retirement Survey data from United States are analyzed to show the comparison.

Il est très naturel d'observer de données manquantes non ignorables dans les études longitudinales lorsque les mêmes unités expérimentales sont observées à plusieurs reprises au fil du temps. Pour modéliser l'issue binaire, un modèle de Markov peut être utilisé accompagné d'un modèle de non-réponse approprié pour la portion manquante des données. Un modèle similaire existe pour de telles données longitinales incomplètes où l'estimation des paramètres de régression s'effectue par maximum de vraisemblance, en additionnant toutes les valeurs possibles des observations manquantes. Cette étude présente une technique d'algorithme EM à des fins d'estimation, qui est simple en termes de calculs et qui produit des estimations efficaces similaires à celles de la méthode d'estimation complexe existante. Les données de la Health and Retirement Survey des États-Unis sont analysées pour montrer la comparaison.

Myrtha Reyna (University of Toronto), Nicholas Mitsakakis (University of Toronto)

Ordinal Regression for the Analysis of Health Utility Data
Régression ordinale pour l'analyse de données de santé publique

Health utility (HU) data are negatively skewed and bounded by 1, with most observations lying close to that bound and some with extremely low levels. HU has been analyzed by linear regression (OLS) which assumes normally distributed errors and lacks a restriction at the upper limit; Ordinal Regression (OR) is not limited by such assumptions and it produces estimates bounded at 1. This study compares the performance of OR to OLS for the prediction of HU given covariates, using simulated and real data of prostate cancer patients. Using various sample sizes and 3 different distributions, HU data were simulated and analyzed by OLS and OR. Models were evaluated by the bias and coverage probability of the estimated mean, and by using the Root Mean Square Error (RMSE) comparing simulated and predicted HU values. Results show that OR provides more accurate estimates regardless of sample size, while OLS bias increases as HU approach zero. Coverage probability is similar in both methods.

Les données de santé publique (SP) sont négativement asymétriques et bornées supérieurement par 1, avec la plupart des observations se situant proche de cette borne et certaines de valeur extrêmement basse. Les données de SP ont été analysées par régression linéaire (RL), méthode qui suppose la normalité des erreurs et qui n'a pas de restriction quant à une borne supérieure; la régression ordinale (RO) n'est pas limitée par ces hypothèses et produit des estimations bornées par 1. Cette étude compare la performance de la RO avec celle de la RL pour la prédiction de données de santé publique conditionnellement à certaines covariables, en utilisant des données simulées et réelles sur le cancer de la prostate. Utilisant diverses tailles échantillonnales et 3 différentes distributions, des données ont été simulées et analysées par la RL et la RO. Les modèles ont été évalués par l'entremise du biais et de la probabilité de couverture des moyennes estimées, et en utilisant la racine de l'erreur quadratique moyenne comparant les valeurs simulées et prédites. Les résultats montrent que la RO procure des estimations plus exactes peu importe la taille échantillonnale, tandis que le biais de la RL augmente quand les données approchent de zéro. La probabilité de couverture est similaire pour les deux méthodes.

Weining Hu (University of British Columbia), David Zheng (University of British Columbia)

Latent Dirichlet Allocation: Can we Push More?
Allocation de Dirichlet latente: Peut-on aller plus loin?

The domain of this project would focus on is topic modelling. As we are living in a data intensive age, there are more information available in various formats which makes it harder focus to extract key contents out. Thus, we need algorithms to help us understand, summarize and search these massive amount of data. Among different types of data, we wish to focus on the application of probabilistic topic modeling on text data, more specifically, the Latent Dirichlet allocation(LDA). We try to push more on the original model to explore two of its extensions: Correlated Topic Model and Dynamic Topic Model.

Ce projet se concentre sur la modélisation de thématique. Comme nous vivons à une ère où les données sont intensives, de plus en plus d'information est disponible, et ce dans divers formats, et il devient difficile de se concentrer et d'extraire de ces données le contenu important. Ainsi, nous avons besoin d'algorithmes pour nous aider à comprendre, résumer et chercher à l'intérieur de ces données. Parmi différents types de données, nous souhaitons nous concentrer sur l'application de la modélisation probabilistique de thématique, sur des données provenant de textes, plus spécifiquement sur l'allocation de Dirichlet latente (ADL). Nous nous intéressons surtout au modèle original et voulons explorer deux de ses extensions : Modèles de thématique corrélés et modèles de thématique dynamiques.

Ken Tam (University of Toronto), Samer Salah (The Princess Margaret Hospital), Nathan Taback (University of Toronto)

Predictors of Thoracic LN Involvement in Colorectal Cancer Patients with Pulmonary Oligometastasis: a Pooled Individual Patients Data Analysis
Prédicteurs de l'action des ganglions lymphatiques thoraciques chez les patients atteints du cancer colorectal avec oligometastase pulmonaire : une analyse de données combinées par patient

This study analyzed 760 colorectal cancer patients with pulmonary oligometastatic disease treated with pulmonary metastasectomy using a pooled dataset of 7 similar studies. Lymph Node dissection is routinely done during the surgery, but whether it should be done remains controversial since it hasn't shown to improve the survival rate of patients. The result from the data shown by fitting a logistic regression was that the patients who underwent Lymph Node dissection had a lower survival rate with an odds ratio of 1.68 and a p-value of 0.01. Factors was also assess in Lymph Node association using a logistic regression in which it shows that synchronous cancer was positive correlated with a odds value of 1.92 and a p-value of 0.02.

Dans cette étude, nous avons analysé les données de 760 patients atteints du cancer colorectal avec oligometastase pulmonaire, étant traités par métastasectomie pulmonaire, à partir d'un jeu de données combinées de 7 études similaires. La dissection des ganglions lymphatiques est une procédure commune pendant la chirurgie, mais demeure une pratique controversée puisqu'elle n'a pas montré augmenter le taux de survie des patients. Les résultats d'une régression logistique ont montré que les patients ayant subi une dissection des ganglions lymphatiques avaient un taux de survie inférieur avec un rapport de cotes de 1.68 et une valeur-p de 0.01. Les facteurs de risque associés aux ganglions lymphatiques ont aussi été étudiés à partir de la régression logistique, qui a montré qu'un cancer simultané était corrélé positivement, avec un rapport de cotes de 1.92 et une valeur-p de 0.02.

Yu-Chung Lin (University of Toronto), Jeremy Lewin (Princess Margaret Cancer Centre), Nathan Taback (University of Toronto)

Evaluation of Perceptions of Adolescents and Young Adults' willingness to participate in clinical trials
Évaluation de la perception de la volonté des adolescents et jeunes adultes à participer aux essais cliniques

Adolescent and Young adults (AYA) with cancer have the lowest cancer clinical trial (CT) enrollment rate of any age demographic. The aim of the project is to understand the differences in perceptions between AYA & non-AYA patients in cancer clinical trials. Furthermore, to explore baseline factors that may influence AYA's decision making on CT participation. Mann-Whitney U test identified attitudinal differences in personal barriers to CT between the two cohorts. AYA patients expressed greater concerns on safety of CT and its potential to affect their long term life goals. Ordered logistic regression models further showed how experiences of being offered CT enrollment and having English as the first language are correlated with positive attitudes towards CT.

Les adolescents et jeunes adultes (AJA) atteints du cancer ont le taux le plus bas de participation de toutes les strates démographiques, dans les essais cliniques sur le cancer. Le but de ce projet est de comprendre les différences dans les perceptions des patients AJA et non-AJA dans les essais cliniques. De plus, nous voulons explorer les facteurs de base pouvant influencer les décisions prises par les AJA concernant leur participation aux essais cliniques. Le test U de Mann-Whitney a identifié des différences d'attitude entre les deux cohortes, quant aux obstacles personnels face aux essais cliniques. Les patients AJA ont exprimé plus de crainte par rapport à la sécurité des essais cliniques et leur potentiel à influencer leur vie à long terme. Des modèles de régression logistique ordinale ont ensuite montré comment le fait de se faire offrir une participation aux essais cliniques et le fait de parler anglais étaient corrélés à une attitude positive envers les essais cliniques.

Shubham Sharma (Dalla Lana School of Public Health, University of Toronto), Rosane Nisenbaum (Centre for Urban Health Solutions, St. Michael's Hospital)

Handling Missing Data in the Quality of Life Interview (QoLI-20) Items
Tenir compte des données manquantes dans les items de l'entrevue sur la qualité de vie (QoLI-20)

Homelessness has been associated with many negative health consequences in the literature. The Coordinated Access to Care for the Homeless (CATCH-H) program aims to help this vulnerable population navigate through healthcare and reintegrate into their community. Our primary outcome was the Lehman Quality of Life Interview (QoLI-20) measured at baseline and 6 months. Our question was whether there is significant association between clinical diagnosis and change in QoLI-20 scores, adjusting for baseline covariates. However, there were non-negligible amounts of missing data in this longitudinal study. Hence, the main aim was to explore multivariate imputation by chained equations (MICE) as an effective tool for handling missing data and compare parameter estimates from multiple regression post-imputation to those done using complete-case analysis.

L'itinérance a été associée à plusieurs conséquences négatives sur la santé dans la littérature. Le programme Accès coordonné aux soins pour les itinérants (CATCH-H) a pour but d'aider cette population vulnérable à naviguer à travers les soins de santé et à les réintégrer dans leur communauté. La variable primaire étudiée est l'entrevue de Lehman sur la qualité de vie (QoLI-20), passée au début du suivi ainsi qu'après 6 mois. Nous voulons savoir s'il y a une association significative entre le diagnostic clinique et le changement dans le score QoLI-20, après ajustement pour des variables mesurées au départ. Cependant, cette étude longitudinale présente des quantités non négligeables de données manquantes. Ainsi, le but principal était d'explorer l'imputation multiple via la méthode MICE en tant qu'outil efficace pour tenir compte des valeurs manquantes et de comparer les estimations des paramètres de la post-imputation avec régression multiple à ceux obtenus à partir de l'analyse sur les cas complets.

Lahiru Wickramasinghe (University of Manitoba), Alexandre Leblanc (University of Manitoba), Saman Muthukumarana (University of Manitoba)

Handling Sparsity in Contingency Tables
Tenir compte des données parcimonieuses dans les tableaux de contingence

Categorical data are often analyzed as a contingency table and sparse contingency tables are very common in practice when there are many cells with small counts and/or zeros. This sparsity is very common in practice when we have large number of classification variables and/or variables with many levels, even the sample size is large. One remedy is to merge or drop categories from the contingency table, which may cause important information to be lost. The sparseness invalidates many classical approaches which are used to analyze contingency tables. We present approaches to smooth sparse contingency tables with ordered data. The smoothing can be carried out by borrowing information from the neighboring cells.

Les données catégorielles sont souvent analysées sous forme de tableaux de contingence et il est assez commun, en pratique, d'utiliser des tableaux de contingence pour données parcimonieuses lorsque les cellules ont un petit nombre d'effectifs (ou même zéro effectif). Cette parcimonie est très commune en pratique lorsque nous avons un grand nombre de variables de classification ou de variables à plusieurs niveaux, même si la taille du jeu de données est grande. Un des remèdes à la situation est de combiner ou encore d'exclure certaines des catégories du tableau de contingence, ce qui peut causer une perte d'information importante. La parcimonie rend invalide l'utilisation de plusieurs méthodes classiques pour l'analyse de tableaux de contingence. Nous présentons des méthodes permettant de lisser les tableaux de contingence parcimonieux avec des données ordonnées. Le lissage peut être entrepris en empruntant de l'information des cellules voisines.

Mitchell Sutton (University of Toronto)

The Effects of Traffic Pollution on Lung Function Development in Children
Les effets de la pollution due au trafic, dans le développement des fonctions pulmonaires des enfants

Background and Introduction: Our focus in this study was to analyze how exposure to traffic pollution effects the lung function development in infants and young children in Toronto.
Methods: Lung function (of 231 subjects) was measured as Lung Clearance Index (LCI), while exposure to traffic pollution estimated average daily exposure to NO2 using a land use regression. The effect was modeled using a random intercept, linear mixed model.
Results: No significant results were found, however, observed a positive effect on LCI (worsening lung function) due to the increased exposure in NO2, measured as a cumulative effect over the first year of a child's life (p-value=0.2674) and when the measurement was matched to the closest LCI test result (p-value=0.125).

Contexte et introduction: Dans cette étude, nous nous concentrons sur l'analyse de l'exposition à la pollution due à la circulation et son effet sur le développement des fonctions pulmonaires chez les bébés et les jeunes enfants de Toronto.
Méthodes: Les fonctions pulmonaires (de 231 individus) ont été mesurées selon l'indice de clairance pulmonaire (ICP), alors que l'exposition à la pollution due à la circulation a été mesurée en termes d'estimation de la moyenne journalière d'exposition au NO2 à partir d'une régression de type LUR. L'effet a été modélisé en utilisant un modèle linéaire mixte avec ordonnées aléatoire.
Résultats: Nous n'avons trouvé aucun résultat significatif mais nous avons tout de même trouvé un effet positif sur l'ICP (empirant les fonctions pulmonaires) dû à l'augmentation de l'exposition au NO2, mesuré comme étant l'effet cumulatif sur la première année de vie de l'enfant (valeur-p=0.2674) et lorsque les observations étaient appariées au plus près ICP (valeur-p=0.125).

Sahar Arshadi (Memorial University of Newfoundland), Taraneh Abarin (Memorial University of Newfoundland)

On Identifiability of the Regression Models with Interaction
À propos de l'identifiabilité des modèles de régression avec interaction

Before any statistical inference on regression model parameters, we need to know whether the parameters of interest are "estimable" or "identifiable". The model is said to be identifiable if all the unknown parameters of the model can be estimated uniquely provided data. In this presentation, we consider different interaction models, both with and without measurement error. We will look at different measurement error models such as Berkson and Classic and apply some remedies for non-identifiability issue, such as instrumental variables as well as replicated and validated data. Key words and phrases: Identifiability, measurement error, interaction

Avant de faire de l'inférence statistique sur les paramètres des modèles de régression, nous devons savoir si les paramètres sont "estimables" ou "identifiables". On dit que le modèle est "identifiable" si tous les paramètres inconnus du modèle peuvent être estimés de façon unique, avec ces données particulières. Lors de cette présentation, nous considérons différents modèles avec interaction, avec ou sans erreur de mesure. Nous examinerons différents modèles d'erreur de mesure, tels que Berkson et Classic, et apporterons quelques solutions au problème de non-identifiabilité, comme les variables instrumentales ou les données validées et répliquées. Mots-clés: Identifiabilité, erreur de mesure, interaction

Mahbuba Sultana (Memorial University of Newfoundland), Mohammad Samsul Alam (University of Dhaka)

Factor Effecting the Nutritional Status of Children Under Five in Bangladesh
Les facteurs influençant l'état nutritionnel d'enfants de moins de cinq ans au Bangladesh

Child malnutrition is a major leading causes of morbidity and mortality among children under five in Bangladesh. The objective of this study is to asses nutritional status of children and find associated risk factors. Bivariate analysis has been performed in this regard for three types of malnutrition - stunted, underweighted and wasted. Based on Z-score, the malnutrition status is categorized into - severe (Z-score greater than 3.0), moderate (3.0 to 2.0) and mild (less than 2.0). Proportional odds model can be developed using the data from Bangladesh Demographic and Health Survey - 2011. This analysis observes that 60% children are mild, 28% moderate and 12% are severely stunted; 64% mild, 26% moderate and 10% are severely underweighted and 81%, 15% and 4% are mild moderate and severely wasted respectively. Risk factors observed are - place of residence, sources of drinking water, parents education, fathers occupation, wealth index, child anemia, vaccination and number of siblings in the family.

La malnutrition infantile est une des principales causes majeures de morbidité et de mortalité parmi les enfants de moins de cinq ans au Bangladesh. L'objectif de cette étude est d'évaluer l'état nutritionnel des enfants et de trouver les facteurs de risque associés. Une analyse bivariée a été effectuée à cet égard pour trois types de malnutrition - le retard de croissance, l'insuffisance pondérale et l'émaciation. Basé sur le score Z, l'état de malnutrition est catégorisé entre: sévère (score Z supérieur à 3.0), modéré (3.0 à 2.0) et léger (inférieur 2.0). Le modèle de cotes proportionnel peut être développé en utilisant les données du Bangladesh Demographic and Health Survey de 2011. Cette analyse constate que 60% des enfants ont un retard de croissance léger, 28% modéré et 12% sévère; 64% des enfants ont une insuffisance pondérale légère, 26% modérée et 10% sévère; et 81%, 15% et 4% des enfants ont une émaciation légère, modérée et sévère, respectivement. Les facteurs de risque observés sont: le lieu de résidence, les sources d'eau potable, l'éducation des parents, l'occupation des pères, l'indice de richesse, l'anémie infantile, la vaccination et le nombre de frères et soeurs dans la famille.

Steven Wu (Simon Fraser University), Dr. Tim Swartz (Simon Fraser University)

Automatically Correcting Play-by-play Substitution Errors in Basketball
Correction automatique des erreurs de substitution 'Play-by-play' au basketball

Basketball play-by-play data offers unique insights that traditional boxscore summaries can't by knowing who is on the court at all times in the game, along with the time remaining and score differential. Canada's U Sports league has play-by-play, however, the logging of substitutions is inconsistent: missing substitutions and unbalanced number of players entering vs. leaving the game are two of the many examples we observe. With ˜400 games per season, an automated framework is necessary for correcting these errors if we want to extract analytics that can be useful to coaches for strategy. We describe the implementation of a software agent that can successfully classify incorrect substitutions and impute missing ones and quantify its effectiveness.

Les données 'Play-by-play' au basketball offrent une information unique que les résumés traditionnels du score dans la boîte ne peuvent offrir, en sachant qui est sur le terrain en tout temps, en plus du temps restant et des différentiels de score. La ligue de sport universitaire du Canada possède des données Play-by-play mais l'enregistrement des substitutions n'est pas cohérent: deux exemples souvent observés sont le manque de certaines substitutions et un nombre de joueurs entrant et sortant du jeu qui n'est pas balancé. Avec environ 400 matchs par saison, un cadre automatisé est nécessaire afin de corriger ces erreurs, si nous désirons extraire et analyser des données qui peuvent être utilisées pour améliorer les stratégies des joueurs. Nous décrivons l'implémentation d'un agent logiciel qui peut classifier avec succès les substitutions incorrectes et imputer celles qui sont manquantes, et nous quantifions son efficacité.

Mehdi Arzandeh (University of Manitoba), Julieta Frank (Department of Agribusiness and Agricultural Economics, University of Manitoba)

The Information Content of the Limit Order Book
L'information contenue dans le carnet d'ordres à cours limité

Trading in futures markets occurs through a computerized system where incoming buy and sell limit orders are stored in a limit order book (LOB). Such orders contain traders intended prices and number of contracts to buy or sell and may therefore reflect traders' new information. We assess the extent to which the LOB contains information that is valuable in determining the price of a futures contract (i.e., price discovery). The level of information of the LOB is measured as its contribution to the transaction price innovation variance. We reconstruct the LOB for six major products and use three different approaches based on an error correction model. The results suggest that the LOB contributes by over 27% to price discovery.

Le commerce des marchés futurs se produit à travers un système informatisé dans lequel les ordres limités d'achat et de vente sont enregistrés dans un carnet d'ordres à cours limités (OCL). Ces ordres contiennent le nombre de contrats à acheter ou vendre et les prix prévus par les commerçants et peuvent donc refléter des informations nouvelles sur les commerçants. Nous évaluons à quel point l'information contenue dans les OCL est pertinente pour déterminer le prix d'un contrat futur (i.e. découverte de prix). Le niveau d'information des OCL est mesuré par sa contribution à la variation de l'innovation des prix des transactions. Nous reconstruisons l'OCL pour les six produits principaux et utilisons trois approches différentes basées sur un modèle corrigé pour l'erreur. Les résultats suggèrent que l'OCL contribue à plus de 27% de la détermination des prix.

Shamsia Sobhan (University of Manitoba), Dr Mahbub Latif (St Luke's International University, Japan)

Joint Modeling in the Presence of Competing Risks: An Application to Diabetes Data
Modélisation conjointe en présence de risques concurrents: une application aux données de diabète

Longitudinal and survival outcomes are often studied in many clinical and medical studies. Separate analysis of such outcomes produces bias and inefficient results, specially when both the outcomes are correlated. Joint modeling approach has become popular for modelling longitudinal and survival outcomes simultaneously. In this study, we are interested in estimating the effect of longitudinal blood glucose level on the time to first diagnosis of one of three diabetes related complications, which are cardiovascular, retinopathy, and chronic kidney disease. The main goal of this study is to see whether there is any difference in the association between the longitudinal blood glucose level and the time of first diagnosis of one of the three diabetes related complications.

Les événements longitudinaux ou liés à la survie sont fréquemment étudiés dans plusieurs études cliniques et médicales. Des analyses séparées de ces événements peuvent mener à un biais ou à des résultats inefficaces, surtout lorsque les événements sont corrélés entre eux. L'approche de modélisation conjointe est devenue populaire pour modéliser les événements longitudinaux et liés à la survie de façon simultanée. Dans cette étude, nous nous intéressons à l'estimation de l'effet du niveau longitudinal de glucose dans le sang sur le temps menant au premier diagnostic de l'un des trois types de complications reliées au diabète, qui sont les complications cardiovasculaires, la rétinopathie et la maladie rénale chronique. Le but premier de cette étude est de vérifier s'il existe une différence dans l'association entre le niveau longitudinal de glucose sanguin et le temps menant au premier diagnostic, pour l'un des trois types de complications reliées au diabète.

Anita Brobbey (University of Calgary), Samuel Wiebe (Department of Clinical Neurosciences, University of Calgary), Meng Wang (Department of Clinical Neurosciences, University of Calgary), Zhiying Liang (Community Health Sciences, University of Calgary), Shane Goodwin (Department of Epidemiology & Biostatistics, Western University), Mark A. Ferro (School of Public Health and Health Systems, University of Waterloo), Kathy N. Speechley (Department of Epidemiology & Biostatistics, Western University), Tolulope T. Sajobi (Community Health Sciences, University of Calgary)

Variations in Disease Severity in Children with Epilepsy
Variations dans la sévérité de l'épilepsie chez les enfants

Severity of epilepsy has been associated with clinical outcomes in individuals with epilepsy. We investigate variations in neurologists' ratings of disease severity across time in children newly diagnosed with epilepsy (HERQULES). Severity of epilepsy was measured using the neurologist reported Global Assessment of Severity of Epilepsy (GASE) scale. Repeated measures latent class analysis (RMLCA) was used to characterize the longitudinal trajectories for severity of epilepsy over a two-year period. Multinomial logistic regression was used to identify predictors among the latent classes of trajectories. RMLCA identified four distinct trajectories of severity of epilepsy; "Early Large Improvement" (12.9%), "Early Moderate Improvement" (46.3%), "Late Moderate Improvement" (26.6%), and "Unchanged" (14.2%). The identified four distinct trajectories of severity of epilepsy, predicted by comorbid cognitive problems, seizure type, and school age.

Il a été montré que la sévérité de l'épilepsie est associée aux issues cliniques chez les individus atteints de la maladie. Nous étudions les variations dans les évaluations de la sévérité de la maladie faites par les neurologistes aux enfants nouvellement diagnostiqués comme étant épileptiques (HERQULES). La sévérité de l'épilepsie a été mesurée à partir de l'échelle de l'évaluation globale de la sévérité de l'épilepsie (GASE). L'analyse de classes latentes à mesures répétées (RMLCA) a été utilisée pour caractériser les trajectoires longitudinales de la sévérité de l'épilepsie sur une période de deux ans. La régression logistique multinomiale a été utilisée pour identifier les facteurs prédictifs parmi les classes latentes des trajectoires. RMLCA a identifié quatre trajectoires distincts pour la sévérité de l'épilepsie; "Grande Amélioration Rapide" (12.9%), "Amélioraton Modérée Rapide" (46.3%), "Amélioration Modérée Tardive" (26.6%) et "Inchangée" (14.2%). Les quatres trajectoires de la sévérité qui ont été identifiées résultent d'un modèle prédictif qui comprend les problèmes cognitifs de comorbidité, le type de crise et l'âge scolaire.

Jacob Mortensen (Simon Fraser University), Luke Bornn (Simon Fraser University)

From Markov Models to Poisson Point Processes: Understanding Player Movement in the NBA
Des modèles de Markov aux processus de Poisson ponctuels: comprendre le mouvement des joueurs dans la NBA

When considering movement in space, a useful tool is a Markov model, where the position of the agent at time $t+1$ depends only on their position at time t. In this paper we build on existing theory to show that as the number of spatial locations in a bounded region approaches infinity, a Markov model can be represented by a Poisson point process, a popular type of spatial model that accounts for correlation between nearby locations. Using SportVu player tracking data provided by the National Basketball Association we show how this model can be used to produce distinct maps of player movement for each team in the NBA.

Lorsqu'on désire considérer le mouvement dans l'espace, le modèle de Markov représente un outil utile, où la position de l'agent au temps $t+1$ dépend seulement de sa position au temps t. Dans cette étude, nous nous basons sur la théorie existante afin de montrer que pour un nombre de locations spatiales tendant vers l'infini dans une région délimitée, un modèle de Markov peut être représenté par un processus de Poisson ponctuel, un type de modèle spatial populaire qui tient compte de la corrélation entre les locations à proximité. À partir des données sportives SportVu du suivi des joueurs, qui sont fournies par la NBA, nous montrons comment ce modèle peut être utilisé pour produire des cartes distinctes du mouvement des joueurs pour chaque équipe dans la NBA.

Thuva Vanniyasingam (McMaster University), Caitlin Daly (McMaster University), Xuejing Jin (McMaster University), Yuan Zhang (McMaster University), Gary Foster (McMaster University), Charles Cunningham (McMaster University), Lehana Thabane (McMaster University)

Investigating the Impact of Design Characteristics on Statistical Efficiency within Discrete Choice Experiments: a Systematic Survey
Étude de l'impact des caractéristiques du plan d'expérience sur l'efficacité statistique dans les expériences à choix discret: une étude systématique

This systematic survey aimed to review simulation studies of discrete choice experiments (DCEs) to determine what design features affect statistical efficiency. Electronic searches were conducted in JSTOR, Science Direct, PubMed and OVID. Screening and data extraction were performed independently and in duplicate. From 371 potentially relevant studies, 9 proved eligible. Statistical efficiency improved when: increasing the number of choice tasks or alternatives; decreasing the number of attributes, attribute levels, or overlaps; incorporating response behaviour or heterogeneity; correctly specifying Bayesian priors; minimizing prior variances; or matching the method to the research question. Studies need to improve reporting of: study objectives, failures, random number generators, starting seeds, and software. These results may help to inform investigators during DCE design creation.

Cette étude systématique a pour but de revoir les études de simulation des expériences à choix discrets afin de déterminer quelles caractéristiques du plan d'expérience influenceront l'efficacité statistique. Les recherches électroniques dans JSTOR, Science Direct, PubMed et OVID ont été menées. La sélection et l'extraction des données ont été performées de façon indépendante et en double. De 371 études potentiellement intéressantes, 9 étaient finalement éligibles. L'efficacité statistique était améliorée lorsque: on augmentait le nombre de choix des tâches ou d'alternatives; on diminuait le nombre d'attributs, de niveaux d'attributs, ou de chevauchements; lorsqu'on incorporait le comportement de la réponse ou de l'hétérogénéité; lorsqu'on spécifiait correctement la distribution bayésienne à priori; lorsqu'on minimisait la variance à priori; ou lorsqu'on utilisait une méthode reliée à la question de recherche. Les études devraient s'améliorer sur les points: les objectifs de recherche, les échecs, les générateurs de nombres aléatoires, les valeurs de départ fixées, et le logiciel. Ces résultats peuvent aider les chercheurs durant la conception d'expériences à choix discrets.

Shirin Moossavi (Department of Medical Microbiology, University of Manitoba), E. Khafipour (University of Manitoba), S. Sepehri (University of Manitoba), B. Robertson (University of California San Diego), L. Bode (University of California San Diego), C.J. Field (University of Alberta), A.B. Becker (University of Manitoba), P.J. Mandhane (University of Alberta), P. Subbarao (University of Toronto), S.E. Turvey (University of British Columbia), D.L. Lefebvre (McMaster University), M.R. Sears (McMaster University), the CHILD Study Investigators (Canadian Healthy Infant Longitudinal Development Study), M.B. Azad (University of Manitoba)

Structural Equation Modeling of Determinants of Human Milk Microbiota in the Canadian Healthy Infant Longitudinal Development (CHILD) Cohort

Modèles d'équations structurelles pour les déterminants du microbiote du lait humain dans la cohorte de "Canadian Healthy Infant Longitudinal Development" (CHILD)

Milk composition was assessed among 395 lactating mothers in the CHILD study. Milk microbiota and environment were modelled as latent constructs. Parameter estimation was achieved by maximum likelihood with bootstrapping. The measurement component had a good fit. The final model included exclusive breastfeeding and maternal body mass index (BMI) as exogenous factors influencing the milk environment, and infant sex, delivery mode, and breastfeeding mode as exogenous factors influencing the milk microbiota. Healthy eating index was modeled as an exogenous factor influencing maternal BMI. This final model had a good fit and accounted for 9.5% and 60% of observed variability in the milk microbiota and milk environment, respectively. The milk environment was not significantly associated with milk microbiota. We identified mode of breastfeeding as the primary factor directly affecting the milk microbiota while controlling for the structural association of other potentially important factors.

La composition du lait provenant de 395 mères recrutées pour l'étude CHILD a été analysée. Le microbiote du lait et l'environnement ont été modélisés comme des éléments latents. Les paramètres ont été estimés par la méthode du maximum de vraisemblance avec bootstrap. Les composantes du modèle présentaient un bon ajustement. Le modèle final comprenait l'allaitement exclusif et l'indice de masse corporelle (IMC) maternel comme facteurs exogènes influençant l'environnement du lait, ainsi que le sexe du nourrisson, le mode d'accouchement et le mode d'allaitement comme facteurs exogènes influençant le microbiote du lait. L'indice d'alimentation saine a été modélisé comme facteur exogène influençant l'IMC maternel. Ce dernier modèle présentait un bon ajustement et représentait 9.5% et 60% de la variabilité observée dans le microbiote et l'environnement du lait, respectivement. L'environnement du lait n'était pas significativement associé au microbiote. Nous avons identifié que le mode d'allaitement était le facteur primaire affectant le microbiote du lait directement, lorsque nous contrôlions pour l'association structurelle des facteurs potentiellement importants.

Social Evening • Soirée Sociale

To cap off the Student Conference, we will be hosting a social event after the conference, starting at around 6:00pm on campus. We will provide free appetizers and have a cash bar for the event, with beer, wine, sangria and simple cocktails available for purchase. You won't want to miss it!

Restaurant: UMSU's Degrees Restaurant is located on the third floor of University Centre, across from IQ's. Degrees is a licensed restaurant that offers an eclectic array of fast but healthy food, with cuisine from Italian to Indian, and blended with staples like hamburgers and falafel. The taste, quality, and friendliness of Degrees will be a pleasant surprise in the world of Campus eating.

Pour terminer la journée, nous tiendrons une soirée sociale après la conférence. La soirée commencera vers 18 heures, sur le campus. Nous offrirons gratuitement des entrées et un bar en espèces sera à votre disposition, avec de la bière, du vin, et la sangria et des cocktails simples disponibles à l'achat. Vous ne voulez pas manquer ça!

Restaurant: Le restaurant Degrees, sur le campus de l'Université du Manitoba, se situe au troisième étage du centre universitaire, juste en face du IQ. Le Degrees est un restaurant avec license qui offre un choix éclectique de nourriture rapide, mais santé, avec un type de cuisine allant de l'italien à l'indien, mélangé à des plats courants comme les hamburgers et falafels. Le goût, la qualité et l'hospitalité que vous retrouverez au Degrees vous surprendrons!

DEGREES

Degrees Restaurant map

304 University Centre

Get Directions

Notes

Notes