Twelfth Annual Canadian Statistics Student Conference

•

Douzième Congrès Canadien des Étudiants en Statistique

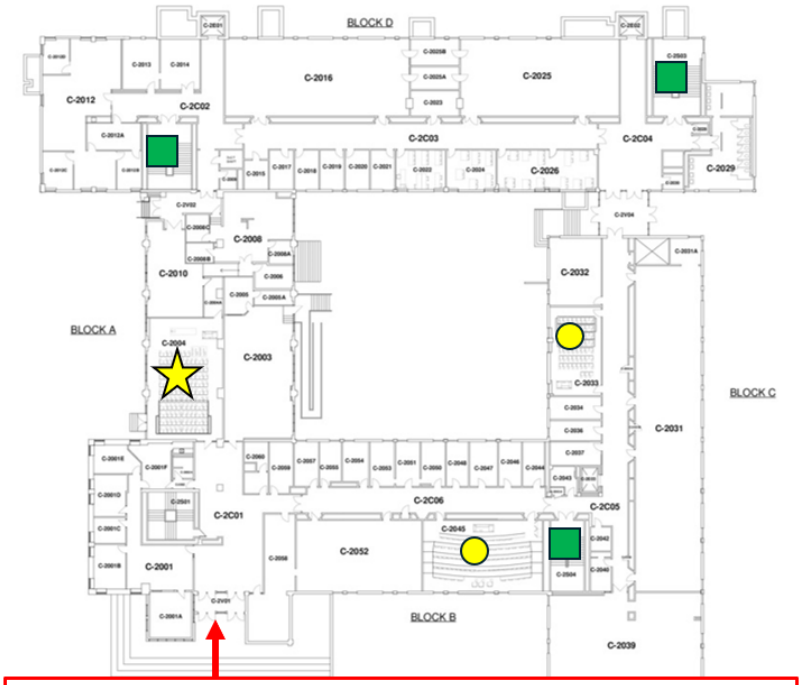
CCÉS CSSC

Saturday • Samedi
June 1 • 1 Juin
2024

- Breakfast
- Registration
- Lunch
- Poster Presentations
- Social Event (18:00)
- Main Building for CSSC

- Déjeuner
- Inscription
- Lunch
- Présentations par affiche
- Événement social (18h00)
- Édifice principale pour CCÉS

Chemistry-Physics Building

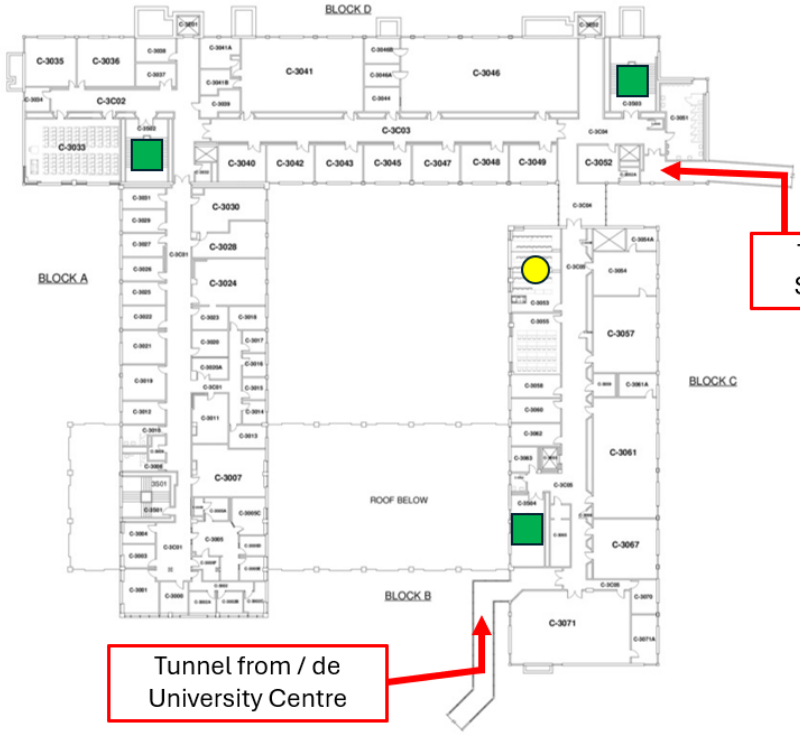


Entrance facing Irwins Road and Prince Philip Drive
Entrée devant Irwins Road et Prince Philip Drive

Chemistry Physics
Level 2 2e étage

- Main Room for CSSC
- Oral Presentation Room
- Stairs

- Salle principale pour le CCÉS
- Salle pour présentation orale
- Escaliers



Tunnel from / de
Science Building

Tunnel from / de
University Centre

Chemistry Physics
Level 3 3e étage

Contents • Table des matières

Welcome • Bienvenue

We are pleased to welcome you to the 12th Annual Canadian Statistics Student Conference, hosted at Memorial University of Newfoundland!

Our organizing committee has worked tirelessly to offer a comprehensive program. This year's conference includes a technical workshop on functional data analysis and a career panel featuring professionals from a wide variety of fields, including education, insurance and ecology! Prof. Richard Cook from the University of Waterloo will deliver a keynote address on identifying dependent selection and observation schemes and how to mitigate their effects.

There will be many opportunities to connect, meet new people, and share ideas, including a networking lunch and an evening social event.

Finally, we would like to acknowledge that the lands on which Memorial University's campuses are situated are in the traditional territories of diverse Indigenous groups, and we acknowledge with respect the diverse histories and cultures of the Beothuk, Mi'kmaq, Innu, and Inuit of this province.

We hope you have a great time at this year's student conference!

Nous sommes heureux de vous accueillir au 12^{e} Congrès canadien des étudiants en statistique, présenté à l'Université Memorial de Terre-Neuve!

Notre comité organisateur a travaillé sans relâche pour vous offrir un programme complet. Cette édition du congrès va inclure un atelier sur l'analyse des données fonctionnelles et une table ronde de professionnels de divers horizons statistiques y compris l'éducation, les assurances et l'écologie! Prof. Richard Cook de l'Université de Waterloo prononcera la conférence plénière sur l'identification des schémas de sélection et d'observation dépendants et comment atténuer leurs effets.

Il y aura plusieurs occasions tout le long de la journée pour établir des liens, rencontrer de nouvelles personnes et échanger des idées, notamment un dîner de réseautage et un événement social en soirée de la conférence.

Finalement, on aimerait reconnaître que les terres sur lesquelles se situe le campus de l'Université Memorial de Terre-Neuve font partie des territoires traditionnels de divers groupes autochtones. Nous nous montrons reconnaissants avec le plus grand respect des diverses histoires et cultures des peuples Béothuk, Mi'kmaq, Innu et Inuit de cette province.

Nous espérons que vous passerez un bon moment à cette édition de notre congrès étudiant!

Sponsors • Commanditaires

We would like to thank all our sponsors who have provided generous support to the Canadian Statistics Student Conference. These contributions have made this event possible.

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

What makes UBC Statistics unique?

UBC Statistics is renowned in Canada for its research excellence and its leadership in the research community. Students are engaged through both courses and research, and develop a strong set of skills, both applied and theoretical. The Department has always valued data-driven research, consulting and collaboration, and has long held communication and computing skills as crucial for success. These values are apparent not only in individual faculty's research programs but also in our undergraduate and graduate curriculum. Our students can prepare themselves for a successful career in areas such as biostatistics, bioinformatics/genomics, data science, statistical computing and mathematical statistics, in academia or in the public or private sector.

Applied Statistics and Data Science Group (ASDa) is the research support and general statistical consulting group of the Department of Statistics. ASDa aspires to engage with users of statistical methods across, and beyond, campus. In addition, ASDa provides opportunities for valuable consulting and writing experience. Between 10 and 20 STAT graduate students are hired on an hourly basis each term to support ASDa's consulting. Hired students will attend consultations, write meeting summaries, teach webinars and workshops, and may even handle their own clients, depending on their experience.

Statistics Graduate Student Association (SGSA) holds various activities and events such as table tennis, end-of-term hangout, and trips for grad students to get involved. Please check out the SGSA website if you would like to know more about life as a graduate student in the department.

Research Highlights

UBC Statistics faculty conduct core statistical methods and theory research across areas including Bayesian statistics, Bioinformatics/Genomics/Genetics, Biostatistics, Data Science, Environmental and Spatial Statistics, Modern Multivariate and Time-Series Analysis, Robust Statistics, and Statistical Learning.

Department faculty participate in numerous research collaborations involving other disciplines, and at various scales. Beyond traditional granting agencies, department members have recent involvement with projects supported by federal agencies (e.g., Department of Oceans and Fisheries, Public Health Agency of Canada), international organizations (e.g., Gates Foundation, World Health Organization), and the private sector (e.g., Scotiabank).

The department is a stakeholder in UBC's nascent AI Methods for Scientific Impact (AIM-SI) research cluster, under the broader auspices of the UBC ICICS Centre for Artificial Intelligence Decision-making and Action (CAIDA).

Forest Products Stochastic Modeling Group: Five faculty members, numerous trainees, and ASDa staff have been involved in this long-standing venture. The work is carried out in partnership with FPinovations, a private, not-for-profit organization that supports the Canadian forest sector's global competitiveness.

Learn More

THE UNIVERSITY OF BRITISH COLUMBIA
Department of Statistics
Faculty of Science

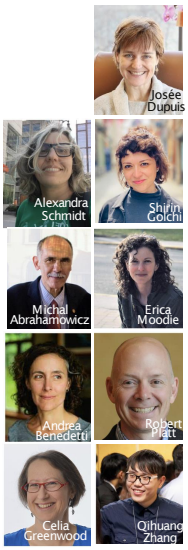
McGill

Department of Mathematics and Statistics
Département de mathématiques et de statistique

805, rue Sherbrooke ouest
Montréal (Québec)
Canada H3A 0B9

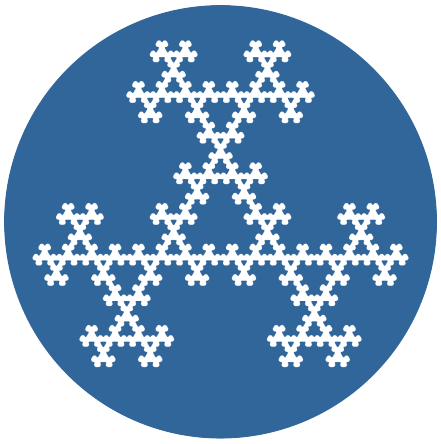
Biostatistics

Bayesian clinical trials
Bayesian computation
Causal inference
Disease mapping
Dynamic linear models
Dynamic treatment regimes
Genomics
Longitudinal data
Meta-analysis
Pharmacoepidemiology
Spatio-temporal processes
Statistical genetics



McGill | Department of **Epidemiology, Biostatistics and Occupational Health**

One of the largest concentrations of PhD statisticians in any Canadian medical school.


AARMS


Société statistique du Canada
Statistical Society of Canada


STUDC
CMS
SMC


MEMORIAL
UNIVERSITY
Conference and
Event Services

Organizers • Organisateurs

Organizing Committee • Comité organisateur

Co-chairs / Co-présidents:
Yuan Bian (University of Western Ontario)
Kyle McRae (Concordia University)

External and Internal Secretary / Secrétaire externe et interne:
Nathan Situ (Carleton University)

Fundraising / Collecte de fonds:
Alexandra Mossman (University of Waterloo)
Jay Sivathayalan (University of Waterloo)

Local Organization / Organisation locale:
Selina Elvayn (Memorial University of Newfoundland)
Sanjay Chandra Roy (Memorial University of Newfoundland)

Scientific Program / Programme scientifique:
Maryam Onifade (University of Ottawa)
Amanda Qiu (University of Victoria)

Sessions Committee / Séance sur les compétences techniques:
Gradon Nicholls (University of Waterloo)
Azizur Rahman (University of Manitoba)

Translation / Traduction:
Mathilde Dicaire-Cartier (Université de Montréal)
Marc Parsons (McGill University)

Support and Thanks • Support et remerciements

SSC President / Président de la SSC: Shirley Mills

SSC Associate Executive Director / Directrice exécutive associée de la SSC: Larysa Valachko

SSC Executive Assistant / Adjointe exécutive de la SSC: Michelle Benoit

SSC Electronic Services Administrator / Administrateur des services électroniques de la SSC: Clayton Forrest

SSC Treasurer / Trésorier de la SSC: Wesley Yung

SSC Meetings Coordinator / Coordinateur des congrès de la SSC: Xin Gao

SSC Local Organizers / Organisateurs locaux de la SSC:
Asokan Variyath, Zhaozhi Fan

SSC Electronic Services Manager / Responsable des services électroniques de la SSC:
Angelo Canty

SSC Public Relations Officer / Résponsable des relations publiques de la SSC:
Rhonda Rosychuk

SARGC Liaison / Liaison CÉDIR: Luke Hagar

Photographer / Photographe: Peter Macdonald

Volunteers / Bénévoles: Kismat Ghimire, Md Istinab Mahie, Sri Sriram

Judges / Juges: Gracia Dong, David Haziza, Kuan Liu, Mélina Mailhot, Nahid Sadr, Denis Talbot, Chi-Kuang Yeh, Jasper Zhang, Qihuang Zhang

Program Overview

Date: Saturday, June 1, 2024
Location: Memorial University of Newfoundland - St. John's Campus

Time	Session	Room	Page
8:00 - 9:00	Registration and Breakfast	R. Gushue Hall	
9:00 - 9:25	Opening Ceremony	C 2004	
9:30 - 10:30	**Student Oral Presentations I**		
	Public Health and Epidemiology	C 2004	35
	Statistical Methods in Genomics	C 2033	39
	Environmental and Ecological Statistics	C 2045	43
	Innovative Algorithms in Data Science	C 3053	47
10:30 - 10:45	Coffee Break	C Atrium	
10:45 - 11:50	**Workshop** *Functional Data Analysis in R*	C 2004	27
12:00 - 13:30	Poster Presentations	IIC Atrium	
	Lunch	R. Gushue Hall	67
13:30 - 14:30	**Student Oral Presentations II**		
	Time Series and Dynamic Models	C 2004	51
	Bayesian Methods and Applications	C 2033	55
	Epidemiological and Clinical Studies	C 2045	59
	Risk Assessment and Management	C 3053	63
14:30 - 14:45	Coffee Break	C Atrium	
14:45 - 15:55	**Careel Panel**	C 2004	29
15:55 - 17:00	**Keynote Speech** *Identifying dependent selection and observation schemes and mitigating their effects*	C 2004	25
17:00 - 17:30	Awards Ceremony	C 2004	
18:00 -	Social Event at The Breezeway	UC 1004	34
Online	Virtual Poster Presentations	CSSC Website	87

Legend:
C: Chemistry - Physics
IIC: Bruneau Centre for Research and Innovation
UC: University Centre

Aperçu du programme

Date: Samedi 1 juin 2024
Lieu: Université Memorial de Terre-Neuve - Campus Saint-Jean

Heure	Séance	Salle	Page
8h00 - 9h00	Inscription et petit déjeuner	R. Gushue Hall	
9h00 - 9h25	Discours d'ouverture	C 2004	
9h30 - 10h30	**Présentations orales étudiantes I**		
	Santé publique et épidémiologie	C 2004	35
	Méthodes statistiques en génomique	C 2033	39
	Statistiques environnementales et écologiques	C 2045	43
	Algorithmes innovants en science des données	C 3053	47
10h30 - 10h45	Pause-café	C Atrium	
10h45 - 11h50	**Atelier** *L'analyse des données fonctionnelles en R*	C 2004	27
12h00 - 13h30	Présentations par affiche	IIC Atrium	
	Déjeuner	R. Gushue Hall	67
13h30 - 14h30	**Présentations orales étudiantes II**		
	Séries chronologiques et modèles dynamiques	C 2004	51
	Méthodes bayésiennes et applications	C 2033	55
	Études épidémiologiques et cliniques	C 2045	59
	Évaluation et gestion des risques	C 3053	63
14h30 - 14h45	Pause-café	C Atrium	
14h45 - 15h55	**Panel sur les carrières**	C 2004	29
15h55 - 17h00	**Conférence plénière** *Identifier les schémas de sélection et d'observation dépendants et atténuer leurs effets*	C 2004	25
17h00 - 17h30	Remise des prix	C 2004	
18h00 -	Événement social à The Breezeway	UC 1004	34
En ligne	Présentations par affiche virtuelles	Site web du CCÉS	87

Légende:
C: Chemistry - Physics
IIC: Bruneau Centre for Research and Innovation
UC: University Centre

Oral Presentations List • Liste des présentations orales

Student Oral Presentations I • Présentations orales étudiantes I

9:30 - 10:30 C 2004

Public Health and Epidemiology / Santé publique et épidémiologie
Chair/Présidente: Jay Sivathayalan

9:30 - 9:45 **Tianyi Pan**
Distributed Lag Nonlinear Models Using Penalized Splines with Application in Air Pollution Epidemiology / Modèles non linéaires à décalage distribué utilisant des splines pénalisés avec application à l'épidémiologie de la pollution de l'air

9:45 - 10:00 **Amin Abed**
Gonorrhea Cluster Detection in Manitoba, Canada: Spatial, Temporal, and Spatio-Temporal Analysis / Détection des clusters de gonorrhée au Manitoba, Canada: analyse spatiale, temporelle et spatio-temporelle

10:00 - 10:15 **Yushu Zou**
Investigating the association between school substance programs and student substance use: accounting for informative cluster size / Étude de l'association entre les programmes scolaires sur les substances et l'usage de substances chez les élèves: prise en compte de la taille informative des scolaire

10:15 - 10:30 **Rachel Lobay**
Retrospective estimation of latent COVID-19 infections over the pandemic in U.S. states / Estimation rétrospective des infections latentes de la COVID-19 pendant la pandémie dans les états américains

9:30 - 10:30 C 2033

Statistical Methods in Genomics / Méthodes statistiques en génomique
Chair/Président: Kyle McRae

9:30 - 9:45 **Qicheng Zhao**
Bayesian Model for Disease-Specific Gene Detection in High-Dimensional Spatially Resolved Transcriptomics / Modèle bayésien pour la détection des gènes spécifiques à une maladie dans la transcriptomique à haute dimension spatiale

9:45 - 10:00 **Yushan Hu**
Single-Cell Sequencing of Lung Macrophages and Monocytes Reveals Novel Therapeutic Targets in COPD / Le séquençage unicellulaire des macrophages et des monocytes pulmonaires révèle de nouvelles cibles thérapeutiques de la MPOC

10:00 - 10:15 **Kyle Gardiner**
BLESS: Bagged Logistic Regression for Biomarker Identification / BLESS: Régression logistique bagged pour l'identification de biomarqueurs

10:15 - 10:30 **Bertrand Sodjahin**
Spike-and-slab based Informative Priors for Bayesian Network Structure Learning in High-dimensional regimes / Apprentissage de la structure des réseaux bayésiens de haute dimension en utilisant les distributions a priori informatives spike-and-slab

9:30 - 10:30 C 2045

Environmental and Ecological Statistics / Statistiques environnementales et écologiques

Chair/Présidente: Selina Elvayn

9:30 - 9:45 **Zirui Dong**

Arsenic Contamination in Nova Scotia's Domestic Well Water: a Spatial-Temporal Statistical Analysis / Contamination par l'arsenic dans l'eau des puits domestiques de la Nouvelle-Écosse: une analyse statistique spatio-temporelle

9:45 - 10:00 **Raphael McDonald**

Hidden bias in model-based stock assessment index standardization methods applied to design-based stratification / Biais camouflé dans la normalization des indices d'évaluations des stocks par modèles appliqués à une stratification fondé sur le plan de sondage

10:00 - 10:15 **Liam Cann**

Impacts of correlation in bioassays / Les impacts de la corrélation dans les bioessais

10:15 - 10:30 **Reshani Abayasekara**

A functional curve registration approach to understanding physical activity levels measured by accelerometers / Une approche de l'enregistrement des courbes fonctionnelles pour comprendre les niveaux d'activité physique mesurés par des accéléromètres

9:30 - 10:30 C 3053

Innovative Algorithms in Data Science / Algorithmes innovants en science des données

Chair/Présidente: Mathilde Dicaire-Cartier

9:30 - 9:45 **Nikola Surjanovic**
autoMALA: Locally adaptive Metropolis-adjusted Langevin algorithm / autoMALA: Algorithme de Langevin ajusté par Metropolis avec adaptation locale

9:45 - 10:00 **Chunlei Ge**
Quick and Simple Kernel Differential Equation Regression Estimators for Data with Sparse Design / Estimateurs rapides et simples de régression par équation différentielle à noyau pour les données à structure éparse

10:00 - 10:15 **Kai Yang**
fastHDMI: Fast Fourier Transform (FFT)-based Mutual Information estimation for High-Dimensional Data / fastHDMI: Estimation de l'information mutuelle basée sur la transformation de Fourier rapide (FFT) pour les données de haute dimension

10:15 - 10:30 **Bingcheng Wang**
Weighted Allocation Probability-adjusted Thompson Sampling (WAPTS): a Novel Algorithm for Sparse and Adaptive Contextual Bandits / L'échantillonnage de Thompson ajusté à la probabilité de l'allocation pondérée (WAPTS): un nouvel algorithme pour les bandits contextuels épars et adaptatifs

Student Oral Presentations II • Présentations orales étudiantes II

13:30 - 14:30 C 2004

Time Series and Dynamic Models / Séries chronologiques et modèles dynamiques
Chair/Présidente: Jay Sivathayalan

13:30 - 13:45 **Roberto Curti**
ts.shiny: Interactive Visualization through a Shiny App Applied to Time Series Data / ts.shiny: visualisation interactive à travers une application Shiny appliquée aux données de séries temporelles

13:45 - 14:00 **Parham Pishrobat**
Introducing Dynamic Regression Model for Hydrological Inference / Introduction d'un modèle de régression dynamique pour l'inférence hydrologique

14:00 - 14:15 **Xize Ye**
Generalized Autoregressive Conditionally Stochastic Heteroskedasticity: Motivation and Applications / Generalized Autoregressive Conditionally Stochastic Heteroskedasticity: motivations et applications

14:15 - 14:30 **Chen Chen**
Longitudinal Cognitive Trajectory Modelling and Phenotyping with Multiple Features Using Health Administrative Data / Modélisation et phénotypage de trajectoires cognitives longitudinales avec des caractéristiques multiples à l'aide de données administratives sur la santé

13:30 - 14:30 C 2033

Bayesian Methods and Applications / Méthodes bayésiennes et applications
Chair/Président: Kyle McRae

13:30 - 13:45 **Wuqian Gao**
Bayesian Z-residuals for Hurdle Models / Résidus Z bayésiens pour les modèles d'obstacles

13:45 - 14:00 **Jingwen Ji**
Improving Toronto's Overnight Shelter Allocation and Utilization using Bayesian Non-Parametric Models / Amélioration de l'allocation et de l'utilisation des refuges de nuit à Toronto en utilisant des modèles bayésiens non-paramétriques

14:00 - 14:15 **Linke Li**
Efficiently Evaluating the Operating Characteristics of Bayesian Clinical Trial with Machine Learning / Évaluation efficace des caractéristiques opérationnelles d'un essai clinique bayésien avec l'apprentissage automatique

14:15 - 14:30 **Muye Nanshan**
A Joint Estimation Approach to Sparse Additive Ordinary Differential Equations / Une approche d'estimation conjointe des équations différentielles ordinaires additives éparse

13:30 - 14:30 C 2045

Epidemiological and Clinical Studies / Études épidémiologiques et cliniques
Chair/Présidente: Selina Elvayn

13:30 - 13:45 **Md Ashiqul Haque**
Model-based algorithms to ascertain smoking in administrative health data: a registry-based validation study / Algorithmes basés sur des modèles pour vérifier le tabagisme dans les données administratives sur la santé: une étude de validation basée sur un registre

13:45 - 14:00 **Jing Wang**
Dose-response relationship for a skewed predictor containing lot of zeros / Relation dose-réponse pour un prédicteur asymétrique contenant beaucoup de zéros

14:00 - 14:15 **Michael Agronah**
Are Microbiome Studies Underpowered? Investigating Power in Differential Abundance Studies / Les études sur le microbiome manquent-elles de puissance? Étude de la puissance dans les études d'abondance différentielle

14:15 - 14:30 **Éloïse Soucy**
Advanced machine learning and classification of ECG data / Apprentissage automatique avancé et classification des électrocardiogrammes

13:30 - 14:30 C 3053

Risk Assessment and Management / Évaluation et gestion des risques
Chair/Présidente: Mathilde Dicaire-Cartier

13:30 - 13:45 **Bartosz Glowacki**
Ruin probability and rare event simulation / Probabilité de ruine et simulation d'événements rares

13:45 - 14:00 **Armin Mohammadioroojeh**
Approximating generalized gamma convolutions and mixture of exponentials via multipoint Padé method / Approximation des convolutions gamma généralisées et des mélanges d'exponentielles via la méthode multipoint de Padé

14:00 - 14:15 **Peiheng Gao**
NLP-based detection of systematic anomalies among the narratives of consumer complaints / Détection des anomalies systématiques par l'analyse basée sur le TALN des récits de plaintes de consommateurs

14:15 - 14:30 **Assane Kholle**
Méthodes de regroupement des variables catégorielles multiniveaux dans un GLM / Grouping methods for multilevel categorical variables in a GLM

Poster List • Liste des affiches

In-person IIC Atrium

Brynn O'Connell
Variable Selection for the Classification of Data with Missing Values / Sélection de variables pour la classification de données avec des valeurs manquantes

Yutong Lu
A statistical framework to integrate large chemical language models for molecular property analysis / Un cadre statistique pour intégrer de grands modèles de langage chimique pour l'analyse des propriétés moléculaires

Ziqian Zhuang
Joint Modeling of Complex Multivariate Adverse Events in Clinical Trial Data / Modélisation conjointe des événements indésirables multivariés complexes dans les données d'essais cliniques

Lina Li
TSMA: a Two-stage Sampling Aggregation Framework to Construct Prediction Models for Unbalanced Case-Control Disease Data from Electronic Medical Record and Genomics (eMERGE) Network / TSMA: un cadre d'agrégation d'échantillonnage en deux étapes pour construire des modèles de prédiction pour des données cas-témoins sur les maladies non équilibrées provenant du réseau Electronic Medical Record and Genomics (eMERGE)

Zixuan Yang
Modelling the Forecasting Error Distributions of Several Fire-Weather Variables / Modélisation de la distribution des erreurs dans la prévision de plusieurs variables forêt-météo

Jesse Ghashti
A bootstrap augmented k-means algorithm for fuzzy partitions / Un algorithme bootstrap augmenté de k-moyennes pour les partitions floues

Tracy Qian
Automatic Model Selection using Wasserstein Generative Adversarial Networks / Sélection automatique de modèles à l'aide de réseaux antagonistes génératifs de Wasserstein

Benjamin Frizzell
Optimal Experimental Design using Simulated Annealing / Conception optimale des expériences à l'aide du recuit simulé

Haochen Ning
IIMI: Advancements of Novel Machine-Learning Toolset for Plant Virus Detection / IIMI: progrès d'un nouvel ensemble d'outils d'apprentissage automatique pour la détection des virus des plantes

Sarah Organ
Vertex cover matroid variable selection for controlling the false discovery rate and improving power with correlated predictors / Sélection de variables dans un matroïde de couverture de sommet pour contrôler le taux de fausse découverte et améliorer la puissance avec des prédicteurs corrélés

Yasaman Shahhosseini
Spatiotemporal fractal based analysis of fMRI time series / L'analyse des séries temporelles de l'IRMf basée sur les fractales spatio-temporelles

Simon Maltby
Spatial Analysis of the Risk Factors for Covid-19 / Analyse spatiale des facteurs de risque de la Covid-19

Pranath Pussella
Simulation for Cricket: a Machine Learning Approach / Simulation pour le cricket: une approche d'apprentissage automatique

Juan Liyau
Intuitive Segmentation for Hyperspectral Fluorescence Imaging in Ophthalmology: an Innovative Machine Learning Tool / Segmentation intuitive pour l'imagerie hyperspectrale de fluorescence en ophtalmologie: un outil innovant d'apprentissage automatique

Yu Shi
A Deep Learning-Driven Out of Distribution Approach for Predicting Patient-Specific Cancer Dependency Maps / Une approche hors distribution basée sur l'apprentissage profond pour prédire des cartes de dépendance du cancer spécifiques aux patients

Feifan Xiang
Enhancing Osteoarthritis Progression Prediction with LSTM: Leveraging Bilateral Knee Data and Prospects for Multimodal Integration in Osteoarthritis Initiative / Amélioration de la prédiction de la progression de l'arthrose avec la MLCT: exploitation des données bilatérales du genou et perspectives d'intégration multimodale dans l'Initiative sur l'arthrose

Yuhang Ou
Actuarial study and statistical analysis of flood insurance claims in Canada / Étude actuarielle et analyse statistique des sinistres d'assurance contre les inondations au Canada

Ankita Shelke
Optimizing Canada's Inflation: a Novel Approach Integrating Machine Learning and Deep Learning Techniques / Optimisation de l'inflation au Canada: une nouvelle approche intégrant les techniques d'apprentissage automatique et d'apprentissage profond

Adrian Neumann
Predicting Real Estate Prices in Edmonton, Alberta / Prédiction des prix immobiliers à Edmonton, Alberta

Solmaz Ghajar
COVID-19 Infection During Pregnancy changes Gene Expression in Umbilical Cord Blood cells / L'infection par la COVID-19 pendant la grossesse modifie l'expression génique dans les cellules sanguines du cordon ombilical

Online / En ligne

CSSC Website / Site web du CCÉS

Yan Song

Developing a Brief PSP Mental Health Screening Tool with Generalized Linear Model and Regularization / Élaboration d'un bref outil de dépistage de la santé mentale dans le cadre du PSP à l'aide d'un modèle linéaire généralisé et d'une régularisation

Minoli Munasinghe

A robust regression model in the presence of missing and censored data / Un modèle de régression robuste en présence de données manquantes et censurées

Shaomeng Yin

Improving Predictive Ability for Student Academic Performance through Imbalanced Data Handling / Amélioration de la capacité prédictive des résultats scolaires des étudiants grâce à un traitement déséquilibré des données

Saeid Moradi

Skin Cancer Detection Using Deep Convolutional Neural Networks / Détection du cancer de la peau à l'aide de réseaux de neurones convolutifs profonds

Azar Taheri Tayebi

Can smartphone apps reveal fishing catch rates and durations? / Les applications mobiles peuvent-elles révéler les taux de capture des poissons et les durées de pêche?

Meira Golberg

Using Multiple Imputation to Deal with Missing Data in the Canadian Longitudinal Study on Aging / Utilisation de l'imputation multiple pour traiter les données manquantes dans l'Étude longitudinale canadienne sur le vieillissement

Xiao Yan

Practical Implementation of Advanced Causal Inference Method: Development of an R Package for Bayesian Marginal Structural Models with Time-Varying Treatment / Implémentation pratique d'une méthode avancée d'inférence causale : développement d'un package R pour les modèles structurels marginaux bayésiens avec traitement variant dans le temps

Keynote Address • Conférence plénière

Identifying dependent selection and observation schemes and mitigating their effects • Identifier les schémas de sélection et d'observation dépendants et atténuer leurs effets

Richard Cook is University Professor in the Department of Statistics and Actuarial Science at the University of Waterloo and Faculty of Mathematics Research Chair. He holds a cross-appointment in the School of Public Health (University of Waterloo). His research interests include the analysis of life history data, the design and analysis of clinical and epidemiological studies, and statistical methods for the analysis of incomplete data. He has published extensively in these areas and written two books with Jerald Lawless (The Statistical Analysis of Recurrent Events, Springer, 2007; Multistate Models for the Analysis of Life History Data, Taylor and Francis, 2018). He served as Director or co-Director for three major graduate training programs in biostatistics in Ontario including the GlaxoSmithKline-UW Pharmaceutical Statistics Graduate Program, the Methodology Team of the CANNeCTIN grant for methodology for clinical trials, and the Biostatistics Training Initiative funded by the Ontario Institute for Cancer Research. Richard is also deeply engaged in collaborative research with other scientists working in transfusion medicine, immunology, and cancer, and consults widely with industry and government organizations. In 2018 he was awarded the Gold Medal of the Statistical Society of Canada and in 2021 he was named Fellow of the Royal Society of Canada.

Richard Cook est professeur au Département de statistique et d'actuariat à l'Université de Waterloo et titulaire de la Chaire de recherche de la Faculté de mathématiques. Il est également professeur à l'École de santé publique (Université de Waterloo). Ses domaines de recherche incluent l'analyse des données de cycles de vie, la conception et l'analyse d'études cliniques et épidémiologiques, ainsi que les méthodes statistiques pour l'analyse de données incomplètes. ll a publié de manière approfondie dans ces domaines et a écrit deux livres avec Jerald Lawless (The Statistical Analysis of Recurrent Events, Springer, 2007; Multistate Models for the Analysis of Life History Data, Taylor and Francis, 2018). Il a conçu et dirigé trois importants programmes de formation aux cycles supérieurs en biostatistique en Ontario, notamment le Programme de statistiques pharmaceutiques GlaxoSmithKline-UW, l'Équipe de méthodologie de la subvention CANNeCTIN pour la méthodologie des essais cliniques, et l'Initiative de formation en biostatistique financée par l'Ontario Institute for Cancer Research. Richard est également profondément impliqué dans la recherche collaborative avec d'autres scientifiques travaillant dans les domaines de la médecine transfusionnelle, de l'immunologie et du cancer, et il consulte largement auprès d'organisations industrielles et gouvernementales. En 2018, il a reçu la Médaille d'or de la Société statistique du Canada, et en 2021, il a été nommé membre de la Société royale du Canada.

Abstract

The proliferation of large health data sets makes it an exciting time for statisticians to study chronic disease processes. However it is often unclear what selection mechanisms lead to inclusion in such data sets, and what factors influence the acquisition of information from individuals or retention. This talk will review some recent work exploring the effect of disease-related selection processes along with our ability to detect them in order to mitigate their effects. In addition, once individuals are recruited to a study, dependent observation schemes (e.g. intermittent visit processes and loss to follow-up) can create biases which make findings uninterpretable. It is therefore important to detect them when present to ensure inferences are valid and generalizable to the desired population. The issues will be discussed in the context of a longitudinal cohort of patients attending a rheumatology clinic but the findings have bearing on analyses of any types of cohort data.
This talk is based on joint work with Jerry Lawless.

Résumé scientifique

La prolifération de vastes ensembles de données de santé permettent aux statisticiens d'étudier les processus de maladies chroniques. Cependant, il n'est pas souvent clair quels mécanismes de sélection conduisent à l'inclusion dans ces bases de données et quels facteurs influent sur l'acquisition d'informations auprès des individus ou sur leur rétention. Cette présentation examinera certains travaux récents explorant l'effet des processus de sélection liés à la maladie ainsi que notre capacité de les détecter afin d'atténuer leurs effets. De plus, une fois que les individus sont recrutés dans une étude, des schémas d'observation dépendants (par exemple, des processus de visites intermittentes et des pertes de suivi) peuvent créer des biais rendant les résultats inexploitables. Il est donc important de les détecter lorsqu'ils sont présents pour garantir que les inférences sont valides et généralisables à la population souhaitée. Ces caractéristiques seront discutées dans le contexte d'une cohorte longitudinale de patients fréquentant une clinique de rhumatologie, mais les conclusions ont des implications pour l'analyse de tout type de données de cohorte.
Cette présentation est basée sur un travail conjoint avec Jerry Lawless.

Workshop • Atelier

Functional Data Analysis in R • Une introduction au calcul parallèle et de haute performance en R

Dr. Jiguo Cao holds the prestigious position of Canada Research Chair in Data Science and works as a Professor in the Department of Statistics and Actuarial Science at Simon Fraser University. Dr. Cao's research spans diverse areas such as functional data analysis (FDA), sports analytics, and machine learning. His expertise in statistical methods extends to addressing real-world challenges across a spectrum of disciplines, including sports, neuroscience, public health, image analysis, genetics, pharmacology, ecology, environment, and engineering. Dr. Cao was honored with the CRM-SSC award in 2021, jointly presented by the Statistical Society of Canada (SSC) and the Centre de Recherches Mathématiques (CRM). This accolade highlights his research excellence and significant achievements. Dr. Cao's impactful research has gained widespread visibility, reflected in his impressive portfolio of over 100 publications in esteemed refereed journals. Additionally, he has delivered 15 distinguished lectures and short courses on FDA, along with presenting 91 invited talks and seminars in various countries, including Australia, Canada, China, and the USA.

Professeur Jiguo Cao occupe le poste prestigieux de Chaire de recherche du Canada en science des données et travaille présentement comme professeur au Département de statistique et de science actuarielle de l'Université Simon Fraser. Les domaines de recherche du Professeur Cao couvrent des domaines divers incluant l'analyse des données fonctionnelles (ADF), l'analyse des données provenant du domaine des sports et l'apprentissage automatique. Son expertise en méthodes statistiques concerne plusieurs disciplines, notamment le sport, les neurosciences, la santé publique, l'analyse d'image numérique, la génétique, la pharmacologie, l'écologie, l'environnement et l'ingénierie. Professeur Cao a reçu le prix CRM-SSC en 2021, présenté conjointement par la Société statistique du Canada (SSC) et le Centre de recherches mathématiques (CRM). Cette distinction souligne son excellence en recherche et ses accomplissements importants. Les recherches marquantes du Professeur Cao ont accru une grande visibilité, témoignée par son portefeuille impressionnante de plus de 100 publications dans des revues scientifiques prestigieuses. De plus, il a présenté 15 conférences distinguées et cours au sujet de l'ADF, ainsi que 91 conférences et séminaires invités dans divers pays, dont l'Australie, le Canada, la Chine et les États-Unis.

Abstract

Functional Data Analysis (FDA) represents a burgeoning statistical discipline dedicated to analyzing curves, images, or multidimensional functions. A distinguishing characteristic of FDA is the treatment of each random function as an individual sample element. Functional data is prevalent in various applications, including longitudinal studies and brain imaging. Throughout this workshop, I will introduce some cutting-edge methods for functional data analysis.

Résumé scientifique

Le domaine statistique de l'analyse de données fonctionnelles (ADF) est en pleine croissance pour l'analyse de trajectoires longitudinales, de courbes, d'images numériques ou de toutes autres variétés. L'ADF traite chaque fonction aléatoire observée sur l'ensemble du domaine comme un élément d'échantillonnage. Les données fonctionnelles peuvent être couramment trouvées dans de nombreuses applications. Celles-ci incluent les données de condition physique provenant de l'appareil portable, la pollution de l'air, les études longitudinales, expressions génétiques temporales et images cérébrales. Ce court cours couvrira les principales méthodes de l'ADF telles que l'analyse en composantes principales fonctionnelles et les modèles de régression linéaires fonctionnels. Les applications de toutes ces méthodes seront démontrées en utilisant des données réelles et le langage de programmation R. L'objectif de ce cours est d'apprendre aux étudiants à utiliser et développer des méthodes ADF pour analyser les données.

Invited Career Speakers • Panélistes invités à la table ronde sur les carrières

Noel Cadigan

Dr. Cadigan joined the Centre for Fisheries Ecosystems Research (CFER) at the Fisheries and Marine Institute of Memorial University of Newfoundland in 2012. He is an associated professor and the Ocean Choice International Research Chair in Stock Assessment and Sustainable Harvest Advice for Northwest Atlantic Fisheries. Dr. Cadigan first started research on stock assessment methods in 1990 when he worked with Fisheries and Oceans Canada (DFO) in Newfoundland. He received a PhD in statistics in 1999 at the University of Waterloo.

Dr. Cadigan's research deals with statistical methods for fish stock assessment and sustainable fisheries management. He has extensive experience in the assessment of Newfoundland fish stocks, and experience with many other Canadian, American, and European stocks. Recently his research has been focused on spatiotemporal models for complex fisheries data, and state-space/spatial stock assessment models.

Le Professeur Cadigan s'est joint au Centre for Fisheries Ecosystems Research (CFER) du Fisheries and Marine Institute de l'Université Memorial de Terre-Neuve en 2012. Il est professeur agrégé et titulaire de la Ocean Choice Chaire de recherche internationale en évaluation des stocks et conseils en matière de récolte durable pour les pêches de l'Atlantique Nord-Ouest. Professeur Cadigan a commencé ses recherches sur les méthodes d'évaluation des stocks en 1990, alors qu'il travaillait pour Pêches et Océans Canada (MPO) à Terre-Neuve. Il obtient un doctorat en statistique en 1999 à l'Université de Waterloo.

Les recherches du Professeur Cadigan portent sur les méthodes statistiques pour l'évaluation des stocks de poissons et la gestion durable des pêches. Il possède une vaste expérience dans l'évaluation des stocks de poissons de Terre-Neuve et de nombreux autres stocks canadiens, américains et européens. Récemment, ses recherches se sont concentrées sur les modèles spatio-temporels pour les données de pêche complexes et les modèles espace-des-états/spatiaux d'évaluation des stocks espace-des-états/spatial.

Jayde Eustace

Jayde Eustace is a senior data scientist on the research and development team at Aviva Insurance. She works closely with actuaries and business stakeholders to develop new use cases for machine learning in the home and auto insurance field. Recent projects include a scoring method for group business, and entity recognition within the company's customer database. Before joining Aviva, Jayde graduated from Memorial University with a B.Sc. Pure Mathematics and M.Sc. Statistics with a focus on extreme value theory. She began her career as a government statistician, but became fascinated with the emerging field of data science after Johns Hopkins University launched their first online course track on the topic back in 2014.

Jayde Eustace est une experte seniore en science des données au sein de l'équipe de recherche et développement chez Aviva Assurance. Elle travaille en étroite collaboration avec les actuaires et les parties prenantes pour développer de nouveaux cas d'utilisation de l'apprentissage automatique dans le domaine de l'assurance habitation et automobile. Parmi ses projets récents figurent une méthode de notation pour les activités collectives et la reconnaissance d'entités dans la base de données client de l'entreprise. Avant de rejoindre Aviva, Jayde a obtenu son baccalauréat en mathématiques pures et sa maîtrise en statistique avec une concentration sur la théorie des valeurs extrêmes à l'Université Memorial. Elle a commencé sa carrière en tant que statisticienne au gouvernement, mais a été fascinée par le domaine émergent de la science des données après que l'Université Johns Hopkins a lancé son premier cours en ligne sur le sujet en 2014.

Bethany White

Bethany White (PhD in Statistics-Biostatistics & MMATH in Statistics, both from the University of Waterloo, and BScH in Mathematics and Statistics from Acadia University) is an Associate Professor, Teaching Stream, in the Department of Statistical Sciences at the University of Toronto. Her research interests relate to the impact of technology-enhanced and simulation activities on student learning and attitudes toward statistics. She also has a pedagogical interest in the quantitative training of life sciences students. She served on the Statistical Society of Canada (SSC) Statistical Education Section Executive Committee between 2013-2016 (President of the Section for 2014-2015) and on the SSC Board of Directors (2017-2021), and has made contributions on the editorial boards of a couple of statistics education journals and on organizing committees for statistics and science education workshops and conferences in Canada and the US.

Bethany White (Ph.D. en statistiques-biostatistiques et MMATH en statistiques, tous deux de l'Université de Waterloo, et BScH en mathématiques et statistiques de l'Université Acadia) est professeure agrégée d'enseignement au Département des sciences statistiques de l'Université de Toronto. Ses intérêts de recherche portent sur l'impact des activités d'enseignement qui incluent des simulations et qui sont assistés par la technologie sur l'apprentissage et les attitudes à l'égard des statistiques des étudiants participants. Elle a également un intérêt pédagogique pour la formation quantitative des étudiants en sciences de la vie. Elle a siégé au comité exécutif de la Section de l'enseignement statistique de la Société statistique du Canada (SSC) entre 2013 et 2016 (présidente de la section pendant 2014-2015) et au conseil d'administration de la SSC (2017-2021). Elle a également contribué aux conseils éditoriaux de certaines revues pédagogiques en statistique et à des ateliers et conférences sur la statistique et l'enseignement scientifique au Canada et aux États-Unis.

A Message from the SARGC

Dear students and recent graduates in statistics-related fields,

The Student and Recent Graduate Committee (SARGC) is a committee of the Statistical Society of Canada (SSC) responsible for organizing activities for students and recent graduates and advocating on their behalf within the SSC. The SARGC recently hosted a comic strip competition. We would like to congratulate Umar Khan, Yuliya Nesterova, and Md. Ashiqul Haque for winning prizes for their comic strips! This year, the SARGC also organized a networking event before the SSC's annual meeting and hosted a monthly showcase for students and recent graduates on our social media platforms. **We will solicit our next round of nominations in the fall** to continue highlighting the excellent work being done by all of you!

In addition, we are responsible for selecting the organizing committee for the Canadian Statistics Student Conference. By the way, **congratulations** to the organizing committee for having organized such a fantastic conference! **To continue organizing these great conferences, we need motivated students and recent graduates just like you!** If you would like to get involved with either the SARGC itself or directly with the CSSC's organizing committee, please get in touch with us via the following form: SARGC Volunteer Form.

If you are also attending the SSC's main conference, we want to **invite you to the Student BBQ**, taking place Tuesday, June 4, from 5 to 7 P.M. in Gushue Hall. Food and beverages will be served free of charge. This is a great opportunity to informally get to know other students! **There will also be participation prizes raffled at the end of the event!**

Finally, please remember to connect with us on social media for general updates and to view SARGC's monthly showcase of students and recent graduates: Twitter, LinkedIn, Facebook, Instagram.

Thank you and enjoy the conference,

The Student and Recent Graduate Committee of the SSC

Message du CÉDIR

Chers étudiants et nouveaux diplômés dans des domaines reliés à la statistique,

Le Comité des étudiants et diplômés récents (CÉDIR) est un comité de la Société statistique du Canada (SSC) responsable de l'organisation d'activités pour les étudiants et les nouveaux diplômés des universités canadiennes travaillant dans des domaines reliés à la statistique et agissant pour représenter leurs intérêts auprès de la SSC. Le CÉDIR a récemment organisé un concours de bande dessinée. Nous félicitons Umar Khan, Yuliya Nesterova et Md. Ashiqul Haque pour avoir remporté des prix pour leurs bandes dessinées! Cette année, le CÉDIR a également organisé un événement de réseautage avant la réunion annuelle de la SSC et a accueilli une vitrine mensuelle pour les étudiants et les nouveaux diplômés sur nos réseaux sociaux. **Nous sollicitons notre prochaine série de nominations cet automne** afin de continuer à mettre en lumière l'excellent travail réalisé par chacun d'entre vous!

Nous sommes aussi responsables de la sélection du comité organisateur du Congrès des étudiants en statistique du Canada (CCÉS). D'ailleurs, un gros **félicitations** au comité organisateur pour avoir organisé une conférence aussi fantastique! **Pour continuer à organiser ces grandes conférences, nous avons besoin d'étudiants motivés et de nouveaux diplômés comme vous!** Si vous souhaitez vous impliquer dans le CÉDIR ou directement dans le comité d'organisation du CCÉS, veuillez nous contacter via le formulaire suivant : Formulaire de bénévolat du CÉDIR.

Si vous assistez également à la conférence principale de la SSC, nous tenons à vous **inviter au BBQ étudiant**, qui aura lieu le mardi 4 juin, de 17h à 19h, au Gushue Hall. La nourriture et les boissons gratuites seront servies. C'est une excellente occasion de faire connaissance avec d'autres étudiants de manière informelle! **Il y aura également des prix de participation tirés au sort à la fin de l'événement!**

Enfin, n'oubliez pas de consulter nos réseaux sociaux pour pouvoir vous informer sur nos mises à jour et pour pouvoir consulter notre vitrine mensuelle des étudiants et des nouveaux diplômés du CÉDIR: Twitter, LinkedIn, Facebook, Instagram.

Merci et bon congrès,

Le Comité des étudiants et diplômés récents de la SSC

Social Activies • Activités sociales

Networking Lunch • Déjeuner réseautage

During the lunch break in Gushue Hall, representatives from Statistics Canada and McGill University's Department of Biostatistics, Epidemiology and Occupational Health will be available to discuss employment opportunities and graduate programs.

Note that lunch will be served in three waves; the color of the sticker on your nametag indicates when you will eat.

- 11:00 - 11:45 (yellow)
- 12:00 - 12:45 (red)
- 12:45 - 13:30 (blue)

Pendant la pause déjeuner à Gushue Hall, des représentants de Statistique Canada et du Département d'épidémiologie, de biostatistique et de santé au travail de l'Université McGill seront disponibles pour discuter respectivement des postes d'emploi et des programmes d'études supérieures.

À noter que le déjeuner sera servi en trois vagues; la couleur de l'autocollant sur votre badge indique quand vous mangerez.

- 11h00 - 11h45 (jaune)
- 12h00 - 12h45 (rouge)
- 12h45 - 13h30 (bleu)

Social Evening • Soirée

Alcoholic and non-alcoholic beverages, fish and chips and vegetarian meals will be provided at no additional cost.

Des boissons alcoolisées et non alcoolisées, du *fish and chips* et des repas végétariens seront fournis sans frais supplémentaires.

Address • Adresse
The Breezeway
1st Floor University Center Room 1004
1 Arctic Avenue
St. John's, NL A1C 5S7

Time • Heure: 18:00

Scientific Abstracts • Résumés scientifiques

Oral Presentations • Présentations orales

Public Health and Epidemiology • Santé publique et épidémiologie

9:30 - 10:30, *C 2004*
Chair • Présidente: Jay Sivathayalan

Tianyi Pan, Alex Stringer, Glen McGee
Distributed Lag Nonlinear Models Using Penalized Splines with Application in Air Pollution Epidemiology
Modèles non linéaires à décalage distribué utilisant des splines pénalisés avec application à l'épidémiologie de la pollution de l'air

Distributed lag nonlinear models (DLNMs) have been widely used for investigating exposure-lag-response relationships. Most DLNMs are designed for Gaussian or Poisson outcomes, and the computational limitations make them slow in large datasets. We propose a novel DLNM with overdispersed outcomes using penalized B-splines. The exposure processes are modelled to account for missingness and outliers. Two identifiability constraints are imposed by reparameterization. We propose an inference method based on profile likelihood, where the overdispersion and smoothing parameters are estimated by Laplace approximate marginal likelihood. We compute the exact gradient and Hessian of the penalized log-likelihood. This leads to an accurate and fast Newton-type optimization. This inference method scales well to large datasets. We apply the proposed methods on the Canadian air pollution daily data from 2001 to 2014, to model the effects of particulate matter and ozone on mortality.

Les modèles non linéaires à décalage distribué (DLNM) ont été abondamment utilisés pour étudier les relations exposition-décalage-réponse. La plupart des DLNM sont conçus pour des résultats gaussiens ou Poisson, et les limitations informatiques les rendent lents dans les grands ensembles de données. Nous proposons un nouveau DLNM avec des résultats hyperdispersés en utilisant des B-splines pénalisées. Les processus d'exposition sont modélisés pour tenir compte des absences et des valeurs aberrantes. Deux contraintes d'identifiabilité sont imposées par reparamétrage. Nous proposons une méthode d'inférence basée sur la vraisemblance de profil, où les paramètres de hyperdispersion et de lissage sont estimés par la vraisemblance marginale approximative de Laplace. Nous calculons le gradient exact et le hessien de la log-vraisemblance pénalisée. Cela mène à une optimisation de type Newton précise et rapide. Cette méthode d'inférence s'adapte bien aux grands ensembles de données. Nous appliquons les méthodes proposées à une base de données quotidiennes sur la pollution de l'air au Canada entre 2001 à 2014 afin de modéliser les effets de la matière particulaire et de l'ozone sur la mortalité.

Amin Abed, Mahmoud Torabi, Zeinab Mashreghi

Gonorrhea Cluster Detection in Manitoba, Canada: Spatial, Temporal, and Spatio-Temporal Analysis

Détection des clusters de gonorrhée au Manitoba, Canada: Analyse spatiale, temporelle et spatio-temporelle

In Canada, Gonorrhea is the second most prevalent sexually transmitted infection. Manitoba reported three times the national average incidence rate in 2018. This study investigates the spatial, temporal, and spatiotemporal patterns of Gonorrhea in Manitoba (2000-2016). Gonorrhea infections were grouped by district using postal codes, linked to census data, to provide demographic details. The study employs Global Moran's I, Kulldorff's spatial and spatiotemporal scan statistics, and seasonal ARIMA to identify infection clusters across Manitoba health authority districts. Spatial analysis reveals clusters in northern Manitoba and central Winnipeg. Seasonal patterns emerge in late summer and fall. Spatiotemporal analysis uncovers a cluster from 2006 to 2014 in northern Manitoba, with a secondary one from 2004 to 2012 in central Winnipeg. The findings inform public health by identifying high-risk clusters and emphasizing the need for localized prevention measures and resource allocation.

Au Canada, la gonorrhée est la deuxième infection sexuellement transmissible la plus répandue. En 2018, le Manitoba a enregistré un taux d'incidence trois fois supérieur à la moyenne nationale. Cette étude analyse les tendances spatiales, temporelles et spatio-temporelles de la gonorrhée au Manitoba de 2000 à 2016. Les cas sont regroupés par district à l'aide des codes postaux et des données du recensement. Les méthodes d'analyse comprennent l'indice global de Moran, les statistiques de balayage spatial et spatio-temporel de Kulldorff et les modèles ARIMA saisonniers. Des regroupements sont observés dans le nord du Manitoba et le centre de Winnipeg, avec des tendances saisonnières en fin d'été et en automne. Ces conclusions éclairent les stratégies de santé publique en identifiant les regroupements à haut risque et en soulignant la nécessité de mesures de prévention ciblées et d'une meilleure allocation des ressources.

Yushu Zou, Aya A. Mitani, Scott T. Leatherdale, Karen A. Patte

Investigating the association between school substance programs and student substance use: accounting for informative cluster size

Étude de l'association entre les programmes scolaires sur les substances et l'usage de substances chez les élèves: prise en compte de la taille informative des scolaire

Objective: To assess the association between school-level substance use programs on student substance use (binge drinking, cannabis, e-cigarette, cigarette use) while accounting for informative cluster size (ICS), where the cluster is the school. Methods: Using cross-sectional survey data from 74075 students at 136 Canadian high schools, we applied multivariate cluster-weighted generalized estimating equations (CWGEE) to examine the marginal associations between school programs and student substance use. We compared results to those from unweighted GEE analyses. Results: Larger schools tended to have lower rates of substance use. CWGEE showed a significant association between cannabis programs and overall substance use (OR: 0.81, CI: 0.70-0.94), unlike the GEE null findings (OR: 0.90, CI: 0.78-1.03). Conclusion: It is crucial to check for ICS in clustered data. Multivariate CWGEE can be used to study associations between school programs and student substance use while accounting for ICS.

Objectif: Évaluer l'association entre les programmes de lutte contre l'usage de substances au niveau des écoles et l'usage de substances chez les élèves (consommation excessive d'alcool, cannabis, cigarette électronique, utilisation de cigarettes) tout en prenant en compte la taille informative des groupes (ICS), où le groupe est l'école. Méthodes: En utilisant des données d'enquête transversale provenant de 74075 élèves dans 136 lycées canadiens, nous avons appliqué des équations d'estimation généralisées pondérées par cluster multivariées (CWGEE) pour examiner les associations marginales entre les programmes scolaires et l'usage de substances chez les élèves, comparant les résultats à ceux des analyses GEE non pondérées. Résultats: Les écoles plus grandes avaient tendance à avoir des taux d'usage de substances plus faibles. L'CWGEE a montré une association significative entre les programmes sur le cannabis et l'usage global de substances (OR: 0,81, IC: 0,70-0,94), contrairement aux résultats nuls de GEE (OR: 0,90, IC: 0,78-1,03). Conclusion: Il est crucial de vérifier la ICS dans les données groupées. L'CWGEE multivariée peut être utilisée pour étudier les associations entre les programmes scolaires et l'usage de substances chez les élèves tout en prenant en compte la ICS.

Rachel Lobay, Maria Jahja, Ajitesh Srivastava, Ryan J. Tibshirani, Daniel J. McDonald
Retrospective estimation of latent COVID-19 infections over the pandemic in U.S. states
Estimation rétrospective des infections latentes de la COVID-19 pendant la pandémie dans les états américains

Accurate estimates of latent COVID-19 infections can improve our understanding of the true size and scope of the pandemic and provide an indication of disease patterns and burden over time. Therefore, we estimate daily incident infections for each U.S. state. Our methods first deconvolve reported COVID-19 cases to their infection onset. We then use a serology-driven model to scale these deconvolved cases to unreported infections. Unlike existing approaches, our approach is state-specific, incorporates several variant-specific incubation periods, and accounts for reinfections. From its application to the gold-standard case data, we find a disease burden that appears earlier and more extensively than indicated by cases alone. In addition, we observe similar epidemic patterns in surges and periods of waning observed in clusters of neighbouring states. Our findings help to better understand the impact of the pandemic in the U.S. at the level of the state.

Des estimations précises des infections latentes de la COVID-19 peuvent améliorer notre compréhension de l'ampleur de la pandémie et fournir une indication des tendances et de la charge de la maladie au fil du temps. Ainsi, nous estimons les infections incidentes quotidiennes pour chaque état des États-Unis. Nos méthodes déconvoluent d'abord les cas de COVID-19 rapportés à leur début d'infection. Ensuite, nous utilisons un modèle basé sur la sérologie pour mettre à l'échelle ces cas déconvolués aux infections non signalées. Contrairement aux approches existantes, notre approche est spécifique à l'état, intègre plusieurs périodes d'incubation spécifiques aux variants et tient compte des ré-infections. À partir de son application aux données de cas de référence, nous constatons une charge de maladie qui semble apparaître plus tôt et plus largement que ne l'indiquent les cas seuls. De plus, nous observons des schémas épidémiques similaires dans les poussées et les périodes de déclin observées dans des groupes d'états voisins. Nos résultats aident à mieux comprendre l'impact de la pandémie aux États-Unis au niveau des états.

Statistical Methods in Genomics • Méthodes statistiques en génomique

9:30 - 10:30, *C 2033*
Chair • Président: Kyle McRae

Yushan Hu, Xiaojian Shao, Li Xing, Xuan Li, Geoffrey M. Nonis, Graeme J. Koelwyn, Xuekui Zhang, Don D. Sin
Single-Cell Sequencing of Lung Macrophages and Monocytes Reveals Novel Therapeutic Targets in COPD
Le séquençage unicellulaire des macrophages et des monocytes pulmonaires révèle de nouvelles cibles thérapeutiques de la MPOC

Macrophages and monocytes orchestrate inflammatory processes in the lungs. However, their role in the pathogenesis of chronic obstructive pulmonary disease (COPD), an inflammatory condition, isn't well known. We determined the characteristics of these cells in lungs of COPD patients and identified novel therapeutic targets. We analyzed the COPD and control lungs RNA sequencing data and found 16 transcriptionally distinct groups of macrophages and monocytes. We performed enrichment analyses to determine the characteristics of macrophages and monocytes from COPD (versus control) lungs and to identify the therapeutic targets, which were then validated using data from a randomized controlled trial of COPD patients. Our findings suggest COPD lungs harbor transcriptionally distinct lung macrophages and monocytes, reflecting a dysfunctional and hyperinflammatory state. Inhaled corticosteroids and other compounds can modulate the transcriptomic profile of these cells in patients with COPD.

Les macrophages et les monocytes orchestrent les processus inflammatoires dans les poumons. Cependant, leur rôle dans la pathogenèse de la maladie pulmonaire obstructive chronique (MPOC), une affection inflammatoire, n'est peu connu. Nous avons déterminé les caractéristiques de ces cellules dans les poumons de patients atteints de MPOC et identifié de nouvelles cibles thérapeutiques. Nous avons analysé les données de séquençage ARN des poumons de patients atteints de MPOC et de témoins, et avons identifié 16 groupes de macrophages et de monocytes transcriptionnellement distincts. Nous avons effectué des analyses d'enrichissement pour déterminer les caractéristiques des macrophages et des monocytes des poumons de patients atteints de MPOC (par rapport aux témoins) et pour identifier les cibles thérapeutiques qui ont ensuite été validées à l'aide de données issues d'un essai contrôlé randomisé de patients atteints de MPOC. Nos résultats suggèrent que les poumons de patients atteints de MPOC abritent des macrophages et des monocytes transcriptionnellement distincts, reflétant un état dysfonctionnel et hyperinflammatoire. Les corticostéroïdes inhalés et d'autres composés peuvent moduler le profil transcriptomique de ces cellules chez les patients atteints de MPOC.

Qicheng Zhao, Qihuang Zhang

Bayesian Model for Disease-Specific Gene Detection in High-Dimensional Spatially Resolved Transcriptomics

Modèle bayésien pour la détection des gènes spécifiques à une maladie dans la transcriptomique à haute dimension spatiale

Identifying disease-indicative genes is critical for deciphering disease mechanisms and continues to attract significant interest. Spatial transcriptomics offers unprecedented insights for the detection of disease-specific genes by enabling within-tissue contrasts. However, this new technology poses challenges for conventional statistical models developed for RNA-seq, as these models often neglect the spatial organization of tissue spots. In this talk, we discuss a new Bayesian shrinkage model to characterize the relationship between high-dimensional gene expressions and the disease status of tissue spots, incorporating spatial correlation among these spots through autoregressive terms. Our model adopts a hierarchical structure to accommodate for the missing data within tissues and is further extended to facilitate the analysis of multiple correlated samples. To ensure the model's applicability to datasets of varying sizes, we carry out two computational frameworks for Bayesian parameter estimation, tailored to both small and large sample scenarios. Simulation studies are conducted to evaluate the performance of the proposed model, and we also apply our model to analyze the data arising from a HER2-positive breast cancer study. This work is supervised by Dr. Qihuang Zhang.

L'identification des gènes indicatifs de la maladie est essentielle pour déchiffrer les mécanismes de la maladie et suscite un grand intérêt. La transcriptomique spatiale offre des perspectives sans précédent pour la détection de gènes spécifiques à une maladie en permettant des contrastes intra-tissulaires. Cependant, cette nouvelle technologie pose des défis aux modèles statistiques conventionnels développés pour l'ARN-seq, car ces modèles négligent souvent la distribution spatiale des taches tissulaires. Dans cet exposé, nous discutons d'un nouveau modèle bayésien de rétrécissement pour caractériser la relation entre les expressions génétiques à haute dimension et l'état pathologique des taches tissulaires, en incorporant la corrélation spatiale entre ces taches tissulaires par le biais de termes autorégressifs. Notre modèle adopte une structure hiérarchique pour tenir compte des données manquantes dans les tissus et est étendu à l'analyse d'échantillons multiples corrélés. Pour garantir l'applicabilité du modèle à des ensembles de données de différentes tailles, nous réalisons deux cadres de calcul pour l'estimation bayésienne des paramètres, adaptés aux scénarios de petits et de grands échantillons. Des études de simulation sont menées pour évaluer les performances du modèle proposé et nous appliquons également notre modèle pour analyser les données provenant d'une étude sur le cancer du sein HER2-positif. Ce travail est supervisé par le Dr. Qihuang Zhang.

Kyle Gardiner, Li Xing, Xuekui Zhang
BLESS: Bagged Logistic Regression for Biomarker Identification
BLESS: Régression logistique bagged pour l'identification de biomarqueurs

The traditional SNP-wise approach in genome-wide association studies is focused on examining the marginal association between each single nucleotide polymorphism with the outcome separately and applying multiple testing adjustments to the resulting p-values to reduce false positives. However, the approach suffers a lack of power in identifying biomarkers. We design an ensemble machine learning approach to aggregate results from logistic regression models based on multiple sub-samples, which helps to identify biomarkers from high-dimensional genomic data. We employ different methods to analyze a genome-wide association study from the Alzheimer's Disease Neuroimaging Initiative. The SNP-wise approach does not identify any significant signal, while our novel approach provides a list of ranked SNPs associated with the cognitive functions of interests.

L'approche traditionnelle SNP dans les études d'association pangénomique se concentre sur l'examen de l'association marginale entre chaque polymorphisme nucléotidique et le résultat séparément et l'application des ajustements pour les tests multiples aux valeurs-p résultantes afin de réduire les faux positifs. Cependant, cette approche souffre d'un manque de puissance dans l'identification des biomarqueurs. Nous concevons une approche d'apprentissage automatique ensembliste pour agréger les résultats des modèles de régression logistique basés sur de multiples sous-échantillons, ce qui aide à identifier les biomarqueurs à partir de données génomiques de haute dimension. Nous utilisons différentes méthodes pour analyser une étude d'association pangénomique de l'Alzheimer's Disease Neuroimaging Initiative. L'approche SNP n'identifie aucun signal significatif, tandis que notre nouvelle approche fournit une liste de SNP classés associés aux fonctions cognitives d'intérêt.

Bertrand Sodjahin, Alex Stringer, Shoja Chenouri

Spike-and-slab based Informative Priors for Bayesian Network Structure Learning in High-dimensional regimes

Apprentissage de la structure des réseaux bayésiens de haute dimension en utilisant les distributions a priori informatives 'spike-and-slab'

Learning high-scoring structures, in the super-exponential space of Graphical Models, is NP-hard. Current standard approaches for addressing such problem are heuristics-based, notably MCMC. In a Bayesian framework, this is an optimization problem which amounts to appropriately specifying both, prior on structure and marginal likelihood of the data given the structure model. Though significant contributions have been made to the latter, the former has garnered less attention. For convenience, an often not well-motivated uniform prior is routinely used. While it works under large samples because the prior term is asymptotically negligible, uniform prior becomes inadequate in High Dimension Low Sample Size (HDLSS) settings. To inform the optimization and hence improve the accuracy of the posterior, in HDLSS regimes, we propose sparsity-inducing spike-and-slab based structure priors. We compare them to uniform prior and to other non-uniform priors, in low and high dimensional scenarios.

L'apprentissage de structures à score élevé, dans l'espace super-exponentiel des modèles graphiques, est NP-difficile. Les approches standard actuelles pour résoudre ce problème sont basées sur des heuristiques, notamment le MCMC. Dans un cadre bayésien, il s'agit d'un problème d'optimisation qui revient à spécifier de manière appropriée à la fois la distribution a priori sur la structure et la vraisemblance marginale des données compte tenu du modèle structurel. Bien que des contributions significatives aient été apportées à ce dernier point, le premier a suscité moins d'attention. Par commodité, une distribution a priori uniforme souvent mal motivée est fréquemment utilisée. Bien qu'il fonctionne pour les grands échantillons parce que le terme de la distribution a priori est asymptotiquement négligeable, une distribution uniforme a priori devient inadéquate dans les contextes de haute dimension et de faible taille.

Environmental and Ecological Statistics • Statistiques environnementales et écologiques

9:30 - 10:30 *C 2045*
Chair • Présidente: Selina Elvayn

Zirui Dong
Arsenic Contamination in Nova Scotia's Domestic Well Water: a Spatial-Temporal Statistical Analysis
Contamination par l'arsenic dans l'eau des puits domestiques de la Nouvelle-Écosse: une analyse statistique spatio-temporelle

Arsenic contamination in Nova Scotia's well water is a significant public health concern due to its association with various health issues. However, our understanding of the relationship between environmental variables, climate change, and arsenic levels remains limited. We conducted a study spanning from 2000 to 2021, cleaning and linking climate and water contamination data to create a predictive model for arsenic exceedance in well water and compared models to assess the impact on the predictability of each environmental variable. Utilising a generalized additive model with logistic regression, we estimated the proportion of well water exceeding the safety threshold of 5 micrograms per litre for arsenic contamination. Our analysis identified Yarmouth, Halifax, and Cape Breton as areas with significant arsenic exceedance. This is likely due to a range of environmental factors, with geographic location and well type being the most significant predictors according to our findings.

En raison de son association avec divers problèmes de santé, la contamination par l'arsenic dans l'eau des puits de la Nouvelle-Écosse est une préoccupation majeure pour la santé publique. Cependant, notre compréhension de la relation entre les variables environnementales, le changement climatique et les niveaux d'arsenic reste limitée. Nous avons mené une étude couvrant la période de 2000 à 2021 en nettoyant et en reliant les données climatiques et de contamination de l'eau pour créer un modèle prédictif de d'excès de l'arsenic dans l'eau des puits et avons comparé les modèles pour évaluer l'impact sur la prévisibilité de chaque variable environnementale. En utilisant un modèle additif généralisé avec régression logistique, nous avons estimé la proportion d'eau de puits dépassant le seuil de sécurité de 5 microgrammes par litre pour la contamination par l'arsenic. Notre analyse a identifié Yarmouth, Halifax et Cape Breton comme des zones présentant un dépassement significatif de l'arsenic. Ceci est probablement dû à une gamme de facteurs environnementaux, la localisation géographique et le type de puits étant les prédicteurs les plus significatifs selon nos résultats.

Raphael McDonald, Ethan Lawler, Cornelia den Heyer, Bradford Hubley, Lingbo Li, Joanna Mills Flemming

Hidden bias in model-based stock assessment index standardization methods applied to design-based stratification

Biais camouflé dans la normalization des indices d'évaluations des stocks par modèles appliqués à une stratification fondé sur le plan de sondage

Many stock assessments around the world rely on science advice to ensure the sustainability of fisheries. While design-based estimators focusing on stratifications have been extensively used to calculate population indices for advice, many processes are switching to model-based approaches. Little attention has been given to the impact of different sampling designs and allocation schemes on new multivariate hierarchical models. Utilizing a spatio-temporal multinomial exponential model, we examine its performance under five different combinations of sampling designs and allocation schemes. Simulations demonstrate unbalanced allocation schemes introduce bias even when underlying assumptions are respected. Potential solutions are identified, including robust statistics and the inclusion of random effects, and applied to a case study of cusk. Our research helps identify often unknown bias in common approaches and proposes solutions to generate improved science advice.

Plusieurs processus d'évaluation des stocks dépendent d'avis scientifiques pour assurer la durabilité des pêches. Bien que les estimations basées sur les plans de sondage aient été largement utilisés pour créer ces avis, beaucoup de processus ont remplacé ces méthodes par des modèles. L'impact des plans de sondages et des systèmes d'allocations sur ces nouveaux modèles n'a reçu que peu d'attention. À l'aide d'un modèle spatio-temporel multinomiale exponentiel, nous examinons l'impact de cinq combinaisons de plans de sondages et systèmes d'allocations sur sa performance. Les simulations démontrent que des systèmes d'allocations non-balancés introduisent un biais même si les assomptions du modèle sont respectées. Les statistiques robustes et certains effets aléatoires sont proposés comme solution. Une étude de cas est réalisée sur la pêche pour le brosme. Notre recherche identifie une source de biais inconnue lors du développement d'avis scientifiques pour l'évaluation des stocks.

Liam Cann, Matthew Stephenson, Connie Stewart
Impacts of correlation in bioassays
Les impacts de la corrélation dans les bioessais

Bioassays play a key role in ensuring that every batch of drug produced is safe and effective for release. They play a critical role in testing the potency of biologic drugs, such as vaccines or monoclonal antibodies. Due to the importance of bioassays in ensuring a safe and effective product, it is critical that the statistical methods used are appropriate. However, it is the current common practice to treat the replicate responses at each dose level as if they are independent despite the fact they are often correlated. In this research, we look at quantitatively assessing the risks of the conventional analysis methods using a simulation study to investigate the impact of correlation on the statistical analysis of bioassays. Specifically, we consider parallelism assessment, model goodness-of-fit, and relative potency estimation. We also make recommendations within the constraints of the current available commercial bioassay analysis software to provide valid statistical inference.

Les bioessais jouent un rôle clé pour garantir que chaque lot de médicaments produit est sûr et efficace pour sa libération. Ils jouent un rôle essentiel pour tester la puissance des médicaments biologiques, tels que les vaccins ou les anticorps monoclonaux. En raison de l'importance des bioessais, il est essentiel que les méthodes statistiques utilisées soient appropriées. Cependant, il est courant de traiter les réponses répliquées à chaque niveau de dose comme si elles étaient indépendantes, malgré le fait qu'elles sont souvent corrélées. Dans cette recherche, nous examinons l'évaluation quantitative des risques des méthodes d'analyse conventionnelles en utilisant une étude de simulation pour étudier l'impact de la corrélation sur l'analyse statistique des bioessais. Nous examinons spécifiquement l'évaluation du parallélisme, l'ajustement du modèle et l'estimation de la puissance relative. Nous formulons également des recommandations dans les limites des logiciels commerciaux d'analyse de bioessais actuellement disponibles pour fournir des inférences statistiques valides.

Reshani Abayasekara, Jackie L. Whittaker, S. Amanda Ali, Osvaldo Espin-Garcia
A functional curve registration approach to understanding physical activity levels measured by accelerometers
Une approche de l'enregistrement des courbes fonctionnelles pour comprendre les niveaux d'activité physique mesurés par des accéléromètres

Accelerometers are wearable devices that objectively measure physical activity (PA). Functional curve registration (FCR) separates vertical variability measured in PA intensity from horizontal variability measured in time, thus placing individuals on a shared time scale. FCR achieves this by estimating an individual-specific inverse warping function that captures horizontal variability and applies functional principal components analysis to capture vertical variability. We applied this approach to 2,700 participants from an osteoarthritis (OA) cohort. Using the area under the registered curve (AURC) to represent participant PA levels and linear mixed-effects models, we assessed the association between AURC and symptomatic and structural OA outcomes. As the first study to apply curve registration to OA research; our results suggest significant potential in utilizing curve registration on raw accelerometer data to quantify PA levels.

Les accéléromètres sont des dispositifs portables qui mesurent objectivement l'activité physique (AP). L'enregistrement des courbes fonctionnelles (ECF) sépare la variabilité verticale mesurée dans l'intensité de l'AP de la variabilité horizontale mesurée dans le temps, plaçant ainsi les individus sur une échelle temporelle commune. L'ECF y parvient en estimant une fonction de déformation inverse spécifique à l'individu qui capture la variabilité horizontale et applique une analyse fonctionnelle des composantes principales pour capturer la variabilité verticale. Nous avons appliqué cette approche à environ 2 700 participants d'une cohorte d'arthrose. En utilisant l'aire sous la courbe enregistrée (AURC) pour représenter les niveaux d'AP des participants et des modèles linéaires à effets mixtes, nous avons évalué l'association entre l'AURC et les résultats symptomatiques et structurels de l'arthrose. En tant que première étude à appliquer l'enregistrement des courbes à la recherche sur l'arthrose, nos résultats suggèrent un potentiel significatif dans l'utilisation de l'enregistrement des courbes sur les données brutes d'accéléromètre pour quantifier les niveaux d'AP.

Innovative Algorithms in Data Science • Algorithmes innovants en science des données

9:30 - 10:30, *C 3053*
Chair • Présidente: Mathilde Dicaire-Cartier

Nikola Surjanovic, Miguel Biron-Lattes, Saifuddin Syed, Trevor Campbell, Alexandre Bouchard-Côté
autoMALA: Locally adaptive Metropolis-adjusted Langevin algorithm
autoMALA: Algorithme de Langevin ajusté par Metropolis avec adaptation locale

Selecting the step size for the Metropolis-adjusted Langevin algorithm (MALA) is necessary in order to obtain satisfactory performance. However, finding an adequate step size for an arbitrary target distribution is a difficult task and even the best step size can perform poorly in specific regions of the space when the distribution is complex. To resolve this issue we introduce autoMALA, a new Markov chain Monte Carlo algorithm based on MALA that automatically sets its step size at each iteration based on the local geometry of the target distribution. We prove that autoMALA has the correct invariant distribution, and our experiments demonstrate that autoMALA is competitive with related state-of-the-art MCMC methods. We find that autoMALA outperforms state-of-the-art samplers on targets with varying geometries and that it tends to find step sizes comparable to optimally-tuned MALA when a fixed step size suffices for the whole domain.

La sélection de la taille d'étapes pour l'algorithme de Langevin ajusté par Metropolis (MALA) est nécessaire pour obtenir des performances satisfaisantes. Cependant, trouver une taille d'étapes adéquate pour une distribution cible arbitraire est une tâche difficile et même la meilleure taille d'étapes peut donner de mauvais résultats dans des régions spécifiques de l'espace lorsque la distribution est complexe. Pour résoudre ce problème, nous présentons autoMALA, un nouvel algorithme de Monte Carlo à chaîne de Markov basé sur MALA qui définit automatiquement sa taille d'étape à chaque itération en fonction de la géométrie locale de la distribution cible. Nous prouvons qu'autoMALA possède la correcte distribution invariante et nos expériences démontrent qu'autoMALA est compétitif par rapport aux méthodes MCMC de pointe. Nous constatons qu'autoMALA surpasse les échantillonneurs modernes sur des cibles à géométrie variable et qu'il tend à trouver des tailles d'étape comparables à MALA optimisé lorsqu'une taille d'étape fixe suffit pour l'ensemble du domaine.

Chunlei Ge, W. John Braun

Quick and Simple Kernel Differential Equation Regression Estimators for Data with Sparse Design

Estimateurs rapides et simples de régression par équation différentielle à noyau pour les données à structure éparse

Local polynomial regression of order at least one often performs poorly in regions of sparse data. Local constant regression is exceptional in this regard, though it is the least accurate method in general, especially at the boundaries of the data. Incorporating information from differential equations which may approximately or exactly hold is one way of extending the sparse design capacity of local constant regression while reducing bias and variance. A nonparametric regression method that exploits first order differential equations is introduced in this paper and applied to noisy mouse tumor growth data. Asymptotic biases and variances of kernel estimators using Taylor polynomials with different degrees are discussed. Model comparison is performed for different estimators through simulation studies under various scenarios which simulate exponential-type growth.

La régression polynomiale locale avec un ordre égal ou supérieur à un donne souvent des résultats médiocres dans les régions où les données sont peu nombreuses. La régression locale constante est exceptionnelle à cet égard, bien qu'elle soit la méthode la moins précise en général, en particulier aux limites des données. L'incorporation d'informations provenant d'équations différentielles qui peuvent être à peu près ou exactement valides est un moyen d'étendre la capacité de la structure éparse de la régression constante locale tout en réduisant le biais et la variance. Une méthode de régression non paramétrique qui exploite les équations différentielles du premier ordre est présentée dans cet article et appliquée à des données bruyantes sur la croissance des tumeurs chez les souris. Les biais et variances asymptotiques des estimateurs à noyau utilisant des polynômes de Taylor de différents degrés sont discutés. Une comparaison des modèles est effectuée pour différents estimateurs employant le biais des études de simulation dans divers scénarios qui simulent une croissance de type exponentiel.

Kai Yang, Masoud Asgharian, Nikhil Bhagwat, J.B. Poline, Celia Greenwood
fastHDMI: Fast Fourier Transform (FFT)-based Mutual Information Estimation for High-Dimensional Data
astHDMI: Estimation de l'information mutuelle basée sur la transformation de Fourier rapide (FFT) pour les données de haute dimension

This paper tackles the challenge of selecting variables from a vast pool for high-dimensional model fitting, crucial for enhancing model accuracy and interpretability. We implement a comprehensive methodology for variable screening, focusing on mutual information estimation through two main techniques: the k-nearest neighbours (kNN) and Kernel Density Estimation (KDE), with KDE efficiency significantly boosted by the Fast Fourier Transform (FFT). Our study zeroes in on univariate variable screening, aiming to pinpoint the most informative variables. We conduct a detailed evaluation of these methods on the Autism Brain Imaging Data Exchange dataset, containing high-dimensional brain imaging fMRI data for cases and controls. We analyze their performance of variable selection based on simulated outcomes and their overall contribution to model predictions based on real data. This rigorous assessment seeks to offer insights into the effectiveness of these techniques for high-dimensional data analysis, aiding more informed decision-making in statistical modelling and machine learning.

Cet article s'attaque au défi que représente la sélection de variables à partir d'un vaste ensemble pour l'ajustement de modèles à haute dimension, ce qui est crucial pour améliorer la précision et l'interprétabilité des modèles. Nous mettons en œuvre une méthodologie complète pour la sélection des variables, en nous concentrant sur l'estimation de l'information mutuelle par le biais de deux techniques principales: la méthode des k plus proches voisins (kNN) et l'estimation de la densité du noyau (KDE), l'efficacité de la KDE étant considérablement renforcée par la transformation de Fourier rapide (FFT). Notre étude se concentre sur le filtrage des variables univariées, dans le but d'identifier les variables les plus informatives. Nous procédons à une évaluation détaillée de ces méthodes sur l'ensemble de données Autism Brain Imaging Data Exchange, qui contient des données IRMf d'imagerie cérébrale à haute dimension pour les cas et les témoins. Nous analysons leurs performances en matière de sélection de variables sur la base de résultats simulés et leur contribution globale aux prédictions du modèle sur la base de données réelles. Cette évaluation rigoureuse vise à donner un aperçu de l'efficacité de ces techniques pour l'analyse de données à haute dimension, en aidant à prendre des décisions plus éclairées en matière de modélisation statistique et d'apprentissage automatique.

Bingcheng Wang, Haochen Song, Pan Chen, Ilya Musabirov, Yi Wang, Ananya Bhattacharjee, Joseph Williams

Weighted Allocation Probability-adjusted Thompson Sampling (WAPTS): a Novel Algorithm for Sparse and Adaptive Contextual Bandits

L'échantillonnage de Thompson ajusté à la probabilité de l'allocation pondérée (WAPTS): un nouvel algorithme pour les bandits contextuels épars et adaptatifs

In statistical inference and personalization, adaptive bandit algorithms are known for their reward maximization nature. Yet, the power of statistical inferences is less studied due to their non-converging Markov Chain probabilities. Among these, contextual bandit algorithms, such as Contextual Thompson Sampling (CTS), further personalize reward maximization, but face greater curse of dimensionality challenge due to adding contexts. Considering the limited data-to-dimension ratio and the need for dynamic, personalizable experimental designs, we introduce a novel algorithm called Weighted Allocation Probability-adjusted Thompson Sampling (WAPTS). WAPTS uses a tunable weighting parameter to iteratively adjust allocation probabilities per intervention, accelerating optimization convergence. It retains CTS's core properties while delegating intervention more responsively. We examined the effect of WAPTS through two real-world deployments with two criteria: the average probability of a good-intervention selection, and the average of potential loss.

Dans le domaine de l'inférence statistique et de la personnalisation, les algorithmes de bandits adaptatifs sont connus pour leur nature de maximisation de la récompense. Cependant, la puissance des inférences statistiques est moins étudiée en raison de la non-convergence des probabilités de la chaîne de Markov. Parmi ceux-ci, les algorithmes de bandits contextuels, tels que l'échantillonnage contextuel de Thompson (CTS), personnalisent davantage la maximisation de la récompense, mais sont confrontés à une plus grande malédiction de la dimensionnalité en raison de l'ajout de contextes. Compte tenu du rapport limité entre les données et les dimensions et de la nécessité de concevoir des modèles expérimentaux dynamiques et personnalisables, nous présentons un nouvel algorithme appelé Weighted Allocation Probability-adjusted Thompson Sampling (l'échantillonnage de Thompson ajusté à la probabilité de l'allocation pondérée ou WAPTS). WAPTS utilise un paramètre de pondération réglable pour ajuster itérativement les probabilités d'allocation par intervention, accélérant ainsi la convergence de l'optimisation. Il conserve les propriétés essentielles du CTS tout en déléguant les interventions de manière plus réactive. Nous avons examiné l'effet de WAPTS à travers deux déploiements réels avec deux critères: la probabilité moyenne d'une sélection de bonne intervention et la moyenne de la perte potentielle.

Time Series and Dynamic Models • Séries chronologiques et modèles dynamiques

13:30 - 14:30, *C 2004*
Chair • Présidente: Jay Sivathayalan

Roberto Curti, Erfan Hoque, Sean Hellingman
ts.shiny: Interactive Visualization through a Shiny App Applied to Time Series Data
ts.shiny: visualisation interactive à travers une application Shiny appliquée aux données de séries temporelles

Addressing the complexities in research and data analysis remains a challenge, despite the advancements in traditional graphical tools. Exploratory data analysis is pivotal, setting the base tone for research, with visualisation playing a key role. For time series, the need for useful graphics is increased: data are inherently complex, high-dimensional, and correlated, generating graphics for such datasets and analysis can be cumbersome and time-consuming. To streamline the creation of graphics for such datasets, the creation of a Shiny App, ts.shiny, introduces a dynamic and interactive visualisation framework for common time series data analyses. Illustrations of the tool with data from Magic: The Gathering card prices highlight its practical implications. This interactive platform allows for real-time market analysis and visualisation, showcasing its potential to make complex data analysis accessible and actionable for a broader audience.

Adresser les complexités de la recherche et de l'analyse de données reste un défi, malgré les avancées dans les outils graphiques traditionnels. L'analyse exploratoire des données est cruciale, car elle établit le ton de base pour la recherche, la visualisation jouant un rôle clé. Pour les séries chronologiques, le besoin de graphiques utiles est accru: les données sont intrinsèquement complexes, multidimensionnelles et corrélées et générer des graphiques pour de tels ensembles de données et analyses peut être fastidieux. Pour rationaliser la création de graphiques pour de tels ensembles de données, la création d'une application Shiny, ts.shiny, introduit un cadre de visualisation dynamique et interactif pour les analyses de séries chronologiques courantes. Des illustrations de l'outil avec des données sur les prix des cartes du jeu Magic: The Gathering mettent en évidence ses implications pratiques. Cette plateforme interactive permet une analyse et une visualisation du marché en temps réel, démontrant son potentiel à rendre l'analyse de données complexe accessible et opérationnelle pour un public plus large.

Parham Pishrobat, Stefan Schrunner, William Welch
Introducing Dynamic Regression Model for Hydrological Inference
Introduction d'un modèle de régression dynamique pour l'inférence hydrologique

This study introduces the Multiple Sliding Window Regression (MSWR) model, a novel framework designed to enhance hydrological inference and prediction using readily collectible climate variables. Stream flow directly results from current and past rainfalls and other climate variables like temperature impose a non-constant effect over time. The MSWR model effectively accounts for the lagged effects of rainfall on streamflow and the variability introduced by temperature fluctuations, thus effectively capturing streamflow's temporal dynamics. The parameters of the MSWR model represent the characteristics of different lagged windows, where each window represents a distinct flow path. As a result, MSWR provides direct interpretability on the properties of each flow path, including their location, spread, and weight over time. Results from both simulation studies and real-world data applications imply that temperature significantly improves the model's fitness and predictive performance.

Cette étude présente le modèle de régression à fenêtre mobile multiple (MSWR), un nouveau cadre conçu pour améliorer l'inférence et la prédiction hydrologique en utilisant des variables climatiques facilement collectables. Le débit d'un cours d'eau résulte directement des précipitations actuelles et passées et d'autres variables climatiques, telles que la température, imposent un effet non constant dans le temps. Le modèle MSWR tient efficacement compte des effets retardés des précipitations sur le débit d'eau et de la variabilité introduite par les fluctuations de température, capturant ainsi la dynamique temporelle du débit d'eau. Les paramètres du modèle MSWR représentent les caractéristiques de différentes fenêtres retardées, chaque fenêtre représente un chemin d'écoulement distinct. En conséquence, le MSWR fournit une interprétation directe des propriétés de chaque chemin d'écoulement, y compris leur emplacement, leur répartition et leur poids dans le temps. Les résultats des études de simulation et des applications de données réelles indiquent que la température améliore significativement l'ajustement et la performance prédictive du modèle.

Xize Ye, Marcos Escobar, Lars Stentoft

Generalized Autoregressive Conditionally Stochastic Heteroskedasticity: Motivation and Applications

Generalized Autoregressive Conditionally Stochastic Heteroskedasticity: Motivations et applications

Typical GARCH models are proven successful in capturing time-varying conditional variance of asset return. Nonetheless, the construction that only 1 innovation drives both the return and variance process make it difficult to reconcile historical return and forward-looking information, such as the VIX (volatility index). Instead, we propose a methodology to add another innovation to GARCH models to allow stochastic volatility, hence resulting in a 2-shock model named Generalized Autoregressive Conditionally Stochastic Heteroskedasticity (GARCSH). In this talk, we discuss the motivation, implementation and financial applications of GARCSH.

Les modèles GARCH typiques se sont avérés efficaces pour capturer la variance conditionnelle variable dans le temps des rendements d'actifs. Néanmoins, la construction selon laquelle une seule innovation détermine à la fois le processus de rendement et de variance rend difficile la conciliation entre les rendements historiques et les informations prospectives, telles que l'indice de volatilité (VIX). À la place, nous proposons une méthodologie pour ajouter une autre innovation aux modèles GARCH afin de permettre une volatilité stochastique, conduisant ainsi à un modèle à 2 chocs nommé Generalized Autoregressive Conditionally Stochastic Heteroskedasticity (GARCSH). Dans cette présentation, nous discutons de la motivation, de la mise en œuvre et des applications financières de GARCSH.

Chen Chen, Zihang Lu, Geoff Anderson, Davide Chicco, Kuan Liu

Longitudinal Cognitive Trajectory Modelling and Phenotyping with Multiple Features Using Health Administrative Data

Modélisation et phénotypage de trajectoires cognitives longitudinales avec des caractéristiques multiples à l'aide de données administratives sur la santé

Dementia is a diverse set of diseases characterized by multiple progressive phenotypes. However, most longitudinal dementia clustering studies were restricted to a single longitudinal dementia feature or limited sample size. To understand the heterogeneity in dementia trajectories, we used a latent class mixed model to jointly model the underlying trajectories of three repeated dementia features: aggressive behavior, cognitive impairment, and activities of daily living. Time-to-death was considered as informative dropout and modeled through Cox model jointly with the multiple dementia features. In addition, multinomial logistic regression was used to identify predictive clinical factors of identified dementia clusters. We analyzed a population-based administrative cohort with 42,774 patients from the Canadian Institute for Health Information. Understanding the diverse patterns of dementia trajectories can provide effective care and optimize healthcare resources.

La démence est un ensemble diversifié de maladies caractérisées par de multiples phénotypes progressifs. Cependant, la plupart des études longitudinales sur la démence employant l'analyse des regroupements se limitaient à une seule caractéristique longitudinale de la démence ou à un échantillon de taille limitée. Pour comprendre l'hétérogénéité des trajectoires de la démence, nous avons utilisé un modèle mixte à classes latentes pour modéliser conjointement les trajectoires sous-jacentes de trois caractéristiques répétées de la démence: comportement agressif, troubles cognitifs et activités de la vie quotidienne. Le temps écoulé jusqu'au décès a été considéré comme un abandon informatif et modélisé par le modèle de Cox conjointement avec les multiples caractéristiques de la démence. De plus, la régression logistique multinomiale a été utilisée pour identifier les facteurs cliniques prédictifs des groupes de démence identifiés. Nous avons analysé une cohorte administrative basée sur la population de 42 774 patients de l'Institut canadien d'information sur la santé. La compréhension des divers modèles de trajectoires de démence peut permettre de fournir des soins efficaces et d'optimiser les ressources de santé.

Bayesian Methods and Applications • Méthodes et applications bayésiennes

13:30 - 14:30, *C 2033*
Chair • Président: Kyle McRae

Wuqian Gao, Tingxuan Wu, Longhai Li, Cindy X. Feng
Bayesian Z-residuals for Hurdle Models
Résidus Z bayésiens pour les modèles d'obstacles

Model diagnosis is a crucial step in checking the goodness-of-fit of a fitted model. Z-residuals were proposed as a diagnostic tool for frequentist model diagnosis, overcoming the limitations of Pearson and deviance residuals. The Z-residual is transformed from the predictive p-values with the normal quantile function. We show that the predictive p-value has a uniform distribution on (0,1) under the true model. Due to the uniformity of the predictive p-values, Z-residuals are normally distributed under the true model. Z-residuals can be extended to various Bayesian models. The study focuses on hurdle models, with developed generic R functions for predictive p-value computation based on fitting outputs of the R package brms. The study aims to enhance predictive p-value computation for improved Z-residual derivation, contributing to better model diagnosis in Bayesian analysis.

Le diagnostic d'un modèle est une étape cruciale pour vérifier sa qualité de l'ajustement. Les résidus Z ont été proposés comme outil de diagnostic pour les diagnostics de modèles fréquentiste, surmontant les limitations des résidus de Pearson et de déviance. Le résidu Z est transformé à partir des valeurs-p prédictives avec la fonction quantile normale. Nous montrons que la valeur-p prédictive a une distribution uniforme sur (0,1) sous le vrai modèle. En raison de l'uniformité des valeurs-p prédictives, les résidus Z sont normalement distribués sous le vrai modèle. Les résidus Z peuvent être étendus à différents modèles bayésiens. L'étude se concentre sur les modèles de seuil, avec le développement de fonctions R génériques pour le calcul des valeurs-p prédictives basé sur les sorties d'ajustement du package R brms. L'étude vise à améliorer le calcul des valeurs-p prédictives pour une meilleure dérivation des résidus Z, contribuant ainsi à un meilleur diagnostic de modèle dans l'analyse bayésienne.

Jingwen Ji, Ruochen Zhao, Ruiying Wang

Improving Toronto's Overnight Shelter Allocation and Utilization using Bayesian Non-Parametric Models

AAmélioration de l'Allocation et de l'Utilisation des Refuges de Nuit à Toronto en Utilisant des Modèles Bayésiens Non Paramétriques

Purpose: The surge in homelessness demands efficient shelter management in Toronto's urban areas. Accurately predicting shelter occupancy rates is crucial for resource allocation. However, estimating the nonlinear relationship between shelter occupancy rates and predictors is challenging with conventional models. Our study employs Bayesian nonparametric regression, which are known for their flexibility and robustness in modeling complex data patterns, to accurately predict occupancy rates. Methods: We developed logistic regression, gradient boosting, Bayesian causal forest, Bayesian additive regression tree, and Gaussian process regression models. Precision, recall, F1 score are used for evaluation. Conclusion: Our model could accurately predict the outcomes. By anticipating nightly demands, Toronto can manage resources effectively, reduce overcrowding, and enhance homeless individuals' well-being.

Objectif: La hausse du taux d'itinérance exige une gestion efficace des refuges dans les zones urbaines de Toronto. Prédire avec précision les taux d'occupation des refuges est crucial pour l'allocation des ressources. Cependant, estimer la relation non linéaire entre les taux d'occupation des refuges et les prédicteurs est un défi avec les modèles conventionnels. Notre étude utilise la régression bayésienne non paramétrique, réputée pour sa flexibilité et sa robustesse dans la modélisation des motifs de données complexes, pour prédire avec précision les taux d'occupation. Méthodes: Nous avons développé des modèles de régression logistique, de boosting, de forêt causale bayésienne, d'arbre de régression additive bayésienne et de régression par processus gaussien. La précision, le rappel et le score F1 sont utilisés pour l'évaluation. Conclusion: Notre modèle a pu prédire avec précision les résultats. En anticipant les demandes nocturnes, Toronto pourrait mieux gérer les ressources, réduire la surpopulation et améliorer le bien-être des personnes sans-abri.

Linke Li, Anna Heath

Efficiently Evaluating the Operating Characteristics of Bayesian Clinical Trial with Machine Learning

Évaluation efficace des caractéristiques opérationnelles d'un essai clinique bayésien avec l'apprentissage automatique

Bayesian statistical methods for clinical trial design have grown in popularity over recent decades, offering the ability to incorporate external information into the planning phase and enhance the cost-effectiveness of trials. However, Bayesian trial designs that adopt non-conjugate probabilistic models usually require the repeated evaluation of the posterior probability statements of effectiveness through the Monte Carlo method. This increases the computational burden and hinders implementation in time-sensitive scenarios, such as the Covid-19 pandemic. To address this, we have adopted machine learning techniques and summary statistics to efficiently estimate the posterior probability of effectiveness. We compared the performance of our methods with the conventional estimation approach using a real-world clinical trial. The results demonstrate that our proposed methods can significantly reduce computational time from hours to minutes without sacrificing accuracy.

Les méthodes bayésiennes pour la conception d'essais cliniques, permettant d'intégrer des informations externes, ont gagné en popularité. Toutefois, ces méthodes, utilisant des modèles non conjugués, nécessitent une évaluation répétée par la méthode de Monte Carlo, augmentant la charge de calcul, ce qui est problématique dans des contextes urgents comme la pandémie de Covid-19. Pour y remédier, nous avons utilisé des techniques d'apprentissage automatique et des statistiques résumées pour estimer efficacement la probabilité postérieure d'efficacité. En comparant nos méthodes à une approche conventionnelle sur un essai clinique réel, nous avons constaté une réduction significative du temps de calcul, passant de heures à minutes, sans perte de précision.

Muye Nanshan, Nan Zhang, Jiguo Cao

A Joint Estimation Approach to Sparse Additive Ordinary Differential Equations
Une approche d'estimation conjointe des équations différentielles ordinaires additives éparse

Ordinary differential equations (ODEs) are widely used to characterize the dynamics of complex systems in real applications. In this article, we propose a novel joint estimation approach for generalized sparse additive ODEs where observations are allowed to be non-Gaussian. The new method is unified with existing collocation methods by considering the likelihood, ODE fidelity and sparse regularization simultaneously. We design a block coordinate descent algorithm for optimizing the non-convex and non-differentiable objective function. The global convergence of the algorithm is established. The simulation study and two applications demonstrate the superior performance of the proposed method in estimation and improved performance of identifying the sparse structure.

Les équations différentielles ordinaires (EDO) sont très utilisées pour caractériser la dynamique des systèmes complexes dans des applications réelles. Nous proposons une nouvelle approche d'estimation conjointe pour les EDO additives éparses généralisées où les observations peuvent ne pas être gaussiennes. La nouvelle méthode provient de méthodes de collocation existantes en considérant simultanément la vraisemblance, la fidélité des EDO et la régularisation éparse. Nous concevons un algorithme de descente de coordonnées par bloc pour optimiser la fonction objective non convexe et non différentiable. La convergence globale de l'algorithme est établie. L'étude de simulation et deux applications démontrent les performances supérieures de la méthode proposée en termes d'estimation et d'amélioration de l'identification de la structure éparse.

Epidemiological and Clinical Studies • Études épidémiologiques et cliniques

13:30-14:30, *C 2045*
Chair • Présidente: Selina Elvayn

Michael Agronah, Benjamin Bolker
Are Microbiome Studies Underpowered? Investigating Power in Differential Abundance Studies
Les études sur le microbiome manquent-elles de puissance? Étude de la puissance dans les études d'abondance différentielle

Identifying microbial taxa that exhibit differential abundance between treatment groups (control/treatment, healthy/diseased, etc.) is important for both basic and applied science. As in research more generally, microbiome studies must have good statistical power to detect taxa with substantially different abundance between treatments; low power leads to poor precision and mistakes via the winner's curse. Several studies have raised concerns about low power in microbiome studies. We analysed seven real case-control microbiome datasets and developed a novel method for simulating microbiome data. We present an innovative approach for estimating the statistical power to detect effects at the level of individual taxon as a function of effect size (fold change) and mean abundance. Our results suggest that many differential abundance studies are indeed underpowered, and illustrate how power varies with effect size and mean abundance.

L'identification des taxons microbiens qui présentent une abondance différentielle entre les groupes de traitement (contrôle/traitement, sain/malade, etc.) est importante tant pour la science fondamentale que pour la science appliquée. Comme dans la recherche en général, les études sur le microbiome doivent avoir une bonne puissance statistique pour détecter les taxons dont l'abondance est substantiellement différente entre les traitements; une faible puissance conduit à une faible précision et à des erreurs via la malédiction du vainqueur. Plusieurs études ont soulevé des préoccupations concernant la faible puissance des études sur le microbiome. Nous avons analysé sept ensembles de données microbiomiques cas-témoins réels et développé une nouvelle méthode de simulation des données microbiomiques. Nous présentons une approche innovatrice pour estimer la puissance statistique de détection des effets au niveau des taxons individuels en fonction de la taille de l'effet (changement de pli) et de l'abondance moyenne. Nos résultats suggèrent que de nombreuses études d'abondance différentielle sont en effet sous-puissantes et illustrent.

Jing Wang, Li Xing, Kyle Gardiner, Jinglan Qiu
Dose-response relationship for a skewed predictor containing lot of zeros
Relation dose-réponse pour un prédicteur asymétrique contenant beaucoup de zéros

Traditional regression models, predicated on the assumption of linearity between explanatory variables and quantitative outcomes, often fall short in the realm of epidemiology, where complex systems yield intricate, non-linear relationships. These non-linear associations can present challenges in understanding the effects of variables and predicting outcomes, especially in disease causality investigations. To address these challenges, various methods have been proposed, including categorizing exposure variables into quantiles, polynomial or fractional polynomial regression, spline modeling, and the application of logistic Box-Cox (LBC) transformations. However, handling excessive zero values for predictors remains a challenge in epidemiological analyses. In this study, we propose a modification to the LBC model to address the presence of many zero values in predictor while retaining the benefits of accurately modeling non-linear relationships and enhancing model interpretability. Using simulations, we compare our revised LBC model with traditional methods such as logistic regression, generalized additive models (GAM), fractional polynomial (FP) regression, and natural splines (NS). Subsequently, we apply the best-performing model to National Health and Nutrition Examination Survey (NHANES) data to explore the relationship between alcohol consumption and hypertension.

Les modèles de régression traditionnels, fondés sur l'hypothèse de la linéarité entre les variables explicatives et les résultats quantitatifs, sont souvent insuffisants dans le domaine de l'épidémiologie, où des systèmes complexes produisent des relations complexes et non linéaires. Ces associations non linéaires peuvent poser des problèmes pour comprendre les effets des variables et prédire les résultats, en particulier dans les enquêtes sur la causalité des maladies. Pour relever ces défis, diverses méthodes ont été proposées, notamment la catégorisation des variables d'exposition en quantiles, la régression polynomiale ou polynomiale fractionnaire, la modélisation spline et l'application de transformations logistiques Box-Cox (LBC). Cependant, le traitement des valeurs nulles excessives pour les prédicteurs reste un défi dans les analyses épidémiologiques. Dans cette étude, nous proposons une modification du modèle LBC pour traiter la présence de nombreuses valeurs nulles dans les prédicteurs tout en conservant les avantages d'une modélisation précise des relations non-linéaires et en améliorant l'interprétabilité du modèle. À l'aide de simulations, nous comparons notre modèle LBC modifié avec les méthodes traditionnelles telles que la régression logistique, les modèles additifs généralisés (GAM), la régression polynomiale fractionnée (FP) et les splines naturelles (NS). Ensuite, nous appliquons le modèle le plus performant aux données de la National Health and Nutrition Examination Survey (NHANES) afin d'explorer la relation entre la consommation d'alcool et l'hypertension.

Md Ashiqul Haque, Nathan C. Nickel, Maxime Turgeon, Lisa M. Lix
Model-based algorithms to ascertain smoking in administrative health data: a registry-based validation study
Algorithmes basés sur des modèles pour vérifier le tabagisme dans les données administratives sur la santé: une étude de validation basée sur un registre

We compared the validity of machine-learning model-based algorithms (MBA) for measuring smoking in Administrative Health Data (AHD) to that of simple rule-based algorithms (RBA). The study included 24,718 adults (≥18) in a clinical registry containing self-reported current-smoking data from 2017 to 2020 in Manitoba, Canada. RBA were defined using indicators of smoking status, particularly diagnosis codes for tobacco use and prescription medications such as varenicline. MBA used comorbid conditions and sociodemographic information along with smoking status indicators. MBA were based on random forest models. Validity measures included sensitivity, specificity, and positive predictive value (PPV), and their 95% confidence intervals (CI). RBA had 27.3% (CI: 24.2-30.7) sensitivity with 96.6% (CI: 96.1-97.0) specificity and 47.2% (CI: 42.9-51.5) PPV. MBA estimated 68.6% (CI: 65.1-71.9) sensitivity besides 76.3% (CI: 75.2-77.3) specificity and 24.3% (CI: 23.2-25.6) PPV. The study results indicate, balancing accurate smoker identification with the risk of false positives is crucial when choosing algorithmic approaches for measuring smoking in AHD.

Nous avons comparé la validité des algorithmes basés sur un modèle d'apprentissage automatique (MBA) pour mesurer le tabagisme dans les données administratives de santé (DAS) à celle des algorithmes basés sur des règles simples (RBA). L'étude a inclus 24 718 adultes (≥18 ans) dans un registre clinique contenant des données autodéclarées sur le tabagisme actuel de 2017 à 2020 au Manitoba, au Canada. Les RBA ont été définies à l'aide d'indicateurs du statut tabagique, en particulier les codes de diagnostic pour le tabagisme et les médicaments délivrés sur ordonnance tels que la varénicline. Les MBA ont utilisé des états de santé comorbides et des informations sociodémographiques ainsi que des indicateurs de l'usage du tabac. Les MBA étaient basées sur des modèles de forêt aléatoire. Les mesures de validité comprenaient la sensibilité, la spécificité et la valeur prédictive positive (VPP), ainsi que leurs intervalles de confiance (IC) à 95 %. La méthode RBA avait une sensibilité de 27,3 % (IC: 24,2-30,7), une spécificité de 96,6 % (IC: 96,1-97,0) et une VPP de 47,2 % (IC: 42,9-51,5). La méthode MBA a estimé une sensibilité de 68,6 % (IC: 65,1-71,9), une spécificité de 76,3 % (IC: 75,2-77,3) et une VPP de 24,3 % (IC: 23,2-25,6). Les résultats de l'étude indiquent qu'il est crucial de trouver un équilibre entre l'identification précise des fumeurs et le risque de faux positifs lors du choix des approches algorithmiques pour mesurer le tabagisme chez les personnes atteintes de la maladie d'Alzheimer.

Éloïse Soucy, Salah-Eddine Adlouni, Sophie Léger-Auffrey
Advanced machine learning and classification of ECG data
Apprentissage automatique avancé et classification des électrocardiogrammes

The classification of objects characterized by high-dimensional variables requires efficient and adapted machine learning approaches. The aim of this study is to explore dimension reduction techniques and classification algorithms for highly correlated variable spaces. The classification, using machine learning algorithms, is performed on 10,646 patient data from Chapman University and Shaoxing People's Hospital, characterized by 11 electrocardiogram (ECG) variables. Dimensionality reduction approaches are used to provide a more efficient representation of the data and preserve complex structures. The UMAP approach is considered and compared with conventional algorithms. Classification is then performed in the reduced-dimensional space by using support vector machine (SVM) algorithm of different kernels. The results show that the SVM algorithm with radial basis function kernel leads to the best results, with a performance of 83.47%.

La classification des objets caractérisés par des variables en dimension élevée, nécessite des approches d'apprentissage machine efficaces et adaptées. Le but de ma recherche est de faire de la classification, par l'entremise d'algorithmes d'apprentissage machine, de 10 646 données de patients provenant de Chapman University et Shaoxing People's Hospital et caractérisées par 11 variables captées par électrocardiogrammes (ECG). Des approches de réduction de dimensionnalité sont utilisées pour permettre d'avoir une représentation plus efficace des données et préserver les structures complexes. L'approche UMAP, basée sur une réduction non-linéaire de la dimension, est considérée. La classification des ECG est ensuite réalisée dans l'espace de dimension réduite par l'algorithme de séparateur à vaste marge (SVM) de différents noyaux. Les résultats montrent que l'algorithme SVM à noyau de fonction de base radial mène aux meilleurs résultats avec une performance de $83,47\%$ pour les pathologies étudiées.

Risk Assessment and Management • Évaluation et gestion des risques

13:30 - 14:30, *C 3053*
Chair • Présidente: Mathilde Dicaire-Cartier

Bartosz Glowacki
Ruin probability and rare event simulation
Probabilité de ruine et simulation d'événements rares

We are interested in the surplus process and the associated ruin probability. The distributions of the time of ruin, the surplus before the time of ruin and the deficit at the ruin are complicated and no analytical expressions exist. Gerber and Shiu (1998) analyzed their joint distribution by considering an expected discounted penalty function. They proved it is a solution of a certain renewal equation. In this paper we study the problem under some specific conditions for surplus model. By using Gerber-Shiu functions and renewal equations we obtain some explicit theoretical formulas. With their help we present algorithms for Monte Carlo simulation to estimate ruin probability and the moments of the deficit. We also present algorithms based on change of measures and importance sampling introduced in Asmussen and Albrecher (2010). Finally, simulation study is conducted to compare crude Monte Carlo simulations with importance sampling.

Nous nous intéressons au processus de surplus et à la probabilité de ruine associée. Les distributions du moment de la ruine, du surplus avant le moment de la ruine et du déficit à la ruine sont compliquées et il n'existe pas d'expressions analytiques. Gerber et Shiu (1998) ont analysé leur distribution conjointe en considérant une fonction de pénalité actualisée attendue. Ils ont prouvé qu'il s'agissait d'une solution d'une certaine équation de renouvellement. Dans cet article, nous étudions le problème sous certaines conditions spécifiques pour le modèle de surplus. En utilisant les fonctions de Gerber-Shiu et les équations de renouvellement, nous obtenons des formules théoriques explicites. Avec leur aide, nous présentons des algorithmes de simulation de Monte Carlo pour estimer la probabilité de ruine et les moments du déficit. Nous présentons également des algorithmes basés sur le changement de mesures et l'échantillonnage d'importance introduits dans Asmussen et Albrecher (2010). Une étude de simulation a aussi été menée pour comparer les simulations de Monte Carlo brutes avec l'échantillonnage d'importance.

Armin Mohammadioroojeh, Alexey Kuznetsov

Approximating generalized gamma convolutions and mixture of exponentials via multipoint Padé method

Approximation des convolutions gamma généralisées et des mélanges d'exponentielles via la méthode multipoint de Padé

We propose an efficient algorithm for approximating distributions of the generalized gamma convolutions and of mixtures of exponentials. In the first case the approximating distribution is of a finite sum of independent gamma random variables and in the second case it is given as a finite mixture of exponential distributions. Our method is based on approximating the Laplace transform of the distribution via multipoint Padé approximation. We will present examples demonstrating outstanding accuracy of this approximation and we will also discuss applications of this technique in actuarial science and in mathematical finance. This talk is based on joint work with Alexey Kuznetsov.

Nous proposons un algorithme efficace pour approximer les distributions des convolutions gamma généralisées et des mélanges d'exponentielles. Dans le premier cas, la distribution d'approximation est une somme finie de variables aléatoires indépendantes gamma, et dans le second cas, elle est donnée comme un mélange fini de distributions exponentielles. Notre méthode est basée sur l'approximation de la transformation de Laplace de la distribution via l'approximation multipoint de Padé. Nous présenterons des exemples démontrant une précision de cette approximation et discuterons également des applications de cette technique en science actuarielle et en finance mathématique. Cette présentation est basée sur un travail conjoint avec Alexey Kuznetsov.

Peiheng Gao, Ricardas Zitikis, Chen Yang, Ning Sun

NLP-based detection of systematic anomalies among the narratives of consumer complaints
Détection des anomalies systématiques par l'analyse basée sur le TALN des récits de plaintes de consommateurs

We develop an NLP-based procedure for detecting systematic nonmeritorious consumer complaints, simply called systematic anomalies, among complaint narratives. While classification algorithms are used to detect pronounced anomalies, in the case of smaller and frequent systematic anomalies, the algorithms may falter due to a variety of reasons, including technical ones as well as natural limitations of human analysts. Therefore, as the next step after classification, we convert the complaint narratives into quantitative data, which are then analyzed using an algorithm for detecting systematic anomalies. We illustrate the entire procedure using complaint narratives from the Consumer Complaint Database of the Consumer Financial Protection Bureau.

Nous développons une procédure basée sur le TALN (Traitement Automatique du Langage Naturel) pour détecter les plaintes systématiques et non fondées de consommateurs, simplement appelées anomalies systématiques, parmi les récits de plaintes. Alors que des algorithmes de classification sont utilisés pour détecter les anomalies prononcées, dans le cas d'anomalies systématiques plus petites et fréquentes, les algorithmes peuvent faiblir en raison de diverses causes, y compris des raisons techniques et des limites naturelles des analystes humains. Par conséquent, après la classification, nous convertissons les récits de plaintes en données quantitatives, qui sont ensuite analysées à l'aide d'un algorithme pour détecter les anomalies systématiques. Nous illustrons l'ensemble de la procédure en utilisant des récits de plaintes de la base de données des plaintes des consommateurs du Bureau de protection des consommateurs en matière financière.

Assane Kholle, Marie-Pier Côté

Grouping methods for multilevel categorical variables in a GLM
Méthodes de regroupement des variables catégorielles multiniveaux dans un GLM

In general insurance (auto or home), it often happens that some segmentation variables are categorical with many levels. We can think of the job held by the main driver or their region of residence. In practice, companies that use, for example, generalized linear models for pricing will often create groupings of levels of the variable based on the value of the estimated coefficient. Using test theory and a simulation study, it was shown that the likelihood ratio statistic, of which the reduced model is the grouping of similar coefficients estimated a posteriori, does not follow a chi-square distribution under the null hypothesis. Furthermore, the Mean Square Error on the parameter estimates with the real coefficients fixed in the simulation is not satisfactory. Other methods for grouping multilevel categorical variables in the literature, such as SMURF (cf. Devriendt et al. (2021)) and MAIDRR (cf. Henckaerts et al.(2022), have been implemented and compared based on Poisson deviance.

Pour la tarification en assurance générale (auto ou habitation), il arrive souvent que certaines variables de segmentation soient des variables catégorielles avec beaucoup de niveaux. On peut penser à l'emploi occupé par le conducteur principal ou à sa région de résidence. En pratique, les compagnies qui utilisent, par exemple, les modèles linéaires généralisés (GLM) pour la tarification vont souvent créer des regroupements de niveaux de la variable en se basant sur la valeur du coefficient (beta) estimé. Il serait possible, par exemple, de faire des tests à priori sur les regroupements, mais à posteriori (après avoir observé les betas), cela n'est pas adéquat. En utilisant la théorie des tests et avec une étude de simulation, on a montré que la statistique du rapport de vraisemblance dont le modèle réduit est le regroupement des coefficients similaires estimés à postériori, ne suit pas une distribution du chi-deux sous l'hypothèse nulle. De plus, l'erreur quadratique moyenne sur les coefficients évaluée avec les vrais paramètres fixés dans la simulation n'est pas satisfaisante. D'autres méthodes de regroupement des variables catégorielles multiniveaux dans la littérature, comme le SMURF (cf. Devriendt et al. (2021) et le MAIDRR (cf. Henckaerts et al. (2022) ont été aussi implémentées et leurs performances comparées sur la base de la déviance de Poisson.

In-person Posters • Affiches en personne

Brynn O'Connell, Brian Franczak

Variable Selection for the Classification of Data with Missing Values
Sélection de variables pour la classification de données avec des valeurs manquantes

We present a comprehensive exploration of explicit variable selection procedures for model-based classification. First, we dissect the intricacies of variable selection, setting the stage for an examination of the Variable Selection for Clustering and Classification (VSCC) approach. With a focus on enhancing classification accuracy and interpretability, we will unveil the details of VSCC, elucidating its performance in model-based classification applications. Next, we will investigate how this approach performs when applied to simulated and real data sets with missing values. Through meticulous evaluation and analysis, we will scrutinize the performance of this variable selection approach when applied to incomplete data sets. We conclude with a comprehensive discussion that sheds light on the implications of the results, offers valuable insights for future research directions, and suggest refinements for variable selection methodologies used in model-based classification applications.

Nous présentons une exploration complète des procédures de sélection explicite des variables pour la classification basée sur un modèle. Tout d'abord, nous examinons les subtilités de la sélection de variables, en préparant le terrain pour un examen de l'approche de la sélection de variables pour le regroupement et la classification (VSCC). En mettant l'accent sur l'amélioration de la précision de la classification et de l'interprétabilité, nous dévoilerons les détails de la VSCC, en élucidant sa performance dans les applications de classification basées sur des modèles. Ensuite, nous étudierons les performances de cette approche lorsqu'elle est appliquée à des ensembles de données simulées et réelles avec des valeurs manquantes. En menant une évaluation et une analyse méticuleuse, nous examinerons la performance de cette approche de sélection des variables lorsqu'elle est appliquée à des ensembles de données incomplets. Nous concluons par une discussion approfondie qui met en lumière les implications des résultats, offre des indications précieuses sur les orientations futures de la recherche et suggère des améliorations pour les méthodologies de sélection des variables utilisées dans les applications de classification basées sur des modèles.

Yutong Lu, Yan Yi Li

A statistical framework to integrate large chemical language models for molecular property analysis

Un cadre statistique pour intégrer de grands modèles de langage chimique pour l'analyse des propriétés moléculaires

Ensemble learning is a method that combines several individual models to enhance overall prediction performance. We propose a novel statistical framework that integrates outputs from multiple models trained on molecular structural data to predict molecular properties. Firstly, we train three unique large chemical language models on a shared dataset for preliminary predictions. Then, we employ a second-level model that takes the first-level results as input to obtain final predictions. Alternatively, we distribute subsets of the dataset to each model using a probabilistic assignment mechanism and then aggregate the subset predictions based on the probability weights. Comparative analyses against traditional deep learning models demonstrate that our statistical fusion method yields superior performance, suggesting its capacity to harness the unique strengths and diverse techniques of various models for comprehensive data interpretation.

L'apprentissage d'ensemble est une méthode qui combine plusieurs modèles individuels afin d'améliorer la performance globale des prédictions. Nous proposons un nouveau cadre statistique qui intègre les résultats de plusieurs modèles formés sur des données structurelles moléculaires afin de prédire les propriétés moléculaires. Tout d'abord, nous entraînons trois grands modèles uniques de langage chimique sur un ensemble de données partagé pour les prédictions préliminaires. Ensuite, nous utilisons un modèle de deuxième niveau qui prend les résultats du premier niveau comme entrée pour obtenir des prédictions finales. Alternativement, nous distribuons des sous-ensembles de l'ensemble de données à chaque modèle à l'aide d'un mécanisme d'affectation probabiliste et agrégeons ensuite les prédictions des sous-ensembles sur la base des poids de probabilité. Des analyses comparatives avec des modèles d'apprentissage profond traditionnels démontrent que notre méthode de fusion statistique est plus performante, ce qui suggère sa capacité à exploiter les forces uniques et les techniques diverses des différents modèles pour une interprétation complète des données.

Ziqian Zhuang, Wei Xu

Joint Modeling of Complex Multivariate Adverse Events in Clinical Trial Data
Modélisation conjointe des événements indésirables multivariés complexes dans les données d'essais cliniques

Adverse events (AE) are harmful outcomes during medical care. The severity and frequency of these events serve as study endpoints in clinical trials, crucial for evaluating treatment safety. Patients may encounter multiple adverse events concurrently, and the recorded data exhibit diverse structures due to varying durations and characteristics of both short-term and long-term AEs. Moreover, AE severity may fluctuate over time due to disease progression or treatment response. Most current analyses focus solely on a single AE, neglecting severity information and failing to distinguish adequately between short-term and long-term AEs. In response, we propose an efficient joint model to assess treatment effects on multiple AE occurrences. This model comprehensively considers AE severities and correlations while effectively addressing structural differences between short-term and long-term AEs. Through simulation studies, this method has demonstrated high accuracy in parameter estimation.

Les événements indésirables (EI) surviennent lors des soins médicaux. Leur gravité et fréquence sont des critères d'évaluation essentiels dans les essais cliniques pour la sécurité des traitements. Les patients peuvent avoir plusieurs EI simultanément, avec des données enregistrées de structures diverses en raison de la durée et des caractéristiques variables des EI à court et long terme. La gravité des EI peut fluctuer avec la progression de la maladie ou la réponse au traitement. La plupart des analyses se concentrent sur un seul EI, négligeant la gravité et sans distinguer les EI à court et long terme. Nous proposons un modèle conjoint pour évaluer les effets du traitement sur plusieurs EI. Ce modèle tient compte des gravités et corrélations des EI tout en traitant les différences structurelles entre les EI. Les études de simulation ont démontré une grande précision dans l'estimation des paramètres.

Lina Li, Kyle Gardiner, Jinglan Qiu, Xuekui Zhang, Li Xing

TSMA: A Two-stage Sampling Aggregation Framework to Construct Prediction Models for Unbalanced Case-Control Disease Data from Electronic Medical Record and Genomics (eMERGE) Network

TSMA: Un cadre d'agrégation d'échantillonnage en deux étapes pour construire des modèles de prédiction pour des données cas-témoins sur les maladies non équilibrées provenant du réseau Electronic Medical Record and Genomics (eMERGE)

The general adoption of Electric Medical Records (EMR) into the health system has become a trend in recent years. Rich clinical information from EMR combined with genomic data presents researchers with an unprecedented opportunity to uncover associations of genomics to human disease and propel advancements in precision medicine. Our work is to tackle the challenge caused by unbalanced EMR Genome-wide Association Study (GWAS) data. We build a novel Two-stage Sampling Aggregation (TSMA) framework, which incorporates the bagging strategy in the ensemble machine learning with resampling techniques to reduce the overrepresentation from the majority class of the unbalanced data. We demonstrate the superior performance of the prediction models built based on the TSMA framework through data application to the three extremely unbalanced cohorts from the Electronic Medical Records and Genomics (eMERGE) Network.

L'adoption généralisée des dossiers médicaux électroniques (DME) dans le système de santé est devenue une tendance ces dernières années. Les riches informations cliniques des DME combinées aux données génomiques offrent aux chercheurs une occasion sans précédent de découvrir les associations entre la génomique et les maladies humaines et de faire progresser la médecine de précision. Notre travail consiste à relever le défi posé par les données déséquilibrées de l'étude d'association à l'échelle du génome (GWAS) des DME. Nous construisons un nouveau cadre d'agrégation d'échantillonnage en deux étapes (TSMA), qui incorpore la stratégie de bagging dans l'apprentissage automatique d'ensemble avec des techniques de rééchantillonnage pour réduire la surreprésentation de la classe majoritaire des données déséquilibrées. Nous démontrons les performances supérieures des modèles de prédiction construits sur la base du cadre TSMA en appliquant les données aux trois cohortes extrêmement déséquilibrées du réseau eMERGE (Electronic Medical Records and Genomics).

Zixuan Yang, Douglas G. Woolford

Modelling the Forecasting Error Distributions of Several Fire-Weather Variables
Modélisation de la distribution des erreurs dans la prévision de plusieurs variables forêt-météo

Local conditions play a crucial role in wildfire occurrence prediction modelling. Key variables include common weather variables as well as output from the Canadian Fire Weather Index System, such as the Fine Fuel Moisture Code. Our work characterizes the one to four days ahead forecasting error distributions of several fire-weather variables using historically observed weather and forecasts from a region in northwestern Ontario, Canada. It is important to understand these forecasting error distributions because those fire-weather variables are to be used to predict fire occurrences. Ignoring the forecasting errors can lead to biased predictions. It can be shown that although the forecasting error distributions seem to be bell shaped, they are not normally distributed because they have heavy tails. By comparing different methods for characterizing forecasting error, we conclude that a finite normal mixture distribution is most appropriate.

Les conditions locales jouent un rôle crucial dans la modélisation de la prévision des incendies de forêt. Les variables clés comprennent les variables météorologiques courantes ainsi que les résultats de la Méthode canadienne de l'indice Forêt-Météo, tels que l'Indice du combustible léger. Notre travail caractérise les distributions des erreurs de prévision d'un à quatre jours à l'avance de plusieurs variables forêt-météo en utilisant les observations météorologiques historiques et les prévisions d'une région du nord-ouest de l'Ontario. Il est important de comprendre ces distributions des erreurs de prévision, car ces variables forêt-météo sont utilisées pour prévoir les incendies. Ignorer les erreurs de prévision peut conduire à des prévisions biaisées. Il peut être démontré que, bien que les distributions des erreurs de prévision semblent avoir une forme de cloche, elles ne sont pas distribuées normalement à cause de leurs queues lourdes. En comparant différentes méthodes de caractérisation des erreurs de prévision, nous concluons qu'une distribution de mélange normal fini est la plus appropriée.

Jesse Ghashti, Jeffrey Andrews, John Thompson
A bootstrap augmented k-means algorithm for fuzzy partitions
Un algorithme bootstrap augmenté de k-moyennes pour les partitions floues

Fuzzy c-means (FCM) algorithms partition data into probabilistic cluster assignments by selecting a so-called fuzzy parameter. Although FCM is built on the efficient and straightforward framework of k-means clustering, its drawbacks include the required a priori knowledge of fuzziness to select the optimal fuzzy parameter, potential local optima entrapment, and sensitivity to initial cluster centres. In this talk, we present a bootstrap augmented k-means clustering algorithm that incorporates bootstraps into the loss function optimization scheme, allowing for probabilistic cluster assignments without tuning parameters and reducing the impact of random initializations with iterated refinement of cluster centres. We show that the proposed algorithm more accurately models the preordained uncertainty of cluster allocations for simulated data. We demonstrate, under satisfied model assumptions, that this augmented algorithm will mathematically match or exceed the performance of FCM variants.

Les algorithmes c-moyennes flous (FCM) répartissent les données en groupes probabilistes en sélectionnant un paramètre de flou. Bien que le FCM soit construit sur le cadre efficace et simple du regroupement k-moyennes, ses inconvénients comprennent la connaissance a priori du niveau de flou nécessaire pour sélectionner le paramètre de flou optimal, la prise potentielle d'optima locaux et la sensibilité aux centres de regroupement initiaux. Dans cet exposé, nous présentons un algorithme de regroupement k-moyennes augmenté par bootstrap qui incorpore ce dernier dans le schéma d'optimisation de la fonction de perte, ce qui permet des attributions probabilistes de regroupements sans paramètres de réglage et de réduire l'impact des initialisations aléatoires grâce à l'affinement itéré des centres des regroupements. Nous montrons que l'algorithme proposé modélise plus précisément l'incertitude préétablie des affectations de regroupements pour des données simulées. Nous démontrons, sous des hypothèses de modèle satisfaites, que cet algorithme augmenté atteindra ou dépassera mathématiquement les performances des variantes FCM. Les algorithmes c-moyennes flous (FCM) répartissent les données en groupes probabilistes en sélectionnant un paramètre de flou. Bien que le FCM soit construit sur le cadre efficace et simple du regroupement k-moyennes, ses inconvénients comprennent la connaissance a priori du niveau de flou nécessaire pour sélectionner le paramètre de flou optimal, la prise potentielle d'optima locaux et la sensibilité aux centres de regroupement initiaux.

Tracy Qian, Max Piasevoli, Michael Guerzhoy

Automatic Model Selection using Wasserstein Generative Adversarial Networks
Sélection automatique de modèles à l'aide de réseaux antagonistes génératifs de Wasserstein

We propose a novel approach for automatic model selection for hierarchical models using Wasserstein Generative Adversarial Networks (WGANs). Model checking and selection can be performed by graphically comparing fake data generated by the proposed model to the actual data. The aim is to select a model that generates fake data with a similar distribution to that of the actual data. The critic component of a WGAN is trained to discriminate data generated by the generator component from the real data. In our propose approach, we use the critic components of WGANs trained on data simulated from candidate models. If the critic component of a WGAN for a candidate model cannot successfully discriminate between synthetic data generated from that model and the real data, that indicates better model fit. We describe an algorithm for model selection using this intuition. We demonstrate that our approach can be used to select appropriate models for synthetic and real social science datasets.

Nous proposons une nouvelle approche pour la sélection automatique de modèles hiérarchiques à l'aide de réseaux antagonistes génératifs de Wasserstein (WGAN). La vérification et la sélection des modèles peuvent être effectuées en comparant graphiquement les données fictives générées par le modèle proposé avec les données réelles. L'objectif est de sélectionner un modèle qui génère des données fictives dont la distribution est similaire à celle des données réelles. La composante critique d'un WGAN est entraînée à distinguer les données générées par la composante générateur des données réelles. Dans l'approche que nous proposons, nous utilisons les composantes critiques des WGANs entraînés sur des données simulées à partir de modèles candidats. Si la composante critique d'un WGAN pour un modèle candidat ne peut pas distinguer avec succès entre les données synthétiques générées par ce modèle et les données réelles, cela indique une meilleure adéquation du modèle. En utilisant cette intuition, nous décrivons un algorithme de sélection de modèle. Nous démontrons que notre approche peut être utilisée pour sélectionner des modèles appropriés pour des ensembles de données synthétiques et réelles dans le domaine des sciences sociales.

Benjamin Frizzell

Optimal Experimental Design using Simulated Annealing
Conception optimale des expériences à l'aide du recuit simulé

Optimal experimental designs seek to optimize an experiment according to some prescribed optimality criterion. Optimal experimental designs are especially useful under constraints due to practical feasibility, ethical issues, or budget limits. Optimal design problems are typically solved using disciplined convex programming (DCP) methods, but DCP may provide inaccurate solutions when the number of experiments is exactly specified (ie. the solution is discretized), and fail when solving for high-dimensional, nonlinear models. Alternatively, this study investigates simulated annealing, a popular probabilistic algorithm used for many other optimization problems. In many cases, simulated annealing can provide a strong approximation to the solution provided by DCP and provides solutions where DCP fails. The simulated annealing method can also be configured to account for exact optimal designs and is overall an effective algorithm for optimal experimental design.

Les plans optimaux des expériences visent à optimiser une expérience en fonction d'un critère d'optimalité donné. Les plans d'expérience optimaux sont particulièrement utiles en cas de contraintes liées à la faisabilité pratique, aux questions éthiques ou aux limites budgétaires. Les problèmes de conception optimale sont généralement résolus à l'aide de méthodes de programmation convexe disciplinée (PCD), mais la PCD peut fournir des solutions inexactes lorsque le nombre d'expériences est exactement spécifié (c'est-à-dire que la solution est discrétisée), et échouer lors de la résolution de modèles non linéaires à haute dimension. Cette étude s'intéresse également au recuit simulé, un algorithme probabiliste très répandu, utilisé pour de nombreux autres problèmes d'optimisation. Dans de nombreux cas, le recuit simulé peut fournir une forte approximation de la solution fournie par la méthode PCD et fournit des solutions là où la méthode PCD échoue. La méthode du recuit simulé peut également être configurée pour prendre en compte les plans optimaux exacts et constitue globalement un algorithme efficace pour les plans d'expériences optimaux.

Haochen Ning, Ian Boyes, Michael Rott, Ibrahim Numanagić, Li Xing, Xuekui Zhang
IIMI: Advancements of Novel Machine-Learning Toolset for Plant Virus Detection
IIMI: Progrès d'un nouvel ensemble d'outils d'apprentissage automatique pour la détection des virus des plantes

Plant virus detection is vital for safeguarding Canada's agriculture, mitigating economic impacts on farmers, and ensuring plant health. Our team developed IIMI, a machine-learning algorithm tailored for precise plant virus identification. IIMI automates detection, reducing manual input and incorporating a mappability profile to minimize false positives from genomic similarities. Our work aims to enhance IIMI as a robust tool for broad agricultural contexts through three initiatives: introducing a known virus as a positive control counters biases from sequencing platforms or technician variability, enhancing model adaptability and accuracy. Employing machine learning to address discrepancies in segment-level labelling of training data, optimizing the training process. Expanding the mappability profile to include diverse host genomes enables personalized profiling and elevates virus detection accuracy.

La détection des virus des plantes est essentielle pour protéger l'agriculture canadienne, atténuer les conséquences économiques pour les agriculteurs et garantir la santé des plantes. Notre équipe a mis au point IIMI, un algorithme d'apprentissage automatique conçu pour l'identification précise des virus des plantes. IIMI automatise la détection, en réduisant la saisie manuelle et en incorporant un profil de la 'mappabilité' pour minimiser les faux positifs dus aux similitudes génomiques. Notre travail vise à améliorer l'IIMI en tant qu'outil robuste pour de vastes contextes agricoles grâce à trois initiatives: l'introduction d'un virus connu en tant que contrôle positif permet de contrer les biais des plateformes de séquençage ou la variabilité des techniciens, et d'améliorer l'adaptabilité et la précision du modèle. L'utilisation de l'apprentissage automatique pour traiter les divergences dans l'étiquetage au niveau des segments des données d'entraînement, optimisant ainsi le processus d'entraînement. L'élargissement du profil de 'mappabilité' à divers génomes d'hôtes permet d'établir des profils personnalisés et d'améliorer la précision de la détection des virus.

Sarah Organ, Hong Gu, Toby Kenney

Vertex cover matroid variable selection for controlling the false discovery rate and improving power with correlated predictors

Sélection de variables dans un matroïde de couverture de sommet pour contrôler le taux de fausse découverte et améliorer la puissance avec des prédicteurs corrélés

Variable selection methods struggle with controlling for false discovery (FDR) while maintaining a high power when the variables are correlated. To address this problem, we rethink variable selection to allow for the selection of surrogate pairs of variables. By selecting surrogate pairs, we are considering that if the pairs of variables are highly correlated, we do not know which of the variables is a true variable, therefore choosing either variable is appropriate. This approach allows us to overcome the problems multicollinearity introduces to standard variable selection. One of the challenges we overcome in this method is how to measure true and false positive rates for methods that select choices of surrogates, rather than picking a single variable. By utilizing our chosen algorithm to measure true and false positives, simulations show our two-stage method maintains an FDR less than 5% while achieving higher power than existing methods when the correlation is greater than 0.5.

Les méthodes de sélection des variables ont du mal à tenir compte des taux de fausses découvertes (FDR) tout en conservant une puissance élevée lorsque les variables sont corrélées. Pour résoudre ce problème, nous reconsidérons la sélection des variables afin de permettre la sélection de paires de substitut de variables. En sélectionnant des paires de substituts, nous considérons que si les paires de variables sont fortement corrélées, nous ne savons pas laquelle des variables est la vraie, et que le choix entre l'une et l'autre est donc approprié. Cette approche nous permet de surmonter les problèmes que la multicolinéarité introduit dans la sélection habituelle des variables. L'un des défis que nous avons relevés avec cette méthode est de savoir comment mesurer les taux de vrais et de faux positifs pour les méthodes qui sélectionnent des choix de substituts, plutôt que de choisir une seule variable. En utilisant l'algorithme que nous avons choisi pour mesurer les vrais et les faux positifs, les simulations démontrent que notre méthode en deux étapes maintient un FDR inférieur à 5% tout en atteignant une puissance supérieure à celle des méthodes existantes lorsque la corrélation est supérieure à 0,5.

Yasaman Shahhosseini, Michelle Miranda, Farouk Nathoo, Cedric Beaulac
Spatiotemporal fractal based analysis of fMRI time series
L'analyse des séries temporelles de l'IRMf basée sur les fractales spatio-temporelles

In this talk, we consider some inference problems about the drift parameter in generalized mean-reverting processes with possible change-points. We generalize the results in recent literature in five ways. First, as opposed to the pre-existing results, the dimensions of the drift parameter are considered unknown. Second, we weaken some assumptions underlying the asymptotic properties of some estimators of the drift parameter. Third, we derive an asymptotic test for detecting a change-point and parameter dimensions. Fourth, we establish the asymptotic power of the proposed test and the distributional risk of the proposed estimators as well as their relative efficiency in the context of the unknown dimension of the parameters. Fifth, we prove that the proposed methods improve the goodness-of-fit and the predictive accuracy. Finally, we present the simulation results which corroborate the theoretical findings and we analyze a financial market data set.

Dans cet exposé, nous étudions des problèmes d'inférence concernant le paramètre de dérive d'un processus de retour à la moyenne généralisé avec des points de rupture éventuels. Nous généralisons de récent résultats de la littérature de cinq façons. Premièrement, contrairement aux résultats existants, dans cette approche, les dimensions du paramètre de dérive sont supposées inconnues. Deuxièmement, nous affaiblissons certains présupposés sous-jacents des propriétés asymptotiques de certains estimateurs du paramètre de dérive. Troisièmement, nous établissons un test asymptotique pour détecter un point de rupture et les dimensions de paramètre. Quatrièmement, nous établissons la puissance asymptotique du test établi et le risque distributionnel asymptotique des estimateurs proposés ainsi que leur efficacité relative dans le contexte où les dimensions du paramètre sont inconnues. Cinquièmement, nous prouvons que les méthodes proposées améliorent la qualité d'ajustement et la précision prédictive. Enfin, nous présentons les résultats de simulation qui corroborent les résultats théoriques et nous analysons les données des marchés financiers.

Simon Maltby, Kyran Cupido
Spatial Analysis of the Risk Factors for Covid-19
Analyse spatiale des facteurs de risque de la Covid-19

While the impact of Covid-19 has been well studied at the provincial and national levels, we still don't understand how the severity of the impact of the pandemic may vary across space. The primary objective of this research project was to detect patterns in the spatial distribution of the underlying risk factors for Covid-19 across the census divisions of Canada. After identifying the risk factors of interest, a spatial weights matrix and Moran's I coefficient were used to determine hot spots and cold spots for these variables across Canada. Then, principal components and model-based clustering were employed to pool together data and identify which regions were closest to each other in terms of risk for Covid-19. The defining characteristics of each cluster and how their frequency varied across Canada were determined. The findings of this study should be compared to how the pandemic affected different regions once data becomes available at the census division level.

Bien que l'impact de la Covid-19 ait été étudié aux niveaux provincial et national, nous ne comprenons toujours pas comment la gravité de l'impact de la pandémie peut varier dans l'espace. L'objectif principal de ce projet de recherche était de détecter des tendances dans la distribution spatiale des facteurs de risque sous-jacents de la Covid-19 à travers les divisions de recensement du Canada. Après avoir identifié les facteurs de risque d'intérêt, une matrice de pondérations spatiales et le coefficient de Moran ont été utilisés pour déterminer les zones chaudes et froides pour ces variables à travers le Canada. Ensuite, des composantes principales et des regroupements basés sur des modèles ont été utilisés pour regrouper les données et identifier quelles régions étaient les plus proches les unes des autres en termes de risque pour la Covid-19. Les caractéristiques déterminantes de chaque groupe et comment leur fréquence variait à travers le Canada ont été déterminées. Les résultats de cette étude devraient être comparés à la manière dont la pandémie a affecté différentes régions une fois que les données seront disponibles au niveau de la division de recensement.

Pranath Pussella, Tianyu Guan, Robert Nguyen
Simulation for Cricket: A Machine Learning Approach
Simulation pour le cricket: une approche d'apprentissage automatique

Cricket is the second most popular sport in the world with a significant presence in Commonwealth countries. Despite its popularity, cricket is underrepresented in the literature, especially in the domain of simulation. Simulation in cricket is challenging because of its complexity, dynamic nature, and data scarcity. In this research, we develop a simulation mechanism for cricket using machine learning techniques. The construction of the simulator is based on the availability of a detailed dataset from Cricket Australia. We employ machine learning to predict the outcome of a delivery, the core element of gameplay, which can further be utilized for scorecard generation and match simulations. Our simulator's potential is demonstrated by employing it to determine the optimal batting position of a given batter in a team in Twenty20 cricket. Additionally, we develop an interactive web platform to enable direct interaction with the simulator and the tool for optimizing batting positions.

Le cricket est le deuxième sport le plus populaire du monde avec une présence significative dans les pays du Commonwealth. Malgré sa popularité, le cricket est sous-représenté dans la littérature, notamment dans le domaine de la simulation. La simulation du cricket est difficile en raison de sa complexité, de sa nature dynamique et de la rareté des données. Dans cette recherche, nous développons un mécanisme de simulation pour le cricket en utilisant des techniques d'apprentissage automatique. La construction du simulateur est basée sur la disponibilité d'un ensemble de données détaillé de Cricket Australia. Nous utilisons l'apprentissage automatique pour prédire le résultat d'un lancer, l'élément central du jeu, qui peut ensuite être utilisé pour générer des feuilles de pointage et simuler des matchs. Le potentiel de notre simulateur est démontré en l'utilisant pour déterminer la position de batte optimale d'un batteur donné dans une équipe de cricket Twenty20. De plus, nous développons une plateforme web interactive pour permettre une interaction directe avec le simulateur et l'outil d'optimisation des positions de batte.

Juan Liyau, You Liang, Na Yu, Aleksandar Popovic, Xun Zhou, Keanu Uchida, Tomasz Tkaczyk, Neeru Gupta, Yeni Yucel

Intuitive Segmentation for Hyperspectral Fluorescence Imaging in Ophthalmology: An Innovative Machine Learning Tool

Segmentation intuitive pour l'imagerie hyperspectrale de fluorescence en ophtalmologie: un outil innovant d'apprentissage automatique

Fluorescence hyperspectral imaging (FHSI) is an essential tool in diagnostic pathology and biomedical research. However, visualizing and analyzing high-dimensional FHSI images remains a challenge. We develop an innovative and open-source desktop application for the visualization and analysis of FHSI images of eye tissue sections. This application offers a suit of functionalities: data preprocessing tools such as normalization, denoising, and superpixel generation; visualization tools such as 2D and 3D spectral-based interactive exploration, region of interest (ROI) selection, average spectral curve calculation and demonstration, and initial identification of endmember signatures; segmentation tools such as Spectral Information Divergence Spectral Angle Mapper (SIDSAM) with optional unmixing and Fuzzy C-means (FCM) clustering; and tissue boundary detection tool such as Sobel edge detector. This application holds significant promise for enhancing the diagnosis of various eye diseases.

L'imagerie hyperspectrale de fluorescence (FHSI) est un outil essentiel en pathologie diagnostique et en recherche biomédicale. Toutefois, la visualisation et l'analyse des images FHSI de haute dimension restent un défi. Nous développons une application de bureau innovante et open source pour la visualisation et l'analyse des images FHSI de sections de tissus oculaires. Cette application offre une série de fonctionnalités: des outils de prétraitement des données tels que la normalisation, le débruitage et la génération de superpixels ; des outils de visualisation tels que l'exploration interactive sur les spectres 2D et 3D, la sélection de la région d'intérêt (ROI), le calcul et la démonstration de la courbe spectrale moyenne, et l'identification initiale des signatures de membres terminaux ; des outils de segmentation tels que le Spectral Information Divergence Spectral Angle Mapper (SIDSAM) avec option de démixage et de regroupement Fuzzy C-means (FCM) ; et un outil de détection des limites des tissus tel que le détecteur de contours Sobel. Cette application présente un potentiel significatif pour améliorer le diagnostic de diverses maladies oculaires.

Yu Shi, Wei Xu, Pingzhao Hu

A Deep Learning-Driven Out of Distribution Approach for Predicting Patient-Specific Cancer Dependency Maps

Une approche hors distribution basée sur l'apprentissage profond pour prédire des cartes de dépendance du cancer spécifiques aux patients

Cancer dependency maps are essential for identifying genes critical to cancer cells growth. Although core biological processes are preserved, significant discrepancies in the distribution between cancer cell line (CCL) models and patient-derived data pose challenges to the direct application of CCL insights in clinical settings. To address this, we introduce an unsupervised domain adaptation algorithm to statistically align feature distributions across distinct data domains. Trained on labeled CCL data, our model achieved a Pearson correlation coefficient of 0.8349 in unseen CCL data. Further validation using unlabeled patient data included correlating predicted dependency scores with clinical characteristics, achieving an Area Under the Curve (AUC) of 0.9489 in predicting the status of ER+/HER2+ patients. This approach not only precisely predicts cancer dependency for patient-specific tumors but also signifies a promising advancement in the generalization of out-of-distribution data.

Les cartes de dépendance du cancer sont essentielles pour l'identification des gènes cruciaux à la croissance des cellules cancéreuses. Bien que les processus biologiques fondamentaux soient préservés, des écarts significatifs dans la distribution entre les modèles de lignées cellulaires cancéreuses (LCC) et les données dérivées des patients posent des défis à l'application directe des connaissances des LCC en milieu clinique. Pour remédier à cela, nous introduisons un algorithme d'adaptation de domaine non supervisé pour aligner statistiquement les distributions des caractéristiques à travers différents domaines de données. Entraîné sur des données LCC étiquetées, notre modèle a atteint un coefficient de corrélation de Pearson de 0,8349 avec des données LCC invisibles. Une validation ultérieure en utilisant des données de patients non étiquetées comprenait la corrélation des scores de dépendance prédits avec les caractéristiques cliniques, atteignant une aire sous la courbe (AUC) de 0,9489 pour prédire l'état des patients ER+/HER2+. Cette approche prédit non seulement avec précision la dépendance au cancer pour des tumeurs spécifiques aux patients, mais représente également une avancée prometteuse dans la généralisation de données hors distribution.

Feifan Xiang, Divya Sharma, Osvaldo Espin-Garcia

Enhancing Osteoarthritis Progression Prediction with LSTM: Leveraging Bilateral Knee Data and Prospects for Multimodal Integration in Osteoarthritis Initiative

Amélioration de la prédiction de la progression de l'arthrose avec la MLCT: exploitation des données bilatérales du genou et perspectives d'intégration multimodale dans l'Initiative sur l'arthrose

Osteoarthritis (OA), as one of the most prevalent types of arthritis, represents a major public health concern both worldwide and in Canada, and is expected to escalate with the aging population. Our project utilizes Long Short-Term Memory (LSTM) neural networks for predicting OA progression in the Osteoarthritis Initiative dataset. The gated structure of LSTM efficiently retains or discards data over time, crucial for chronic conditions like OA with complex temporal patterns. This advantage, alongside overcoming the vanishing gradient problem, places LSTMs favoured against other ML models and traditional statistical methods. Our LSTM model, trained on over 8 years of clinical visits, achieved an 85% accuracy, with class imbalance addressed through downsampling and weighted loss functions. Prospective future work will consider separate knee analyses and multimodal data integration, enhancing predictive accuracy and deeper understanding of OA trajectories.

L'arthrose (OA), en tant que l'un des types d'arthrite les plus courants, constitue un problème majeur de santé publique à l'échelle mondiale et au Canada. Elle est susceptible de s'aggraver avec le vieillissement de la population. Notre projet utilise des réseaux neuronaux à mémoire à long et court terme (MLCT) pour prédire la progression de l'OA à partir des données de l'Initiative sur l'Arthrose. La structure à portes des MLCT permet de retenir ou d'éliminer efficacement les informations au fil du temps, ce qui est crucial pour les conditions chroniques telles que l'OA qui présentent des motifs temporels complexes. Cet avantage, conjugué à la résolution du problème des gradients qui disparaissent, confère aux MLCT un avantage par rapport aux autres modèles d'apprentissage machine et méthodes statistiques traditionnelles. Notre modèle MLCT, formé sur plus de 8 ans de données de visites cliniques, a atteint une précision de 85%, en traitant le déséquilibre des classes par sous-échantillonnage et des fonctions de perte pondérées. Les travaux futurs envisagent des analyses distinctes pour chaque genou et une intégration de données multimodales, dans le but d'améliorer la précision prédictive et de mieux comprendre les trajectoires de l'OA.

Yuhang Ou, Xikui Wang

Actuarial study and statistical analysis of flood insurance claims in Canada
Étude actuarielle et analyse statistique des sinistres d'assurance contre les inondations au Canada

In our actuarial study on flood insurance claims in Canada, we leverage regression analysis to identify key factors influencing claim frequencies and severities, such as geographical location and climatic conditions. Time series analysis is employed to detect patterns and forecast future trends in flood insurance claims, taking into account seasonal effects and long-term changes possibly related to climate change and urbanization. These statistical methodologies enable a nuanced understanding and prediction of flood-related risks, aiding in the refinement of insurance pricing strategies and risk management practices. By integrating these advanced analytical techniques, the study offers crucial insights for improving the sustainability of the flood insurance market and enhancing community resilience against flood risks.

Dans notre étude actuarielle sur les sinistres d'assurance contre les inondations au Canada, nous exploitons l'analyse de régression pour identifier les facteurs clés influençant les fréquences et les gravités des sinistres, tels que la localisation géographique et les conditions climatiques. L'analyse de séries chronologiques est utilisée pour détecter les tendances et prévoir les tendances futures des sinistres d'assurance contre les inondations, en tenant compte des effets saisonniers et des changements à long terme potentiellement liés au changement climatique et à l'urbanisation. Ces méthodologies statistiques permettent une compréhension nuancée et une prédiction des risques liés aux inondations, contribuant à affiner les stratégies de tarification des assurances et les pratiques de gestion des risques. En intégrant ces techniques analytiques avancées, l'étude offre des informations cruciales pour améliorer la durabilité du marché de l'assurance contre les inondations et renforcer la résilience des communautés face aux risques d'inondation.

Ankita Shelke, Erfanul Hoque

Optimizing Canada's Inflation: A Novel Approach Integrating Machine Learning and Deep Learning Techniques

Optimisation de l'inflation au Canada: une nouvelle approche intégrant les techniques d'apprentissage automatique et d'apprentissage profond

Inflation is the economic phenomenon marked by the rate of increase in prices over time, affecting the cost of living and the financial stability of a nation. It must be effectively controlled for the economic well-being of any country. In the Canadian context, very few researchers utilized advanced machine learning (ML) and deep learning (DL) tools and other macroeconomic variables to forecast inflation. To address this gap, we propose a novel approach that combines the strengths of dynamic regression, ML and DL techniques to provide more accurate inflation forecasts by capturing complex patterns and inflation dynamics. Various macroeconomic indicators such as unemployment rate, interest rates, oil prices, money supply, etc. are used to forecast Canada's inflation. The results demonstrate that the proposed approach outperforms traditional forecasting models. These findings provide useful insights to policymakers and economists, in the pursuit of maintaining Canada's inflation at 2%.

L'inflation est un phénomène économique caractérisé par une augmentation des prix au fil du temps, ce qui entraîne une hausse du coût de la vie et perturbe la stabilité financière d'un pays. Elle doit être contrôlée de manière efficace pour le bien-être économique de tout pays. Dans le contexte canadien, très peu de chercheurs ont utilisé des outils avancés d'apprentissage automatique (ML) et d'apprentissage profond (DL) et d'autres variables macroéconomiques pour prévoir l'inflation. Pour combler cette lacune, nous proposons une nouvelle approche qui combine les forces de la régression dynamique, des techniques de ML et de DL pour fournir des prévisions d'inflation plus précises en capturant les motifs complexes et la dynamique de l'inflation. Divers indicateurs macroéconomiques tels que le taux de chômage, les taux d'intérêt, le prix du pétrole, la masse monétaire, etc. sont utilisés pour prévoir l'inflation au Canada. Les résultats démontrent que l'approche proposée surpasse les modèles de prévision traditionnels. Ces résultats fournissent des informations utiles aux décideurs politiques et aux économistes dans la poursuite du maintien de l'inflation du Canada à 2%.

Adrian Neumann, Stuart Dovey
Predicting Real Estate Prices in Edmonton Alberta
Prédiction des prix immobiliers à Edmonton, Alberta

This project will aim to investigate the determining factors of real estate prices in Edmonton, Alberta. As real estate remains a significant investment, this analysis aims to offer insights for prospective buyers navigating the surging Edmonton real estate market. Statistical techniques including transformations, LASSO, forward selection, backward selection, and all subsets were used to find the best multiple regression model. Diagnostic evaluations were performed to ensure the model's accuracy, including residual analysis, multicollinearity checks, and outlier identification. Tree based models will be used to compare the performance of the proposed multiple regression model. The proposed predictive model addresses a complex and relevant market by uncovering the underlying interactions and relationships within the data. The study holds its significance through aiding potential home buyers and sellers with a reasonable price estimate based on the physical qualities of the house.

Ce projet vise à investiguer les facteurs déterminants des prix immobiliers à Edmonton, en Alberta. L'immobilier restant un investissement important, cette analyse vise à offrir des informations aux acheteurs potentiels naviguant sur le marché immobilier d'Edmonton en plein essor. Des techniques statistiques incluant des transformations, LASSO, la sélection ascendante, la sélection descendante et tous les sous-ensembles ont été utilisées pour trouver le meilleur modèle de régression multiple. Des évaluations diagnostiques ont été effectuées pour garantir la précision du modèle, comprenant une analyse des résidus, des vérifications de multicolinéarité et l'identification des valeurs aberrantes. Des modèles basés sur les arbres de décision seront utilisés pour comparer la performance du modèle de régression multiple proposé. Le modèle prédictif proposé aborde un marché complexe et pertinent en mettant en lumière les interactions et les relations sous-jacentes dans les données. L'étude revêt son importance en aidant les futurs acheteurs et vendeurs de maisons à estimer un prix raisonnable basé sur les caractéristiques physiques de la propriété.

Solmaz Ghajar, Nastaran Hajizadeh, Fatemeh Nezarat, Sara Khademioureh, Saumyadipta Pyne, Irina Dinu

COVID-19 infection During Pregnancy changes Gene Expression in Umbilical Cord Blood cells

L'infection par la COVID-19 pendant la grossesse modifie l'expression génique dans les cellules sanguines du cordon ombilical

The COVID-19 pandemic has affected all people, including pediatrics. The infection in pregnant mothers induces inflammation in the placenta which causes differential gene expression in the fetus and may result in long-term effects in offspring. In this study, the effects of maternal COVID-19 on the metabolic pathways of stem cells were examined using KEGG-LEGACY gene sets. We employed The Linear Combination Test to compare gene expression profiles of cord blood cells from 8 cases (women who had been exposed to COVID-19 during pregnancy) and 8 controls (who had not been exposed to the virus) using a dataset from the GEO database (accession number: GSE195938). Fourteen gene sets (contributing to translation of cell proteins) of 187 gene sets were differently expressed (p-value < 0.05). Finding the compromised molecular signalling pathways during maternal COVID-19 can aid in understanding of the condition, drug discovery, and decreasing the long-term toll of disease on unborn children.

La pandémie de COVID-19 a grandement affecté le monde, y compris les pédiatres. L'infection chez les mères enceintes induit une inflammation du placenta provoquant une expression génique différentielle chez le fœtus et peut entraîner des effets à long terme sur la descendance. Dans cette étude, les effets de la COVID-19 maternel sur les voies métaboliques des cellules souches ont été examinés en utilisant les ensembles de gènes KEGG-LEGACY. Le test de combinaison linéaire a été utilisé pour comparer les profils d'expression génique des cellules sanguines du cordon ombilical de 8 cas (femmes exposées à la COVID-19 pendant la grossesse) et de 8 témoins (qui n'avaient pas été exposées au virus) en utilisant un ensemble de données de la base de données GEO (numéro d'accès : GSE195938). Quatorze ensembles de gènes (contribuant à la traduction des protéines cellulaires) sur 187 ensembles de gènes étaient exprimés différemment (valeur-p $< 0{,}05$). Identifier les voies de signalisation moléculaire compromises pendant le COVID-19 maternel peut aider à comprendre la maladie, la découverte de médicaments et à réduire le coût à long terme de la maladie sur les enfants à naître.

Online Posters • Affiches en ligne

Yan Song, Mateen Shaikh, R. Nicholas Carleton, Gregory Anderson

Developing a Brief PSP Mental Health Screening Tool with Generalized Linear Model and Regularization

Élaboration d'un bref outil de dépistage de la santé mentale dans le cadre du PSP à l'aide d'un modèle linéaire généralisé et d'une régularisation

Public Safety Personnel (PSP) often face traumatic events, which can lead to significant mental health challenges. Conventional clinical assessments encounter barriers like stigma and limited accessibility, highlighting the necessity for accessible and effective interventions. Completing numerous questionnaires, even in their abbreviated forms, poses a time-consuming hurdle for PSPs due to the diverse spectrum of mental health issues they may encounter. Additionally, redundancy emerges when related questions appear across instruments. This study aims to develop a brief and comprehensive self-assessment tool tailored to the PSP community. Using a generalized linear model with regularization, the study investigates the impact of selected questions from existing instruments on the validity of assessing general mental health status. The performance of a tool with far fewer questions into risk categories is compared to the original long-form surveys which were designed to pinpoint specific mental-health disorders.

Le personnel de sécurité publique (PSP) est souvent confronté à des événements traumatisants qui peuvent entraîner des problèmes de santé mentale importants. Les évaluations cliniques conventionnelles se heurtent à des obstacles tels que la stigmatisation et l'accessibilité limitée, ce qui souligne la nécessité d'interventions accessibles et efficaces. Remplir de nombreux questionnaires, même sous leur forme abrégée, est un obstacle qui fait perdre du temps aux PSP en raison de la diversité des problèmes de santé mentale qu'ils peuvent rencontrer. En outre, la redondance apparaît lorsque des questions apparentées sont posées dans différents instruments. Cette étude vise à développer un outil d'auto-évaluation bref et complet adapté à la communauté des PSP. À l'aide d'un modèle linéaire généralisé avec régularisation, l'étude examine l'impact des questions sélectionnées dans les instruments existants sur la validité de l'évaluation de l'état général de la santé mentale. Les performances d'un outil comportant beaucoup moins de questions sur les catégories de risque sont comparées à celles des enquêtes longues originales conçues pour mettre en évidence des troubles mentaux spécifiques.

Minoli Munasinghe, Mateen Shaikh, Erfanul Hoque
A robust regression model in the presence of missing and censored data
Un modèle de régression robuste en présence de données manquantes et censurées

This work addresses the challenges posed by missing and censored data in regression modeling, presenting an innovative approach based on the Expectation Maximization (EM) algorithm. Incomplete data often challenges the accuracy and reliability of regression analyses, particularly in real-world applications where missingness and censoring are common phenomena. The proposed approach introduces the Expectation Maximization (EM) algorithm for conducting regression analysis in the presence of missing data, left censored, right censored, and interval censored data scenarios based on bivariate normal distribution. Extensive simulation studies are conducted to evaluate the performance of the proposed approach under varying scenarios such as different sample sizes, different missing percentages, and different correlations, offering a versatile solution for researchers engaged in regression modeling.

Ce travail aborde les défis posés par les données manquantes et censurées dans la modélisation de la régression, en présentant une approche innovatrice basée sur l'algorithme espérance-maximization (EM). Les données incomplètes remettent souvent en question la précision et la fiabilité des analyses de régression, en particulier dans les applications du monde réel où les données manquantes et la censure sont des phénomènes courants. L'approche proposée introduit l'algorithme espérance-maximization (EM) pour effectuer des analyses de régression en présence de données manquantes, de scénarios de données censurées à gauche, censurées à droite et censurées par intervalle, sur la base d'une distribution normale bivariée. Des études de simulation approfondies sont menées pour évaluer les performances de l'approche proposée dans divers scénarios tels que différentes tailles d'échantillons, différents pourcentages de données manquantes et différentes corrélations, offrant ainsi une solution polyvalente aux chercheurs engagés dans la modélisation de la régression.

Shaomeng Yin, Mateen Shaikh

Improving Predictive Ability for Student Academic Performance through Imbalanced Data Handling

Amélioration de la capacité prédictive des résultats scolaires des étudiants grâce à un traitement déséquilibré des données

This study addresses the challenge of enhancing predictive ability in student academic performance, focusing on a three-class classification problem. The primary obstacle is effectively utilizing both current- and post-enrollment information while class labels are imbalanced. We employ XGBoost, Random Forest, Random Under-Sampling, Random Over-Sampling, SMOTE, and ADASYN to address these issues. Combinations of these strategies are evaluated for their predictive ability in the imbalanced dataset.

Cette étude aborde le défi de l'amélioration de la capacité prédictive des résultats scolaires des étudiants, en se concentrant sur un problème de classification à trois classes. Le principal obstacle est l'utilisation efficace des informations actuelles et postérieures à l'inscription alors que les étiquettes de classe sont déséquilibrées. Nous utilisons XGBoost, Random Forest, Random Under-Sampling, Random Over-Sampling, SMOTE et ADASYN pour résoudre ces problèmes. Les combinaisons de ces stratégies sont évaluées pour leur capacité prédictive dans l'ensemble de données déséquilibré.

Saeid Moradi, Mateen Shaikh

Skin Cancer Detection Using Deep Convolutional Neural Networks
Détection du cancer de la peau à l'aide de réseaux de neurones convolutifs profonds

Heterogenous forms of skin cancer have emerged as one of the most prevalent forms of cancer, underscoring the significance of early detection and precise diagnosis for effective treatment. This study analyzes the HAM10000 dataset, comprising 10015 skin lesion instances across an imbalanced set of seven different categories of pigmented skin lesions. Rather than addressing the imbalanced data through the classifier, the imbalance is addressed through preprocessing including resampling, and data augmentation. Furthermore, a classifier can be developed quicker using a pretrained model. This work compares the performance of six different pre-trained Convolutional Neural Networks (CNNs): VGG16, VGG19, ResNet50, MobileNet, MobileNetV2, and MobileNetV3.

Les formes hétérogènes de cancer de la peau sont devenues l'une des formes de cancer les plus répandues, soulignant l'importance d'une détection précoce et d'un diagnostic précis pour un traitement efficace. Cette étude analyse l'ensemble de données HAM10000, qui comprend 10 015 instances de lésions cutanées dans un ensemble déséquilibré de sept catégories différentes de lésions cutanées pigmentées. Plutôt que de traiter les données déséquilibrées par le biais du classificateur, le déséquilibre est traité par le biais d'un prétraitement comprenant le rééchantillonnage et l'augmentation des données. En outre, un classificateur peut être développé plus rapidement à l'aide d'un modèle pré-entraîné. Ce travail compare les performances de six réseaux de neurones convolutifs (CNN) pré-entraînés différents : VGG16, VGG19, ResNet50, MobileNet, MobileNetV2 et MobileNetV3.

Meira Golberg

Using Multiple Imputation to Deal with Missing Data in the Canadian Longitudinal Study on Aging

Utilisation de l'imputation multiple pour traiter les données manquantes dans l'Étude longitudinale canadienne sur le vieillissement

Longitudinal studies in cognitive health often face challenges related to missing data, and choosing an appropriate imputation method is crucial for robust analyses. This study investigates the impact of missing data and imputation methods on modelling the outcome of interest, executive function, using two waves of longitudinal data collected from the Canadian Longitudinal Study on Aging. Executive function is a composite score calculated based on five cognitive tests, each of which might be subject to missingness. Results reveal selective attrition effects on variables associated with missingness, such as rurality and social isolation. Multiple imputation proves effective in addressing biases introduced by complete case analysis. The results from our analysis demonstrate the need for careful consideration in imputation strategies. The study underscores the importance of understanding attrition mechanisms, conducting sensitivity analyses, and aligning multiple imputation approaches with the distributional characteristics of the outcome variable. This research contributes to methodological discussions in cognitive health studies and informs researchers on the complexities of imputing missing data in longitudinal settings.

Les études longitudinales sur la santé cognitive sont souvent confrontées à des problèmes liés aux données manquantes, et le choix d'une méthode d'imputation appropriée est crucial pour des analyses solides. Cette étude examine l'impact des données manquantes et des méthodes d'imputation sur la modélisation du résultat d'intérêt, la fonction exécutive, en utilisant deux vagues de données longitudinales recueillies dans le cadre de l'Étude longitudinale canadienne sur le vieillissement. La fonction exécutive est un score composite calculé à partir de cinq tests cognitifs, chacun d'entre eux pouvant être sujet à des données manquantes. Les résultats révèlent des effets d'attrition sélective sur les variables associées à l'omission, telles que la ruralité et l'isolement social. L'imputation multiple s'avère efficace pour remédier aux biais introduits par l'analyse des cas complets. Les résultats de notre analyse démontrent la nécessité d'examiner attentivement les stratégies d'imputation. L'étude souligne l'importance de comprendre les mécanismes d'attrition, d'effectuer des analyses de sensibilité et d'aligner les approches d'imputation multiple sur les caractéristiques de distribution de la variable de résultat. Cette recherche contribue aux discussions méthodologiques dans les études sur la santé cognitive et informe les chercheurs sur les complexités de l'imputation des données manquantes dans les contextes longitudinaux.

Xiao Yan, Kuan Liu

Practical Implementation of Advanced Causal Inference Method: Development of an R Package for Bayesian Marginal Structural Models with Time-Varying Treatment
Implémentation pratique d'une méthode avancée d'inférence causale : Développement d'un package R pour les modèles structurels marginaux bayésiens avec traitement variant dans le temps

Observational studies offer a viable, efficient, and low-cost design to readily gather evidence on exposure effects. Although more practical, exposure mechanism is nonrandomized and causal inference methods are required to draw causal conclusions. Popular approaches used in health research are predominantly frequentist methods. Bayesian approaches have unique estimation features that are useful in many settings, however, there is a general lack of open-access software packages to carry out these analyses. Our project seeks to address this gap by developing a user-friendly R package named "BayesMSM" for the implementation of the Bayesian Marginal Structural Models for longitudinal data with continuous or binary outcome. We will demonstrate the use of this package with simulated data.

Les études observationnelles offrent un design viable, efficace et peu coûteux pour recueillir rapidement des preuves sur les effets d'exposition. Bien que plus pratique, le mécanisme d'exposition n'est pas randomisé et des méthodes d'inférence causale sont nécessaires pour tirer des conclusions causales. Les approches populaires utilisées dans la recherche en santé sont principalement des méthodes fréquentistes. Les approches bayésiennes possèdent des caractéristiques d'estimation uniques qui sont utiles dans de nombreux contextes. Cependant, il existe un manque général de logiciels en libre accès pour effectuer ces analyses. Notre projet vise à combler cette lacune en développant un package R convivial nommé « BayesMSM » pour la mise en œuvre de modèles structurels marginaux bayésiens pour les données longitudinales avec des résultats continus ou binaires. Nous démontrerons l'utilisation de ce package avec des données simulées.