Contents • Table des Matières

1 Welcome • Bienvenue

The Department of Mathematics and Statistics, the Department of Clinical Epidemiology and Biostatistics and the Graduate Program in Statistics at McMaster University welcome you to Hamilton, Ontario. With a population of about 500,000 people, the New City of Hamilton comprises the old City of Hamilton and the adjacent smaller towns of Ancaster, Dundas, Flamborough, Glanbrook and Stoney Creek. Hamilton is located in the westernmost corner of Lake Ontario. Museums, galleries, theatres, parks, gardens, hiking trails and restaurants with great variety of international cuisines are spread around the city and its vicinity. There are extensive networks of hiking trails in the Westdale Ravine behind McMaster and nearby in the Dundas Valley. Ask for a trail guide at the registration desk. Within one hour or so by car you can be in the nearby communities of Niagara Falls, Niagara-on-the-Lake, St. Catharines, Kitchener, Waterloo, St. Jacobs, Guelph or the great City of Toronto. We hope that you find the conference professionally rewarding and that you have a great time in our community.

Les départements de mathématiques et de statistiques, d'épidémiologie clinique et de biostatistique et le programme d'étude supérieure en statistique de McMaster University vous souhaitent la bienvenue à Hamilton, Ontario. Avec une population d'environ 500 000 personnes, la nouvelle ville d'Hamilton comprend l'ancienne ville d'Hamilton et les villes voisines plus petites d'Ancaster, de Dundas, de Flamborough, de Glanbrook et de Stoney Creek. Hamilton est situé dans la partie la plus à l'ouest du lac Ontario. Des musées, des galeries, des théâtres, des parcs, des jardins, des sentiers de randonnée et des restaurants de cuisines internationales variées sont dispersés dans la ville et à proximité. Il y a un grand réseau de sentiers de randonnée à Westdale Ravine derrière McMaster et à proximiter dans Dundas Valley. Demandez un guide des sentiers au bureau d'inscription. moins d'une heure de voiture, vous pouvez être dans les villes voisines de Niagara Falls, de Niagara-on-the-Lake, St. Catharines, Kitchener, Waterloo, St. Jacobs, Guelph et la région urbaine de Toronto. Nous espérons que vous trouvez la conférence très enrichissante du point de vue professionnel et que vous aurez beaucoup de plaisir dans notre communauté.

2 Sponsors • Commanditaires

SSC 2002 thanks the sponsors of the meeting for their kind contributions. In particular, we thank the the Centre de recherches mathématiques, the Fields Institute and the Pacific Institute for the Mathematical for their financial support. We also thank McMaster University, the Faculty of Science and the Department of Mathematics and Statistics for the use of resources and financial support.

SSC 2002 remercie les commanditaires du congrès pour leurs aimables contributions. En particulier, nous remercions le Centre de recherches mathématiques, le Fields Institute et le Pacific Institute for the Mathematical Sciences pour leur aide financière. Nous remercions également McMaster University, la Faculté des sciences et le Département de mathématiques et de statistique pour l'utilisation

de leurs ressources et leur appui financier.

3 Exhibitors • Exposants

John Wiley & Sons Canada Limited
McGraw Hill
Nelson Thomson Learning
Pearson Education Canada

The book displays are located in the USC Marketplace and will be available for viewing and purchasing from 8:30 to 17:00 from Monday through Wednesday.

Le salon des exposants est au USC Marketplace. Les livres pourront être consultés ou achetés de 8h30 à 17h00 de lundi à mercredi.

4 Organizers • Organisateurs

Local Arrangements Committee • Comité organisateur local

Chair/*Président* :	Peter Macdonald
Members/*Membres* :	N. Balakrishnan
	Angelo Canty
	Aaron Childs
	Charlie Goldsmith
	Román Viveros-Aguilera

All are faculty members in the Department of Mathematics and Statistics except for Charlie Goldsmith whose main affiliation is with the Department of Clinical Epidemiology and Biostatistics. The Local Organizing Committee thanks Wendy Read from Housing & Conference Services for coordinating the booking of rooms and furniture for the conference.

Tous sont professeurs dans le Département de mathématiques et de statistique excepté Charlie Goldsmith dont l'affiliation principale est le Département d'épidémiologie clinique et de biostatistique. Le Comité organisateur local remercie Wendy Read des Services de conférences et d'hébergement pour avoir coordonné la réservation des salles et des équipements pour la conférence.

Local Arrangements Assistants • Assistants pour l'organisation locale

Connie Oosterlinck (Department of Mathematics and Statistics) and several graduate student volunteers coordinated by Vaneeta Grover.

Connie Oosterlinck (Département de mathématiques et de statistique) et une équipe de volontaire composée d'étudiants inscrits aux cycles supérieurs supervisée par Vaneeta Grover.

Programme Committee • Comité du programme

Chair/*Président* :	Bruce Smith (Dalhousie University)
Members/*Membres* :	Mik Bickis (University of Saskatchewan) – Biostatistics Section • *Groupe de biostatistique*
	Bovas Abraham (University of Waterloo) – Business and Industrial Statistics Section • *Groupe de statistique industrielle et de gestion*
	Narasimha Prasad (University of Alberta) – Survey Methods Section • *Groupe de méthodologie d'enquête*
	Susan Murphy (University of Michigan) – IMS Sessions
	Patricia Whitridge (Statistics Canada) Survey Methods Contributed Paper Sessions Chair • *Responsables des communications libres en méthodologie d'enquête*
	Rolf Turner (University of New Brunswick) – Contributed Paper Sessions Chair • *Responsable des communications libres*

Translation • Traduction

Chair/*Président* :	Jean-François Angers (Université de Montréal)
Members/*Membres* :	Stéphane Courchesne (Université de Montréal)
	Peter Macdonald (McMaster University)
	André Montpetit (Université de Montréal)

Software Development • Développement informatique

Jasmin Lapalme (Université de Montréal)
Daniel Ouimet (Université de Montréal)

5 General Information • Information générale

Registration • Inscription

The Registration Desk is located in the Marketplace of the University Student Centre (USC Marketplace) and will be open as follows:

Date	**Time**
Sunday May 26	8:00 – 22:00
Monday May 27	7:30 – 18:00
Tuesday May 28	7:30 – 18:00
Wednesday May 29	8:00 – 12:00

The opening mixer on Sunday, the poster session on Sunday and Monday and all coffee breaks will take place at this location.

L'inscription aura lieu à la place du marché du Centre étudiant (USC Marketplace) aux heures suivantes :

Jour	**Heure**
Dimanche 26 mai	*8h00 – 22h00*
Lundi 27 mai	*7h30 – 18h00*
Mardi 28 mai	*7h30 – 18h00*
Mercredi 29 mai	*8h00 – 12h00*

La soirée de dimanche, la séance d'affichage de dimanche et lundi et toutes les pauses café auront lieu au même endroit.

Rooms • Salles

The table below contains a summary of the rooms used for the conference. A "B" in front of a room number indicates a basement room. A campus map is provided in the outside back cover for your convenience. The McMaster campus is fairly compact, the rooms for the sessions are 2-3 min away from the University Student Centre, while the Commons building and residences are within a 5-min. walk.

Le tableau ci-dessous contient un sommaire des salles utilisées pour la conférence. Un "B" devant un numéro de salle indique qu'elle est au sous-sol. Une carte de campus est fournie sur la couverture extérieure arrière pour votre convenance. Le campus de McMaster est assez compact, les salles pour les sessions sont à 2 ou 3 minutes du Centre étudiant (USC), alors que les édifices communaux ("Commons") et les résidences sont à 5 minutes de marche.

Building • Édifice	**Short • Abrégé**	**Rooms/Areas • Salle/ Place publique**
Chester New Hall	CNH	104
Commons	C	Skylight Room, Orchid Room, Market
Faculty Club	AH	Great Hall
Kenneth Taylor Hall	KTH	B135, B105, Celebration Hall
Refectory		Patio
Togo Salmon Hall	TSH	B105, B106, B218, 122
University Student Centre	USC	Marketplace, La Piazza, 203, 206, 207, 213, 214

Workshops • Ateliers

Workshops organized by the three sections of the SSC will be held on Sunday May 26 from 9:00 to 17:00. The rooms for the workshops are as follows.

Les ateliers organisés par les trois sections de la SSC auront lieu dimanche le 26 mai de 9h00 à 17h00. Les salles pour ces ateliers sont les suivantes.

Workshop • Atelier	**Place • Endroit**	**DPA • DPI**
Biostatistics • *Biostatistique*	TSH B128	N
Business and Industrial Statistics • *Statistique industrielle et gestion*	TSH B105	Y
Survey Methods • *Méthodologie d'enquête*	KTH B135	N

DPA: data projector availability • *DPI : disponibilité d'un projecteur d'image-écran*

Check the scientific program and abstract sections for further details. There will be coffee breaks from 10:00 to 10:30 and from 15:00 to 15:30 in the USC Marketplace. Your registration to a workshop includes lunch to be served in a buffet style from 12:00 to 13:30 in Celebration Hall in the basement of the KTH building.

Vérifiez le programme scientifique et la section des résumés pour plus de détails. Il y aura les pauses café de 10h00 à 10h30 et de 15h00 à 15h30 au USC Marketplace. Votre inscription à un atelier inclut un buffet servi entre 12h00 à 13h30 dans la salle Celebration Hall dans le sous-sol de l'édifice KTH.

Food Sevices • Restauration

Your conference registration includes daily lunch from Monday through Wednesday and the banquet on Monday evening. Lunch will be served in a buffet style from 12:00 to 13:30 in the Commons Market (second floor of Commons building). Coffee breaks will take place in the USC Marketplace from 10:00 to 10:30 and from 15:00 to 15:30 Monday through Wednesday. The following campus food services are available Monday through Wednesday of the conference:

La Piazza in the USC building besides the Marketplace	7:30 – 20:00
Hotdog stand outside Gilmor Hall	11:00 – 16:00
MAC Express in the lobby of the J.H. Engineering building	7:45 – 15:15
Café 2000 in the lobby of the Applied Health Sciences building	7:30–21:00
Checkers Food Court in Level 1 of the McMaster Hospital	7:00 – 18:00

La Piazza is closest to the conference rooms, a variety of vendors are available including Tim Hortons coffee and pastries. Committee members attending breakfast meetings should sign up for breaksfast at the registration desk if they are not staying in the McMaster Residence. Anyone staying in the residence will have a meal card for breakfast.

Votre inscription à la conférence inclut le diner de lundi à mercredi et le banquet du lundi soir. Un buffet sera servi de 12h00 à 13h30 à la place du marché de l'édifice Commons (au deuxième étage). Les pauses café auront lieu au USC Marketplace de 10h00 à 10h30 et de 15h00 à 15h30 du lundi au mercredi. Les aires de restauration suivantes seront ouvertes de lundi à mercredi pendant la conférence :

La Piazza dans l'édifice USC près de la place du marché	7h30 – 20h00
Kiosque à hotdogs près du Gilmor Hall	11h00 – 16h00
MAC Express dans le hall d'entrée de l'édifice J.H. Engineering	7h45 – 15h15
Café 2000 dans le hall d'entrée de l'édifice Applied Health Sciences	7h30–21h00
Checkers Food Court au sous-sol de l'hôpital McMaster	7h00 – 18h00

La Piazza est le restaurant le plus proche des salles de conférence, plusieurs concessions sont disponibles incluant un Tim Hortons. Les membres de comités assistant à des réunions à l'heure du déjeuner doivent s'inscrire pour le déjeuner au comptoir d'inscription s'ils ne restent pas au résidence de McMaster. Tout ceux qui restent au résidence auront une carte de repas pour le déjeuner.

Women in Statistics Dinner • Diner pour les femmes en statistique

The Canadian Section of the Caucus for Women in Statistics and the SSC Committee for Women in Statistics invite those interested to attend dinner on Sunday May 26 at The Mandarin Restaurant (1508 Upper James, tel. (905) 383-6000). The restaurant is an all-you-can-eat Chinese and Canadian buffet with a large selection of fresh and appetizing food. This is a pay event, meeting place is the SSC Registration Desk in the USC Marketplace shortly after the workshops end. The event will give you an opportunity to meet old friends and make new acquaintances.

La section canadienne du Caucus des femmes en statistique et le Comité de la SSC des femmes en statistiques invitent ceux et celles intéressés à un souper dimanche le 26 mai au restaurant Mandarin (1508 rue Upper James, téléphone (905) 383-6000). Le restaurant est un buffet chinois et canadien avec un grand choix de nourriture fraîche et appétissante. Cet événement est non inclus dans les frais d'inscription. Le point de rencontre est le comptoir d'inscription au USC Marketplace peu de temps après la fin des ateliers. Cet événement vous donnera une occasion de rencontrer de vieux amis et de faire de nouvelles connaissances.

Opening Mixer/Poster Session • Soirée d'ouverture / Session d'affichage

The Opening Mixer/Poster Session will be held from 18:00 to 22:00 on Sunday May 26 in the Marketplace of the USC. Tasty hors d'oeuvres and a cash bar will be available. It is a great time to relax and chat, all conference attendees and companions are warmly invited to attend. The Registration Desk will also operate in this location. The posters will be left on display until 16:00 on Monday May 27.

La soirée d'ouverture et la session d'affichage aura lieu le dimanche 26 mai de 18h00 à 22h00 au USC Marketplace. Des amuse-gueules et un bar payant seront disponibles. C'est le moment idéal pour se détendre et discuter. Tous les participants à la conférence et leurs compagnons sont chaleureusement invités à y participer. Le comptoir d'inscription sera également ouvert à cet endroit. Les affiches seront laissées en montre jusqu'au lundi 27 mai à 16h00.

Banquet • Banquet

All registered delegates are invited to the Conference Banquet, Monday May 26, 18:30, at The Auditorium of the Royal Botanical Gardens (RBG, 680 Plains Road West, Burlington, ON, tel. (905) 527-1158). A banquet ticket for a four-course dinner is provided in the registration package. A shuttle service will be provided, leaving McMaster at 18:00 from the parking lot near the USC building. The shuttle service will include the main hotels, leaving from the Ramada Hotel at 18:00 with stops at the Royal Connaught, Sheraton, Visitors Inn and Admiral Inn hotels.

For driving to the RBG from McMaster, take the east direction of Main Street west. Turn left at Dundurn Street then left at York Boulevard. This street becomes Plains Road West after the second traffic light, the RBG building is on the right-hand-side at the next traffic light. From downtown Hamilton, take King Street West (towards the west, your only choice), turn left at Hess Street North, then left at York Boulevard and proceed as indicated above. The total driving time should be around 12-15 min.

Tous les délégués inscrits sont invités au banquet du congrès le lundi 26 mai à 18h30, à l'auditorium des Jardins botaniques royaux (RBG, 680 Plain Road West, Burlington, ON, téléphone (905) 527-1158). Un billet de banquet pour un dîner de quatre services est inclus avec votre pochette d'inscription. Un service de navette sera fourni. Il quittera McMaster à 18h00 à partir du stationnement près de

l'édifice USC. Le service de navette inclura aussi les hôtels principaux, partant de l'hôtel Ramada à 18h00 avec arrêt aux hôtels Connaught, Sheraton, Visitors Inn et Admiral Inn.

Pour conduire jusqu'au RBG en partant de MU, prenez la direction est sur la rue Main West. Tournez à gauche à la rue Dundurn puis à gauche au boulevard York. Cette rue devient Plains Road West après le deuxième feu de circulation. L'édifice du RBG est du côté droit au prochain feu de circulation. Du centre-ville d'Hamilton, prenez la rue King West (vers l'ouest, votre seul choix possible), tournez à gauche à la rue Hess North, puis à gauche au boulevard York et procédez comme indiqué ci-dessus. Le temps de conduite total devrait être environ 12-15 minutes.

Barbeque • Barbecue

A barbeque dinner will be held on Tuesday May 28 from 18:00 to 20:30 at the Refectory patio (Commons building in case of rain). Barbeque, cash bar and live music will be provided. Attendance is by ticket only, students can get a free ticket at time of registration, others are invited to purchase tickets for $25. The ticket includes one drink, salad, hot vegetables, a choice of New York striploin steak or vegetarian kebabs, dessert and coffee.

Un dîner au barbecue sera tenu le mardi 28 mai de 18h00 à 20h30 au patio de réfectoire (l'édifice Commons en cas de pluie). Le barbecue, un bar payant et la musique seront fournis. Pour assister à ce barbecue, vous avez besoin d'un billet. Les étudiants peut obtenir un billet gratuit lors de l'inscription. Les autres sont invités à acheter des billets au coût de 25$ chacun. Le billet inclut une consommation, de la salade, des légumes chauds, et un choix entre un entrecôte ou des brochettes végétariennes, le dessert et le café.

Waterloo Statistics Alumni Reception • Reception des anciens diplômés en statistique de Waterloo

University of Waterloo Math alumni attending the 2002 SSC Annual Meeting are warmly invited to a reception in the Great Hall of the Faculty Club from 16:30 to 18:30 on Tuesday May 28. Light refreshments and a cash bar will be available as you mingle with fellow alumni and UW faculty. The event is organized by the Alumni Office of UW's Faculty of Mathematics.

Les anciens diplômés en mathématiques de Waterloo assistant au congrès annuel 2002 de la SSC sont chaleureusement invités à une réception dans le grand Hall du Faculty Club de 16h30 à 18h30 le mardi 28 mai. Les rafraîchissements légers et un bar payant seront disponibles pendant que vous discuterez avec vos anciens camarades et les professeurs de UW. L'événement est organisé par le Bureau des anciens étudiants de la Faculté de mathématiques de Waterloo University.

Job Fair • Salon de l'emploi

Interviews for those participating in the Job Fair will be conducted in rooms 203, 206, 207, 213 and 214 located in the second floor of the USC building. Check the Job Fair Desk and bulletin board in the USC Marketplace for schedules and further updates.

Des entrevues pour ceux participant au Salon de l'emploi auront lieu dans les salles 203, 206, 207, 213 et 214 situés dans le deuxième étage de l'édifice USC. Vérifiez auprès du comptoir du Salon et du tableau d'affichage situé au USC Marketplace pour l'horaire et les mises à jour.

Transportation and Parking • Transport public et stationnement

A 3-day bus pass for Monday through Wednesday can be purchased at Registration Desk at a discounted conference fee of $3.50. The general single bus ride costs $2. Buses run efficiently in the McMaster-Downtown Hamilton corridor. Parking at McMaster costs $8.50 per day. Ask the receptionist at the entrance booth for the nearest available parking lot. Parking on weekends is free. McMaster Residence guests get free parking. For those needing transportation to the Toronto airport, Airways Transit offers a door-to-door service. A 3-day advanced booking is required, you may book by phone at (905) 689-4460. As a conference delegate, you get the discounted rate of $32 per person for a one-way trip, but you need to advise the Airways Transit receptionist that you are attending the SSC conference at the time of booking to get the conference rate. GO Transit offers a bus round service from downtown Hamilton to downtown Toronto. The one-way rate is $8.20, buses run every 20 min at rush hour, every 30 min during the rest of the day and every hour between 21:00 and 23:00. You may board the buses at several designated stops on King Street West between Downtown Hamilton and Dundurn Street.

Un laissez-passer d'autobus pour trois jours (de lundi à mercredi) peut être acheté au bureau d'inscription à un prix spécial pour le congrès de 3.50$. Les aller simple en autobus coûte 2$. Les autobus sont efficaces dans le couloir de McMaster et le centre-ville d'Hamilton. Se garer à McMaster coûte 8.50$ par jour. Demandez au préposé à la guérite pour le stationnement disponible le plus près. Se garer les week-ends est gratuit. Pour ceux qui ont besoin d'un transport vers l'aéroport de Toronto, Airways Transit offre un service porte-à-porte. Une réservation trois jours à l'avance est exigée. Vous pouvez réserver par téléphone au (905) 689-4460. En tant que délégué au congrès, vous avez droit à un tarif spécial de 30$ par personne pour un aller simple. Pour bénéficier de ce rabais, vous devez aviser la réceptionniste d'Airways Transit que vous participer au congrès de la SSC au moment de votre réservation. GO Transit offre un service d'autobus entre le centre-ville d'Hamilton et celui de Toronto. Le prix d'un aller simple est 8.20$. Les autobus sont à tous les 20 minutes pendant la période de pointe, à tous les 30 minutes le reste de la journée et à toutes les heures entre 21h00 et 23h00. Vous pouvez monter à bord de l'autobus à plusieurs arrêts indiqués sur la rue King West entre le centre-ville d'Hamilton et la rue Dundurn.

Internet Access • Accès à l'internet

The Computer Lab located in KTH B121 will be available from 8:00 to 16:00 from Monday through Wednesday for internet access. Only web-based e-mail is possible. Ask the room attendant for instructions.

Le laboratoire informatique est situé au KTH B121 et il est ouvert, pour accéder à l'internet, de 8h00 à 16h00 de lundi à mercredi. Seulement le courriel avec interface web est disponible. Demandez au préposé pour plus d'informations.

Office • Bureau

The Conference Office is located in room 122 in the main floor of the TSH building. Telephone, storage and limited photocopying are provided in the office.

Le Bureau du congrès est situé à la salle 122 à l'étage principal de l'édifice TSH. Accès au téléphone, consigne et polycopie (accès limité) est disponible sur place.

6 Committees and Meetings • Comités et reéunions

Day	Time	Place	Meeting
Saturday	18:00-23:00	USC 220	SSC Executive (D. Brillinger)
Sun	9:00–11:00	Skylight Room	Finance (M. Alvo)
Sun	11:00–12:00	Skylight Room	Publications (J. Braun)
Sun	12:00–17:00	Skylight Room	SSC Board of Directors
Sun	12:00–17:00	Orchid Room	Statistics Chairs (L.-P. Rivest)
Mon	7:15–8:15	Skylight Room	Biostatistics Executive I (M. Bickis)
Mon	7:15–8:15	Skylight Room	Survey Methods Executive (N. Prasad)
Mon	7:15–8:15	Skylight Room	Public Relations (J. Braun)
Mon	12:15–13:15	Skylight Room	CJS Editorial Board (R. Lockhart)
Mon	12:15–13:15	Skylight Room	Implementation of Accreditation I (E. Enns)
Mon	17:00–18:00	TSH B128	Biostatistics Section AGM
Mon	17:00–18:00	TSH B105	Business and Industral Statistics Section AGM followed by BISS Executive (B. Abraham)
Mon	17:00–18:00	KTH B135	Survey Methods Section AGM
Tue	7:15–8:15	Skylight Room	Implementation of Accreditation II (K. McRae)
Tue	7:15–8:15	Skylight Room	Research Committee (J. Petkau)
Tue	12:15–13:15	Skylight Room	Bilingualism (J.-F. Angers)
Tue	12:15–13:15	Skylight Room	Statistical Education (S. Brown)
Tue	12:15–13:15	Skylight Room	Women in Statistics (N. Ghazzali)
Tue	17:00–18:00	TSH B105	SSC AGM
Wed	7:15–8:15	Skylight Room	Biostatistics Executive II (M. Bickis)
Wed	7:15–8:15	Skylight Room	Professional Development (S. Bartlett)
Wed	7:15–8:15	TSH B105	SORA AGM
Wed	12:15–13:15	Skylight Room	Program (P. Cabilio)
Wed	17:30–19:30	Skylight Room	SSC Board of Directors
Wed	19:30–21:00	Skylight Room	SSC Executive (D. Brillinger)

Jour	**Heure**	**Endroit**	**Réunion**
Samedi	18h00-23h00	USC 220	Comité exécutif (D. Brillinger)
Dimanche	9h00–11h00	Skylight Room	Finances (M. lvo)
Dimanche	11h00–12h00	Skylight Room	Publications (J. Braun)
Dimanche	12h00–17h00	Skylight Room	Conseil d'administration
Dimanche	12h00–17h00	Orchid Room	Directeurs de statistique (L.-P. Rivest)
Lundi	7h15–8h15	Skylight Room	Exécutif, Biostatistiques I (M. Bickis)
Lundi	7h15–8h15	Skylight Room	Exécutif, Méthodes d'enquêtes (N. Prasad)
Lundi	7h15–8h15	Skylight Room	Relations publiques (J. Braun)
Lundi	12h15–13h15	Skylight Room	Comité éditorial de la Revue CJS (R. Lockhart)
Lundi	12h15–13h15	Skylight Room	Implémentation de l'accréditation I (E. Enns)
Lundi	17h00–18h00	TSH B128	Assemblée générale, Biostatistiques
Lundi	17h00–18h00	TSH B105	Assemblée générale et exécutif, Statistique industrielle et de gestion (B. Abraham)
Lundi	17h00–18h00	KTH B135	Assemblée générale, Méthodes d'enquête
Mardi	7h15–8h15	Skylight Room	Implémentation de l'accréditation II (K. McRae)
Mardi	7h15–8h15	Skylight Room	Recherche (J. Petkau)
Mardi	12h15–13h15	Skylight Room	Bilinguisme (J.-F. Angers)
Mardi	12h15–13h15	Skylight Room	Formation statistique (S. Brown)
Mardi	12h15–13h15	Skylight Room	Promotion de la femme en statistique (N. Ghazzali)
Mardi	17h00–18h00	TSH B105	Assemblée générale, SSC
Mercredi	7h15–8h15	Skylight Room	Biostatistics Executive II (M. Bickis)
Mercredi	7h15–8h15	Skylight Room	Perfectionnement professionnel (S. Bartlett)
Mercredi	7h15–8h15	TSH B105	Assemblée générale, SORA
Mercredi	12h15–13h15	Skylight Room	Programme (P. Cabilio)
Mercredi	17h30–19h30	Skylight Room	Conseil d'administration
Mercredi	19h30–21h00	Skylight Room	Comité exécutif (D. Brillinger)

7 Outline of Events • Des événements

Day	Time	Place	Event	DPA
Sun	9:00–17:00	TSH B128	Biostatistics Section Workshop	N
Sun	9:00–17:00	TSH B105	Business and Industrial Statistics Section Workshop	Y
Sun	9:00–17:00	KTH B135	Survey Methods Section Workshop	N
Sun	18:00–22:00	USC Marketplace	Reception	
Sun	18:00–22:00	USC Marketplace	Poster Session	
Mon	8:30–10:00	CNH 104	Session 1: Welcome and SSC Presidential Invited Address	N
Mon	8:30–16:00	USC Marketplace	Poster Session	N
Mon	10:30–12:00	TSH B105	Session 2: Directional Statistics	N
Mon	10:30–12:00	TSH B128	Session 3: Statistics and Water Quality	N
Mon	10:30–12:00	TSH B106	Session 4: Probability	N
Mon	10:30-12:00	KTH B135	Session 5: Case Studies I – Treatment of Missing Data	Y
Mon	13:30–15:00	TSH B105	Session 6: Phylogenetics	N
Mon	13:30–15:00	TSH B128	Session 7: Joint Industry/Survey Methods Session	N
Mon	13:30–15:00	KTH B105	Session 8: Statistical Inference	N
Mon	13:30–15:00	TSH B106	Session 9: Applications of Statistics	Y
Mon	13:30–15:00	KTH B135	Session 10: NSERC Open Meeting	Y
Mon	15:30–17:00	TSH B128	Session 11: Applied Survey Methods	Y
Mon	15:30–17:00	TSH B106	Session 12: Nonparametric Methods and Density Estimation	Y
Mon	15:30–17:00	TSH B105	Session 13: Multiscale Behavior in Stochastic Models	N
Mon	15:30–17:00	KTH B135	Session 14: Statistics for Microarray Data Analysis	Y
Mon	18:00	Royal Botanical Gardens	Banquet	
Tue	8:30–10:00	TSH B106	Session 15: Official Statistics	Y
Tue	8:30–10:00	KTH B135	Session 16: Meta-Analysis and Clinical Trials	Y
Tue	8:30–9:15	TSH B128	Session 17: Pierre Robillard Award Session	Y
Tue	9:15–10:00	TSH B128	Session 18: CJS Award Session	Y
Tue	8:30–10:00	TSH B105	Session 19: Data Mining in Drug Discovery	Y

DPA: data projector availability

Day	Time	Place	Event	DPA
Tue	10:30–12:00	TSH B105	Session 20: Split Plot Experiments in Industry	Y
Tue	10:30–12:00	TSH B106	Session 21: Stochastic Modeling, Time Series and Spatial Statistics	Y
Tue	10:30–12:00	TSH B128	Session 22: Statistics and Public Policy	Y
Tue	10:30–12:00	KTH B135	Session 23: Statistics and Brain Mapping	Y
Tue	13:30–15:00	TSH B105	Session 24: Theoretical Aspects of Monte Carlo	Y
Tue	13:30–15:00	KTH B135	Session 25: Environmetrics I	Y
Tue	13:30–15:00	TSH B128	Session 26: Longitudinal Data - Theory and Applications	Y
Tue	13:30–15:00	TSH B106	Session 27: Theoretical Survey Methods	Y
Tue	15:30–17:00	TSH B106	Session 28: Multivariate Methods	N
Tue	15:30–17:00	KTH B105	Session 29: Stochastic Modeling	N
Tue	15:30–17:00	TSH B105	Session 30: Spatial Sampling	Y
Tue	15:30–17:00	KTH B135	Session 31: Statistical Methods for Occupational Risk Assessment	Y
Tue	15:30–17:00	TSH B128	Session 32: Statistical Accreditation, Progress Report and Discussion	N
Tue	16:30–18:30	Faculty Club, Great Hall	University of Waterloo Alumni Reception	
Tue	17:00–18:00	TSH B105	Annual General Meeting	
Tue	18:30 – 20:30	Refectory Patio	Barbeque	
Wed	8:30–10:00	TSH B105	Session 33: Business and Industrial Statistics Section Special Invited Address	Y
Wed	8:30–10:00	TSH B128	Session 34: Statistical Inference for Mixture Models	Y
Wed	8:30–10:00	KTH B135	Session 35: Survival Analysis	N
Wed	8:30–10:00	TSH B106	Session 36: Robust Methods, Outlier Detection and Bootstrapping	N
Wed	10:30–12:00	TSH B105	Session 37: Financial Modeling	Y
Wed	10:30–12:00	KTH B135	Session 38: Environmetrics II	N
Wed	10:30–12:00	TSH B128	Session 39: Case Studies II – Cervical Cancer	Y
Wed	10:30–12:00	TSH B106	Session 40: New Research Findings in Analysis Methods for Survey Data	Y

DPA: data projector availability

Day	**Time**	**Place**	**Event**	**DPA**
Wed	13:30–15:00	TSH B106	Session 41: Statistical Indexes	N
Wed	13:30–15:00	TSH B105	Session 42: Stochastic Operations Research	Y
Wed	13:30–15:00	KTH B135	Session 43: Survival Analysis of Case-Control and Case-Cohort Data	Y
Wed	13:30–15:00	TSH B128	Session 44: Probability and Statistical Inferences	N
Wed	15:30–17:00	KTH B135	Session 45: Point Processes and Applications	N
Wed	15:30–17:00	TSH B128	Session 46: Linear Models and Design	N
Wed	15:30–17:00	TSH B106	Session 47: Statistical Inference	Y
Wed	15:30–17:00	TSH B105	Session 48: Statistics in Finance and Marketing	Y

DPA: data projector availability

Jour	**Heure**	**Endroit**	**Événement**	DPI
Dimanche	9h00–17h00	TSH B128	Atelier sur la biostatistique	N
Dimanche	9h00–17h00	TSH B105	Atelier en statistique industrielle	Y
Dimanche	9h00–17h00	KTH B135	Atelier sur la méthodologie d'enquête	N
Dimanche	18h00–22h00	USC Marketplace	Réception	
Dimanche	18h00–22h00	USC Marketplace	Séance d'affichage	
Lundi	8h30–10h00	CNH 104	Séance 1 : Bienvenue et allocution de l'invité du président de la SSC	N
Lundi	8h30–16h00	USC Marketplace	Séance d'affichage	
Lundi	10h30–12h00	TSH B105	Séance 2 : La statistique directionnelle	N
Lundi	10h30–12h00	TSH B128	Séance 3 : La statistique et la qualité de l'eau	N
Lundi	10h30–12h00	TSH B106	Séance 4 : Probabilité	N
Lundi	10h30-12h00	KTH B135	Séance 5 : Études de cas I – Traitement des données manquantes	Y
Lundi	13h30–15h00	TSH B105	Séance 6 : Phylogénétique	N
Lundi	13h30–15h00	TSH B128	Séance 7 : Séance conjointe des groupes de statistique industrielle et de méthodologie d'enquête	N
Lundi	13h30–15h00	KTH B105	Séance 8 : Inférence statistique	N
Lundi	13h30–15h00	TSH B106	Séance 9 : Applications statistiques	Y
Lundi	13h30–15h00	KTH B135	Séance 10 : Réunion publique du CRSNG	Y
Lundi	15h30–17h00	TSH B128	Séance 11 : Méthodes d'enquête : applications	Y
Lundi	15h30–17h00	TSH B106	Séance 12 : Méthodes non paramétriques et estimation de densité	Y
Lundi	15h30–17h00	TSH B105	Séance 13 : Comportement multi-échelle dans les modèles stochastiques	N
Lundi	15h30–17h00	KTH B135	Séance 14 : La statistique pour l'analyse des données des microarrays	Y
Lundi	18h00	Royal Botanical Gardens	Banquet	
Mardi	8h30–10h00	TSH B106	Séance 15 : La statistique officielle	Y
Mardi	8h30–10h00	KTH B135	Séance 16 : Méta-analyse et essais cliniques	Y
Mardi	8h30–9h15	TSH B128	Séance 17 : Allocution du lauréat du prix Pierre Robillard	Y
Mardi	9h15–10h00	TSH B128	Séance 18 : Allocution du lauréat du prix de la Revue canadienne de statistique	Y
Mardi	8h30–10h00	TSH B105	Séance 19 : Forage de données dans la découverte des médicaments	Y

DPI : disponibilité d'un projecteur d'image-écran

Jour	**Temps**	**Endroit**	**Événement**	DPI
Mardi	10:30–12:00	TSH B105	Séance 20 : Expériences avec les parcelles subdivisées dans l'industrie	Y
Mardi	10:30–12:00	TSH B106	Séance 21 : Modélisation stochastique, séries chronologiques et statistique spatial	Y
Mardi	10:30–12:00	TSH B128	Séance 22 : Statistique et ordre public	Y
Mardi	10:30–12:00	KTH B135	Séance 23 : La statistique et le mappage du cerveau	Y
Mardi	13:30–15:00	TSH B105	Séance 24 : Aspects théoriques de la méthode de Monte Carlo	Y
Mardi	13:30–15:00	KTH B135	Séance 25 : Mésométrie I	Y
Mardi	13:30–15:00	TSH B128	Séance 26 : Données longitudinales - theorie et applications	Y
Mardi	13:30–15:00	TSH B106	Séance 27 : Méthodes d'enquête - théorie	Y
Mardi	15:30–17:00	TSH B106	Séance 28 : Méthodes multidimensionels	N
Mardi	15:30–17:00	KTH B105	Séance 29 : Modélisation stochastique	N
Mardi	15:30–17:00	TSH B105	Séance 30 : L'échantillonnage spatial	Y
Mardi	15:30–17:00	KTH B135	Séance 31 : Les Méthodes Statistiques pour mesurer le risque professionnel	Y
Mardi	15:30–17:00	TSH B128	Séance 32 : Accreditation statistique. Séance sur les progrès accomplis et discussion	N
Mardi	16:30–18:30	Faculty Club Great Hall	Réception - University of Waterloo alumni	
Mardi	17:00–18:00	TSH B105	Assemblée générale annuelle de al SSC	
Mardi	18:30–20:30	Refectory Patio	Barbeque	
Mercredi	8:30–10:00	TSH B105	Séance 33 : Groupe de statistique industrielle et de gestion allocution sur invitation spéciale	Y
Mercredi	8:30–10:00	TSH B128	Séance 34 : Inférence statistique pour les modèles de mélange	Y
Mercredi	8:30–10:00	KTH B135	Séance 35 : Analyse de survie	N
Mercredi	8:30–10:00	TSH B106	Séance 36 : Méthodes robustes, détection de valeurs aberrantes et rééchantillonnage	N
Mercredi	10:30–12:00	TSH B105	Séance 37 : Modélisation financière	Y
Mercredi	10:30–12:00	KTH B135	Séance 38 : Mésométrie II	N
Mercredi	10:30–12:00	TSH B128	Séance 39 : Études de cas II – cancer du col de l'utérus	Y
Mercredi	10:30–12:00	TSH B106	Séance 40 : Nouveaux résultats de recherche dans les méthodes d'analyse pour les donnés d'enquête	Y

DPI : disponibilité d'un projecteur d'image-écran

Jour	**Heure**	**Endroit**	**Événement**	DPI
Mercredi	13h30–15h00	TSH B106	Séance 41 : Index statistiques	N
Mercredi	13h30–15h00	TSH B105	Séance 42 : Recherche opérationnelle stochastique	Y
Mercredi	13h30–15h00	KTH B135	Séance 43 : Analyse de survie des données d'études cas-témoins et d'études cas-cohorte	Y
Mercredi	13h30–15h00	TSH B128	Séance 44 : Probabilité et inférence statistique	N
Mercredi	15h30–17h00	KTH B135	Séance 45 : Processus ponctuels et applications	N
Mercredi	15h30–17h00	TSH B128	Séance 46 : Modèles linéaires et plan d'expérience	N
Mercredi	15h30–17h00	TSH B106	Séance 47 : Inférence statistique	Y
Mercredi	15h30–17h00	TSH B105	Séance 48 : La statistique en finance et en marketing	Y

DPI : disponibilité d'un projecteur d'image-écran

8 Scientific Programme • Programme Scientifique

Sunday, May 26th/Dimanche 26 mai

9:00–17:00 Biostatistics Workshop/Atelier sur la biostatistique TSH B128

Design and analysis of cluster randomization trials
Conception et analyse d'essais de randomisation par groupes
Allan Donner, University of Western Ontario and/et and Neil Klar, Cancer Care Ontario

9:00–17:00 Workshop on Industrial Statistics/Atelier en statistique industrielle TSH B105

Design and analysis of computer experiments for engineering
Conception et analyse d'expériences informatiques pour le domaine de l'ingénierie
Jerome Sacks, Duke University and/et William J. Welch, Univers ity of Waterloo

9:00–17:00 Workshop on Survey Methodology/Atelier sur la méthodologie d'enquête KTH B135

Handling missing data
Traitement de données manquantes
Karla Nobrega and/et David Haziza, Statistics Canada/Statistique Canada

18:00 Poster Session/Séance d'affichage USC Marketplace

Luc Adjengue and/et Soumaya Yacout, École Polytechnique de Montréal
Parameter estimation for a production process
Estimation des paramètres d'un procédé de fabrication

Ejaz Ahmed, Sana S. Buhamra and/et Noriah M. Al-Kandari, University of Regina
Inference concerning quantile for left truncated and right censored data
Inférence concernant les quantiles pour des données tronquées à gauche et censurées à gauche

Gemai Chen, University of Calgary and/et Jinhong You, University of Regina
A modified likelihood ratio test for a mixed treatment effect
Une test du maximum de vraisemblance modifié pour l'effet d'un traitement mixte

Yun Hee Choi and/et David E. Matthews, University of Waterloo
A little known property of the multivariate survivor function
Une propriété peu connue de la fonction de survie multivariée

Laura Cowen and/et Carl J. Schwarz, Simon Fraser University
Adjusting radio-telemetry data for tag failure: a case study
Ajustement de données radio-télémétriques pour des temps de panne : une étude de cas

Sujay Datta, Northern Michigan University
Improved sequential estimation under LINEX Loss: asymptotics and simulation studies
Estimation séquentielle améliorée sous la perte LINEX : étude de l'asymptotique et de simulations

Audrey Fu, University of British Columbia
Extreme values in random precipitation fields
Valeurs extrêmes dans le domaine des précipitations aléatoires

Yuejiao Fu, University of Waterloo
Use of computer image features for discriminating between pathologic nuclear grade groups for breast ductal carcinoma in situ
Utilisation des propriétés d'images informatiques pour discriminer les différents groupes pathologiques nucléaires pour le Ductal Carcinoma in Situ des seins

Isabelle Gaboury, Centre d'étude systématique Thomas C. Chalmers
Analysis of 2×2 tables with both completely and partially observed data
Analyse de tables 2×2 avec données complètement et partiellement observées

Sohee Kang and/et John Hsieh, University of Toronto
Prediction of Ontario HIV/AIDS diagnosis incidence using a back projection method
Prévision de l'incidence du diagnostic du VIH/SIDA en Ontario par la méthode de projection dans le passé

Bashir Khan, Saint Mary's University and/et S.E. Ahmed, University of Regina
Improved estimation of coefficient vectors in regression models when the constraints are of a stochastic nature.
Estimation améliorée du vecteur des coefficients dans un modèle de régression quand les contraintes sont de nature stochastiques

Ying MacNab and/et Zhenguo Qiu, University of British Columbia
Bayesian hierarchical modelling of neonatal mortality: the Neonatal Health Services in Canada Project
Modélisation bayésienne hiérarchique pour la moralité néo-natale : le projet de services de santé néo-nataux au Canada

Catherine Njue, Cancer Care Manitoba.
Efficiency of testing procedures for multivariate longitudinal data
Convergence des procédures de tests pour des données longitudinales multivariées

Gelila Tilahun, Andrey Feuerverger, University of Toronto and/et Peter Hall, Australian National University
The dating of medieval documents
La datation de documents médiévaux

Guangyong Zou and/et Allan Donner, University of Western Ontario, and/et Neil Klar, Cancer Care Ontario

Interim analyses of cluster randomization trials with binary outcomes

Analyses intérimaires d'essais aléatoires de groupes ave réponses dichotomiques

Monday, May 27th/Lundi 27 mai

8:30–10:00 Session/Séance 1 **CNH 104**

Welcome and SSC Presidential Invited Address/Bienvenue et allocution de l'invité du président de la SSC
Special Session/Conférence spéciale
Organizer and Chair/Responsable et président : David Brillinger, University of California, Berkeley

Stephen Stigler, University of Chicago.
Risk and Revolution: Casanova, Napoleon, and the Loterie de France
Révolution et risque : Casanova, Napoléon et la Loterie de France

10:30–12:00 Session/Séance 2 **TSH B105**

Directional Statistics/La statistique directionnelle
Invited Paper Session/Conférences sur invitation
Organizer and Chair/Responsable et président : Peter Kim, University of Guelph

10:30 Jean-François Angers, Université de Montréal, and/et Peter Kim, University of Guelph
Symmetry and Bayesian function estimation on Riemanian manifolds
Symétrie et estimation bayésienne d'une fonction sur une variété riemanienne

11:00 Ted Chang, University of Virginia
Regression models for Stieffel manifolds
Modèles de régression pour les comparaisons multiples de Stieffel

11:30 Tilmann Gneiting, University of Washington
Covariance models for dynamic space-time processes
Modèles de covariance pour des processus spatio-temporels dynamiques

10:30–12:00 Session/Séance 3 **TSH B128**

Statistics and Water Quality/La statistique et la qualité de l'eau
Invited Paper Session/Conférences sur invitation **Biostatistics/Biostatistique**
Organizer and Chair/Responsable et président : William Ross, Health Canada

10:30 Dan Krewski, University of Ottawa and/et William Ross, Health Canada
Managing health risks from drinking water
Gestion du risque sur la santé de l'eau potable

10:40 Abdel El-Shaarawi, National Water Research Institute
Monitoring and assessment of the quality of Canadian fresh waters with emphasis on the Great Lakes
Surveillance et évaluation de la qualité de l'eau douce au Canada avec emphase sur les grands lacs

11:20 Reza Modarres, George Washington University and/et Jade Freeman, EPA
Analysis of censored environmental data with Box-Cox transformations
Analyse de données censurées sur l'environnement à l'aide de transformations de Box-Cox

10:30–12:00 Session/Séance 4 **TSH B106**

Probability/Probabilité
Invited paper session/Conférences sur invitation **IMS**
Organizer and Chair/Responsable et président : Tom Salisbury, York University

10:30 Siva Athreya, Indian Statistical Institute, Martin Barlow, Richard Bass, and/et Edwin Perkins
Uniqueness for a class of degenerate stochastic differential equations arising from super-processes.
Sur l'unicité pour une classe d'équations stochastiques différentielles dégénérées provenant d'un superprocessus

11:00 Rami Atar, Technion, and/et K. Burdzy
On Neumann eigenfunctions for some planar domains
Sur le problème des valeurs propres de Neumann dans certains domaines planaires

11:30 Ilie Grigorescu, University of Miami and/et Min Kang, Northwestern University
Scaling limit for a Fleming-Viot type system
Limite d'échelle pour un système de type Fleming-Viot

10:30–12:00 Session/Séance 5 **KTH B135**

Case Studies I - Treatment of Missing Data/Études de cas I - Traitement des données manquantes
Organizer and Chair/Responsable et président : Peggy Ng, York University

10:30 Patricia Whitridge, Statistics Canada/Statistique Canada
Analyzing Health Data with Missing Values
Analyse de données relatives à la santé avec des valeurs manquantes

10:50 Yin Liu, Peng Zhang, Mei-ting Chiang, Francicso Navaro Aguirre, York University

11:10 Yaqing Chen, Lena Zhang, Sun Tao, York University

11:30 Élyse Picard and Étienne Chassé-Saint-Laurent, Université de Montréal

13:30–15:00 Session/Séance 6 **TSH B105**

Phylogenetics/Phylogénétique
Invited Paper Session/Conférences sur invitation
Organizer and Chair/Responsable et président : Chris Field, Dalhousie University

13:30 Ed Susko, Dalhousie University
Testing for rate variation in phylogenetic subtrees
Tests pour la variation du taux dans les arbres phylogénétiques

14:15 Bret Larget, Donald L. Simon, Duquesne University, and/et Joseph B. Kadane, Carnegie Mellon University
Bayesian phylogenetic inference from animal mitochondrial genome arrangements
Inférence phylogénétique bayésienne à partir d'arrangements du génome mitochondrique d'animaux

13:30–15:00 Session/Séance 7 **TSH B128**

Joint Industry Survey Methods Session/Séance conjointe des groupes de statistique industrielle et de méthodologie d'enquête
Invited Paper Session/Conférences sur invitation **BISS/Survey Methods**
Organizer/Responsable : Bovas Abraham, University of Waterloo and/et Narasimha Prasad, University of Alberta
Chair/président : Narasimha Prasad, University of Alberta

13:30 Joseph Farruggia, ACNielsen Canada
Statistical applications in marketing reseach
Applications statistiques dans des recherches en marketing

14:15 Fernando Camacho, DAMOS
Analysis of life experiments with interventions
Analyse des expériences sur le temps de vie avec interventions

13:30–15:00 Session/Séance 8 **KTH B105**

Statistical Inference/Inférence statistique
Invited Paper Session/Conférences sur invitation
Organizer and Chair/Responsable et président : Marc Moore, École polythecnique de Montréal

13:30 Kris Klaassen, University of Amsterdam
Bonus-malus in acceptance sampling on attributes
Système bonus-malus dans l'échantillonnage pour l'acceptation des attributs

14:00 Boris Levit, Queen's University
On sequential edges recovery in image processing
Sur la détection séquentielle des bordures dans le traitement d'images

14:30 William Strawderman, Rutgers University, Dominique Fourdrinier, University of Rouen, and/et Martin Wells, Cornell University
On estimation of restricted mean vectors
Sur l'estimation d'un vecteur moyen restreint

13:30–15:00 Session/Séance 9 **TSH B106**

Applications of Statistics/Applications statistiques
Contributed Paper Session/Communications libres
Chair/Président : Rolf Turner, University of New Brunswick

13:30 Douglas Dover and/et Subhash Lele , University of Alberta
Elicited data and incorporation of expert opinion in ecological studies
Données illicites et l'inclusion de l'opinion des experts dans des études écologiques

13:45 Francois WATIER, Université de Sherbrooke and/et Jean VAILLANCOURT, Université du Québec à Hull
Optimal mean-variance portfolio strategy in a multiperiod setting
Stratégie optimale du portefeuille de moyenne-variance dans un contexte multipériode

14:00 Yongmin YU and/et Román VIVEROS-AGUILERA, McMaster University
CUSCORE charts for detecting sine wave signals in an autocorrelated process
Diagramme de CUSCORE pour détecter les signaux des ondes sinus dans un processus autocorrélé

14:15 Pierre DUTILLEUL, Bernard PELLETIER, Guillaume Larocque, and/et James W. Fyles, Université McGill
Multi-scale redundancy analysis of multivariate spatial data: I. Methodological Aspects
Analyse de redondance multi-échelles de données spatiales multivariées: I. Aspects Méthodologiques

14:30 Bernard PELLETIER, Pierre DUTILLEUL, Guillaume Larocque, and/et James W. Fyles, Université McGill
Multi-scale redundancy analysis of multivariate spatial data : II. Case study of farm management in Malawi
Analyse de redondance multiéchelles de données spatiales multivariées : II. étude de cas dans la gestion des exploitations agricoles au Malawi

13:30–15:00 Session/Séance 10 **KTH B135**

NSERC Open Meeting/Réunion publique du CRSNG
Judie FOSTER, NSERC/CSNRG
Bruce SMITH, Dalhousie University
How to complete a winning NSERC proposal
Comment remplir une demande CRSNG gagnante

15:30–17:00 Session/Séance 11 **TSH B128**

Applied Survey Methods/Méthodes d'enquête : applications
Contributed Paper Session/Communications libres
Chair:/Président : Patricia WHITRIDGE, Statistics Canada/Statistique Canada

15:30 Martin ST-PIERRE and/et Yves BÉLAND, Statistics Canada/Statistique Canada
Imputation of proxy respondents in the Canadian Community Health Survey
Imputation des répondants par procuration dans l'Enquête sur la santé dans les collectivités canadiennes

15:45 David MACNEIL and/et Stuart PURSEY, Statistics Canada/Statistique Canada
Dealing with industry misclassifications in the Unified Enterprise Survey
Traiter des mauvaises classifications d'industries dans le sondage sur les entreprises unifiées

16:00 Jean-Francois DUBOIS and/et Michelle Simard, Statistics Canada/Statistique Canada
Weighting challenges for the Longitudinal Survey of Immigrants to Canada (LSIC)
Pondération de l'Enquête longitudinale auprès des immigrants au Canada (ELIC)

16:15 Steven Matthews and/et Hélène Bérard,Statistics Canada/Statistique Canada
The outlier detection and treatment strategy for the monthly Wholesale and Retail Trade Survey of Statistics Canada
La stratégie de détection et de traitement des valeurs aberrantes pour l'Enquête mensuelle sur le commerce de gros et de détail de Statistique Canada

16:30 Fritz Pierre and/et Yves Béland, Statistics Canada/Statistique Canada
On response errors in the Canadian Community Health Survey
Étude sur les erreurs de réponse dans le cadre de l'Enquête sur la santé dans les collectivités canadiennes

15:30–17:00 Session/Séance 12 **TSH B106**

Nonparametric Methods and Density Estimation/Méthodes non paramétrique et estimation de densité
Contributed Paper Session/Communications libres
Chair/Président : Paul Cabilio, Acadia University

15:30 Gengsheng Qin, Georgia State University and/et Xiao-hua Zhou, Indiana University
A better confidence interval for the difference between two binomial proportions of paired data
Un meilleur intervalle de confiance pour la différence entre deux proportions binomiales de données pairées

15:45 Mayer Alvo, Université d'Ottawa and/et Paul Smrz, University of Newcastle, Australia
An arc model for analyzing ranking data
Un modèle basé sur l'arc pour l'analyse de données ordonnées

16:00 Yogendra Chaubey and/et Pranab K. Sen, Concordia University
Another look at the kernel density estimator for non-negative support
Une nouvelle vision de l'estimation de la densité par la méthode du noyau pour un support non négatif

16:15 Alan Ker, University of Arizona
Nonparametric estimation of possibly similar densities
Estimation non paramétrique de densités possiblement similaires

16:30 Majid Mojirsheibani, Carleton University
An empirical Csorgo-Horvath type CLT for Lp norms of density estimates
Une approche de Csorgo-Horvath empirique du théorème de la limite centrale pour des estimations de densités avec la norme Lp

16:45 R. Karunamuni and/et T. Alberts, University of Alberta
A semiparametric method of boundary correction for kernel density estimation
Une méthode semi-paramétrique pour la correction des frontières dans l'estimation d'une densité par la méthode du noyau

15:30–17:00 Session/Séance 13 **TSH B105**

Multiscale Behavior in Stochastic Models in Population Biology and Physics

Comportement multi-échelle dans les modèles stochastiques dans la biologie de la population et la physique

Invited Paper Session/Conférences sur invitation **Bernoulli Society**
Organizer/Responsable : Donald Dawson, Carleton University
Chair/Président : Reg Kulperger, University of Western Ontario

15:30 Ed Waymire, Oregon State University
Multiplicative cascades and partial differential equations
Cascades multiplicatives et équations différentielles partielles

16:15 Donald Dawson, Carleton University
Hierarchical approach to multiscale phenomena in stochastic population models
Approche hiérarchique d'un phénomène à échelle multiple dans un modèle de population stochastique

15:30–17:00 Session/Séance 14 **KTH B135**

Statistics for Microarray Data Analysis/La statistique pour l'analyse des données des microréseaux

Invited Paper Session/Conférences sur invitation
Organizer and Chair/Responsable et président : Shelley Bull, Samuel Lunenfeld Research Institute

15:30 Michael Newton and C.M. Kendziorski, University of Wisconsin at Madison
On mixture model calculations for gene expression analysis
Sur les calculs reliés aux modèles de mélanges de lois dans des analyses pour l'expression des gènes

16:00 Hugh Chipman, University of Waterloo and/et Rob Tibshirani, Stanford University
Hybrid Hierarchical Clustering With Applications To Microarray Data
Regroupements hiérarchiques hybride avec applications à des données de microréseaux

16:30 Terry Speed and/et Yongchao Ge, University of California, Berkeley
Multiple testing in large-scale gene expression experiments
Tests multiples dans des plans d'expérience à grande échelle sur l'expression des gènes

18:30–22:00 Banquet **Royal Botanical Gardens**

Tuesday, May 28th/Mardi 28 mai

8:30–10:00 Session/Séance 15 **TSH B106**

Official Statistics/La statistique officielle
Invited Paper Session/Conférences sur invitation
Caucus for Women in Statistics and/et Committee on Women in Statistics/Comité sur les femmes en statistique
Organizer and Chair/Responsable et président : Nadia Ghazzali, Université Laval and/et Cyntha Struthers, University of Waterloo

8:30 Sylvie Michaud, Statistics Canada/Statistique Canada
Methodological issues in producing income statistics
Problèmes méthodologiques des statistiques sur le revenu

9:00 Louise Bourque, l'Institut de la statistique du Québec
Citizen security and personal information confidentiality: governmental statistical organizations approach.
La sécurité et la confidentialité des renseignements personnels ou sensibles sur les citoyens : l'approche des organismes statistiques gouvernementaux

9:30 Denise Lievesley, UNESCO Institute for Statistics
Improving the quality of cross-national data - meeting the challenge
Amélioration de la qualité des données transnationales - relever le défi

8:30–10:00 Session/Séance 16 **KTH B135**

Meta-analysis and Clinical Trials/Méta-analyse et essais cliniques
Contributed Paper Session/Communications libres
Chair/Président : Mik Bickis, University of Saskatchewan

8:30 Emma Bartfay, Queen's University
Internal validation versus external validation: A case study
Validation interne contre validation externe : une étude de cas

8:45 Nicholas Barrowman, Thomas C. Chalmers Centre for Systematic Reviews
Statistical methods for detecting non-statistical bias
Méthodes statistiques pour la détection du biais non statistique

9:00 Shagufta Sultan, Health Canada and/et Robert Platt, McGill University
Probabilistic assessment of study quality in meta-analysis
Évaluation probabiliste de la qualité d'étude en méta-analyse

9:15 Gerarda Darlington, University of Guelph, and/et Neil Klar, Cancer Care Ontario
Assessing change: applications of analysis of covariance to data from cluster randomization trials
Évaluer le changement : applications d'une analyse de covariance de données provenant d'essais aléatoires groupés

9:30 Jianhua Liu, Dongsheng Tu and/et Joe Pater, Queen's University
A comparative analysis of quality of life data from a clinical trial in patients with advanced breast cancer
Une analyse comparative de données sur la qualité de vie d'une étude clinique faite sur des patients ayant un cancer du sein avancé

9:45 Gregory Pond, Princess Margaret Hospital
Using likelihood methods for phase II clinical trials
Utilisation de méthode de vraisemblance pour des essais cliniques de phase II

8:30–9:15 Session/Séance 17 **TSH B128**

Pierre Robillard Award Winner Lecture/Allocution du lauréate du prix Pierre Robillard
Chair/Président : Michael Evans, University of Toronto

9:15–10:00 Session/Séance 18 **TSH B128**

Canadian Journal of Statistics Award Winner Lecture/Allocution du lauréate du prix de la Revue canadienne de statistique
Chair/président: Jack Kalbfleisch, University of Michigan

8:30–10:00 Session/Séance 19 **TSH B105**

Data Mining in Drug Discovery/Forage de données dans la découverte des médicaments
Invited Paper Session/Conférences sur invitation
Organizer and Chair/Responsable et président : Hugh Chipman, University of Waterloo

8:30 Raymond Lam, GlaxoSmithKline, William Welch, University of Waterloo and/et Stanley Young
Design and analysis of large chemical databases
Plan d'expérience et analyse d'un large jeu de données chimiques

9:00 Marcia Wang, Hugh A. Chipman and/et William J. Welch, University of Waterloo
Tree-averaging models for high throughput screening data
Modèles de moyennage par arbre pour des données de filtrage à débit élevé

9:30 David Cummins and Richard E. Higgs, Eli Lilly and Company
Molecular diversity: statistical perspectives and approaches
Diversité moléculaire : approches et perspectives statistiques

10:30–12:00 Session/Séance 20 **TSH B105**

Split Plot Experiments in Industry/Expériences avec les parcelles subdivisées dans l'industrie

Invited Paper Session/Conférences sur invitation **BISS**

Organizer and Chair/Responsable et président : John Brewster, University of Manitoba

10:30 Geoff Vining, Virginia Tech

Lack-of-fit tests for industrial split-plot experiments

Manque d'ajustement des tests pour les expériences industrielles en subdivision de parcelles

11:15 Robert McLeod, University of Manitoba

Design and analysis of blocked fractional factorial split-plot designs

Plan d'expérience et analyse de plans d'expérience factoriel fractionnaire avec subdivision de parcelles par bloc

10:30–12:00 Session/Séance 21 **TSH B106**

Stochastic Modeling, Time Series and Spatial Statistics/Modélisation stochastique, séries chronologiques et statistique spatial

Contributed Paper Session/Communications libres

Chair/Président : Ed Susko, Dalhousie University

10:30 Dingan Feng, York University

A class of stochastic conditional duration models for financial transaction data

Une classe de modèles stochastiques conditionnels sur la durée pour des données de transactions financières

10:45 Theodoro Koulis, University of Waterloo

Modeling sea ice concentrations with the biased voter model

Modélisation des concentrations de glaces avec le modèle de l'électeur biaisé

11:00 Rachel MacKay, University of British Columbia

Estimation of the order of a hidden Markov model

L'estimation de l'ordre d'une chaîne de Markov cachée

11:15 Paramjit Gill, Okanagan University College

Bayesian modelling for social networks data

Modélisation bayésienne de données de réseaux sociaux

11:30 Jeffrey Picka, University of Maryland

Structural analysis of stationary germ grain models

Analyse de la structure de modèles stationnaires de germes de grain

11:45 Janusz Kawczak and/et Stanislav Molchanov, UNCC

Effective estimation in CLT for the Markov chains with Doeblin condition

Estimation efficace du théorème de la limite centrale pour les chaînes de Markov sous la condition de Doeblin

10:30–12:00 Session/Séance 22 **TSH B128**

Statistics and Public Policy/Statistique et ordre public
Invited Paper Session/Conférences sur invitation
Organizer and Chair/Responsable et président : David Brillinger, University of California, Berkeley

Miron Straf, National Academy of Sciences
Information for public policies
Informations d'ordre public

10:30–12:00 Session/Séance 23 **KTH B135**

Statistics and Brain Mapping/La statistique et le mappage du cerveau
Invited Paper Session/Conférences sur invitation **Biostatistics/Biostatistique**
Organizer and Chair/Responsable et président : Keith Worsley, McGill University

10:30 Pedro Valdes, Eduardo Martinez and Nelson Trujillo, Cuban Neuroscience Center
Image fusion for concurrently recorded spontaneous EEG and fMRI
Fusion d'images spontanées obtenues de façon concurrentielle par EEG et fMRI

11:30 Moo Chung, University of Wisconsin-Madison
How to smooth surface data ?
Comment lisser des données de surface?

13:30–15:00 Session/Séance 24 **TSH B105**

Theoretical Aspects of Monte Carlo/Aspects théoriques de la méthode de Monte Carlo
Invited Paper Session/Conférences sur invitation **IMS**
Organizer and Chair/Responsable et président : Neal Madras, York University

13:30 Duncan Murdoch, University of Western Ontario
Perfect sampling algorithms: connections
Algorithmes d'échantillonage parfait : les liens

14:00 John Yuen, York University and/et G. Roberts, Lancaster University
Optimal scaling of random walk Metropolis algorithms
Échelle optimale pour l'algorithme de marche aléatoire de Métropolis

14:30 Radu Craiu, University of Toronto
Getting perfect more quickly: speed-up methods for perfect sampling algorithms
Comment devenir parfait au plus vite : des méthodes d'accélération pour les algorithmes d'échantillonage parfait

13:30–15:00 Session/Séance 25 **KTH B135**

Environmetrics I/Mésométrie I
Invited Paper Session/Conférences sur invitation
Organizer and Chair/Responsable et président : Abdel El-Shaarawi, National Water Research Institute

13:30 Elena Naumova, Tufts University
Statistical framework for a waterborne infectious outbreak
Cadre statistique pour la détection, la description et la prédiction des épidémies de maladies transmises par l'eau

14:00 Megu Ohtaki, Hiromi Kawasaki, and/et Kenichi Satoh, Hiroshima University
Visualization of time and geographical distribution of cancer mortality in Japan
Visualisation de la distribution du cancer au Japon dans le temps et géographiquement

14:30 Teresa Alpuim, University of Lisbon and/et Abdel El-Shaarawi, National Water Research Institute
Linear regression with correlated residuals. An application to monthly temperature measurements in Lisbon.
Régression linéaire avec des erreurs corrélés. Une application à une série de températures mensuelles à Lisbonne.

13:30–15:00 Session/Séance 26 **TSH B128**

Longitudinal Data - Theory and Applications/Données longitudinales - theorie et applications
Contributed Paper Session/Communications libres
Chair/Président : K.C. Carriere, University of Alberta

13:30 Francois Bellavance, l'École des Hautes Études Commerciales, Serge Tardif, Université de Montréal, and/et Constance van Eeden, University of British Columbia
A nonparametric procedure for the analysis of balanced crossover designs
Tests non paramétriques pour l'analyse de données issues de plans croisés équilibrés

13:45 Grace Yi and/et Mary Thompson, University of Waterloo
A likelihood-based method for analyzing longitudinal binary responses with informative dropouts
Une méthode basée sur la vraisemblance pour l'analyse longitudinale de réponses dichotomiques avec abandons informatifs

14:00 John Holt, and/et O. Brian Allen , University of Guelph
Performance of nonparametric estimates in Poisson time series with small counts
Performance des estimateurs non paramétriques pour des séries chronologiques de Poisson avec faibles fréquences

14:15 Wenjiang Fu, Michigan State University
Consistent estimation in age-period-cohort analysis
Estimateurs convergents dans une analyse de cohorte âge-période

14:30 Shenghai Zhang and/et Mary E. Thompson, Unversity of Waterloo
Finite Sample Properties in Using GEE
Propriétés des échantillons finis par l'utilisation des équations d'estimation généralisées

14:45 Andrea Benedetti and/et Michal Abrahamowicz, McGill University
Searching for thresholds in a simulation study
Recherche de seuils dans une étude de simulations

13:30–15:00 Session/Séance 27 **TSH B106**

Theoretical Survey Methods/Méthodes d'enquête - théorie
Contributed Paper Session/Communications libres
Chair/Président : Wesley Yung, Statistics Canada/Statistique Canada

13:30 Patrick Farrell and/et Sarjinder Singh, Carleton University
Recalibration of higher-order calibration weights
Recalibration des poids de calibration d'ordre élevé

13:45 Yong You, Statistics Canada/Statistique Canada and/et J.N.K. Rao, Carleton University
Benchmarking hierarchical Bayes small area estimators with application in census undercoverage estimation
Standardisation des estimateurs de Bayes hiérarchiques pour de petits domaines avec des applications à la sous-estimation dans un recensement

14:00 Abdellatif Demnati, Statistics Canada/Statistique Canada, and/et J.N.K. Rao, Carleton University
Linearization variance estimators for survey data
Estimateurs de la variance par linéarisation applicables aux données d'enquêtes

14:15 Emile Allie, Statistics Canada/Statistique Canada
Unrounding procedures for systematically rounded survey income data
Procédures exactes pour des données sur le revenus arrondies systématiquement

14:30 Sarjinder Singh, Carleton University and/et Raghunath Arnab, University of Durban-Westville
Estimation of variance from missing data
Estimation de la variance avec des données manquantes

15:30–17:00 Session/Séance 28 **TSH B106**

Multivariate Methods/Méthodes multidimensionels
Contributed Paper Session/Communications libres
Chair/Président : R.P. Gupta, Dalhousie University

15:30 Victor Nzobounsana and/et Dhorne Thierry, Université Rennes 2 - Haute Bretagne
Sensitivity to the criterion in generalized canonical correlation analysis.
Influence du critère sur les analyses canoniques généralisées

15:45 Lehana Thabane, McMaster University and/et Steve Drekic, University of Waterloo
Hypothesis testing and power calculations under the generalized Bessel-type model
Tests d'hypothèses et calcul de puissance sous un modèle du type Bessel généralisé

16:00 François Perron, Université de Montréal and/et Eric Marchand, University of New Brunswick
Improving on the MLE of a bounded mean for spherical distributions
Sur l'estimation d'un paramètre de position borné pour des distributions à symétrie sphérique

16:15 Alexander de Leon and/et K.C. Carriere, University of Alberta
A generalization of the Mahalanobis distance to mixed quantitative and qualitative multivariate data
Une généralisation de la distance de Mahalanobis pour jumeler des données multidimensionnelles quantitatives et qualitatives

16:30 Abdulkadir Hussein and/et Edit Gombay, University of Alberta
Sequential comparison of two treatments by means of parametric tests
Comparaisons séquentielles de deux traitements par des tests paramétriques basés sur les moyennes

16:45 Veeresh Gadag and/et K. Jauakumar, Memorial University of Newfoundland
On a class of multivariate generalized exponential distributions.
Sur une classe de distributions exponentielles multidimensionnelles généralisées

15:30–17:00 Session/Séance 29 **KTH B105**

Stochastic Modeling/Modélisation stochastique
Invited Paper Session/Conférences sur invitation
Organizer and Chair/Responsable et président : Bruno Remillard, l'École des Hautes Études Commerciales

15:30 Jean Vaillancourt, Université du Québec à Hull and/et Susie Fortier, Statistique Canada/Statistics Canada
Associated random variables in the testing of Poisson-Voronoï tessalations
Variables aléatoires associées dans certains tests sur les mosaïques de Poisson-Voronoï

16:00 Christiane Lemieux, University of Calgary
Randomized quasi-Monte Carlo methods for multivariate integration
Méthodes quasi-Monte Carlo randomisées pour l'intégration multidimensionnelle

16:30 Miklós Csörgő, Barbara Szyszkowicz and/et Qiying Wang, Carleton University
Invariance principles for Studentized partial sum processes
Principes d'invariance pour des processus studentisés de sommes partielles

15:30–17:00 Session/Séance 30 **TSH B105**

Spatial Sampling/L'échantillonnage spatial
Invited Paper Session/Conférences sur invitation **Survey Methods/Méthodes d'enquête**
Organizer and Chair/Responsable et président : Subhash Lele, University of Alberta

15:30 Steve Cumming, Boreal Ecosystems Research Ltd, and/et Subhash Lele, University of Alberta
Forest structure and forest birds: a balanced, model-based design for multivariate glms.
Structure de la forêt et oiseaux : un plan d'expérience balancé basé sur un modèle pour des modèles linéaires généralisés multidimensionnels

16:00 Steve Thompson, Pennsylvania State University
On optimal spatial designs
Sur les plans d'expérience spatiaux optimaux

16:30 James Zidek, University of British Columbia, Nhu D. Le, BC Cancer Agency, and/et Li Sun, Ericsson Berkeley Research Center
Designs for predicting the extremes of spatial processes
Plan d'expérience pour prédire les valeurs extrêmes dans un processus dans l'espace

15:30–17:00 Session/Séance 31 **KTH B135**

Statistical Methodology for Occupational Risk Assessment/ Les Méthodes statistiques pour mesurer le risque professionnel
Invited Paper Session/Conférences sur invitation **CSEB/Biostatistics/Biostatistique**
Organizer and Chair/Responsable et président : Yang Mao, Health Canada

15:30 Nhu Le, BC Cancer Agency
Exposure assessment for studying relationships between air pollution and chronic diseases: A case-control study of lung cancer patients in British Columbia.
Estimation de l'exposition pour l'étude des relations entre les polluants atmosphérique et les maladies chroniques : une étude cas-témoin pour les patients atteint d'un cancer des poumons en Colombie-Britannique

16:00 Mik Bickis, University of Saskatchewan, Ugis Bickis and/et Tom Beardall, Phoenix OHC
Application of statistical principles to the determination of occupational health risk
Application des principes statistiques à la détermination du risque professionel pour la santé

16:30 Kyle Steenland, National Institute for Occupational Safety and Health, James A. Deddens and/et Siva Sivaganesan, University of Cincinnati
Issues in exposure-reponse models for occupational risk assessment
Les modèles dose-réponse pour mesurer le risque professionnel

15:30–17:00 Session/Séance 32 **TSH B128**

Statistics Accreditation, Progress report and discussion session/Accréditation statistique. Séance sur les progrès accomplis et discussion
Organizer and Chair : Ernest Enns, University of Calgary

Wednesday, May 29th/Mercredi 29 mai

8:30–10:00 Session/Séance 33 **TSH B105**

Business and Industry Section Special Invited Address/Groupe de statistique industrielle et de gestion : Allocution sur invitation spéciale
Invited Paper Session/Conférences sur invitation **BISS**
Organizer/Responsable : Bovas Abraham, University of Waterloo and/et Román Viveros-Aguilera, McMaster University
Chair/président: Bovas Abraham, University of Waterloo

John MacGregor, McMaster University.
The changing nature of data and its implications for applied statistics
Le changement dans la nature des données et ses implications en statistique appliquée

8:30–10:00 Session/Séance 34 **TSH B128**

Statistical Inference for Mixture Models/Inférence statistique pour les modèles de mélange
Invited Paper Session/Conférences sur invitation
Organizer/Responsable: Jiahua Chen, University of Waterloo and/et Jack Kalbfleisch, University of Michigan
Chair/président: Jack Kalbfleisch, University of Michigan

8:30 Jing Qin and/et Glenn Heller, Memorial Sloan-Kettering Cancer Center
A mixture model approach for finding informative genes in microarray studies
Une approche de mélange de lois pour trouver les gènes informatifs dans des études de microréseaux

9:00 Mary Lesperance, University of Victoria and/et Bruce Lindsay, Penn State University
On computing information in semiparametric mixture models
Sur le calcul de l'information contenue dans des mélanges semi-paramétriques

9:30 Cindy Fu, Jiahua Chen*, University of Waterloo and/et Jack Kalbfleisch, University of Michigan
A modified likelihood ratio test for a mixed treatment effect
Un test du maximum de vraisemblance modifié pour l'effet d'un traitement mixte

8:30–10:00 Session/Séance 35 **KTH B135**

Survival Analysis/Analyse de survie
Contributed Paper Session/Communications libres
Chair/Président : Judy-Anne Chapman, University of Waterloo

8:30 Yingwei Peng, Memorial University of Newfoundland
Cure models for survival data with a nonsusceptible fraction
Modèles de traitement pour des analyses de survie avec une fraction non susceptible

8:45 Karen Kopciuk, Samuel Lunenfeld Research Institute and/et David E. Matthews, University of Waterloo
Flexible regression models for three state progressive processes
Modèles de régression flexibles pour les processus progressifs à trois états

9:00 Wenqing He, Samuel Lunenfeld Research Institute Mt. Sinai Hospital and/et Jerry F. Lawless, University of Waterloo
Bivariate location-scale models for log lifetimes
Modèles de position-échelle bidimensionnel pour les logarithmes des durées de vie

9:15 Xuewen Lu, Agriculture and Agri-Food Canada and/et Radhey Singh, University of Guelph
A Class of Estimators for the Parameters in the Location-Scale Model with Censored Data
Une classe d'estimateurs pour les paramètres dans un modèle de position-échelle des données censurées

9:30 Thierry Duchesne, James E. Stafford, and/et Fasil Woldegeorgis, University of Toronto
Smooth and aggregated incidence rate estimation for interval-censored HIV data
Estimation lisse et agrégée du taux d'incidence du VIH pour des données censurées par intervalle

9:45 Sudhir Paul, University of Windsor and/et Uditha Balasoorya, Nayang Technological University
Some over-dispersed life time models and associated tests
Quelques modèles de temps de vie surdispersés et tests correspondants

8:30–10:00 Session/Séance 36 **TSH B106**

Robust Methods, Outlier Detection and Bootstrapping/Méthodes robustes, détection de valeurs aberrantes et rééchantillonnage.
Chair/Président : Christian Léger, Université de Montréal

8:30 Pierre Duchesne and/et Roch Roy, HEC-Montréal
Robust tests for independence of two time series
Tests robustes d'indépendance entre deux séries chronologiques

8:45 Jean-François Boudreau and/et Christian Léger, Université de Montréal
Bootstrap adaptive least trimmed squares
Moindres carrés tronqués adaptatifs par rééchantillonnage

9:00 Sharmila Banerjee and/et Boris Iglewicz, Temple University
Outliers in large data sets
Valeurs aberrantes dans de grands jeux de données

9:15 Ka Ho Wu and/et Siu Hung Cheung, Chinese University of Hong Kong
Bootstrapping simultaneous prediction intervals for autoregressive time series models
Bootstrap simultané d'intervalles de prédiction pour des modèles de séries chronologiques autorégressives

9:30 Pascal Croteau, Robert Cléroux and/et Christian Léger, Université de Montréal
Bootstrap confidence intervals for periodic replacement policies
Intervalles de confiance bootstrap pour les remplacements préventifs périodiques

10:30–12:00 Session/Séance 37 **TSH B105**

Financial Modeling/Modélisation financière
Invited Paper Session/Conférences sur invitation
Organizer and Chair/Responsable et président : Gary Parker, Genesis Development

10:30 Genevieve Gauthier, l'École des Hautes Études Commerciales, Jin-Chuan Duan, University of Toronto, Jean-Guy Simonato and/et Sophia Zaanoun, l'École des Hautes Études Commerciales
Maximum likelihood estimation of credit risk models
L'estimation par la méthode du maximum de vraisemblance pour des modèles de risque de crédit

11:15 Tom Hurd, McMaster University
Pricing derivatives in incomplete markets
Les prix dérivés dans des marchés incomplets

10:30–12:00 Session/Séance 38 **KTH B135**

Environmetrics II/Mésométrie II
Invited Paper Session/Conférences sur invitation
Organizer and Chair/Responsable et président : Abdel El-Shaarawi, National Water Research Institute

10:30 Cliff Spiegelman and/et Eun Sug Park, Texas A&M University
Nearly nonparametric multivariate density estimates that incorporate marginal parametric density information for the environment
Estimations de densités multivariées non paramétriques approximatives qui incluent de l'information sur la densité marginale paramétrique dans l'environnement

11:00 Timothy Haas, University of Wisconsin at Milwaukee
Nonlinear spatio-temporal statistics via Monte Carlo methods implemented in a Javaspaces distributed computer
Statistiques spatio-temporels non linéaires via des méthodes de Monte Carlo implémentées dans des ordinateurs avec JavaSpaces

11:30 Peter Guttorp, University of Washington, Doris Damian, Worcester Polytechnic Institute and/et Paul D. Sampson, University of Washington
Bayesian analysis of nonstationary spatial covariance
Analyse bayésienne de la covariance spatiale non stationnaire

10:30–12:00 Session/Séance 39 **TSH B128**

Case Studies II - Cervical Cancer/Études de cas II - cancer du col de l'utérus
Organizer and Chair/Responsable et président : Peggy Ng, York University

10:30 Al Covens, University of Toronto and/et Edmee Franssen, Toronto Sunnybrook Health Science Center

10:45 Xiaofei Shi and/et Wei Xu, Dalhousie University

11:00 Eshetu Atenafu and/et Sohee Kang, University of Toronto

11:15 Christine Calzonetti, Simo Goshev, Rongfang Gu, Shahidul Mohammad Islam, Amanda Lafontaine, Marcus Loreti, Maria Porco, William Volterman, Qihao Xie, McMaster University

11:30 Baktiar Hasan, Mark Kane, Melanie Laframboise, Michael Maschio, Andy Quigley, University of Guelph

11:45 Sumanth Sharatchandran, Sophia Lee, Shirin Yazdanian, Noa Rozenblit, York University

Poster Session/Séance d'affichage- Luz Palacios, Alberto Nettel-Aguirre, University of Calgary

10:30–12:00 Session/Séance 40 **TSH B106**

New research findings in analysis methods for survey data/Nouveaux résultats de recherche dans les méthodes d'analyse pour les donnés d'enquête
Invited Paper Session/Conférences sur invitation **Survey Methods/Méthodes d'enquête**
Organizer and Chair/Responsable et président : Georgia Roberts, Statistics Canada/ Statistique Canada

10:30 Christian Boudreau and/et Jerry Lawless, University of Waterloo
Proportional hazards models and ignorable sampling
Modèles à risques proportionnels et plans d'échantillonnage non informatifs

10:55 Brajendra Sutradhar, Memorial University and/et Milorad Kovacevic, Statistics Canada/Statistique Canada
Analysis of longitudinal survey data in the presence of missing outcomes
Analyse de données de sondage longitudinal en présence de réponses manquantes

11:20 Roland Thomas, Carleton University and/et Andre Cyr, Statistics Canada/Statistique Canada
Applying item response theory methods to complex survey data
Application de la théorie des éléments de réponse aux données d'enquête complexe

11:45 David Binder, Statistics Canada/Statistique Canada
Discussion

13:30–15:00 Session/Séance 41 **TSH B106**

Statistical Indexes/Index statistiques
Invited Paper Session/Conférences sur invitation **Survey Methods/Méthodes d'enquête**
Organizer and Chair/Responsable et président : Susana Bleuer, Statistics Canada/ Statistique Canada

13:30 Lenka Mach and/et Abdelnasser Saïdi, Statistics Canada/Statistique Canada
Development of the labour cost index at statistics Canada
Développement de l'Indice des coûts de main-d'oeuvre à Statistique Canada

14:00 Janice Lent, U.S. Bureau of Transportation Statistics and/et Alan Dorfman, U.S. Bureau of Labor Statistics
Using a weighted average of the Jevons and Laspeyres indexes to approximate a superlative index
Utilisation des indices podérés de Jevons and Laspeyres pour approximer un indice superlatif

14:30 Jack Triplett, The Brookings Institution
IT, Hedonic Price Indexes, and Productivity: International Comparability Issues
Technologie de l'information, indices de prix hédonique et productivité : Questions de comparaison internationale

13:30–15:00 Session/Séance 42 **TSH B105**

Stochastic Operations Research/Recherche opérationnelle stochastique
Invited Paper Session/Conférences sur invitation **CORS**
Organizer and Chair/Responsable et président : David Stanford, U. Western Ontario

13:30 Doug Down, McMaster University
Exact asymptotics for polling models
Asymptotique exacte pour des modèles de regroupements

14:00 Evelyn Richards and/et John A. Kershaw, University of New Brunswick
Using simulation to evaluate robustness of forest operations plans
Utilisation de simulations pour évaluer la robustesse des plans sur les opérations en forêts

14:30 Reg Kulperger, W.J. Braun and/et D.A. Stanford, University of Western Ontario
Modelling forest fires stochastically
Modélisation stochastique des feux de forêt

13:30–15:00 Session/Séance 43 **KTH B135**

Survival Analysis of Case-Control and Case-Cohort Data/Analyse de survie des données d'études cas-témoins et d'études cas-cohorte
Invited Paper Session/Conférences sur invitation **Biostatistics/Biostatistique**
Organizer and Chair/Responsable et président : Michal Abrahamowicz,McGill University

13:30 Jack Siemiatycki, Université de Montréal
The nature of and the methodological challenges in epidemiological case-control studies
La nature et les défis méthodologiques dans les études épidémiologique cas-contrôles.

14:00 Shaw-Hwa Lo, Columbia University and/et Kani Chen, Hong Kong University of Science and Technology
Some recent developments of case-cohort analysis with Cox's regression model
Développements récents dans l'analyse de cas-cohorte avec un modèle de régression de Cox

14:30 Karen Leffondré, Michal Abrahamowicz, McGill University, and/et Jack Siemiatycki, Université de Montréal
Comparison of Cox's model versus logistic regression for case-control data with time-varying exposure: a simulation study
Comparaison des modèles de Cox et de régression logistique pour des données cas-témoins dont l'exposition varie au cours du temps : une étude de simulation

13:30–15:00 Session/Séance 44 **TSH B128**

Probability and Statistical Inference/Probabilité et inférence statistique
Contributed Paper Session/Communications libres
Chair/Président : Mary Lesperance, University of Victoria

13:30 Adrienne Kemp, University of St Andrews, Scotland
True odds with a biased coin
Vrais probabilités avec une pièce de monnaie biaisée

13:45 Andre Dabrowski, Université d'Ottawa, H.G. Dehling, Bochum, T. Mikosch, Copenhagen and/et O. Sharipov, Uzbek Academy of Sciences
Poisson limits for U-statistics
Limites poissoniennes pour les U-statistiques

14:00 Mahmoud Zarepour, Université d'Ottawa and Dragan Banjevic, University of Toronto
A note on maximum autoregressive processes of order one
Une note sur le processus autorégressif maximum d'ordre un

14:15 Murray Jorgensen,University of Waikato and Roger Littlejohn, Invermay Agricultural Research Centre, New Zealand,
Evaluation of the asymptotic variance-covariance matrix for finite mixture distributions
L'estimation de la matrice des variances-covariances asymptotiques dans des modèles de mélanges finis

14:30 Tony Almudevar, Acadia University
A formula for the density of solutions to estimating equations
Une formule pour la densité des solutions d'une équation d'estimation

15:30–17:00 Session/Séance 45 **KTH B135**

Point Processes and Applications/Processus ponctuels et applications
Invited Paper Session/Conférences sur invitation
Organizer/Responsable: Reg Kulperger and/et John Braun, U. of Western Ontario
Chair/Président : John Braun, University of Western Ontario

15:30 Mark Bebbington, Massey University and/et Kostya Borovkov, U. of Melbourne
Stress release and transfer models for earthquakes
Stabilité du relâchement de tension et modèles de transferts pour des tremblements de terre.

16:00 David Vere-Jones, Victoria University of Wellington
Point process models and probability forecasts for earthquakes
Modèles basés sur les processus ponctuels et prévision de la probabilité d'un tremblement de terre

16:30 David Brillinger, University of California, Berkeley
Analyses of bivariate time series in which the components are sampled at different instants
Analyses de séries chronologiques bidimensionnelles dans lesquelles les composantes sont échantillonnées à des temps différents

15:30–17:00 Session/Séance 46 **TSH B128**

Linear Models and Design/Modèles linéaires et plan d'expérience
Contributed Paper Session/Communications libres
Chair/Président : Tony Almudevar, Acadia University

15:30 Denis Larocque, École des Hautes Études Commerciales and/et Isabelle Bussières, Université du Québec à Trois-Rivières
Aligned rank test for the bivariate randomized block model
Test de rangs alignés pour le plan de blocs aléatoires bivarié

15:45 B. M. Golam Kibria, Florida International University and/et A.Saleh, Carleton University
Effect of W, LR, and LM tests on the performance of preliminary test ridge regression estimators
Effet des tests de W, LR, et LM sur la performance des estimateurs de régression ridge pour des tests préliminaires

16:00 Jiaqiong Xu, Bovas Abraham and Stefan Steiner, University of Waterloo
Multivariate outlier detection
Détection de valeurs aberrantes dans un contexte multidimensionnel

16:15 Saumendranath Mandal and/et K.C. Carriere, University of Alberta
Optimal designs for model discrimination and fixed efficiency
Plans d'expérience optimaux pour la discrimination des modèles avec convergence fixe

16:30 Bipin Niraula, Tribhuvan University, Nepal
Biostatistics in south Asia region
Biostatistiques dans le sud de l'Asie

15:30–17:00 Session/Séance 47 **TSH B106**

Statistical Inference/Inférence statistique
Contributed Paper Session/Communications libres
Chair/Président : Radhey Singh, University of Guelph

15:30 Yuliya Martsynyuk, Carleton University
On weighted least squares and related estimators in linear functional error-in-variables models
Sur les moindres carrés pondérés et les estimateurs reliés dans les modèles linéaires fonctionnels avec erreur dans les variables linéaires

15:50 Xiaoming Sheng, A. Biswas and/et K.C. Carriere, University of Alberta
Incorporating inter-item correlations for item response data analysis
Incorporation de corrélations inter-items pour l'analyse de données de réponses par item

16:10 Changchun Xie, Anthony Desmond and Radhey Singh, University of Guelph
Hierarchical quasi-likelihoods and their applications to hierarchical generalized linear models (hglms) and survival models with frailty
Quasi-vraisemblance hiérarchique et ses applications aux modèles linéaires généralisés et modèles de survie avec effets aléatoires

16:30 Ying MacNab, University of British Columbia
Hierarchical Bayesian estimation in conditional autoregressive disease mapping models
Estimation bayésienne hiérarchique dans des modèles autorégressif conditionnels pour le marquage d'une maladie

15:30–17:00 Session/Séance 48 **TSH B105**

Statistics in Finance and Marketing/La statistique en finance et en marketing
Invited Paper Session/Conférences sur invitation **BISS**
Organizer and Chair/Responsable et président : Alison Burnham, GE Capital

15:30 Daymond Ling, CIBC
Persuading people to trust a model
Persuader les gens à faire confiance à un modèle

16:00 Anthony Percaccio, Municipal Property Assessment Corporation
AVM: how to value over 3 million properties . . . every month!
AVM : Comment évaluer 3 millions de propriétés ...tous les mois!

16:30 Mark Merritt, TransUnion of Canada
Credit scoring and the use of statistics in the world of consumer lending
Score de crédit et utilisation des statistiques dans le monde du prêt aux consommateurs

9 Abstracts • Resumés

Workshops/Ateliers

Sunday, May 26th/Dimanche 26 mai, 9:00-17:00 **TSH B128**

Design and analysis of cluster randomization trials
Conception et analyse d'essais de randomisation par groupes
Allan Donner, University of Western Ontario and/et and Neil Klar, Cancer Care Ontario

This course presents a systematic and unified treatment of comparative trials which randomize intact social units, or clusters of individuals, to different intervention groups. Such trials have become particularly widespread in the evaluation of nontherapeutic interventions, including lifestyle modification, educational programmes and innovations in the provision of health care. Their increasing popularity over the last two decades has led to an extensive body of methodology and a growing, but somewhat scattered, literature that cuts across several disciplines in the statistical, social and medical sciences. We will integrate this material into a full day course which emphasizes applications to health research. The overall prerequisite for the course is knowledge of the fundamentals of biostatistics and familiarity with the basic principles of design and analysis of clinical trials. The sequence of topics presented will be based on the recently published text entitled Design and Analysis of Cluster Randomization Trials in Health Research by Allan Donner and Neil Klar 2000, (Arnold Publishing Company, London, 2000).

Ce cours présente un traitement systématique et unifié des essais comparatifs qui randomisent des unités sociales intactes,ou des groupes d'individus, à différents groupes d'intervention. De tels essais sont devenus particulièrement fréquents dans l'évaluation d'interventions non thérapeutiques, y compris les changements de style de vie, les programmes éducatifs et les innovations dans la prestation des soins de santé. Leur popularité croissante au cours des deux dernières décennies a mené à un vaste répertoire de méthodologies et à une littérature croissante, mais quelque peu éparpillée, qui couvre plusieurs disciplines des sciences statistiques, sociales et médicales. Nous intégrerons ce matériel à un cours d'une journée qui met l'accent sur les applications dans le domaine de la recherche sur la santé. Les qualifications préalables pour le cours sont une connaissance des principes de biostatistique et des principes de base de la conception et de l'analyse d'essais cliniques. La séquence des sujets présentés sera fondée sur un ouvrage récemment publié, qui s'intitule Design and Analysis of Cluster Randomization Trials in Health Research (Conception et analyse d'essais de randomisation de groupes dans la recherche sur la santé), par Allan Donner et Neil Klar (Arnold Publishing Company, London, 2000).

Sunday, May 26th/Dimanche 26 mai, 9:00-17:00 **TSH B105**

Design and analysis of computer experiments for engineering
Conception et analyse d'expériences informatiques pour le domaine de l'ingénierie
Jerome Sacks, Duke University and/et William J. Welch, University of Waterloo

Computer models are now widespread in engineering: they save development time and cost relative to physical experiments. Electrical engineers use circuit simulators and mechanical engineers have finite-element models, for example. The workshop will provide an overview of the experimental design and statistical modelling strategies that have been developed specifically for computer models. Approximation of slow-to-compute computer models by fast statistical surrogates and the visualization of input-output relationships are the key ideas here. Building on this, the workshop will develop strategies for optimization of product designs and the validation of computer models against empirical data.

Les modèles informatiques sont maintenant largement répandus dans le domaine de l'ingénierie: ils permettent aux développeurs d'économiser temps et argent dans l'exécution d'expériences physiques. Par exemple, les ingénieurs en électricité utilisent des simulateurs de circuit et les ingénieurs mécanique disposent de modèles à éléments finis. L'atelier offrira un aper cu des stratégies de conception d'expériences et de modélisation statistique qui ont été créées spécialement pour les modèles informatiques. L'approximation des modèles informatiques lents par des substituts statistiques rapides et la visualisation des relations entrée-sortie sont ici les idées clés. S'appuyant là-dessus, on élaborera, dans le cadre de l'atelier, des stratégies pour optimiser la conception de produits et la validation de modèles Les modèles informatiques sont maintenant largement répandus dans le domaine de l'ingénierie: ils permettent aux développeurs d'économiser temps et argent dans l'exécution d'expériences physiques. Par exemple, les ingénieurs en électricité utilisent des simulateurs de circuit et les ingénieurs mécanique disposent de modèles à éléments finis. L'atelier offrira un aper cu des stratégies de conception d'expériences et de modélisation statistique qui ont été créées spécialement pour les modèles informatiques. L'approximation des modèles informatiques lents par des substituts statistiques rapides et la visualisation des relations entrée-sortie sont ici les idées clés. S'appuyant là-dessus, on élaborera, dans le cadre de l'atelier, des stratégies pour optimiser la conception de produits et la validation de modèlnformatiques à la lumière de données empiriques.

Sunday, May 26th/Dimanche 26 mai, 9:00-17:00 **KTH B135**

Handling Missing Data
Traitement de données manquantes
Karla Nobrega and/et David Haziza, Statistics Canada/Statistique Canada

Most research studies (observational or experimental) have some level of nonresponse. This one-day workshop introduces attendees to the concepts, implications and methods to handle nonresponse. Survey and epidemiological study frameworks will be used to illustrate differences in nonresponse mechanisms, in methods dealing with nonresponse and in estimation in the presence of nonresponse. Issues such as unit and item survey nonresponse will be discussed along with compliance, intent to treat and complete case analysis in both observational and experimental epidemiological studies. Much of the discussion will be at an introductory level focusing on both survey and health research examples.

La plupart des études de recherche (empiriques ou expérimentales) comportent un certain niveau de non-réponse. Cet atelier d'une journée introduit aux participants les concepts, implications et

méthodes pour traiter la non-réponse. On utilisera les cadres de travail des enquêtes et études épidémiologiques pour illustrer les différences dans les mécanismes de non-réponse, dans les méthodes traitant la non-réponse et dans l'estimation en présence de non-réponse. On discutera de sujets tels que la non-réponse par unité et par item de même que le suivi du protocol, l'intention de traiter et l'analyse des cas complets dans les études épidémiologiques observationnelles et expérimentales. Une grande partie de la discussion sera à un niveau d'introduction se concentrant sur des exemples d'enquêtes et santé.

Poster Session/Séance d'affichage

Sunday, May 26th/Dimanche 26 mai, 18:00 **USC Marketplace**

Parameters estimation for a production process
Estimation des paramètres d'un procédé de fabrication
Luc Adjengue et/and Soumaya Yacout, École Polytechnique de Montréal

In many industrial processes, items are produced during production cycles and inspections about the actual state of the process can take place only at the end of each cycle. Although control charts may generally be used to monitor the quality of the products during a production cycle, in many cases, it may be of interest to estimate the (unobservable) states of the production process as well. In this presentation, we use a Partially Observable Markov Decision Process (POMDP) to model such production processes. A general method based on a modified Expectation-Maximization (EM) algorithm to obtain maximum likelihood estimates of the parameters of interest is presented. A practical application is considered where the core process is a Markov process representing the (unobservable) states of the production process, whereas non conforming items observed in random samples drawn from the production constitute the observation process. A simulation study is performed in order to evaluate the performance of the proposed method with respect to the number and the length of production cycles.

Dans plusieurs procédés industriels de fabrication, les articles sont produits durant des cycles de production et les contrôles sur l'état réel du procédé ne peuvent être effectués qu'à la fin de chaque cycle. Bien que des cartes de contrôle soient généralement utilisées pour surveiller la qualité de la production durant chaque cycle, dans certains cas, il peut également être intéressant d'estimer les états (non observables) du procédé. Dans cette présentation, nous utilisons des processus markoviens de décision pour modéliser de tels procédés. Une méthode générale basée sur une modification de l'algorithme EM sur l'obtention des estimateurs de vraisemblance maximale est présentée. Une application pratique dans laquelle le processus principal est markovien et représente les états (non observables) du procédé de fabrication, tandis que les unités non conformes observées dans des échantillons prélevés dans la production constitue le processus d'observation. Une étude de simulation est menée afin d'évaluer la performance de la méthode proposée quant au nombre et à la longueur des cycles de production.

Sunday, May 26th/Dimanche 26 mai, 18:00 **USC Marketplace**

Inference concerning quantiles for left truncated and right censored data
Inférence concernant les quantiles pour des données tronquées
Ejaz Ejaz Ahmed, Sana S. Buhamra et/and Noriah M. Al-Kandari, University of Regina

The problem of both testing and estimating the quantile function when the data are left truncated and right censored (LTRC) is considered. The aim of this communication is bi-fold. First, a large sample test statistic to test for the quantile function under the LTRC model is defined and its null and non-null distributions are derived. A Monte Carlo simulation study is conducted to assess the properties of the proposed test statistic in a practical setting. Secondly, in the spirit of shrinkage principle in parameter estimation as exposed by many researchers, we propose estimators assuming an uncertain prior non-sample information on the value of the quantiles. The asymptotic mean squared error of the estimators are derived and compared with the usual estimator.

Le problème de tester et d'estimer la fonction des quantiles quand les données sont tronquées à gauche et censuré à droite (LTRC) est considéré. Le but de cette présentation est double. D'abord, un test statistique basé sur un grand échantillon pour tester la fonction de quantile sous le modèle de LTRC est définie et ses distributions sous les hypothèses nulle et alternatives sont obtenues. Une étude de simulations de Monte-Carlo est faite pour évaluer les propriétés du test statistique proposée dans un contexte pratique. Deuxièmement, dans l'esprit du principe de rétrécissement dans l'estimation de paramètre, comme exposée par beaucoup de chercheurs, nous proposons des estimateurs assumant une information a priori subjective (qui ne dépend pas de l'échantillon) sur la valeur des quantiles. L'erreur quadratique moyenne asymptotique des estimateurs est obtenue et comparée à l'estimateur habituel.

Sunday, May 26th/Dimanche 26 mai, 18:00 USC Marketplace

Jackknife estimation for smooth functions of the parametric component in partially linear models

Estimation jackknife de fonctions lisses des composantes paramétriques dans un modèle partiellement linéaire

Gemai Chen, University of Calgary et/and Jinhong You, University of Regina

To reduce the bias of the naive estimator for a smooth function of the linear parameter in a partially linear regression model, we consider two types of jackknife estimators: the Miller estimator and the Hinkley estimator. We show that the three estimators are asymptotically normal and equivalent, but the jackknife estimators have much smaller biases than the naive estimator.

Pour réduire le biais de l'estimateur naïf d'une fonction lisse du paramètre linéaire dans un modèle de régression partiellement linéaire, nous considérons deux types d'estimateurs jackknife : l'estimateur de Miller et l'estimateur de Hinkley. Nous démontrons que les trois estimateurs sont asymptotiquement normaux et équivalent. Toutefois, les estimateurs par le jackknife ont des biais plus petits que celui de l'estimateur naïf.

Sunday, May 26th/Dimanche 26 mai, 18:00 USC Marketplace

A little known property of the multivariate survivor function

Une propriété peu connue de la fonction de survie multidimensionnelle

Yun Hee Choi et/and David E. Matthews, University of Waterloo

For a nonnegative random variable, T, with finite mean, it is well known that the integral of the survivor function is equal to the mean. We demonstrate that this univariate property has a useful, but little known, multivariate analog with respect to the joint survivor function of k nonnegative random variables. Via examples, we illustrate the benefits of understanding and exploiting this property of the multivariate survivor function.

Il est bien connu que l'intégrale de la fonction de survie d'une variable aléatoire non négative est égale à sa moyenne, si celle-ci est un nombre fini. Nous démontrons l'équivalent multidimensionnel de cette propriété. Ce résultat concernant l'intégrale de la fonction de survie conjointe de k variables aléatoires non négatives est forte utile, mais peu connu. Par le biais d'exemples, nous illustrons les avantages de comprendre et d'utiliser cette propriété de la fonction de survie multidimensionnelle.

Sunday, May 26th/Dimanche 26 mai, 18:00 USC Marketplace

Adjusting radio-telemetry data for tag failure: a case study.
Ajustement de données radio-télémétriques pour des temps de panne : une étude de cas

Laura Cowen et/and Carl J. Schwarz, Simon Fraser University

Mark-recapture techniques are often employed to estimate survival rates and test hypotheses concerning survival processes. In these studies, animals are marked with distinctive tags and released. Succeeding recaptures provide information for the estimation of both survival and catchability.

Radio-tags have the advantage of high catchability rates but the disadvantage of tag-failure due to battery stoppage. This may cause survival estimates to be conservative, as animals that are not recaptured could either be dead or be experiencing battery failure.

As outlined in Cowen and Schwarz (SSC 2001 poster) an adjustment to survival estimates can be made when tag-failure curves are known. This project applies these methods to estimate Chinook salmon (Oncorhynchus tshawytscha) survival through dams on the Columbia River.

Les techniques de capture-recapture sont souvent utilisés pour estimer des taux de survie et pour tester des hypothèses au sujet des processus de survie. Dans ces études, des animaux sont identifiés par des étiquettes distinctives et libérés. La nombre de recaptures fournit des informations pour l'estimation de la survie et du taux de recapture.

Les radio-étiquettes ont l'avantage de fournir des taux de recapture élevés mais elles ont l'inconvénient que l'étiquette tombe en panne dû à l'interruption du fonctionnement de la batterie. Ceci peut rendre des estimations du taux de survie conservatrices, car les animaux qui ne sont pas recapturés pourraient être soit morts ou soit avoir éprouvé une panne de batterie.

Conformément à Cowen et à Schwarz (affiche de la SSC, 2001) un ajustement des estimations des taux de survie peut être fait quand des courbes des pannes reliées aux étiquettes sont connues. Ce projet applique ces méthodes au taux de survie des saumons de Chinook (tshawytscha d'Oncorhynchus) passant par les barrages sur la rivière Colombia.

Sunday, May 26th/Dimanche 26 mai, 18:00 USC Marketplace

Improved sequential estimation under LINEX loss: asymptotics and simulation studies
Estimation séquentielle améliorée sous la perte LINEX : étude de l'asymptotique et de simulations

Sujay Datta, Northern Michigan University

This presentation is about estimation of parameters under an asymmetric loss function. To be specific, we consider the LINEX loss function which penalizes unequally for overestimation and underestimation—linearly in one case and exponentially in the other. This loss function has been widely discussed in the literature and frequently used in economics and other areas. Zellner(1986, JASA) showed how to derive the Bayes estimators of parameters under the LINEX loss starting with various prior distributions. Here we adopt his basic technique to derive 'improved' estimators of location or scale parameters of a number of distributions (in the sense that they have lower (frequentist)

risks than the 'usual' estimators such as the sample mean or the sample variance). Using these 'improved' estimators, we address the problem of fixed-precision estimation (i.e. bounded-risk or point estimation, fixed-width interval estimation, etc.) and show that if the sample size is pre-determined, often these problems do not have solutions that work for all possible values of the unknown parameters. One possible remedy is to resort to sequential (i.e. one-at-a-time) sampling schemes governed by appropriately defined stopping rules. We put forward such sampling schemes and study the performance of the resulting estimators(by large-sample asymptotics as well as moderate-sample simulations)

Cette présentation porte sur l'estimation des paramètres sous une fonction de perte asymétrique. Pour être spécifique, nous considérons la fonction de perte LINEX qui pénalise de fa con inégale pour la surestimation et la sous-estimation, linéairement dans un cas et exponentiellement dans l'autre. Cette fonction de perte a été largement présentée dans la littérature et fréquemment utilisée dans les sciences économiques et autres champs d'étude. Zellner(1986, JASA) a montré comment dériver les estimateurs de Bayes des paramètres sous la fonction de perte LINEX en commen cant par diverses distributions a priori.

Ici nous adoptons sa technique de base pour obtenir les estimateurs "améliorés" des paramètre de position ou d'échelle d'un certain nombre de distributions (dans le sens qu'ils ont des risques (fréquentistes) inférieurs que les estimateurs "habituels" tels que la moyenne de l'échantillon ou la variance échantillonnale). En utilisant ces estimateurs "améliorés", nous considérons le problème de l'estimation à précision fixe (c'est-à-dire risque borné ou estimation ponctuelle, ou estimation avec intervalle fixe, etc.) et nous prouvons que si la dimension de l'échantillon est prédéterminée, souvent ces problèmes n'ont pas des solutions qui fonctionnent pour toutes les valeurs possibles des paramètres inconnus. Une solution possible est de recourir à des méthodes d'échantillonnage séquentiel(c'est-à-dire un à la fois) régis par une règle d'arrêt appropriée. Nous proposons de tels méthodes d'échantillonnage et étudions la performance des estimateurs résultants (par l'asymptotique pour de grands échantillons et des simulations pour des échantillons de taille moyenne).

Sunday, May 26th/Dimanche 26 mai, 18:00 **USC Marketplace**

Extreme values in random precipitation fields
Valeurs extrêmes dans le domaine des précipitations aléatoires

Audrey Fu, University of British Columbia

We are interested in estimating the return values of the extremes for 312 regions across Canada based on the daily precipitation data simulated from the first version of the Canadian Global Coupled Model (CGCM1). In order to account for the spatial correlation, we assume the annual maximum precipitations follow a multivariate log normal distribution and use the Bayesian approach to estimate the mean vector and the covariance matrix. Thus the prior information is updated by the data.

An empirical Bayes method is used. First the mean vector given the covariance matrix has the multivariate normal distribution as the conjugate prior, while the covariance matrix has the inverted Wishart distribution as the conjugate prior. Then algorithms for estimating the hyperparameters are developed from the EM algorithm.

To check the appropriateness of the multivariate log normal distribution, we borrow the idea of cross validation and construct confidence ellipsoids for the predicted values.

Having obtained the posterior multivariate distribution of the transformed extremes, we approximate it by a multivariate normal distribution and marginalize it to calculate the return values (percentiles) at each site. These percentiles are transformed back to the original scale to be consistent with the data.

The major advantage of this multivariate approach is that we are not only able to calculate the return values, but also able to obtain the joint exceeding probability which can have a significant impact on policy making. Moreover, the return values estimated for adjacent regions will cohere.

Nous sommes intéressé à estimer les valeurs extrêmes pour 312 régions à travers le Canada basé sur des données quotidiennes de précipitation simulées à partir de la première version du modèle global canadien de couplage (CGCM1). Afin d'expliquer la corrélation spatiale, nous supposons que les précipitations maximum annuels suivent une distribution log-normale multidimensionnelle, et nous employons une approche bayésienne pour estimer le vecteur moyen et la matrice de covariance. Ainsi, l'information a priori est mise à jour par les données.

Une méthode de Bayes empirique est utilisée. Dans un premier temps, nous supposons une distribution normale multidimensionnele conjuguée a priori pour le vecteur moyen étant donné la matrice de covariance, alors que la distribution a priori sur la matrice de covariance est un distribution conjuguée de Wishart inverse . Des algorithmes pour estimer les hyper-paramètres sont développés à partir de l'algorithme EM.

Pour contrôler la pertinence du choix de la distribution log-normale multidimensionnelle, nous empruntons l'idée de la validation croisée et construisons des ellipsoïdes de confiance pour les valeurs prédites.

Après avoir obtenu la distribution multidimensionnelle a posteriori des extrêmes transformés, nous l'approximons par une distribution normale multidimensionnelle et nous la marginalisons afin de pouvoir calculer les valeurs résultantes (percentiles) à chaque site. Ces percentiles sont transformés de nouveau à l'échelle initiale pour être conforme aux données.

Le principal avantage de cette approche multidimensionnelle est que nous pouvons non seulement calculer les valeurs résultantes, mais nous pouvons également obtenir la probabilité conjointe d'excès qui peut avoir un impact significatif sur la prise de décision. D'ailleurs, les valeurs résultantes estimées pour des régions limitrophes sont cohérentes.

Sunday, May 26th/Dimanche 26 mai, 18:00 **USC Marketplace**

Use of computer image features for discriminating between pathologic nuclear grade groups for breast ductal carcinoma in situ

Utilisation des propriétés d'images informatiques pour discriminer les différents groupes pathologiques nucléaires pour le Ductal Carcinoma in Situ des seins

Yuejiao Fu, University of Waterloo

Ductal carcinoma in situ (DCIS) of the breast is more common with the prevalent use of mammographic screening. Pathologic grading is a major determinant of treatment, although there have been difficulties with reproducibility and handling mixed grades within a patient. All current grading systems entail subjectivity. We aim to develop a quantitative, objective grading system from a computer image; computer grading should be prognostically predictive of disease progression. We have pathologic grade by the Van Nuys system and computer images (39 features) for 82 patients. In descriptive analyses, we concluded that the assumption of normality was appropriate for all factors. For these patients, the number of cells assessed varied from 11 to 20; there were also other sources of variance. We considered both factor means by patient and factor means/standard error of means. Using Fisher linear discriminant analysis, we found the weighting strategy affected the selection of factors having significant association with pathologic grade. Using factor means/standard error of means, three factors led to the correct jackknifed classification of 62.5 percent of grade 1/2, 38.9 percent of grade 2, 42.9 percent of grade 2/3, and 65.2 percent of grade 3 tumours. We expect improved classification when we ultimately obtain data for 200 cells per patient.

Le carcinome canalaire in situ (DCIS) du sein est plus commun avec l'utilisation d'un test de dépistage mammographique. L'évaluation pathologique est un déterminant majeur du traitement, bien qu'il y ait certaines difficultés avec la reproductibilité et la manipulation de plus d'une évaluation pour un patient. Tous les systèmes d'évaluation actuels dépendent d'une forme de subjectivité. Nous visons à développer un système d'évaluation quantitatif et objectif à partir d'une image informatique; l'évaluation de l'ordinateur devrait prédire un prognostic de la progression de la maladie. Nous avons les classes pathologiques par images de système informatique de van Nuys (39 dispositifs) pour 82 patients. Par des analyses descriptives, nous avons conclu que l'hypothèse de normalité était appropriée pour tous les facteurs. Pour ces patients, le nombre de cellules évaluées varie de 11 à 20; il y avait également d'autres sources de variation. Nous avons considéré deux facteurs soit la moyenne par patient et la moyenne/variance pour les moyennes.

En utilisant l'analyse discriminante linéaire de Fisher, nous avons trouvé que la stratégie avec poids affecte la sélection des facteurs ayant une association significative avec la classe pathologique. En utilisant le facteur moyenne/variance pour les moyennes, trois facteurs ont mené à une bonne classification par jackknife pour 62,5 pour cent de la classe 1/2, 38,9 pour cent de la classe 2, 42,9 pour cent de classe 2/3, et 65,2 pour cent de la catégorie 3 de tumeurs. Nous nous attendons à une amélioration de la classification quand nous obtiendrons des données de 200 cellules par patient.

Sunday, May 26th/Dimanche 26 mai, 18:00 **USC Marketplace**

Analysis of 2x2 tables with both completely and partially observed data
Analyse de tables 2X2 avec données complètement et partiellement observées

Isabelle Gaboury, Centre d'étude systématique Thomas C. Chalmers

Nonresponse, which poses difficult questions in inference, is usually the consequence of human behavior. It would seem appealing to discard incomplete observations and analyse only complete data. However, this method gives interesting results only where the amount of missing data is small. When facing a potential loss of information with categorical data, trying to make use of the partially observed data presents an efficient solution. It is important to develop models which take account of the missing data mechanism (i.e. missing completely at random (MCAR), or missing at random (MAR)) and then reflect the implication of the nonresponse in the data. With this poster, I will introduce four different families of models that fit categorical missing data, and compare their efficiency on a data set of a clinical trial.

La non-réponse est généralement la conséquence du comportement humain et pose d'importants problèmes en inférence statistique. Une solution simple serait d'éliminer les observations incomplètes et d'analyser seulement la partie complète de l'ensemble de données. Cependant, cette méthode s'avère efficace que lorsque l'ensemble de données ne contient qu'un petit nombre de données manquantes. Dans le cas contraire, on peut alors tenter d'utiliser l'information partielle des données catégorisées. C'est pourquoi il est important de développer des modèles qui tiennent compte des mécanismes de données manquantes et qui reflètent l'implication de la non-réponse dans les données. Avec cette affiche, je présenterai quatre différentes familles de méthodes modélisant les données manquantes catégorisées, et comparerai leur performance à l'aide d'un ensemble de données provenant d'un essai clinique.

Sunday, May 26th/Dimanche 26 mai, 18:00 **USC Marketplace**

Prediction of Ontario HIV/AIDS diagnosis incidence using the back-projection method

Prévision de l'incidence du diagnostic du VIH SIDA en Ontario par la méthode de projection dans le passé

Sohee Kang et/and John Hsieh, University of Toronto

AIDS is caused by HIV infection. Incidence of HIV infection in a population is a crucial indicator of the HIV/AIDS epidemic development. The time of diagnosis of AIDS symptoms as well as the time of sero-positive test give the valuable information on this epidemic, and both are observable. However, time of HIV infection is not observable. Back-projection is one of the methods for reconstructing HIV incidence curve from AIDS incidence data. The key assumption of back-projection method is that the distribution function of the incubation period, which is a time interval from the HIV infection to AIDS diagnosis, is known. A number of cohort studies suggests that Weibull distribution is appropriate for modeling the incubation period distribution. The hazard rate should be modified with different calendar period since the new treatment which has introduced after 1987 substantially prolonged the incubation period. The recent improvements of the back-projection method are firstly modeling the incubation distribution with different calendar period, and secondly using not only the date of AIDS diagnosis, but also the date of first positive HIV test. We used this improved method on annual HIV diagnoses incidences in Ontario among men having a sexual contact with men (MSM). The estimated HIV incidence curve gives the annual HIV infections in Ontario among MSM. The high peak of infections occurred in 1985 and dropped quickly after that and there is another small mode in 1995. The projected HIV/AIDS diagnosis incidence up to 2005 gives the useful estimates for medical care planning and policy.

SIDA provoqué par l'infection par le VIH. L'incidence de l'infection par le VIH dans une population est un indicateur crucial du développement de l'épidémie du VIH/SIDA. La période du diagnostic des symptômes du SIDA aussi bien que le moment du test qui confirme la séropositivité donnent de l'information importante sur cette épidémie, et toutes les deux sont observables. Cependant, le temps de l'infection par le VIH n'est pas observable. La projection dans le passé est un des méthodes pour reconstruire la courbe d'incidence du VIH à partir des données d'incidence du SIDA.

L'hypothèse principale de la méthode de projection dans le passé est que la fonction de distribution de la période d'incubation, qui est l"intervalle de temps entre le moment de l'infection par le VIH jusqu'au diagnostic du SIDA, est connue. Un certain nombre d'études de cohortes suggère que la distribution de Weibull soit appropriée pour modéliser la distribution de la période d'incubation. Le taux de risque devrait être modifiée selon la période du calendrier puisque le nouveau traitement présenté après 1987 a sensiblement prolongé la période d'incubation. Les améliorations récentes de la méthode de projection dans le passé modélisent premièrement la distribution d'incubation selon la période du calendrier, et utilisent deuxièmement non seulement la date du diagnostic du SIDA, mais également la date du premier test positif du VIH.

Nous avons utilisé cette méthode améliorée sur des incidences annuelles de diagnostic du VIH en Ontario parmi les hommes ayant un contact sexuel avec d'autres hommes (MSM). La courbe estimée de l'incidence du VIH donne les infections annuelles par le VIH en Ontario parmi les MSM. Le sommet de la courbe des infections s'est produit en 1985 et a diminué rapidement ensuite. Il y a un autre petit mode en 1995. L'incidence des diagnostic de VIH/SIDA prévue jusqu'en 2005 donne les estimations utiles pour la planification et la politique des soins médicaux.

Sunday, May 26th/Dimanche 26 mai, 18:00 **USC Marketplace**

Improved estimation of coefficient vector in regression model when the constraints are of stochastic nature.

Estimation améliorée du vecteur des coefficients dans un modèle de régression quand les contraintes sont de nature stochastiques

Bashir Khan, Saint Mary's University et/and S.E. Ahmed, University of Regina

We consider the estimation of coefficient vector Θ in the general linear model $y = X(\Theta) + \epsilon$, with some prior information about Θ that are stochastic in nature and are given by $h = H(\Theta) + v$. The $(qx1)$ vector v of unobservable random components (the error of the prior information) is assumed to be normally distributed with mean δ and covariance matrix $\sigma^2\Omega$ where Ω may reflect subjective prior information. The (q x p) matrix H is a known hypothesis design matrix of rank q that expresses the structure of hypotheses about the individual or linear combinations of parameters in Θ. We shall take Θ as fixed, which implies that h must be random. The elements of (q x 1) vector h are the estimates of $H(\Theta)$, where $H(\Theta)$ consists of q linear combinations of the components of Θ, and these estimates are obtained from previous samples. We use the method of mixed estimation to estimate Θ and incorporate prior knowledge of coefficients in regression analysis and the sample information contained in the linear statistical model. This prior knowledge about the regression coefficients is formulated in terms of the prior estimates of the parameters that are assumed to be unbiased and have a given moment matrix.

Nous avons considérer l'estimation du vecteur de coefficients Theta dans le modèle linéaire général y = X(Theta)+ epsilon, avec quelque information a priori sur Theta qui sont stochastiquement dans la nature et donnée par h = H(Theta) + v. Le vecteur v de dimension (q X 1) des composantes aléatoires non observables (erreur de l'information a priori) est supposé de distribution normale avec moyenne delta et matrice de covariance sigma carré (Omega majuscule) où Omega majuscule peut refléter l'information subjective a priori. La matrice H (q X p) est une matrice de design connue de rang q qui exprime la structure des hypothèses au sujet des paramètres ou des combinaisons linéaires des paramètres de Theta.

Nous prendrons thêta comme fixé, ce qui implique que h doit être aléatoire. Les éléments du vecteur h (q X 1) sont les estimations de H(Theta), où H(Theta) se compose des combinaisons linéaires des q composantes de Theta, et ces estimations sont obtenues à partir d'échantillons précédents. Nous employons la méthode d'estimation mixte pour estimer Theta et pour incorporer l'information a priori sur les coefficients dans l'analyse de régression et l'information de l'échantillon contenue dans le modèle linéaire. Cette information a priori au sujet des coefficients de régression est formulée en termes d'estimations a priori des paramètres. On assume qu'ils sont non biaisés et ont une matrice des moments connue.

Sunday, May 26th/Dimanche 26 mai, 18:00 **USC Marketplace**

Bayesian hierarchical modelling of neonatal mortality: the neonatal health services in Canada project

Modélisation bayésienne hiérarchique pour la mortalité néo-natale : le projet de services de santé néo-nataux au Canada

Ying MacNab et/and Zhenguo Qiu, University of British Columbia

This presentation introduces the Neonatal Health Services in Canada Project. The study examines the impact of geography, local health access and health care systems on variations in outcomes, practices and resource use in neonatal intensive care units (NICUs) across Canada. The assessment

of NICU outcomes and resource utilization is considered from the viewpoint of statistical modeling. Methodological issues relating to Bayesian hierarchical modeling, Bayesian computation, analysis involving clustered observations and estimation of multi-level effects are discussed. A variety of models are applied to data of 20,000 neonates in 17 Canadian NICUs. We present a multilevel analysis of neonatal mortality variations among the 17 NICUs.

Cette présentation présente le projet canadien de services de santé néo-nataux. L'étude examine l'impact de la géographie, de l'accès local au système de santé et les variations des résultats en tenant compte du système de santé, des pratiques et de l'utilisation de ressource dans les services de soins intensifs néo-nataux (NICUs) à travers le Canada. L'évaluation des résultats de NICU et de l'utilisation de ressource est considérée du point de vue de la modélisation statistique. Les conséquences méthodologiques reliées à la modélisation bayésienne hiérarchique, au calcul bayésien, à l'analyse comportant des observations groupées et à l'estimation des effets à plusieurs niveaux sont discutées. Une variété de modèles sont appliquées aux données de 20 000 nouveau-nés dans 17 NICUs canadiens. Nous présentons une analyse à plusieurs niveaux des variations de la mortalité néo-natale parmi les 17 NICUs.

Sunday, May 26th/Dimanche 26 mai, 18:00 **USC Marketplace**

Efficiency of testing procedures for multivariate longitudinal data
Convergence des procédures de tests pour des données longitudinales multidimensionnelles

Catherine Njue, Cancer Care Manitoba

The multivariate longitudinal design, in which multiple characteristics are measured over time on the same subject, is typical of many agricultural, biological, clinical and medical studies. In such studies, it is important to account for both cross-sectional and longitudinal correlations. In some situations, it may be reasonable to express the within-subject variance-covariance matrix as the Kronecker product of two matrices. In this presentation, we will describe the linear model for multivariate longitudinal data with a Kronecker product covariance structure. We will investigate the asymptotic relative efficiency of hypothesis tests for the mean vector that exploits the Kronecker product structure. We will also consider the converse situation, that is, the loss from imposing the Kronecker product structure. The measures of asymptotic relative efficiency used can be applied to compare competing test statistics with limiting non-central Chi-square distributions utilising a suitable Pitman alternative. We introduce a likelihood ratio test of the Kronecker product pattern and define an index that measures how far a given matrix departs from this pattern. Results of simulation studies designed to estimate efficiency will be presented. Our results demonstrate that important gains in efficiency can be achieved using the Kronecker product covariance model, which takes into account and separates cross-sectional and longitudinal correlations.

Un plan d'expérience longitudinal multivarié, dans lequel de multiples caractéristiques sont mesurées dans le temps sur un même sujet, est typique de beaucoup d'études agricoles, biologiques, cliniques et médicales. Dans de telles études, il est important d'expliquer les corrélations tant en coupe que longitudinales.

Dans quelques situations, il peut être raisonnable d'exprimer la matrice de variance-covariance intrasujet comme le produit de Kronecker de deux matrices. Dans cette présentation, nous décrirons le modèle linéaire pour des données longitudinales multivariées avec une structure de covariance basée sur le produit de Kronecker. Nous étudierons l'efficacité relative asymptotique des tests d'hypothèses pour le vecteur moyen qui exploite la structure du produit de Kronecker. Nous considérerons également la situation inverse, c.-à-d., la perte qu'impose la structure du produit de Kronecker. Les mesures

d'efficacité relative asymptotique utilisées peuvent être appliquées pour comparer des statistiques concurrents avec des distributions non centrales limites de khi-deux en utilisant une alternative de Pitman. Nous présentons un test basé sur le rapport de vraisemblance pour la configuration du produit de Kronecker et définissons un index qui mesure à quelle distance une matrice donnée s'écarte de cette configuration. Les résultats d'études de simulation con cues pour estimer la convergence seront présentés. Nos résultats démontrent que des gains importants dans la convergence peuvent être réalisés en utilisant le modèle de covariance du produit de Kronecker, qui tient compte et sépare les corrélations en coupe et longitudinales.

Sunday, May 26th/Dimanche 26 mai, 18:00 **USC Marketplace**

The dating of medieval documents
La datation de documents médiévaux

Gelila Tilahun, Andrey Feuerverger, University of Toronto et/and Peter Hall, Australian National University

The dating of undated medieval documents is one of the major works being done by the DEEDS (Documents of Essex England Data Set) project group at the University of Toronto. These documents deal with the transfer of property. I will present statistical methods for the dating of medieval documents based on their similarities to other documents for which the dates are known. Statistical ideas are used to define "similarity" by smoothing among distance measures and frequency of word and phrase patterns. In addition, word and phrase frequencies are used to estimate the probable date of an undated document.

Dater des documents médiévaux non datés est un des travaux majeurs fait par le groupe de recherche DEEDS (Documents of Essex England Data Set) à l'université de Toronto. Ces documents traitent le transfert de la propriété. Je présenterai des méthodes statistiques pour dater des documents médiévaux basé sur leurs similitudes à d'autres documents pour lesquels les dates sont connues. Des idées statistiques sont utilisées pour définir "similitude" en lissant certaines mesures de distance et fréquence des configurations de mot et d'expression. De plus, les fréquences de mots et d'expression sont employées pour estimer la date probable d'un document non daté.

Sunday, May 26th/Dimanche 26 mai, 18:00 **USC Marketplace**

Interim analyses of cluster randomization trials with binary outcomes
Analyses intérimaires d'essais aléatoires de groupes avec réponses dichotomiques

Guangyong Zou et/and Allan Donner, University of Western Ontario et/and Neil Klar, Cancer Care Ontario

A cluster randomization trial is one in which clusters of individuals are randomly allocated to different intervention arms. Due to logistic, ethical and practical reasons this design has gained popularity among health researchers, with extensive body of methodology formed over the past two decades. However the use of formal interim analysis procedures for monitoring such trials have received relatively little attention.

Our main objective is to examine the performance of group sequential methods for monitoring cluster randomization trials with binary outcomes.

Fourteen test statistics, ranging from simple cluster-level analyses to extensions of logistic regression adjusted for clustering, are discussed and placed in a unified framework. The sequentially computed test statistics are shown to have independent increment structures in large samples, even with correlation misspecification.

The simulation results reveal that direct adjusted chi-square test and a Wald test resulting from a bias-corrected robust variance approach are appropriate for monitoring trials with small number of clusters. The O'Brien-Fleming boundaries are preferred to the Pocock ones both in terms of preserving overall Type I error rates and achieving higher overall power.

Une étude de groupes aléatoires est une étude dans laquelle les groupes d'individus sont aléatoirement assignées à différentes méthodes d'intervention. En raison d'arguments logistiques, morales et pratiques, ce plan d'expérience a gagné en popularité parmi des chercheurs de la santé, avec une panoplie de méthodologies modifiées pendant les dernières deux décennies. Cependant l'utilisation des procédures formelles intérimaire d'analyse pour surveiller de telles études ont suscité relativement peu d'attention.

Notre objectif principal est d'examiner l'exécution des méthodes séquentielles de groupe pour surveiller des études de groupes aléatoires ayant des résultats dichotomiques.

Quatorze tests statistiques, s'étendant des analyses simples au niveau des groupes à la généralisation de la régression logistique ajustée pour les regroupements, sont discutés et placés dans un cadre unifié. Les tests statistiques séquentiellement calculés montrent des structures indépendantes d'incrémentation pour de grands échantillons, même avec une mauvaise spécification de la corrélation.

Les résultats des simulations indiquent que le test direct du khi-deux ajusté et le test de Wald résultant d'une approche pour la variance robuste corrigée pour le biais sont appropriés pour surveiller les études ayant un petit nombre de groupes. Les bornes de O'Brien-Fleming sont préférées à celles de Pocock en termes de la préservation du taux d'erreur globales de type I et à l'obtention d'une puissance globale plus élevée.

Session 1: Welcome and SSC Presidential Invited Address/Allocution de l'invité du président de la SSC

Monday, May 27th/Lundi 27 mai, 8:30 **CNH 104**

Risk and revolution: Casanova, Napoleon, and the Loterie de france
Révolution et risque : Casanova, Napoléon et la Loterie de France

Stephen Stigler, University of Chicago.

Just how risk-adverse was Robespierre? How did the sans-culottes lose their culottes? In the eighteenth century in France, citizens and royalty faced a multitude of risks, from sexually transmitted disease to decapitation. The Loterie de France operated under various names and in various places from August 1758 until May 1836, and was based upon principles slightly but interestingly different from modern Canadian Lotto games. Some unusual data sources on this Loterie provide a window on how financial risk was addressed in that tumultuous time, both by the government and by citizens. The story sheds light upon how the emerging calculus of probabilities affected the perception of risk, and how this understanding interacted with political events. An unusually diverse cast of characters is involved, including Casanova, d'Alembert, and Bonaparte, as well as some modern statistical technique, much of it created in Canada.

Dans quelle mesure Robespierre avait-il un risque défavorable? Comment les sans-culottes ont-ils perdus leurs culottes? Au dix-huitième siècle en France, les citoyens et la royauté ont fait face à une multitude de risques: des maladies transmissibles sexuellement à la décapitation. Le Loterie de France a fonctionné sous divers noms et dans divers endroits à partir d'août 1758 jusqu'à mai 1836, et a été basé sur des principes légèrement mais tout de même étonnamment différents des jeux modernes de loteries canadiennes. Quelques sources de données peu communes sur cette ancienne loterie fournissent une fenêtre sur la fa con dont le risque financier a été considéré dans ce temps tumultueux, par le gouvernement et par les citoyens. L'histoire fait la lumière sur la fa con dont le

calcul des probabilités a affecté la perception du risque, et la fa con dont cette compréhension a eu des effets sur des événements politiques. Un groupe exceptionnellement diversifié de personnalités est impliqué, y compris Casanova, d'Alembert, et Bonaparte, ainsi que certaines techniques statistiques modernes, une grande partie créée au Canada.

Session 2: Directional Statistics/La statistique directionnelle

Monday, May 27th/Lundi 27 mai, 10:30 **TSH B105**

Symmetry and Bayesian function estimation on Riemanian manifolds
Symétrie et estimation bayésienne d'une fonction sur une variété riemanienne

Jean-François Angers, Université de Montréal, et/and Peter Kim, University of Guelph

This presentation is about Bayesian curve fitting on compact Riemannian manifolds. The approach is to combine Bayesian methods along with aspects of spectral geometry associated with the Laplace-Beltrami operator on Riemannian manifolds. Although frequentist and Bayesian nonparametric curve estimation in Euclidean space abound, to date, no attempt has been made with respect to Bayesian curve estimation on a general Riemannian manifold. The Bayesian approach to curve fitting is very natural for manifolds because one can elicit very specific prior information on the possible symmetries in question. One can then establish Bayes estimators that possess built-in symmetries. Alternatively, one can diffuse away some of the prior information in which case a connection with smoothing splines on manifolds is obtained. A detailed analysis for the 2D-sphere is provided.

Cette conférence portera sur l'ajustement d'une fonction sur une variété riemannienne compacte. L'approche développée combine les méthodes bayésiennes d'estimation fonctionnelle et la géométrie spectrale associée à l'opérateur de Laplace-Beltrami sur les variétés riemanniennes. Même si plusieurs techniques d'estimation non paramétrique (fréquentiste et bayésienne) de fonction dans les espaces euclidiens existent, aucune, jusqu'à maintenant, n'a été adaptée pour les variétés riemanniennes générales. Il est naturel d'utiliser l'approche bayésienne pour l'estimation de fonction sur les variétés car celle-ci nous permet d'incorporer les différentes symétries possibles via les densités a priori. Nous pouvons aussi montrer que les estimateurs de Bayes obtenus possèdent les symétries désirées. Nous discutons aussi du lien qui existe entre les splines de lissage sur les variétés et les estimateurs de Bayes obtenus à partir d'une loi a priori vague. Un exemple sur une sphère à deux dimensions sera aussi présenté.

Monday, May 27th/Lundi 27 mai, 11:00 **TSH B105**

Regression models for Stieffel manifolds
Modèles de régression pour les variétés de Stieffel

Ted Chang, University of Virginia

Prentice introduced a regression model for Stieffel manifolds which he used on a data set of vector cardiograms first analyzed by Downs in his seminal paper on orientation statistics. This talk will discuss the physical meaning of the regression parameters in the Prentice model and formulate hypothesis tests for orientation statistics of physical interest.

The distributions used in the paper by Downs are matrix Fisher. However the matrix Fisher distribution has somewhat disturbing properties under marginalization. The matrix Fisher assumption can be replaced by an assumption of group invariance and the assumption of group invariance does behave properly under marginalization.

Prentice a présenté un modèle de régression pour les variétés de Stieffel qu'il a utilisées sur un jeu de données de cardiogrammes analysé précédemment par Downs dans son article fondamentale sur les statistiques d'orientation. Cette présentation discutera du sens physique des paramètres de régression dans le modèle de Prentice et formulera des tests d'hypothèses pour des statistiques d'orientation intéressant du point de vue physique.

Les distributions utilisées dans l'article de Downs sont la matrice de Fisher. Cependant la distribution de la matrice de Fisher a des propriétés en quelque peu inquiétantes sous la marginalisation. L'hypothèse de la matrice de Fisher peut être remplacée par une hypothèse d'invariance de groupe et cette hypothèse se comporte correctement sous la marginalisation.

Monday, May 27th/Lundi 27 mai, 11:30 **TSH B105**

Covariance models for dynamic space-time processes
Modèles de covariance pour des processus spatio-temporels dynamiques
Tilmann Gneiting, University of Washington

Geostatistical approaches to spatio-temporal prediction in environmental science, meteorology, and related disciplines rely on suitable analytic covariance models. This talk proposes general classes of nonseparable, stationary covariance functions for space-time processes. The approach is based on classical results in Fourier analysis but avoids closed form Fourier inversion. The model parameters are easily interpretable and include a space-time interaction parameter. Special emphasis is put on the modeling of dynamic processes such as prefered wind directions or ocean currents, and strategies for model fitting are illustrated using wind data from Ireland.

Les approches géostatistiques pour des prévisions spatio-temporelles en science environnementale, météorologie, et les disciplines associées se fondent sur les modèles analytiques appropriés de covariance. Cet entretien propose les classes générales de fonctions non séparables et de fonctions avec covariances stationnaires pour des processus d'espace-temps. L'approche est basée sur des résultats classiques d'analyses de Fourier mais évite l'inversion de Fourier de façon analytique. Les paramètres du modèles sont facilement interprétables et incluent un paramètre d'interaction espace-temps. Une considération particulière est mise sur la modélisation des processus dynamiques comme la direction préférée du vent ou des courants de l'océan, et des stratégies pour l'ajustement d'un modèle sont illustrées en utilisant des données sur le vent en Irlande.

Session 3: Statistics and Water Quality/La statistique et la qualité de l'eau

Monday, May 27th/Lundi 27 mai, 10:30 **TSH B128**

Managing health risks from drinking water
Gestion du risque sur la santé de l'eau potable
Dan Krewski, University of Ottawa et/and William Ross, Health Canada

Recent events in Ontario and elsewhere have highlighted the vulnerability of communities to failures in managing their supply of drinking water. This short presentation outlines an analysis, provided to the Walkerton Inquiry, of the scientific basis for drinking water risk assessment and strategies for managing these risks. While zero risk is an unattainable goal, the analysis identifies enhancements to management systems that can minimize health risks from drinking water in Ontario.

Les récents événements survenus en Ontario et ailleurs ont accentué la vulnérabilité des communautés face aux échecs du contrôle de leur approvisionnement en eau potable. Cette courte présentation décrit une analyse, fournie par l'enquête Walkerton, des bases scientifiques pour l'évaluation des risques reliés à l'eau potable et des stratégies pour contrôler ces risques. Comme l'obtention du risque nul est un but inaccessible, l'analyse identifie des améliorations quant aux systèmes de gestion. Ceux-ci peuvent minimiser les risques qui ont trait à la santé dus à l'eau potable en Ontario.

Monday, May 27th/Lundi 27 mai, 10:40 **TSH B128**

Monitoring and assessment the quality of Canadian fresh waters with emphases on the Great Lakes

Surveillance et évaluation de la qualité de l'eau douce au Canada avec emphase sur les Grands Lacs

Abdel El-Shaarawi, National Water Research Institute

Water contamination and protection have been of major concern to the Canadian public and governments. Contaminants are discharged into waters from municipal, industrial, agricultural and atmospheric sources. Exposure to these contaminants may adversely affect human health and or aquatic life. Governmental agencies have established quantitative standards and criteria whose goals are to protect human health and the environment. Data colleted routinely by various monitoring programs are used to measure current status and trends towards the achievement of these goals. The objectives of this talk are to (i) provide an overview of some Canadian water quality monitoring programs for the Great Lakes, Turkey Lakes, effluent quality and drinking waters and (ii) discuss the statistical issues involved in the analysis of their data. The endpoints objectives of these programs range from determining current status, detecting and estimating trends, defining problem areas, ensuring compliance with regulations and setting quantitative water quality standards.

La contamination et la protection de l'eau ont été un des soucis majeurs pour les gens et le gouvernement canadien. Des contaminants sont déchargés dans l'eau par les sources municipales, industrielles, agricoles et atmosphériques. L'exposition à ces contaminants peut compromettre la santé des humains et/ou la vie aquatique. Les agences gouvernementales ont établi des normes et des critères quantitatifs dont les buts sont essentiellement la protection de la santé humaine et de l'environnement. Les données recueillies fréquemment par divers programmes de contrôle sont utilisées pour mesurer l'état actuel et les tendances vers l'accomplissement de ces buts. Les objectifs de cette conférence sont (i) fournir une vue d'ensemble de quelques programmes de contrôle canadiens de qualité de l'eau pour les Grands Lacs, les lacs Turkey, influençant la qualité des eaux potables et (ii) discuter des sujets statistiques impliqués dans l'analyse de telles données. Les objectifs finaux de ces programmes sont entre autres de déterminer l'état actuel, de détecter et d'estimer des tendances, de définir des domaines problématiques, d'assurer la conformité aux règlements et de fixer des normes quantitatives de qualité de l'eau.

Monday, May 27th/Lundi 27 mai, 11:20 **TSH B128**

Analysis of censored environmental data with Box-Cox transformations

Analyse de données censurées sur l'environnement à l'aide de transformations de Box-Cox

Reza Modarres, George Washington University et/and Jade Freeman, EPA

We present a method for estimating a mean vector from a bivariate skewed distribution that includes some unobserved data below the detection limits. The method uses a bivariate Box-Cox

transformation, of which the parameters are found by maximizing the likelihood function over a fixed power transformation set. The Expectation-Maximization algorithm is used to maximize the likelihood and obtain MLE's for the mean vector and covariance matrix. Given a transformation, we use the form of the mean vector in the original scale coupled with the invariance property of the MLE's to obtain the estimates in the original scale. The asymptotic normality of the MLE's and the delta-method for the functions of asymptotically normal vectors provide a confidence region for the mean vector in the original scale. The performance of the MLE method in selecting the correct power transformation and the coverage rate of the confidence region as a function of censoring proportion, correlation structure, size of power transformation set, and sample size are investigated in a simulation study. The MLE method gives reliable results for finding effective transformations for highly skewed data sets. Furthermore, the delta-method confidence region provides good coverage for the bivariate mean vector. The method is applied to a data set of pollutant measurements taken for monitoring water quality.

Nous présentons une méthode pour estimer le vecteur moyen d'une distribution bidimensionnelle asymétrique qui inclut quelques données non observées sous des limites de détection. La méthode utilise une transformation bidimensionnelle de Box-Cox, dont les paramètres sont trouvés en maximisant la fonction de vraisemblance sous un ensemble de transformations de puissance fixée. L'algorithme d'Espérance-Maximisation (EM) est employé pour maximiser la probabilité et pour obtenir les estimateurs du maximum de vraisemblance (MLE) pour le vecteur moyen et la matrice de covariance. Pour une transformation donnée, nous utilisons la forme du vecteur moyen dans l'échelle initiale jointe à la propriété d'invariance des MLE pour obtenir les estimateurs dans l'échelle initiale. La normalité asymptotique des MLE et la méthode delta pour les fonctions de vecteurs asymptotiquement normaux fournissent une région de confiance pour le vecteur moyen dans l'échelle initiale. La performance de la méthode du MLE pour choisir la transformation de puissance appropriée et le taux de couverture de la région de confiance comme fonction de la proportion d'observations censurées, de la structure de corrélation, de la taille de l'ensemble des transformation de puissance, et de la dimension de l'échantillon sont étudiées dans une étude de simulation. La méthode du MLE donne des résultats fiables pour trouver des transformations pertinentes pour des données fortement aymétriques. De plus, la région de confiance obtenue par la méthode delta fournit la bonne couverture pour le vecteur moyen bidimensionnel. La méthode est appliquée à un jeu de données de mesures de polluant prises pour surveiller la qualité de l'eau.

Session 4: Probability/Probabilité

Monday, May 27th/Lundi 27 mai, 10:30 **TSH B106**

Uniqueness for a class of degenerate stochastic differential equations arising from super-processes

Sur l'unicité pour une classe d'équations stochastiques différentielles dégénérées provenant d'un superprocessus

Siva Athreya, Indian Statistical Institute, Martin Barlow, Richard Bass, et/and Edwin Perkins

We will consider diffusions corresponding to the generator

$$Lf(x) = \sum_{i=1}^{d} x_i \gamma_i(x) \frac{\partial^2}{\partial {x_i}^2} f(x) + b_i(x) \frac{\partial}{\partial x_i} f(x),$$

$x \in [0,\infty)^d$, for continuous $\gamma_i, b_i : [0,\infty)^d \to R$ with γ_i nonnegative.

We establish uniqueness for the corresponding martingale problem under certain non-degeneracy conditions on b_i, γ_i. We will begin by briefly explaining the motivation for studying such diffusions, followed by a discussion on why standard techniques fail and conclude with a brief description of the proof.

Nous allons considérer les équations de diffusion correspondant au générateur

$$Lf(x) = \sum_{i=1}^{d} x_i \gamma_i(x) \frac{\partial^2}{\partial {x_i}^2} f(x) + b_i(x) \frac{\partial}{\partial x_i} f(x),$$

pour $\gamma_i, b_i : [0,\infty)^d \to R$ continu et γ_i nonnégatif.

Nous établissons le problème de l'unicité des martingales correspondantes sous certaines conditions de non-dégénérescence sur b_i, γ_i. Nous allons débuter par une brève explication des motivations pour l'étude de telles équations de diffusion. Nous allons ensuite poursuivre en élaborant sur les raisons qui font que les techniques standards ne fonctionnent pas bien, et nous conclurons par une description sommaire de la preuve.

Monday, May 27th/Lundi 27 mai, 11:00 **TSH B106**

On Neumann eigenfunctions for some planar domains
Sur le problème des valeurs propres de Neumann dans certains domaines planaires
Rami Atar, Technion et/and K. Burdzy

A "lip domain" is a planar set lying between the graphs of two Lipschitz functions with constant 1. We show that the second Neumann-Laplacian eigenvalue is simple in every lip domain except the square. The arguments use properties of "mirror couplings" of reflecting Brownian motions, that may be of independent interest. These include the construction of such a coupling in piecewise smooth domains as a strong solution to an SDE, and the behavior of a coupled pair conditioned not to couple.

Un "domaine de lèvres" est un ensemble planaire se trouvant entre les graphiques de deux fonctions de Lipschitz avec 1 comme constante. Nous prouvons que la deuxième valeur propre de Neumann-Laplace est simple dans chaque "domaine de lèvres" à l'exception du carré. Les arguments emploient des propriétés des "couplages par miroir" pour refléter les mouvements browniens qui peuvent aussi être d'intérêt. Ces arguments incluent la construction d'un couplage tel que les domaines sont lissés par morceaux comme solution au SDE et le comportement d'une paire de couples n'est pas conditionné sur un couple.

Monday, May 27th/Lundi 27 mai, 11:30 **TSH B106**

Scaling limit for a Fleming-Viot type system
Limite d'échelle pour un système de type Fleming-Viot
Ilie Grigorescu, University of Miami et/and Min Kang, Northwestern University

We consider a system of N Brownian particles evolving independently in a domain D. As soon as one particle reaches the boundary it is killed and one of the other particles splits into two independent particles. The model is introduced by Burdzy, Holyst, Ingerman and March (1996). As N approaches infinity, we prove the hydrodynamic limit for the empirical measure process and determine the exact law of the tagged particle. In addition, we show that any finite number of labelled particles become independent in the limit.

Nous considérons un système brownien de N particules évoluant indépendamment dans un domaine D. Dès qu'une particule atteint la frontière, elle est tuée et une des autres particules se divise en deux particules indépendantes. Le modèle a été présenté par Burdzy, Holyst, Ingerman et March (1996). Quand N tend vers l'infini, nous pouvons déduire la limite hydrodynamique du processus empirique et déterminons la loi exacte de la particule étiquetée. De plus, nous prouvons que, à la limite, tout nombre fini de particules étiquetées deviennent indépendantes.

Session 6: Phylogenetics/Phylogénétique

Monday, May 27th/Lundi 27 mai, 13:30 **TSH B105**

Testing for rate variation in phylogenetic subtrees
Tests pour la variation du taux dans les arbres phylogénétiques

Ed Susko, Dalhousie University

It has long been recognized that the rates of molecular evolution vary amongst sites in proteins. The usual model for rate heterogeneity assumes independent rate variation according to a rate distribution. In such models the rate at a site, while random, is assumed fixed throughout the evolutionary tree. We present methods that can be useful in detecting whether different rates occur in two different subtrees of the larger tree and where these differences occur. Parametric bootstrapping and orthogonal regression methodologies are used to test for rate differences and to make statements about the general differences in the rates at sites. Confidence intervals based on the conditional distributions of rates at sites are then used to detect where the rate differences occur. Such methods will be helpful in studying the phylogenetic, structural and functional bases of changes in evolutionary rates at sites, a phenomenon that has important consequences for deep phylogenetic inference.

Il y a longtemps que l'on a identifié que les taux d'évolution moléculaire changent parmi des sites de protéines. Le modèle habituel pour l'hétérogénéité des taux suppose la variation indépendante des taux selon une distribution quelconque. Dans de tels modèles, le taux à un site, tout en étant aléatoire, est supposée fixe dans tout l'arbre évolutionnaire. Nous présentons les méthodes actuelles qui peuvent être utiles pour détecter si des taux différents se produisent dans deux sous-arbres de l'arbre le plus grand et où ces différences se produisent.

Le rééchantillonnage paramétrique et les méthodes de régression orthogonales sont employés pour déterminer des différences de taux et pour faire des rapports au sujet des différences générales dans les taux aux sites. Des intervalles de confiance basés sur les distributions conditionnelles des taux aux sites sont alors employés pour détecter où les différences de taux se produisent. De telles méthodes seront utiles dans les études sur la phylogénétique, sur la base des changements structurelles et fonctionnelles dans les taux d'évolution aux sites, un phénomène qui a des conséquences importantes pour l'inférence phylogénétique avancée.

Monday, May 27th/Lundi 27 mai, 14:15 **TSH B105**

Bayesian phylogenetic inference from animal mitochondrial genome arrangements
Inférence phylogénétique bayésienne à partir d'arrangements du génome mitochondrique d'animaux

Bret Larget, Donald L. Simon, Duquesne University, et/and Joseph B. Kadane, Carnegie Mellon University

The determination of evolutionary relationships is a fundamental problem in evolutionary biology. Genome arrangement data is potentially more informative than DNA sequence data for the inference

of evolutionary relationships among distantly related taxa. Under a simple model in which gene inversion is the sole mechanism by which genomes may rearrange, we describe a Bayesian approach for estimating phylogenies (evolutionary trees) from mitochondrial genome arrangement data using Markov chain Monte Carlo methods. We apply this method to the full mitochondrial genomes of several animals.

La détermination des relations dans l'évolution est un problème fondamental dans la biologie de l'évolution. Les données d'agencement du génome sont potentiellement plus instructives que les données des séquences d'ADN pour l'inférence des relations dans l'évolution parmi des taxons reliés mais lointain.

Sous un modèle simple dans lequel l'inversion du gène est l'unique mécanisme par lequel les génomes peuvent se modifier, nous décrivons une approche bayésienne pour estimer les phylogénies (arbres de l'évolution) des données mitochondriales dans l'agencement du génome en utilisant des méthodes de Monte-Carlo par chaîne de Markov. Nous appliquons ces méthodes à des génomes mitochondriales complets de plusieurs animaux.

Session 7: Joint Industry Survey Methods Session / Session conjointe des groupes de statistique industrielle et de méthodologie d'enquête

Monday, May 27th/Lundi 27 mai, 13:30 **TSH B128**

Statistical applications in marketing reseach
Applications statistiques dans des recherches en marketing

Joseph Farruggia, ACNielsen Canada

How do companies measure the success of their products? How do they track and monitor their sales? Who are their main competitors? Which products respond better to trade promotion, consumer promotion or media? All these issues are critical to a company's success. ACNielsen provides companies in the consumer packaged goods, durable goods, pharmaceutical, and entertainment industries answers to these questions. In the Modeling and Analytics Department, specialists work as partners in client businesses, drawing on Retail, Consumer Panel, Media, customized Research and other data to provide actionable answers to a broad range of business challenges. A full range of analytical approaches are used, from straight forward analytical reports to sophisticated modeling techniques. Today's talk will describe how our data is collected and how we use it to develop statistical models.

Comment les compagnies mesurent-elles le succès de leurs produits? Comment est-ce qu'elles suivent et surveillent leurs ventes? Qui sont leurs principaux concurrents? Quels produits répondent le mieux à la promotion faite aux commerces, aux consommateurs ou aux médias? Toutes ces réponses sont critiques au succès d'une compagnie.

ACNielsen fournit les réponses à ces questions pour certaines compagnies oeuvrant dans les domaines de marchandises préemballées, de marchandises durables, autant pour des industries pharmaceutiques que de divertissement. Dans le département de modélisation et d'analyse, les spécialistes travaillent comme des associés des entreprises-clients, se prononçant sur le détail, sur ce qui a trait aux consommateurs, sur la médiatitisation, la recherche personnalisée et autres données pour fournir des réponses exigibles à un large éventail de défis reliés aux affaires. Une gamme complète d'approches analytiques sont utilisées, autant de simple rapports d'analyses que des techniques sophistiquées de modélisation. La présentation d'aujourd'hui décrira comment nos données sont rassemblées et comment nous les employons pour développer les modèles statistiques.

Monday, May 27th/Lundi 27 mai, 14:15 **TSH B128**

Analysis of life experiments with interventions
Analyse des expériences sur le temps de vie avec interventions
Fernando Camacho, DAMOS

Life time experiments sometimes are subject to interference that may affect the outcome and bias the results. In this paper a hazard intervention model is proposed to analyze these experiments and "filter out" the effect of the undesired interferences. This model has the form h(t) = h0(t) + hI(t) where h0(t) is the hazard function if no interferences are present and hI(t) is the "hazard perturbation" due to the interference. An example is presented to illustrate the estimation of the intervention and the removal of this effect from the probability of survival.

Les expériences sur des temps de vie sont parfois sujettes à l'interférence, ce qui peut affecter les réponses et ainsi biaiser les résultats. Dans cette présentation un modèle d'intervention de risque est proposé pour analyser ces expériences et "filtrer" l'effet des interférences non désirées. Ce modèle a la forme h(t) = h0(t) + hI(t) où h0(t) est la fonction de risque si aucune interférence n'est présente et hI(t) est la "perturbation du risque" due à l'interférence. Un exemple est présenté pour illustrer l'estimation de l'intervention et le retrait de cet effet pour la probabilité de survie.

Session 8: Statistical Inference/Inférence statistique

Monday, May 27th/Lundi 27 mai, 13:30 **KTH B105**

Bonus-malus in acceptance sampling on attributes
Système bonus-malus dans l'échantillonnage pour l'acceptation des attributs
Kris Klaassen, University of Amsterdam

Credit is introduced in acceptance sampling on attributes and a credit based acceptance sampling system is developed that is very easy to apply in practice. The credit of a producer is defined as the total number of items accepted since the last rejection. In our sampling system the sample size for a lot depends via a simple function on the lot size, the credit, and the chosen guaranteed upper limit on the outgoing quality. The higher the credit, the smaller the sample size will be. Typically, it will be much smaller than in isolated lot inspection. This Bonus-Malus acceptance sampling system will be discussed together with the simple continuous sampling plan that will be derived from it.

Le crédit est présenté dans l'échantillonnage d'acceptation-rejet sur des attributs et un système d'acceptation-rejet basé sur le crédit est développé. Ce système est très facile d'application en pratique. Le crédit d'un producteur est défini comme le nombre total des éléments acceptés depuis le dernier rejet. Dans notre système d'échantillonnage la dimension du lot dépend par l'intermédiaire d'une fonction simple, de la taille de ce lot, le crédit, et la limite supérieure garantie choisie sur la qualité. Plus le crédit est haut, plus la dimension de l'échantillon sera petite. Typiquement, elle sera beaucoup plus petite que dans l'inspection de ce lot isolé. Ce système bonus-malus d'échantillonnage par acceptation-rejet sera discuté en même temps que le plan simple de prélèvement continu dérivé de celui-ci.

Monday, May 27th/Lundi 27 mai, 14:00 **KTH B105**

On sequential edges recovery in image processing
Sur la détection séquentielle des bordures dans le traitement d'images
Boris Levit, Queen's University

Most experts seem to agree that edge detection is one of the crucial tasks in Image Processing. Although a human eye solves this problem seemingly effortlessly, developing corresponding efficient algorithms is far from completion.

Some experiments suggest that the efficiency of the humane eye rests in the fact that it doesn't scan a picture linearly, i.e. pixel by pixel. Rather, it performs a sequential search for the most noteworthy and telling fragments.

In this talk, the edge detection will be discussed in the simplest case of black-on-white fragments, while the edges are supposed to be smooth and periodic. Under these assumptions, upper and lower bounds for the accuracy of edge detection will be presented. When the observation time becomes large, the (logarithmic) asymptotics of these bounds coincide to the order. Moreover, these asymptotics agree to the constant, in the benchmark case for which the noise is absent.

La plupart des experts semblent d'accord que la détection de bordure est une des tâches cruciales dans le traitement d'images. Bien qu'un œil humain résolve ce problème apparemment sans effort, développer des algorithmes efficaces correspondants est loin d'être accompli.

Quelques expériences suggèrent que l'efficacité de l'oeil humain repose dans le fait qu'il ne balaye pas une image linéairement, c'est-à-dire, pixel par pixel. Il exécute plutôt une recherche séquentielle des fragments les plus remarquables et informatifs.

Dans cette présentation, la détection des bordures sera discutée dans le cas le plus simple de fragments noirs sur blancs, alors que les bordures sont censées être lisses et périodiques. Sous ces hypothèses, des bornes supérieures et inférieures pour l'exactitude de la détection des bordures seront présentées. Quand le temps d'observation devient grand, l'asymptotique (logarithmique) de ces bordures coïncident à l'ordre. D'ailleurs, ces asymptotiques sont en accord sur la constante, dans le cas standard où le bruit est absent.

Monday, May 27th/Lundi 27 mai, 14:30 **KTH B105**

On estimation of restricted mean vectors
Sur l'estimation d'un vecteur moyen restreint

William Strawderman, Rutgers University, Dominique Fourdrinier, Université de Rouen, et/and Martin Wells, Cornell University

We consider estimation of the mean vector of a spherically symmetric distribution with unknown scale when the mean vector is restricted to lie in a cone. A main example is the restriction of the mean vector to a polyhedral cone. This example includes such common restrictions as ordered parameters and general linear inequality constraints. We show that certain types of shrinkage estimators have the strong robustness property that they improve over the MLE for the normal model uniformly over the class of all spherically symmetric distributions. This work is joint with Dominique Fourdrinier and Martin Wells.

Nous considérons l'estimation du vecteur moyen d'une distribution sphérique symétrique avec paramètre d'échelle inconnu quand le vecteur moyen est limité à se situer dans un cône. L'exemple principal est la restriction du vecteur moyen à un cône polyèdre. Cet exemple inclut des restrictions communes telles que des paramètres ordonnés et des contraintes linéaires générales d'inégalité. Nous prouvons que certains types d'estimateurs de réduction ont la forte propriété de robustesse qu'ils améliorent les résultats obtenus par la méthode MLE pour le modèle normal uniformément dans la classe de toutes les distributions sphériques symétriques. Ce travail est conjoint avec le Dominique Fourdrinier et Martin Wells.

Session 9: Applications of Statistics/Applications statistiques

Monday, May 27th/Lundi 27 mai, 13:30 **TSH B106**

Elicited data and incorporation of expert opinion in ecological studies
Données illicites et l'inclusion de l'opinion des experts dans des études écologiques

Douglas Dover et/and Subhash Lele, University of Alberta

The determination or prediction of locations where mountain shrews may be present is of interest. However, it is both difficult and costly to do a complete survey of shrew presence/absence. Prediction can be carried out using a simple logistic regression model that includes habitat covariates. The performance of the logistic regression approach can be improved by incorporating expert opinion. Expert opinion is obtained in terms of elicited data (Lele and Das 2000, Lele 2002) and predictions are elicited from experts for all locations. These expert guesses are then combined with the observed data in a hierarchical model. The gains from the hierarchical approach are substantial: Estimation is improved (lower MSE for habitat covariates), prediction as measured by ROC is better, and it is possible to distinguish between experts adding information and experts adding noise.

An illustration of this methodology is given using an experiment conducted in Montana predicting the presence/absence of the mountain shrew population.

Nous nous intérresson A la détermination ou la prévision des emplacements où les musaraignes de montagne peuvent être présents. Cependant, il est difficile et coûteux de faire une étude complète de la présence/absence de musaraignes. La prévision peut être effectuée en utilisant un modèle simple de régression logistique qui inclut des covariables sur l'habitat. La performance de l'approche de régression logistique peut être améliorée en incorporant l'opinion des experts. L'opinion des experts est obtenue en terme de données obtenues (Lele et Das 2000, Lele 2002) et des prévisions sont obtenues par des experts pour tous les emplacements. Ces hypothèses d'experts sont alors combinées avec les données observées dans un modèle hiérarchique. Les gains de l'approche hiérarchique sont substantiels : l'estimation est améliorée (EQM inférieur pour les covariables d'habitat), la prévision mesurée par ROC est meilleure, et il est possible de distinguer selon les experts en ajoutant de l'information et en ajoutant un bruit sur les experts. Une illustration de cette méthodologie est donnée en utilisant une expérience réalisée au Montana pour prévoir la présence/absence de la population de musaraignes de montagne.

Monday, May 27th/Lundi 27 mai, 13:45 **TSH B106**

Optimal mean-variance portfolio strategy in a multiperiod setting
Stratégie optimale du portefeuille de moyenne-variance dans un contexte multipériode

François Watier, Université de Sherbrooke et/and Jean Vaillancourt, Université du Québec à Hull

We propose a solution to a multiperiod Markowitz ("mean-variance") type problem when the investor's portfolio consists of a single stock and bond and where only fairly general conditions are imposed on these assets. Among the advantages of the proposed solution one finds that it is general enough to allow for the incorporation of time dependence in modelling the rate of return, as well as dependence, if one so wishes, on exogeneous variables, such as economic factors that might have the property to improve substantially our ability to assess future rate of return. Finally, a wide class of examples are presented, the examples being chosen according to the following criteria : simplicity of form of the solution, reduced complexity of the statistical estimation of parameters, computionnal accuracy and efficiency in real-time calculations.

Nous proposons une solution à un problème multipériodique de type Markowitz ("mean-variance") lorsque l'investisseur détient un portefeuille avec un titre sans risque (obligation) et un titre risqué (action) où des conditions d'ordre général sont associées à ces titres. Parmi les avantages de la solution proposée, notons qu'elle est suffisamment générale pour permettre une dépendance temporelle à l'intérieur de la modélisation du rendement ainsi qu'une dépendance, si désirée, sur des variables exogènes telles que des indicateurs économiques qui pourraient présenter la propriété d'accroître notre capacité d'améliorer l'estimation des rendements futurs. Enfin nous présentons un éventail d'exemples ; ces exemples sont choisies en fonction des critères suivants : simplicité de la forme de la solution, complexité réduite de l'estimation statistique des paramètres, précision computationnelle et efficacité des calculs en temps réels.

Monday, May 27th/Lundi 27 mai, 14:00 **TSH B106**

CUSCORE charts for detecting sine wave signals in an autocorrelated process
Diagramme de CUSCORE pour détecter les signaux des ondes sinus dans un processus autocorrélé

Yongmin Yu et/and Roman Viveros, McMaster University

CUSUM control charts are widely used in many industries to monitor processes with the objective of improving process quality and productivity. However CUSUM charts are only efficient for detecting set changes in a process parameter such as the mean. When periodic special causes are present, such as when harmonic cycling about the target occurs, CUSCORE charts outperform the CUSUMs for detecting this kind of signals. Actually CUSUMs are special cases of CUSCOREs. In practical use, to judge when to declare an out-of-control state, the decision interval CUSUM (DI CUSUM) is widely used. We developed a decision interval CUSCORE (DI CUSCORE) as an extension of the DI CUSUM and apply it to both independent and autocorrelated processes. Our simulations are related to the frequency and phase angle of the sine wave signal as well as the autocorrelation parameters. We use Fourier analysis and maximum likelihood methods to estimate these values. This work is motivated by data from the lumber industry, related to thicknesses of boards in a sawing process. These data sets have the feature that they are highly autocorrelated and show some periodic special causes that can be modelled by sine wave signals. The problem seems well suited for the development and application of CUSCORE charts. The talk is based on joint work with Roman Viveros.

Les diagrammes CUSUM sont largement répandus dans beaucoup d'industries pour surveiller des processus avec l'objectif d'améliorer la qualité et la productivité de ces processus. Cependant les diagrammes CUSUM sont seulement efficaces pour détecter des changements d'un paramètre de processus tel que la moyenne. Quand des phénomènes spéciaux périodiques sont présents, comme un cycle harmonique autour du paramètre cible se produit, les diagrammes CUSCORE surpassent les diagrammes CUSUM pour détecter ce genre de signaux.

En fait des diagrammes CUSUM sont des cas spéciaux des diagrammes CUSCORE. En pratique, pour déclarer un état hors de contrôle, l'intervalle de décision CUSUM (DI CUSUM) est largement répandu. Nous avons développé un intervalle de décision CUSCORE (DI CUSCORE) comme généralisation des DI CUSUM et nous l'appliquons à des processus indépendants et autocorrélés. Nos simulations sont liées à la fréquence et à la phase de l'angle sinus de l'onde du signal aussi bien qu'aux paramètres d'autocorrélation. Nous employons des méthodes d'analyse de Fourier et du maximum de vraisemblance pour estimer ces valeurs. Ce travail est motivé par des données de l'industrie du bois de charpente, liées aux épaisseurs des panneaux dans un processus de sciage. Ces données ont la particularité d'être fortement autocorrélées et montrent quelques phénomènes spéciaux périodiques qui peuvent être modélisées par des signaux d'onde sinus. Le problème semble bien adapté au développement

et à l'application des diagrammes CUSCORE. La présentation est basée sur un travail réalisé conjointement avec Roman Viveros.

Monday, May 27th/Lundi 27 mai, 14:15 **TSH B106**

Multi-scale redundancy analysis of multivariate spatial data: I. Methodological Aspects
Analyse de redondance multi-échelles de données spatiales multivariées: I. Aspects Méthodologiques

Pierre Dutilleul, Bernard Pelletier, Guillaume Larocque, et/and James W. Fyles, Université McGill

In this study, we investigate an approach that combines redundancy analysis (RDA; van den Wollenberg, 1977) and geostatistical tools (Isaaks and Srivastava, 1989), to estimate the portion of variation (R-squared) in a response spatial data set (Y) that is explained by a set of predictors (X) at multiple scales. In general terms, the approach is based on the prediction by cokriging or kriging of the spatial components of each variable that correspond to the spatial structures of a nested semivariogram linear model of (co)regionalization, followed by the estimation of the portion of variation in Y that is explained by X using the (co)kriged predicted values at each spatial structure in scale-specific RDA's. Following this general approach, we developed a number of procedures. Options in this development included the prewhitening and the orthogonalization of the (co)kriged predicted values prior to the multi-scale RDA, in order to remove the bias caused by spatial autocorrelation and to better meet the assumptions of the linear model of (co)regionalization. We assessed the multi-scale RDA procedures in two ways. First, we calculated theoretically the bias in the estimation of R-squared at each scale, with or without use of an effective sample size depending on the procedure. Secondly, the mean square error was evaluated from the R-squared's estimated for data sets simulated in a Monte Carlo study. In this comparison, we included another procedure in which the estimates of the coefficients of the linear model of coregionalization are used to calculate R-squared at each scale (Goovaerts, 1994; Wackernagel, 1998). Scenarios ranged from intrinsic correlation (i.e. correlations between response and predictor are the same at all spatial scales) to extreme scale dependency (i.e. correlations vary from one scale to another both in magnitude and in sign). Theoretical and simulation results indicate that in the case of intrinsic correlation, most of the procedures are efficient in estimating R-squared. When correlations differ in magnitude but not in sign among scales, the procedure using kriged and prewhitened responses with cokriged and prewhitened predictors is the most efficient. When correlations change of magnitude and sign among spatial scales, part of the information specific to each scale seems to be lost in the cross-semivariograms of the nested model, which affects the efficiency of all the procedures. In closing, we will discuss how the multi-scale RDA procedures can be used to identify the underlying spatial and correlation structures of real data sets. Details about the application of the multi-scale RDA to a case study are presented in the companion contributed paper.

References

Goovaerts, P. (1994). Study of spatial relationships between two sets of variables using multivariate geostatistics. Geoderma, 62: 93-107.

Isaaks, E. H., & Srivastava, R. M. (1989). Applied Geostatistics. New York: Oxford University Press.

van den Wollenberg, A.L. (1977). Redundancy analysis: An alternative for canonical correlation analysis. Psychometrika, 42: 207-219.

Wackernagel, H. (1998). Multivariate Geostatistics. Berlin: Spinger-Verlag.

Dans cette étude, nous évaluons une approche combinant l'analyse (canonique) de redondance (ACR; van den Wollenberg, 1977) et des outils géostatistiques (Isaaks et Srivastava, 1989) afin

d'estimer la portion de variation (R-carré) d'un ensemble de variables dépendantes spatiales (Y) qui est expliquée par un ensemble de variables explicatives (X) à de multiples échelles. En termes généraux, cette approche se base sur la prédiction par cokrigeage et krigeage des composantes de chaque variable correspondant aux structures spatiales d'un modèle linéaire de (co)régionalisation hiérarchisé ajusté aux semivariogrammes, suivie de l'estimation de la portion de variation dans Y expliquée par X à l'aide d'ACRs spécifiques à chaque échelle et utilisant les valeurs prédites par (co)krigeage à chaque structure. Dans cette approche générale, nous avons développé un nombre de procédures. Les options dans ce développement incluaient le préblanchiment et l'orthogonalisation des variables (co)krigées avant l'ACR multiéchelles, afin d'éliminer le biais causé par l'autocorrélation spatiale et de satisfaire les présupposés du modèle linéaire de (co)régionalisation. Nous avons évalué les procédures de l'ACR multiéchelles de deux façons. Premièrement, nous avons calculé de façon théorique le biais dans l'estimation du R-carré à chaque échelle, avec ou sans utilisation d'une taille d'échantillon effective selon la procédure. Deuxièmement, l'erreur quadratique moyenne a été évaluée à partir des R-carrés estimés pour des jeux de données simulés dans une étude de Monte-Carlo cette comparaison inclut également une autre procédure dans laquelle les coefficients estimés du modèle linéaire de corégionalisation sont utilisés pour le calcul du R-carré à chaque échelle (Goovaerts, 1994; Wackernagel, 1998). Les scénarios étudiés allaient de la corrélation intrinsèque (c'est-à-dire les corrélations entre réponse et prédicteur sont les mêmes d'une échelle à l'autre) à une dépendance extrême vis-à-vis de l'échelle (c'est-à-dire les corrélations varient d'une échelle à l'autre tant au niveau du signe que de la magnitude). Les résultats théoriques et par simulation indiquent que dans le cas de la corrélation intrinsèque, la plupart des procédures sont efficaces dans l'estimation du R-carré. Quand les corrélations diffèrent en magnitude mais non en signe entre les échelles, la procédure utilisant des réponses krigées et préblanchies avec des prédicteurs cokrigés et préblanchis est la plus efficace. Lorsque les corrélations changent de signe et en magnitude, une partie de l'information spécifique à chaque échelle semble être perdue dans le semivariogramme croisé du modèle hiérarchisé, ce qui affecte l'efficacité de l'ensemble des procédures. En terminant, nous discuterons de la manière dont les procédures de l'ACR multiéchelles peuvent être utilisées afin d'identifier les patrons de structure spatiale et de corrélation sous-jacents à des jeux de données réelles. Les détails de l'application de l'ACR multiéchelles à une étude de cas sont présentés dans la communication suivante.

Références

Goovaerts, P. (1994). Study of spatial relationships between two sets of variables using multivariate geostatistics. Geoderma, 62: 93-107.

Isaaks, E. H., & Srivastava, R. M. (1989). Applied Geostatistics. New York: Oxford University Press.

van den Wollenberg, A.L. (1977). Redundancy analysis: An alternative for canonical correlation analysis. Psychometrika, 42: 207-219.

Wackernagel, H. (1998). Multivariate Geostatistics. Berlin: Spinger-Verlag.

Monday, May 27th/Lundi 27 mai, 14:30 **TSH B106**

Multi-scale redundancy analysis of multivariate spatial data: II. case study of farm management in Malawi

Analyse de redondance multiéchelles de données spatiales multivariées : II. étude de cas dans la gestion des exploitations agricoles au Malawi

Bernard Pelletier, Pierre Dutilleul, Guillaume Larocque, et/and James W. Fyles, Université McGill

In smallholder farming systems of sub-Saharan Africa, assessing the performance of management practices developed to improve soil quality is made difficult by the heterogeneity of the biophysical

environment and the diversity of strategies adopted by farmers. Soil properties observed in these farming systems are affected by a multitude of interacting environmental factors and management practices that may also vary at different scales (Pelletier, 2000). In this study, we propose an analytical procedure that takes into account the multivariate, multi-scale and spatially heterogeneous nature of the data collected in these complex agroecosystems. In a watershed of central Malawi, soil properties, management practices and biophysical conditions were observed on a series of small plots located on farmers' fields. Semivariograms were used to identify the main scales at which each random variable was spatially structured. A linear model of coregionalization (LMC) including a nugget effect and two spherical structures was fitted to the sample auto- and cross-semivariograms. For each random variable, the spatial components corresponding to the three nested structures of the LMC were predicted by kriging and co-kriging. The predicted spatial components were then used in a redundancy analysis (RDA) performed at multiple scales (Dutilleul et al., 2002), to estimate the portion of variation in soil properties that could be explained by the predictor variables. Biplots of the first few axes of the RDA were also used to explore the nature of the relationships between soil properties and predictor variables and to assess whether they differed among spatial scales. This study shows that the nature and strength of the relationships between soil properties, management and biophysical variables were different among scales and that such a scale dependency should have implications for management decisions in these farming systems.

References

Pelletier, B. (2000). Management Practices, Soil Quality and Maize Yield in Smallholder Farming Systems of Central Malawi. Unpublished Ph.D. Thesis, Department of Natural Resource Sciences, McGill University, Canada.

Dutilleul, P., Pelletier, B., Larocque, G., & Fyles, J. W. (2002). Multi-Scale Redundancy Analysis of Multivariate Spatial Data: I. Methodological Aspects. Contributed paper, Annual Meeting of the Statistical Society of Canada, Hamilton, Canada.

Dans les petites exploitations agricoles de l'Afrique subsaharienne, l'évaluation de la performance des pratiques de gestion développées dans le but d'améliorer la qualité du sol est rendue difficile par l'hétérogénéité de l'environnement biophysique et la diversité des stratégies adoptées par les paysan(ne)s. Les propriétés du sol observées dans ces systèmes d'exploitation agricole sont affectées par une multitude de facteurs environnementaux et de pratiques agronomiques qui interagissent et peuvent varier à des échelles multiples (Pelletier, 2000). Dans cette étude, nous proposons une procédure d'analyse qui tient compte de la nature multidimensionnelle, multiéchelles et hétérogène des données récoltées dans ces agro-écosystèmes complexes. Dans un bassin versant du centre du Malawi, les propriétés des sols, les pratiques agronomiques paysannes et les caractéristiques biophysiques du milieu ont été observées sur un nombre de petites parcelles localisées dans les champs. Des semivariogrammes ont été utilisés pour identifier les principales échelles auxquelles les variables aléatoires étaient structurées spatialement. Un modèle linéaire de corégionalisation (MLC) incorporant un effet pépite et deux modèles sphériques a été ajusté aux semivariogrammes (autos et croisés) empiriques. Pour chaque variable aléatoire, les composantes spatiales correspondant aux trois structures hiérarchisées du MLC ont été prédites par krigeage et cokrigeage. Ces composantes spatiales prédites ont été utilisées ensuite dans une analyse (canonique) de redondance (ACR) multiéchelles (Dutilleul et al., 2002) afin d'estimer la portion de la variation des propriétés du sol qui est expliquée par les variables prédictrices. La représentation graphique des premiers axes de l'ACR a également été utilisée afin d'explorer la nature des relations entre les propriétés du sol et les variables prédictrices, et évaluer si ces relations diffèrent entre les échelles spatiales. Cette étude démontre que la nature et la magnitude des relations entre propriétés du sol, pratiques agronomiques et caractéristiques biophysiques du milieu diffèrent d'une échelle à l'autre et qu'une telle dépendance spatiale devra être prise en compte dans les décisions concernant la gestion de ces petites exploitations agricoles.

Références

Pelletier, B. (2000). Management Practices, Soil Quality and Maize Yield in Smallholder Farming Systems of Central Malawi. Thèse de doctorat non-publiée, Département des sciences des ressources naturelles, Université McGill, Canada.

Dutilleul, P., Pelletier, B., Larocque, G., & Fyles, J. W. (2002). Analyse de redondance multiéchelles de données spatiales multidimensionnelles : I. Aspects Méthodologiques. Communication, Congrès annuel de la Société Statistique du Canada, Hamilton, Canada.

Session 11: Applied Survey Methods/Méthodes d'enquête : applications

Monday, May 27th/Lundi 27 mai, 15:30 **TSH B128**

Imputation of proxy respondents in the Canadian Community Health Survey
Imputation des répondants par procuration dans l'Enquête sur la santé dans les collectivités canadiennes

Martin St-Pierre et/and Yves Béland, Statistics Canada/Statistique Canada

Between September 2000 and October 2001, the Canadian Community Health Survey (CCHS) collected a lot of information on the health of Canadians. The sample contains over 130,000 respondents distributed among 136 health regions in Canada in order to produce reliable estimates at the health region level. Among the respondents, a small percentage are proxy respondents, that is, another person in the household has answered to the questions of the survey on behalf of the selected person in the sample.

Given the private or personal subject of certain topics in the survey, several questions could not be asked through the intermediary of another person. Therefore, a non-negligible amount of information is missing for these respondents. Considering the scale of the situation in some health regions, an imputation of the missing data using a nearest neighbor approach has been developed. This article presents in details the imputation strategy as well as results of simulations done to verify his efficiency.

Entre septembre 2000 et octobre 2001, l'Enquête sur la santé dans les collectivités canadiennes (ESCC) a recueilli beaucoup d'information sur la santé des Canadiens. L'échantillon contient plus de 130 000 répondants répartis de façon à produire des estimations fiables pour 136 régions sociosanitaires au Canada. Parmi les répondants, un faible pourcentage sont des répondants par procuration, c'est-à-dire, qu'une autre personne dans le ménage a répondu aux questions de l'enquête à la place du répondant choisi. Étant donné la caractère privé ou personnel de certains sujets de l'enquête, plusieurs questions ne pouvaient être demandées par l'intermédiaire d'une autre personne. Par conséquent, une quantité non négligeable d'information est manquante pour ces répondants. Étant donné l'ampleur de la situation dans certaines régions sociosanitaires, une imputation des données manquantes utilisant une approche basée sur le plus proche voisin a été développée. Cet article présente en détails la stratégie d'imputation ainsi que les résultats de simulations effectuées pour vérifier l'efficacité de celle-ci.

Monday, May 27th/Lundi 27 mai, 15:45 **TSH B128**

Dealing with industry misclassifications in the Unified Enterprise Survey
Traiter des mauvaises classifications d'industries dans le sondage sur les entreprises unifiées

David MacNeil et/and Stuart Pursey, Statistics Canada/Statistique Canada

The Unified Enterprise Survey is a survey that brings together many of the industry surveys of Statistics Canada that were formerly isolated from each other. This integration provides an opportunity to use businesses that are discovered, during data collection, to be "misclassified by industry". This paper describes the amount of industry misclassifications that have been found; what can be done during data collection and data processing after the correct classification is determined; and proposes several methods that can be used during estimation to improve industry estimates.

L'enquête unifiée sur les entreprises est une étude qui rassemble plusieurs enquêtes sur des industries faites par Statistiques Canada et qui ont été autrefois isolées les uns des autres. Ce jumelage fournit une occasion d'utiliser les entreprises qui sont découvertes, pendant la collecte de données, "comme étant mal classées en tant qu'industrie". Cette présentation décrit la quantité de fausses classifications d'industries qui ont été trouvées; ce qui peut être fait pendant la collecte des données et l'analyse des données après que la classification correcte soit déterminée. Nous proposons aussi plusieurs méthodes qui peuvent être employées pendant l'estimation pour améliorer des estimations faites par les industries.

Monday, May 27th/Lundi 27 mai, 16:00 **TSH B128**

Weighting challenges for the Longitudinal Survey of Immigrants to Canada (LSIC)
Pondération de l'Enquête longitudinale auprès des immigrants au Canada (ELIC)

Jean-François Dubois et/and Michelle Simard, Statistics Canada/Statistique Canada

As in other surveys, LSIC is facing non-response. In order to get proper population estimates, the survey weights have to be corrected by using a non-response adjustment. A more efficient estimator is obtained if this adjustment is calculated within given classes. This is especially true if the response pattern differs from one class to the next. In most surveys, the adjustment is calculated and then applied to the survey weights to get the final weights. For LSIC, calculating the adjustment will not be as straightforward as there is a significantly higher unresolved rate than most surveys. Unresolved units are units that could not be traced nor contacted during the collection period. We are aware that some groups of immigrants land in Canada but then go to the USA or go back to their original countries after a certain period for various reasons. This leads to think that there are two main reasons why there is unresolved units: 1) the immigrant is no longer in Canada so even the best sources of tracing don't permit to find him; 2) the immigrant is really in Canada but operational constraints prevent us from finding him. The paper presents some new approaches to adjust the responding units based on various models that predicts an estimated rate of inscope immigrants in the unresolved portion. Comparison of several methods to create the classes of adjustment will be presented. Different strategies will be discussed and evaluated.

Comme dans bien des enquêtes, l'ELIC fait face à la non-réponse. Une des façons d'obtenir des estimations adéquates consiste à calculer un facteur d'ajustement de non-réponse. Afin d'obtenir un estimateur plus efficace, l'ajustement est calculé à l'intérieur de classes données. Ceci est particulièrement pertinent si on croit que le mécanisme de réponse diffère d'une classe à l'autre. Dans la plupart des enquêtes, l'ajustement est calculé puis appliqué aux poids de sondage afin d'obtenir les poids finaux. Pour l'ELIC, le calcul de l'ajustement ne sera pas aussi simple puisque que nous observons un taux plus important d'unités non résolues que la plupart des enquêtes. Les unités non résolues sont des unités n'ayant pu être dépistées ou contactées au cours de la période de collecte. Nous somme conscients que certains groupes d'immigrants arrivent au Canada mais finissent par aller aux É.-U. ou dans leur pays d'origine pour diverses raisons. Ceci porte à penser qu'il existe deux raisons principales expliquant la présence d'unités non-résolues : 1) l'immigrant n'est plus au Canada et donc même les meilleures sources de dépistage ne permettent pas de le retracer; 2) l'immigrant est

effectivement au Canada mais ne peut être retrouvé à cause de contraintes opérationnelles. L'article présente de nouvelles approches pour ajuster les unités répondantes basées sur divers modèles permettant de prédire un taux estimé d'immigrants ciblés à l'intérieur de la portion non résolue. Des comparaisons de diverses méthodes permettant de créer les classes d'ajustement seront présentées. Différentes stratégies seront discutées et évaluées.

Monday, May 27th/Lundi 27 mai, 16:15 **TSH B128**

The outlier detection and treatment strategy for the Monthly Wholesale and Retail Trade Survey of Statistics Canada

La stratégie de détection et de traitement des valeurs aberrantes pour l'Enquête mensuelle sur le commerce de gros et de détail de Statistique Canada

Steven Matthews and/et Hélène Bérard, Statistics Canada/Statistique Canada

Statistics Canada conducts a major survey known as the Monthly Wholesale and Retail Trade Survey (MWRTS), which produces estimates based on monthly data collected for sales and inventories at various geographic and industry levels. The sales trend is used as an important economic indicator, and the monthly sales estimates form a substantial portion of the monthly estimates for the Gross Domestic Product (GDP). The MWRTS is currently being redesigned, in part to produce estimates according to the new North American Industry Classification System (NAICS) and to take full advantage of the availability of administrative data from the Goods and Services Tax (GST) program. Although many improvements will be implemented, such as reduction of frame misclassification by use of administrative data and innovative sample update processes, influential missclassified units will continue to occur and specific procedures must be put in place to treat them. This talk presents the overall strategy developed to reduce the effect of these influential units. The results from an empirical study that compares the efficiency of four methods to identify and treat outliers are presented and the implementation of these methods in the context of a monthly production will be discussed.

L'Enquête mensuelle sur le commerce de gros et de détail (EMCGD) produit des estimations mensuelles ventilées par régions géographiques et groupes industriels en utilisant des données collectées pour les ventes et les inventaires. La tendance des ventes d'un mois à l'autre constitue un important indicateur économique, de plus les estimations mensuelles des ventes forment une portion substantielle des estimations mensuelles pour le produit intérieur brut (PIB). L'EMCGD est actuellement dans un processus de remaniement en partie pour produire des estimations suivant le nouveau système de classifications des industries de l'Amérique du Nord (SCIAN), et pour bénéficier de la disponibilité de données administratives provenant du programme des taxes sur les produits et services (TPS). Dans la nouvelle enquête, plusieurs procédures seront mises en place afin de réduire les erreurs de classification dans la base de sondage dont une meilleure utilisation des données administratives ainsi que l'introduction d'un processus innovateur pour la mise à jour de l'échantillon. Toutefois, malgré ces nouvelles procédures, des erreurs de classification subsisteront et des procédures spéciales doivent être élaborées afin de traiter les unités influentes qui sont mal classées. Cette communication présentera la stratégie développée pour réduire l'effet de ces unités influentes. Les résultats d'une étude empirique qui compare l'efficacité de quatre méthodes pour identifier et traiter les valeurs aberrantes seront présentés de même que l'intégration de ces méthodes dans le contexte d'une production mensuelle.

Monday, May 27th/Lundi 27 mai, 16:30 **TSH B128**

On response errors in the Canadian Community Health Survey
Étude sur les erreurs de réponse dans le cadre de l'Enquête sur la santé dans les collectivités canadiennes
Fritz Pierre et/and Yves Béland, Statistics Canada/Statistique Canada

To remedy the principle statistical gaps relating to health determinants, health status and health system utilization at the level of the health regions, Statistics Canada has implemented a new survey, the Canadian Community Health Survey (CCHS). The CCHS is comprised of two distinct surveys: a health region-level survey in the first year with a total sample of 130,000, for which data collection finished in October 2001, and a provincial-level survey in the second year with a total sample of 30,000. The first objective of the regional survey was to produce cross-sectional estimates for 136 health regions across the country. The household sample for the CCHS is drawn from two overlapping sampling frames: the areal frame set up for the Labour Force Survey, which includes computer-assisted personal and telephone interviews (CAPI and CATI), and the frame of randomly composed telephone numbers within which computer-assisted telephone interviews (CATI) were carried out.

This article illustrates the extent of response errors in the regional component of the CCHS. More specifically, it takes a first look at the link between the form of declaration (in person or by proxy) and the prevalence of certain health problems. It also shows the impact of the mode of collection (CAPI or CATI) on different health determinants.

Dans le but de remédier aux principales lacunes statistiques en ce qui a trait aux déterminants de la santé, à l'état de santé et à l'utilisation du système de santé de la population canadienne à l'échelle des régions socio-sanitaires, Statistique Canada a élaboré une nouvelle enquête, l'Enquête sur la santé dans les collectivités canadiennes(ESCC). L'ESCC est une enquête constituée de deux composantes: une composante régionale la premièer année , auprès d'un échantillon de plus 133 000 répondants dont la collecte s'est terminée en octobre 2001 et une composante provinciale la deuxième année au près de 30 000 répondants. Le but premier de la composante régionale était de produire des estimations transversales pour 136 régions socio-sanitaires du Canada. L'échantillon de ménages de l'ESCC est sélectionné à partir de deux bases de sondage chevauchantes: la base aréolaire mise en place pour l'Enquête sur la Population Active à l'intérieur de laquelle des interviews personnelles et téléphoniques assistées par ordinateur(IPAO et ITAO) ont été effectuées et la base de composition aléatoire de numéros de téléphone où des interviews téléphoniques assistées par ordinateur(ITAO) ont été effectuées.

Le présent article illustre la portée des erreurs de réponse dans la composante régionale de l'ESCC. Plus précisement, il étudie dans un premier temps, le lien entre la forme de déclaration(en personne ou par procuration) et la prévalence de certains problèmes de santé. Il expose également l'impact du mode de collecte(IPAO ou ITAO) sur différents déterminants de la santé.

Session 12:Nonparametric Methods and Density Estimation/ Méthodes non paramétrique et estimation de densité

Monday, May 27th/Lundi 27 mai, 15:30 **TSH B106**

A better confidence interval for the difference between two binomial proportions of paired data
Un meilleur intervalle de confiance pour la différence entre deux proportions binomiales de données pairées
Gengsheng Qin, Georgia State University et/and Xiao-hua Zhou, Indiana University

Motivated by a study on comparing sensitivities and specificities of two diagnostic tests in a paired design when the sample size is small, we first derived an Edgeworth expansion for the studentized difference between two binomial proportions of paired data. This Edgeworth expansion helps us understand why the usual Wald interval for the difference has poor coverage performance in the small sample size. Based on the Edgeworth expansion, we then derived a transformation based confidence interval for the difference. The new interval removes the skewness in the Edgeworth expansion. It is easy to compute, and its coverage probability converges to the nominal level at a rate of $O(n^{-1/2})$. Simulation results indicate that the new interval has the coverage probability that is very close to the nominal level even for sample sizes as small as 10. Simulation results also indicate this new interval has better average coverage accuracy than the best existing interval in the finite samples.

Motivé par une étude sur la comparaison des sensibilités et des spécificités de deux tests de diagnostic dans un plan d'expérience pairé lorsque la dimension de l'échantillon est petite, nous avons d'abord déduit un développement en série d'Edgeworth pour studentiser la différence entre deux proportions binomiales de données pairées. Cette série d'Edgeworth nous aide à comprendre pourquoi l'intervalle de Wald pour la différence a une faible performance de couverture quand la taille de l'échantillon est petite.

Basé sur la série d'Edgeworth, nous avons alors déduit un intervalle de confiance pour la différence basée sur une transformation. Le nouvel intervalle enlève l'asymétrie dans la série d'Edgeworth. Il est facile à calculer, et sa probabilité de couverture converge au niveau nominal à un taux de $O(n^{-1/2})$.

Les résultats de simulations indiquent que le nouvel intervalle a une probabilité de couverture qui est très près du niveau nominal même pour des dimensions d'échantillon aussi petites que 10. Les résultats de simulations indiquent également que ce nouvel intervalle a une meilleure exactitude dans la couverture moyenne que le meilleur intervalle existant pour des échantillons finis.

Monday, May 27th/Lundi 27 mai, 15:45 **TSH B106**

An arc model for analyzing ranking data
Un modèle basé sur l'arc pour l'analyse de données ordonnées

Mayer Alvo, Université d'Ottawa et/and Paul Smrz, University of Newcastle, Australia

Distance based models for ranking data have been in use for some time. The model based on Spearman distance is exponential in nature whereas the one based on Kendall distance is not. The latter is somewhat difficult to work with. In this paper, we propose a new model based on the arc connecting two rankings. It is shown that this model is close to Kendall's and has nice properties.

Les modèles basés sur la distance entre les données ordonnées existe depuis longtemps. Celui basé sur la distance de Spearman est exponentiel tandis que celui basé sur la distance de Kendall ne l'est pas. Dans cette présentation, on propose un nouveau modèle basé sur la distance mesurée par l'arc entre deux ordonnancements. Ce dernier possède des propriétés intéressantes et ressemble beaucoup au modèle basé sur la distance de Kendall.

Monday, May 27th/Lundi 27 mai, 16:00 **TSH B106**

Another look at the kernel density estimator for non negative support
Une nouvelle vision de l'estimation de la densité par la méthode du noyau pour un support non négatif

Yogendra Chaubey et/and Pranab K. Sen, Concordia University

This paper uses a generalization of of the technique used for estimating various functionals of the distribution function introduced in Chaubey and Sen (1996) and derives the usual Kernel density estimator proposed by Rosenblatt (1956) and Parzen (1958). This approach is seen to offer an

alternative approach of specifying asymmetric kernels for estimating the densities with non-negative support which take into account the whole data in contrast to the approach used in Prakasa-Rao and Bagai (1996) which may omit some observations depending on the location of the point where the density is desired. The properties of the alternative estimator are further investigated.

Cette présentation utilise une généralisation de la technique utilisée pour estimer diverses fonctionnelles pour des distributions présentée dans Chaubey et Sen (1996) et nous dérivons l'estimateur habituel de la densité par la méthode du noyau proposé par Rosenblatt (1956) et Parzen (1958). Cette approche peut offrir une approche alternative pour spécifier les noyaux asymétriques dans l'estimation des densités avec un support non négatif et tient compte la totalité des données contrairement à l'approche utilisée en Prakasa-Rao et Bagai (1996) qui peut omettre quelques observations selon l'emplacement du point où la densité est considérée. Les propriétés de notre estimateur sont étudiées plus en profondeur.

Monday, May 27th/Lundi 27 mai, 16:15 **TSH B106**

Nonparametric estimation of possibly similar densities
Estimation non paramétrique de densités possiblement similaires
Alan Ker, University of Arizona

Often it is necessary to estimate a set of densities $f_1, ...f_Q$ which are thought to be of similar structure. In a parametric framework, similarity may be imposed by assuming $f_i \in G_\Theta \; \forall \; i = 1, ..., Q$ where $G_\Theta = \{f_\theta : \theta \in \Theta\}$ while G_Θ is also assumed to be known. A class of nonparametric methods, inspired by the work of Hjort and Glad (1995), are developed that offer greater efficiency if the set of densities are similar while not losing any if the set of densities are quite dissimilar. Both theoretical properties and finite sample performance are found to be very promising. The developed estimator is relatively easy to implement and may be combined with semiparametric and bias reduction techniques.

Souvent il est nécessaire d'estimer un ensemble de densités $f_1, \ldots, f_q$ que nous pensons être de structure semblable. Dans un cadre paramétrique, la similitude peut être imposée en assumant f_i est élément de G_Θ pour tout $i = 1, \ldots, Q$ où $G_\Theta = (f_\theta | \theta$ élément de $\Theta)$ tandis qu'on assume également G_θ est connu. Une classe de méthodes non paramétriques, inspiré du travail de Hjort et Glad (1995), est développée qui offre une meilleure convergence si l'ensemble des densités est relativement semblable tandis que rien n'est perdu si les densités sont plutôt différentes. Des propriétés théoriques et des performances pour un échantillon fini s'avèrent très prometteuses. L'estimateur développé est relativement facile à mettre en application et peut être combiné avec des techniques semi-paramétriques et de réduction de biais.

Monday, May 27th/Lundi 27 mai, 16:30 **TSH B106**

An empirical Csorgo-Horvath type CLT for LP norms of density estimates
Une approche de Csorgo-Horvath empirique du théorème de la limite centrale pour des estimations de densités avec la norme Lp
Majid Mojirsheibani, Carleton University

We consider empirical Lp norms of kernel density estimates. This corresponds to replacing integrals by sums of non-iid random variables, which are generally easier to compute. We also give central limit theorems for such estimates.

Nous considérons la classe des normes empiriques Lp pour l'estimation de densité par la méthode du noyau. Ceci correspond à substituer des intégrales par des sommes de variables aléatoires non i.i.d.

qui sont généralement plus facile à calculer. Nous donnons également des théorèmes de limite centrale pour de telles estimations.

Monday, May 27th/Lundi 27 mai, 16:45 **TSH B106**

A semiparametric method of boundary correction for kernel density estimation
Une méthode semi-paramétrique pour la correction des frontières dans l'estimation d'une densité par la méthode du noyau
R. Karunamuni et/and T. Alberts, University of Alberta

We propose a new estimator for boundary correction for kernel density estimation. Our method is based on local Bayes techniques of Hjort(1998). The resulting estimator is semiparametric type estimator: a weighted average of an intial guess and the ordinary reflection method estimator. The proposed estimator is seen to perform quite well compared with the other existing well-known estimators for densities which have the shoulder condition.

Nous proposons un nouvel estimateur pour la correction aux frontières pour l'estimation d'une densité par la méthode du noyau. Notre méthode est basée sur des techniques locales bayésiennes de Hjort(1998). L'estimateur résultant est un estimateur semi-paramétrique de type : une moyenne pondérée d'une idée intiale et de l'estimateur ordinaire obtenu par la méthode de réflexion. L'estimateur proposé performe bien comparé aux autres estimateurs existants bien connus pour les densités qui ont les "épaules" mal conditionnées.

Session 13: Multiscale Behavior in Stochastic Models in Population Biology and Physics / Comportement multi-échelle dans les modèles stochastiques dans la biologie de la population et la physique

Monday, May 27th/Lundi 27 mai, 15:30 **TSH B105**

Multiplicative cascades and partial differential equations
Cascades multiplicatives et équations différentielles partielles
Ed Waymire, Oregon State University

A fundamental unsolved problem of contemporary applied mathematics is to decide whether smooth, physically reasonable solutions to the Navier-Stokes equations of fluid motion will exist for all time given smooth, physically reasonable initial data; e.g. see C.L.Fefferman (2000) paper on the Clay Institute Millenium Problems at http://www.clayinstitute.org. As remarked by Fefferman, "Standard methods of PDE appear inadequate to settle the problem. Instead, we probably need some deep, new ideas." In this talk we shall develop some new ideas from probability originating in work by Y. LeJan and A.S.Sznitman (1997), Probab. Theory Relat. Fields, 109, 343-366. In particular we extend these ideas to representations of Fourier transforms of solutions to broad classes of partial differential equations, which include Navier-Stokes equations in two and higher dimensions, in terms of expected values of products of given initial data and forcings over nodes of a multitype branching random walk in Fourier space. We explore new approaches to obtain unique, global, regular solutions based on such representations. This talk is based on joint work with Rabi Bhattacharya, Indiana University, and Larry Chen, Scott Dobson, Ronald Guenther, Mina Ossiander, Enrique Thomann, and Chris Orum at Oregon State University, partially supported by a Focussed Research Grant from the National Science Foundation.

Un problème fondamental non résolu des mathématiques appliquées contemporaines est de décider si les solutions lisses et physiquement raisonnables aux équations de Navier-Stokes du mouvement

d'un liquide existeront toujours si on considère des données initiales physiquement raisonnables; par exemple voir l'article de C.L. Fefferman (2000) sur les problèmes du millénaire du Clay Institute à http://www.clayinstitute.org.

Comme remarqué par Fefferman, "les méthodes standard de EDP semblent insatisfaisantes pour régler le problème. Au lieu de cela, nous avons probablement besoin de quelques idées substantielles et nouvelles". Dans cette présentation nous développerons quelques nouvelles idées des probabilités provenant du travail de Y. LeJan et A.S. Sznitman (1997), Probab. Theory Relat. Fields, 109, 343–366. Nous étendrons ces idées aux représentations des transformées de Fourier comme solutions à de grandes classes d'équations différentielles partielles, qui incluent des équations de Navier-Stokes en dimensions deux ou plus, en termes de valeurs prévues des produits à partir de données initiales et restreint au-dessus des noeuds d'une marche aléatoire à embranchement multitypes dans l'espace de Fourier. Nous explorons de nouvelles approches pour obtenir des solutions globales uniques et régulières basées sur de telles représentations. Cette présentation est basée sur un travail commun avec Rabi Bhattacharya, Indiana University, ainsi que Larry Chen, Scott Dobson, Ronald Guenther, Mina Ossiander, Enrique Thomann et Chris Orum du Oregon State University, et partiellement supporté par Focussed Research Grant de la National Science Foundation.

Monday, May 27th/Lundi 27 mai, 16:15 **TSH B105**

Hierarchical approach to multiscale phenomena in stochastic population models
Approche hiérarchique d'un phénomène à échelle multiple dans un modèle de population stochastique

Donald Dawson, Carleton University

We describe an approach to modeling phenomena in different space and time scales. This is based on an extension of the mean-field models commonly used in scientific applications and involves looking at the small or large scale behaviors in a range of essentially different space and times scales and relations among the different scales. A number of examples from population genetics, both neutral and with selection, as well as from interacting particle systems such as branching particle systems and voter models will be used to illustrate these ideas.

Nous décrivons une approche pour modéliser des phénomènes dans différents espaces et échelles de temps. Ceci est basé sur une généralisation du "champ moyen" : modèles généralement utilisés dans les applications scientifiques. Ils impliquent de regarder les comportements à petite ou à grande échelle dans différents espaces et échelles de temps ainsi que les relations existant entre ces différentes échelles. Un certain nombre d'exemples sur la génétique des populations, modèle neutre ou avec sélection, tout comme les systèmes d'interaction des particules tels que les systèmes de particules de branchement et les modèles de votes seront employés pour illustrer ces idées.

Session 14: Statistics for Microarray Data Analysis/La statistique pour l'analyse des données des microréseaux

Monday, May 27th/Lundi 27 mai, 15:30 **KTH B135**

On mixture model calculations for gene expression analysis
Sur les calculs reliés aux modèles de mélanges de lois dans des analyses pour l'expression des gènes

Michael Newton and C.M. Kendziorski, University of Wisconsin at Madison

I will review the formulation of mixture models for gene expression analysis and describe experiences with the mixture methodology from a series of expression studies done in Madison. The methodology as we have implemented it involves several layers of mixing – over discrete patterns of differential expression and over quantitative underlying expression levels – and represents the first empirical Bayesian approach to expression data analysis. Model assumptions yield both parametric and semiparametric inferences about differential expression. I will try to demonstrate the utility of reporting the odds of various forms of differential expression, and I will discuss model fitting, model checking, and the role of an interesting arithmetic-geometric mean ratio.

Je passerai en revue la formulation des modèles de mélanges pour l'analyse d'expression de gènes et décrirai des expériences avec la méthodologie sur les mélanges d'une série d'études d'expression de gènes faites à Madison. La méthodologie, comme nous l'avons mise en application, implique plusieurs niveaux de mélanges – sur des configurations discrètes d'expression différentielle et sur des niveaux d'expression quantitatif sous-jacents au modèle – et représente la première approche bayésienne empirique à l'analyse de données d'expression de gènes. Les hypothèses du modèle engendrent des inférences paramétriques et semi-paramétriques au sujet de l'expression différentielle. J'essayerai de démontrer l'utilité de quantifier la pertinence de diverses formes d'expression différentielle, et je discuterai de l'ajustement de modèles, le contrôle de modèles, et le rôle intéressant du rapport des moyennes arithmétiques et géométriques.

Monday, May 27th/Lundi 27 mai, 16:00 **KTH B135**

Hybrid hierarchical clustering with applications to microarray data
Regroupements hiérarchiques hybride avec applications à des données de microréseaux

Hugh Chipman, University of Waterloo and/et Rob Tibshirani, Stanford University

Bottom-up hierarchical clustering algorithms are an important analysis tool for microarray data, organizing the rows and columns of the data matrix so as to reveal patterns among samples and/or genes. Hierarchical methods are especially useful, because interest simultaneously focuses on many small clusters (e.g. repeat samples) and a few large clusters (e.g. different prognosis groups). Bottom-up clustering, which successively joins objects, is good at identifying small clusters, but can provide sub-optimal performance for identifying a few large clusters. Conversely, top-down methods are good at identifying a few large clusters, but weaker at many small clusters. We seek to combine the strengths of both approaches, modifying top-down procedures with information gained from a preliminary bottom-up clustering. A useful concept is that of a mutual cluster, which is a group of objects which, collectively, are closer to each other than to any other object. We shall illustrate how this technique and others can be used to produce more effective hierarchical clusterings, with examples using microarray data.

Des algorithmes de regroupements hiérarchiques ascendants sont un outil important pour l'analyse de données de microréseaux, organisant les rangées et des colonnes de la matrice de données afin

d'indiquer des modèles dans les échantillons et/ou les gènes. Les méthodes hiérarchiques sont particulièrement utiles, car l'intérêt est à la fois porté sur plusieurs petits regroupements (par exemple des échantillons de répétition) et quelques grands regroupements (par exemple différents groupes de pronostic). Les regroupement ascendants, qui joigne successivement des objets, sont également bon pour identifier de petits regroupements, mais peuvent fournir des performances non optimales pour identifier quelques grands regroupements. Réciproquement, les méthodes descendantes sont bonnes pour identifier quelques grands regroupements, mais plus faibles à pour identifier plusieurs petits regroupements. Nous cherchons à combiner les forces de ces deux approches en modifiant les procédures descendantes compte tenu de l'information obtenue par les regroupements ascendants préliminaires. Un concept clé est celui du regroupement mutuel, qui est essentiellement un groupe d'objets qui collectivement, sont plus près les uns des autres que n'importe quel autre objet. Nous illustrerons comment cette technique et d'autres peuvent être employées pour produire des regroupements hiérarchiques plus efficaces avec des données de microréseaux.

Monday, May 27th/Lundi 27 mai, 16:30 **KTH B135**

Multiple testing in large-scale gene expression experiments
Tests multiples dans des plans d'expérience à grande échelle sur l'expression des gènes
Terry Speed et/and Yongchao Ge, University of California Berkeley

Large-scale gene expression experiments using techniques such as SAGE or oligonucleotide or cDNA microarrays typically involve tens of thousands of genes. Such experiments are frequently comparative, so that we make tens of thousands of comparisons, seeking the "significant" ones. Under such circumstances care needs to be taken to adjust appropriately for the number of tests taken. Not surprisingly, Bonferroni is usually too severe, but the tests are correlated so that no simple alternative exists. In this talk I will outline some of the approaches which have been taken to this issue and mention some outstanding problems.

Les expériences à grande échelle pour l'identification d'un gène utilisant des techniques telles que SAGE, l'oligonucléotide ou encore les microréseaux de cADN impliquent généralement des dizaines de milliers de gènes. De telles expériences sont fréquemment basées sur des comparaisons, de sorte que nous faisons des milliers de comparaisons, recherchant celles qui sont "significatives". Sous de telles conditions,il faut prendre soin d'ajuster convenablement les méthodes au nombre de tests effectués. La méthode proposée par Bonferroni est habituellement trop sévère, mais comme les tests sont corrélés aucune alternative simple existe. Dans cette présentation je tracerai les grandes lignes de certaines des approches qui ont été adaptées à ce problème et je mentionnerai quelques problèmes en suspens.

Session 15: Official Statistics/La statistique officielle

Tuesday, May 28th/Mardi 28 mai, 8:30 **TSH B106**

Methodological issues in producing income statistics
Problèmes méthodologiques des statistiques sur le revenu
Sylvie Michaud, Statistics Canada/Statistique Canada

The Income Statistics Division's mandate is to produce information on the financial well-being of Canadians and their families. To do so, information is collected, mostly through surveys, on income, expenditures, wealth and pensions. Because of the quantitive nature of the data, distributed in a highly skewed fashion, the use of sound statistical methods in various parts of the process is very

important. This is done at the sampling stage, for the edit and imputation processes, for outlier detection, at the weighting processes or at the analysis stage. The presentation will briefly describe some of the surveys that are done in income statistics division, and the role the statistics play in producing accurate data.

Le mandat de la Division des statistiques sur le revenu doit produire de l'information sur le bien-être financier des Canadiens et de leurs familles. Pour ce faire, l'information est recueillie, la plupart du temps par des sondages, sur le revenu, les dépenses, la richesse et les pensions. En raison de la nature quantitative des données fortement asymétriques, l'utilisation des méthodes statistiques appropriées dans diverses parties du processus est très importante. Ceci est fait à l'étape de la collecte des données, pour la modification ou l'imputation des données, la détection des valeurs aberrantes se fait à l'étape de la pondération des processus et de l'analyse. La présentation décrira brièvement certaines des études qui sont faites dans la Division des statistiques de revenu, et le rôle que les statistiques jouent en produisant des données précises.

Tuesday, May 28th/Mardi 28 mai, 9:00 **TSH B106**

Citizen security and personal information confidentiality: governmental statistical organizations approach.

La sécurité et la confidentialité des renseignements personnels ou sensibles sur les citoyens : l'approche des organismes statistiques gouvernementaux

Louise Bourque, l'Institut de la statistique du Québec

In most countries, government statistical organizations have the legislated right to gather information on their citizens. In consequence, official statistical organizations have become major collectors of data on individuals, businesses, institutions, etc. The information they request concern subject matter that is diverse and often sensitive: education; health and welfare; culture; commercial, industrial or financial activity; work and employment, etc. To counterbalance the powers of statistical organizations, legislators have created laws to protect the confidentiality of this information. In consequence, the statistical organizations are bound by policies, procedures, ethical guidelines, etc., aimed at assuring the confidentiality and security of their data.This talk will give a general overview of the policies and regulations currently in effect to protect the confidentiality and security of dataheld by Canadian, Quebec and European government statistical organizations.

Les organismes statistiques gouvernementaux de la plupart des pays au monde ont des pouvoirs de recueillir des informations sur les citoyens en vertu de lois. Ces organismes sont en conséquence de grands cueilleurs d'informations tant auprès des personnes que des entreprises, des institutions, etc. Les informations demandées portent sur des sujets très divers, et parfois sensibles : éducation, santé et bien-être, culture, activité commerciale, industrielle ou financière, travail et emploi, etc.

Pour équilibrer les pouvoirs des organismes statistiques, les législateurs ont établi l'obligation juridique de protéger la confidentialité des renseignements détenus par ceux-ci. En conséquence, ces organisations se dotent de politiques, procédures, règles d'éthique, etc., qui visent à assurer la confidentialité et la sécurité des renseignements qu'ils détiennent.

La conférence donnera un aperçu général des politiques et règles qui sont en vigueur en matière de confidentialité et sécurité des informations confidentielles détenues par les organismes statistiques gouvernementaux canadien, québécois et européens.

Tuesday, May 28th/Mardi 28 mai, 9:30 **TSH B106**

Improving the quality of cross-national data - meeting the challenge
Amélioration de la qualité des données transnationales - relever le défi
Denise Lievesley, UNESCO Institute for Statistics

The collection of comparable data must be conceptually well-anchored and is heavily dependent upon the use of standardised classifications of key variables. The development and maintenance of such classifications is expensive and time-consuming but without this work "comparability is only skin deep". It is essential to ensure that differences between countries are real and are not an artefact of the data collection method or a reflection of differences in administrative systems. The challenge we face is how to develop classifications which have conceptual clarity and which are relevant to a wide range of different countries. Improving the quality of cross-national data when many countries have weak statistical systems and when a culture of data integrity is often lacking is a significant challenge. Denise Lievesley will discuss some of these challenges from her perspective of directing a statistical unit within the UN system.

La collecte de données comparables doit être conceptuellement bien ancrée et dépend fortement de l'utilisation des classifications normalisées des variables principales. Le développement et l'entretien de telles classifications sont très dispendieux et prennent énormément de temps mais, sans ce travail la "comparabilité est seulement superficielle". Il est essentiel de s'assurer que les différences entre les pays soient véritables et qu'elles ne dépendent pas de la méthode de collecte des données ou d'une différence entre les systèmes administratifs. Le défi auquel nous devons faire face est de développer un système de classifications qui est conceptuellement clair et qui est appropriée pour de nombreux pays différents. L'amélioration de la qualité des données transnationales lorsque beaucoup de pays ont un système statistique faible et lorsqu'une culture d'intégrité de données manque souvent, est un grand défi. Denise Lievesley discutera certains de ces défis à titre de directrice d'une unité statistique de l'ONU.

Session 16: Meta-analysis and Clinical Trials/Méta- analyse et essais cliniques

Tuesday, May 28th/Mardi 28 mai, 8:30 **KTH B135**

Internal validation versus external validation: a case study
Validation interne contre validation externe : une étude de cas
Emma Bartfay, Queen's University

Studies relating to factors that have prognostic significance in many malignant diseases are abundant in the literature. To evaluate the accuracy of the outcome predictions, one approach is to develop prognostic systems and assess whether the results can be generalized onto future patients. This is called system validation, which can be broadly classified into internal and external validation. Although external validation has been advocated by many, studies that solely rely on internal validation remained widely used. The purpose of this presentation is to demonstrate that there are circumstances when internal validation can lead to misleading conclusions. I utilized two data sets obtained from the National Cancer Institute of Canada (NCIC) clinical trials of patients with limited stage small-cell lung cancer. Internal validation was evaluated using bootstrapping on the original data set. External validation was conducted by applying the frozen models onto our second data set. Accuracy was quantified by calibration curves, Brier scores and c-index. The results showed that internal validation

can lead to different conclusions when compared to external validation. I concluded that external validation should be included in system validation, when possible.

Les études concernant les facteurs qui sont significatifs dans le pronostique de beaucoup de maladies malignes sont abondantes dans la littérature. Pour évaluer l'exactitude des prévisions des résultats, une approche est de développer les systèmes de pronostiques et d'évaluer si les résultats peuvent être généralisés sur de futurs patients. Ceci s'appelle la validation de système et peut être classifiée dans la validation interne et externe.

Bien que la validation externe ait été préconisée par plusieurs, les études qui se fondent seulement sur la validation interne sont encore énormément utilisées. Le but de cette présentation est de démontrer qu'il existe des circonstances où la validation interne peut mener à des conclusions fallacieuses.

J'ai utilisé deux jeux de données fournis par l'Institut national des épreuves cliniques sur le cancer au Canada (NCIC) limités à des patients ayant un cancer du poumon sur de petites cellules. La validation interne a été effectuée en utilisant le rééchantillonnage sur le jeu de données initial. La validation externe a été faite en appliquant les modèles "gelés" sur notre deuxième jeu de données. La précision a été mesurée par des courbes de calibration, des scores de Brier et des indices-c. Les résultats ont prouvé que la validation interne peut mener à différentes conclusions une fois comparée à la validation externe. J'en conclu que la validation externe devrait être incluse dans la validation de système si possible.

Tuesday, May 28th/Mardi 28 mai, 8:45 **KTH B135**

Statistical methods for detecting non-statistical bias
Méthodes statistiques pour la détection du biais non statistique
Nicholas Barrowman, Thomas C. Chalmers Centre for Systematic Reviews

In health care research, bias generated by statistical methods is often secondary in importance to other sources of bias. Poorly conducted ("low quality") clinical trials, for example, may overestimate treatment effects. Comparing effect estimates of low and high quality trials that are otherwise clinically comparable permits estimation of systematic differences in effect estimates. Conveniently, groups of comparable trials are available from meta-analyses. However questions remain about optimal statistical methods for combining information across meta-analyses. In this talk, I will introduce several methods, including one I have developed, compare their performance, and describe future research directions.

Dans le domaine de la recherche en santé, le biais généré par les méthodes statistiques est souvent secondaire comparé à d'autres sources de biais. Les essais cliniques de mauvaise qualité, par exemple, peuvent surestimer l'effet du traitement. Comparer les estimés d'essais cliniques de bonne et mauvaise qualité qui sont cliniquement comparables permet l'estimation des différences systématiques des estimés de l'effet. Des groupes d'essais cliniques comparables sont facilement disponibles dans les méta-analyses. Cependant, des questions demeurent encore quant aux méthodes statistiques optimales pour combiner l'information entre méta-analyses. Dans cet exposé, j'introduirai quelques méthodes, y compris une que j'ai développée, je comparerai leur performance, et finalement, je donnerai des pistes pour les recherches futures.

Tuesday, May 28th/Mardi 28 mai, 9:00 **KTH B135**

Probabilistic assessment of study quality in meta-analysis
Évaluation probabiliste de la qualité d'étude en méta-analyse
Shagufta Sultan, Health Canada et/and Robert Platt, McGill University

Meta-analysis is a set of statistical procedures designed to accumulate experimental and correlational results across independent studies that address a related set of research questions. It uses the summary statistics from individual studies as the data points. The quality of studies in meta-analysis can vary widely. This variation is usually described by assigning quality scores to studies. A natural way to incorporate quality scores into an analysis is to weight each study by its quality score in addition to its size or inverse of the variance of the estimate for that study. This method has been used by several authors but has been criticized due to its subjectivity. To improve on this and other existing approaches, Tritchler (1999) introduced a probability model for assessing the effect of quality on a summary effect measure. He derived a number of summarization methods and compared them for simulated data and also applied them on an actual set of studies for fixed effect models.

We performed two evaluations of Tritchler's approach. We compared the probability model to other methods for quality adjustment in a simulation study. In this study we used a distribution of quality scores that adequately reflects the quality scores in meta-analytic data. In general, the probability model gives little bias, less mean-squared error and better coverage probabilities for confidence intervals than do other methods. In addition, we compared the probability model to several other methods for quality adjustment using an empirical study of 31 meta-analyses. We show that adjusting for study quality using Tritchler's approach gives less biased estimates, with better estimates of precision, compared to other quality adjustment approaches. Under the most strenuous simulation conditions, his method performs reasonably well while most of the other methods break down. In our empirical study, while there were only small differences between methods, Tritchler's approach had the largest impact relative to ignoring quality.

La méta-analyse est un ensemble d'opérations statistiques qui permet de compiler des résultats expérimentaux et corrélationnels tirés d'études indépendantes portant sur un ensemble de thèmes de recherche connexes. Elle utilise les statistiques sommaires de chaque étude comme observations. La qualité des études soumises à une méta-analyse peut varier considérablement. Pour décrire cette variation, on attribue ordinairement aux études des cotes de qualité. Une façon naturelle d'incorporer ces cotes de qualité dans une analyse consiste à pondérer chaque étude en fonction de sa cote de qualité et de sa taille ou en fonction inverse de la variance de ses valeurs estimées. Cette méthode a été employée par plusieurs auteurs, mais on lui a reproché sa subjectivité. Pour améliorer cette démarche et d'autres méthodes existantes, Tritchler (1999) a élaboré un modèle probabiliste pour évaluer l'incidence de la qualité sur une mesure de réduction des données. Il en a déduit un certain nombre de méthodes de réduction, les a comparées à l'aide de données simulées et les a appliquées à un ensemble d'études réelles, dont il s'est servi comme modèles à effets fixes.

Nous avons procédé à deux évaluations de la démarche de Tritchler. Nous avons comparé le modèle probabiliste à d'autres méthodes de correction en fonction de la qualité au moyen d'une étude de simulation. Dans cette étude, nous avons utilisé une distribution de cotes de qualité semblable à celles qui caractérisent les données méta-analytiques. En général, le modèle probabiliste donne, en regard des autres méthodes, un biais négligeable, une erreur quadratique réduite et un bon niveau pour les intervalles de confiance. Nous avons également comparé le modèle probabiliste à plusieurs autres méthodes de correction en fonction de la qualité à l'aide d'une étude empirique de 31 méta-analyses. Nous démontrons que la correction en fonction de la qualité des études par la méthode de Tritchler produit des valeurs estimées peu biaisées et de bonnes estimations de la précision par rapport à d'autres démarches. Dans les conditions de simulation les plus rigoureuses, la méthode de Tritchler continue de fournir un rendement relativement bon, tandis que la plupart des autres démarches cessent de produire des résultats valables. Notre étude empirique n'a certes révélé que de légères différences entre les méthodes, mais c'est la démarche de Tritchler qui a eu la plus profonde incidence sur les écarts de qualité.

Tuesday, May 28th/Mardi 28 mai, 9:15 **KTH B135**

Assessing change: applications of analysis of covariance to data from cluster randomization trials

Évaluer le changement : applications d'une analyse de covariance de données provenant d'essais aléatoires groupés

Gerarda Darlington, University of Guelph, et/and Neil Klar, Cancer Care Ontario

Randomized trials are often designed to assess an intervention's ability to change patient knowledge, behaviour or health. The study outcome will then need to be measured twice for each subject - prior to random assignment and following implementation of the intervention. In this talk we consider methods of analysing change when data are obtained from cluster randomization trials where the unit of allocation is a family, school or community. Attention focuses on mixed effects linear regression extensions of analysis of covariance to account for dependencies among cluster members. Particular attention is given to measuring the precision of the estimated intervention effect when there are different individual-level and cluster-level associations between the baseline and follow-up assessments. Algebraic relationships are derived in the special case where there are a fixed number of subjects per cluster while simulation studies are used in the more realistic case where there is variability in cluster size. The discussion is illustrated using data from a school-based smoking prevention trial.

Des expériences aléatoires sont souvent conçues pour évaluer la capacité de changer la connaissance, le comportement ou la santé d'un patient par une certaine méthode d'intervention. Les résultats de l'étude devront alors être mesurés deux fois sur chaque sujet - avant l'affectation aléatoire et après la mise en place de la méthode d'intervention. Dans cette conférence, nous considérons des méthodes d'analyse pour la mesure du changement quand des données sont obtenues à partir des tirages aléatoires par grappes où l'unité de regroupement est une famille, une école ou une communauté. L'attention est mise sur la mesure des effets mixtes pour la régression linéaire dans une analyse de covariance pour expliquer des dépendances parmi des membres d'un même groupe. Une attention plus particulière est donnée à la mesure de précision de l'effet estimé de l'intervention quand il y a différents niveaux d'associations parmi les individus et les grappes entre les mesures de départ et celles obtenues après l'intervention. Des liens algébriques sont dérivés dans le cas spécial où il y a un nombre fixe de sujets par grappe tandis que des études de simulations sont utilisées dans le cas plus réaliste où il y a de variabilité dans la taille des grappes. Ceci est illustré en utilisant des données d'une étude sur la prévention du tabagisme en milieu scolaire.

Tuesday, May 28th/Mardi 28 mai, 9:30 **KTH B135**

A comparative analysis of quality of life data from a clinical trial in patients with advanced breast cancer

Une analyse comparative de données sur la qualité de vie d'une étude clinique faite sur des patients ayant un cancer du sein avancé

Jianhua Liu, Dongsheng Tu et/and Joe Pater, Queen's University

Quality of life (QOL) is rapidly becoming an important outcome measure in randomized clinical trials of cancer treatments. The analysis of quality of life data is, however, complicated by the correlation between repeated measurements and large amount of missing assessments. Many statistical methods have been proposed to solve these problems. The Quality of Life Committee of the National Cancer Institute of Canada Clinical Trials Group (NCIC CTG) recently proposed an approach that combines cross-sectional analysis with a global test calculated from the proportion of patients whose quality of life were improved or deteriorated significantly based on some clinical criteria during the course of the study. This approach seems easier for clinicians to understand but its statistical properties

have not been investigated. In this talk, we present some results from a study which compares this approach with some other summary-measure based approaches and also approaches based on statistical modelings by performing a comprehensive analysis on a quality of life dataset from a clinical trial conducted by NCIC CTG in patients with advanced breast cancer.

La qualité de la vie (QV) devient rapidement une mesure importante dans des essais cliniques aléatoires pour des traitements reliés au cancer. L'analyse des données sur la qualité de vie est cependant compliquée par la corrélation entre les mesures répétées et la grande quantité d'estimations manquantes. On a proposé beaucoup de méthodes statistiques pour résoudre ces problèmes. Le Comité sur la Qualité de vie du groupe d'essais cliniques du Canada à l'Institut national de cancer (NCIC CTG) a récemment proposé une approche qui analyse en section-croisée avec un test global calculé à partir de la proportion de patients dont la qualité de vie a été améliorée ou a détériorée significativement basé sur quelques critères cliniques pendant l'étude. Il semble plus facile pour des cliniciens de comprendre cette approche mais ses propriétés statistiques n'ont pas été étudiées. Dans cette présentation, nous présentons quelques résultats d'une étude qui compare cette approche à quelques autres approches basées sur des mesure sommaires et également à des approches basées sur des modélisations statistiques en effectuant une analyse complète des données sur la qualité de vie d'un essai clinique conduit par le NCIC CTG sur des patients avec un cancer de sein avançé.

Tuesday, May 28th/Mardi 28 mai, 9:45 **KTH B135**

Using likelihood methods for phase II clinical trials
Utilisation de méthode de vraisemblance pour des essais cliniques de phase II

Gregory Pond, Princess Margaret Hospital

Phase II clinical trials are used to study possible efficacy of a new treatment and whether further research is warranted. The dominant statistical theory for these trials is Neyman-Pearson (NP) decision theory. With NP methods, investigators must conclude the treatment is potentially effective and worthy of further research or ineffective and unworthy of further research. Routinely, this decision is based on a single endpoint, response rate. Trials are designed to control erroneous conclusions in identically repeated trials. Exact error rates are estimated but unrealistic due to practical conduct and one may accept a hypothesis which is less credible than its competing hypothesis. Likelihood methods are proposed for phase II clinical trial design and reporting. Data analysis can occur at anytime, not just after an exact number of patients are accrued, providing more flexibility and shortening trial duration. One can calculate a measure of evidential strength, unavailable using NP methods, which depends only on observed results and not on trials which are not carried out. All endpoints, not just response rate, can be objectively analysed by investigators within each trial.

Des essais cliniques de phase II sont employés pour étudier l'efficacité possible d'un nouveau traitement et si d'autres recherches sont justifiées. La théorie statistique dominante pour ces essais est la théorie de la décision de Neyman-Pearson (NP). Avec les méthodes de NP, les chercheurs doivent conclure si le traitement est potentiellement efficace et nécessite plus de recherche ou si le traitement est inefficace et les recherches doivent se terminer. Habituellement, cette décision est basée sur un simple point d'arrêt : le taux de réponse.

Des essais sont con cus pour contrôler des conclusions incorrectes dans des essais identiquement répétés. Des taux exacts d'erreur sont estimés mais peu réalistes pour des raisons pratiques. Un chercheur peut accepter une hypothèse qui est moins probable qu'une hypothèse concurrente. On propose des méthodes basées sur la vraisemblance pour le design et les résultats d'essais cliniques de phase II. L'analyse des données peut se produire à n'importe quel moment, pas nécessairement après qu'un nombre exact de patients aient participé, fournissant plus de flexibilité et réduisant la durée

de l'essai. On peut calculer une mesure de force, non disponible si on utilise la méthode de NP, qui dépend seulement des résultats observés et non des essais qui n'ont pas été effectuées. Tous les points d'arrêt, pas uniquement les taux de réponse, peuvent être objectivement analysés par des chercheurs pour chaque essai.

Session 19: Data Mining in Drug Discovery/Forage de données dans la découverte des médicaments

Tuesday, May 28th/Mardi 28 mai, 8:30 **TSH B105**

Design and analysis of large chemical databases
Plan d'expérience et analyse d'un grand jeu de données chimiques

Raymond Lam, GlaxoSmithKline, William Welch, University of Waterloo et/and Stanley Young

The drug discovery paradigm has changed in two important ways. The human genome project is giving us many more new biological targets for drug discovery. Hundreds of unknown disease genes are expected to turn up in the next few years. Combinatorial chemistry and the availability of commercial compounds have made millions of compounds available for drug screening. It is no longer possible to test all available compounds for every new target of potential biological importance.

In this talk I will describe novel statistical methods for design and analysis of large chemical databases. The design problem is to choose a representative set of thousands of chemical compounds from a library that may have hundreds of thousands to millions of compounds, for assay against a biological target (screening). The analysis problem is to find regions of a high dimensional space where active compounds reside. These methods improve the efficiency and effectiveness of the drug discovery process for reducing drug screening costs and time.

Le paradigme de découverte de médicament a changé de deux façons importantes. Le projet de génome humain nous donne beaucoup plus de nouvelles cibles biologiques pour la découverte de médicament. On s'attend à ce que des centaines de gènes de maladie inconnus soient identifiés dans les années à venir. La chimie combinatoire et la disponibilité de composés commerciaux ont fait des millions des composés disponibles pour le développement de médicaments. Il n'est plus possible de tester tous les composés disponibles pour chaque nouvelle cible potentiellement importante d'un point de vue biologique.

Dans cette présentation je décrirai de nouvelles méthodes statistiques pour la conception d'un plan d'expérience et l'analyse de grandes bases de données chimiques. Le problème du plan d'expérience est de choisir un ensemble représentatif de milliers de composés chimiques d'une population qui peut avoir des centaines de milliers jusqu'à plusieurs millions de composés, pour l'analyse contre une cible biologique (filtrage). Le problème de l'analyse est de trouver des régions d'un espace à dimension élevée où résident les composés actifs. Ces méthodes améliorent la performance et l'efficacité du procédé de découverte de médicament pour réduire des coûts et le temps d'études des médicaments.

Tuesday, May 28th/Mardi 28 mai, 9:00 **TSH B105**

Tree-averaging models for high throughput screening data
Modèles de moyennage par arbre pour des données de filtrage à débit élevé

Marcia Wang, Hugh A. Chipman et/and William J. Welch, University of Waterloo

In drug discovery, high throughput screening (HTS) is used to assay large numbers of compounds against a biological target. A research pharmaceutical company might have of the order 1 million compounds available, and the human genome project is generating many new targets. Hence, there is a need for a more efficient strategy: smart or virtual screening.

In smart screening, a representative sample (experimental design) of compounds is selected from a collection and assayed against a target. A model is built relating activity to explanatory variables describing compound structure. The model is used to predict activity in the remainder of the compound collection and only the more promising compounds are screened. There is much previous work showing that classification and regression trees are very competitive in terms of prediction accuracy. This talk will concentrate on tree-averaging strategies, including bagging, boosting, and some new methods based on subsets of explanatory variables.

Dans la découverte de médicaments, le filtrage à débit élevé(HTS) est employé pour tester un grand nombre de composés en comparaison à une cible biologique. Une centre de recherche pharmaceutique peut avoir environ un million de composés disponibles, et le projet sur le génome humain produit en plus de nouvelles cibles. Par conséquent, il y a un besoin pour développer une stratégie plus efficace: filtrage judicieux ou virtuel. Dans le filtrage judicieux, un échantillon représentatif (plan d'analyse expérimental) de composés est choisi parmi un ensemble et il est analysé en le comparant à un modèle cible. Un modèle est établi reliant l'activité en fonction des variables explicatives concernant la structure d'un composé. Le modèle est ensuite employé pour prédire l'activité pour l'ensemble des composés non analysés et seuls les composés les plus prometteurs sont examinés. Plusieurs expériences ont montré que la classification et les arbres de régression sont très concurrentiels en termes d'exactitude des prévisions. Cette présentation se concentrera sur des stratégies sur le moyennage des arbres de régression, y compris l'ensachage, l'amplification, et quelques nouvelles méthodes basées sur des sous-ensembles de variables explicatives.

Tuesday, May 28th/Mardi 28 mai, 9:30 **TSH B105**

Molecular diversity: statistical perspectives and approaches
Diversité moléculaire : approches et perspectives statistiques

David Cummins and Richard E. Higgs, Eli Lilly and Company

In recent years drug discovery has undergone major changes, partially due to the joint application of high throughput screening (or HTS) with combinatorial chemistry synthesis. For most pharmaceutical companies the potential number of compounds available from in-house collections and combinatorial synthesis exceeds the capacity of HTS operations. Thus, one must choose a method for selecting a finite subset of compounds to fill allocated screening capacity. Under this restriction it becomes apparent that testing two molecular analogs may be equivalent to testing the same hypothesis twice, which comes at the expense of testing a different hypothesis. Ideally, screening a finite subset would lead to the same conclusions as if the entire library had been screened. We observe that diversity analysis has been oversold and over-hyped as a tool for increasing the rate of discovery of active molecules ("hits"). Rather than focusing on hit find rates, we focus on maintaining the quality of information obtained from hits identified when only a subset of a large library can be tested. The idea is to make the subset as information rich as possible in order to enhance subsequent lead optimization efforts. With this in mind, we view diversity analysis as a tool for managing screening libraries in order to increase the number of SAR series (rather than merely increasing the number of raw hits) identified in HTS and hence improve the chances of identifying a series that has acceptable ADME and toxicity properties. We begin with a discussion of what is reasonable to expect from diversity analysis. We give a brief review of diversity methods reported in the literature and offer a reasonable metric for assessing an analysis. We then present a statistically motivated method we have used (JCICS Vol. 37, No. 5, pp.

861-870). We conclude with examples illustrating the method and a substantial simulation experiment designed to assess its effectiveness and compare it with random selection and other popular subset selection methods.

Ces dernières années, la découverte de médicaments a subi plusieurs changements majeurs, partiellement dûs à l'application commune du filtrage élevé de débit (ou HTS) avec la synthèse combinatoire de chimie. Pour la plupart des compagnies pharmaceutiques, le nombre potentiel par de composés recueillis à l'interne et la synthèse combinatoire excède la capacité d'exécutions de HTS. Ainsi, on doit choisir une méthode pour choisir un sous-ensemble fini de composés pour utiliser la capacité maximale possible d'exécutions. Sous cette restriction, il devient évident que le test de deux molécules analogues peut être équivalent à évaluer la même hypothèse deux fois, qui s'exécuterait aux dépens d'évaluer une hypothèse différente. Dans le meilleur des cas, examiner un sous-ensemble fini mènerait aux mêmes conclusions que si tous les cas possibles avaient été examiné. Nous observons que l'analyse de la diversité a été surexploité comme un outil pour augmenter le taux de découverte des molécules actives ("hits"). Plutôt que de se concentrer sur des taux de découverte élevé de "hits", nous nous concentrons sur la qualité de l'information obtenue à partir des "hits" identifiés quand seulement un sous-ensemble de la population totale peut être testé. L'idée est de faire que le sous-ensemble soit le plus informatif possible afin de maximiser les efforts d'optimisation éventuels. Avec cette nouvelle approche, nous considérons l'analyse de la diversité comme un outil pour parvenir à examiner efficacement la population de composés afin d'augmenter le nombre de séries SAR (plutôt que simplement augmenter le nombre brut de "hits") identifiées dans HTS et par conséquent améliorer les chances d'identifier une série qui a des propriétés acceptables d'ADME et de toxicité. Nous débutons par une discussion sur ce qui est raisonnable d'obtenir par des analyses de diversité. Nous donnons un bref résumé des méthodes de diversité présentes dans la littérature et offrons un métrique raisonnable pour évaluer une analyse. Nous présentons alors une méthode statistique que nous avons utilisée (JCICS vol. 37, numéro 5, pp 861-870). Nous concluons avec des exemples illustrant la méthode et une expérience substantielle de simulations conçue pour évaluer son efficacité et pour la comparer avec la sélection aléatoire et d'autres méthodes populaires pour la sélection de sous-ensembles.

Session 20: Split Plot Experiments in Industry/Expériences avec les parcelles subdivisées dans l'industrie

Tuesday, May 28th/Mardi 28 mai, 10:30 **TSH B105**

Lack-of-fit tests for industrial split-plot experiments
Manque d'ajustement des tests pour les expériences industrielles en subdivision de parcelles

Geoff Vining, Virginia Tech

This talk considers test for lack-of-fit for industrial experiments within a split-plot context. It is well known that split-plot experiments result in at least two different error terms, which complicates the analysis. However, the literature has little or nothing to say about the resulting consequences on tests for lack-of-fit. The split-plot structure of the experiment necessarily leads to two possible lack-of-fit terms: at the whole-plot and at the subplot levels. Each of these lack-of-fit tests require different error terms, both of which need to be based on pure error. This talk outlines how to construct an industrial split-plot experiment to detect lack-of-fit and how to analyze its results.

Cette présentation considère un test pour le manque d'ajustement pour des expériences industrielles dans un contexte de plan d'expérience en subdivision de parcelles. Il est bien connu que les expériences selon un plan en subdivision de parcelles ont comme conséquence au moins deux termes d'erreur

différents, qui en compliquent l'analyse. Cependant, la littérature a peu ou rien à dire au sujet des conséquences résultantes sur des tests pour le manque d'ajustement. La structure de l'expérience impliquant un plan en subdivision de parcelles mène nécessairement à deux termes possibles pour le manque d'ajustement: au niveau global et au niveau sous-global. Chacun de ces tests pour le manque d'ajustement exige des termes d'erreur différents, et tous doivent être basés sur l'erreur pure. Cette présentation proposera une méthode pour construire un plan d'expérience en subdivision de parcelles pour le domaine industriel dans le but de détecter le manque d'ajustement et comment analyser les résultats.

Tuesday, May 28th/Mardi 28 mai, 11:15 **TSH B105**

Design and analysis of blocked fractional factorial split-plot designs
Plan d'expérience et analyse de plans d'expérience factoriel fractionnaire avec subdivision de parcelles par bloc
Robert McLeod, University of Manitoba

In industrial experiments fractional factorial designs are widely used for screening purposes. If some of the factors are hard-to-vary and others are easy-to-vary, restrictions on randomization may lead to the use of fractional factorial split-plot (FFSP) designs. Bingham and Sitter (1999) and others have considered the choice of optimal FFSP designs, using the minimum aberration (MA) criterion. A number of authors have also considered the choice of optimal blocked fractional factorial (BFF) designs. In this talk, several approaches to constructing blocked fractional factorial split-plot (BFFSP) designs will be considered. We will also discuss the optimality of BFFSP designs with respect to the MA criterion and the precision of estimated factorial effects.

Dans les expériences industrielles, les plans factorielles fractionnaires sont largement répandues pour le filtrage. Si certains des facteurs sont "difficiles à modifier" et d'autres sont "facile à changer", des restrictions sur la randomisation peuvent mener à l'utilisation d'un plan avec subdivision des parcelles factoriel fractionnaire (FFSP). Bingham et Sitter (1999) et d'autres ont considéré le choix des plans optimaux de FFSP, en utilisant le critère de l'aberration minimales (MA). Un certain nombre d'auteurs ont également considéré le choix des plans factorielles fractionnaires à blocs optimaux.

Dans cette présentation, plusieurs approches pour construire des plans d'expérience avec subdivision des parcelles factoriels fractionnaires par bloc (BFFSP) seront considérées. Nous discuterons également de l'optimalité des plans BFFSP en ce qui concerne le critère de MA et la précision des effets factoriels estimés.

Session 21: Stochastic Processes, Time Series and Spatial Statistics/Processus stochastique, séries chronologiques et statistique spatial

Tuesday, May 28th/Mardi 28 mai, 10:30 **TSH B106**

A class of stochastic conditional duration models for financial transaction data
Une classe de modèles stochastiques conditionnels sur la durée pour des données de transactions financières
Dingan Feng, York University

We proposes a class of stochastic conditional duration (SCD) models for financial transaction data, which extends both the autoregressive conditional duration (ACD) model (Engle and Russell, 1998)

and the existing stochastic conditional duration model (Bauwens and Veredas, 1999). The proposed models are in nature a class of linear non-Gaussian state space models, where the observation equation for the duration process takes an additive form of a latent process and noise. The latent process is driven by an autoregressive component to characterize the transition property and an error term associated with the observed duration process. The inclusion of such an error term allows the latent process to be inter-temporally correlated with the duration process. The Monte Carlo maximum likelihood (MCML) method is employed for consistent and efficient parameter estimation with an application using the IBM transaction data. Our analysis suggests that inclusion of the inter-temporal correlation is appealing for better out-of-sample forecasting performance.

Nous proposons une classe des modèles stochastiques conditionnels sur la durée (SCD) pour les données de transactions financières, qui généralisent le modèle conditionnel autorégressif de durée (ACD) (Engle et Russell, 1998) et le modèle conditionnel stochastique de durée (Bauwens et Veredas, 1999) existant.

Les modèles proposés sont par nature une classe des modèles linéaires non gaussiens d'états spatials, où l'équation des observations pour le processus de durée prend une forme additive d'un processus latent et d'un bruit. Le processus latent est régit par une composante autorégressive pour caractériser la propriété de transition et un terme d'erreur associé au processus de durée observé. L'inclusion d'un tel terme d'erreur permet au processus latent d'être corrélé à plusieurs niveaux temporels avec le processus de durée. La méthode du maximum de vraisemblance de Monte-Carlo (MCML) est utilisée pour l'estimation convergente et efficace des paramètres avec une application sur les données de transaction d'IBM. Notre analyse suggère que l'inclusion de la corrélation intertemporelle est une solution pour une meilleure performance sur les prévisions de données non incluses dans l'échantillon.

Tuesday, May 28th/Mardi 28 mai, 10:45 **TSH B106**

Modeling sea ice concentrations with the biased voter model
Modélisation des concentrations de glaces avec le modèle de l'électeur biaisé

Theodoro Koulis, University of Waterloo

Many researchers now agree that climate change is a real threat to our survival and to our ecosystem. The role of seasonal sea ice formation at the poles is complex and closely linked to the Earth's climate. It is thought that the amount of sea ice can have a significant effect on the energies transferred between the atmosphere and the ocean. Understanding the seasonal sea ice process at the poles is therefore of great interest to scientists. Sea ice concentration data sets derived from Earth orbiting satellites are readily available and contain observations that span several decades. This data, which is both spatial and temporal in nature, can be quite difficult to analyze. The methods of analysis for this type of data can be computationally intensive. We present the biased voter model as a candidate for describing the sea ice process. The model, which is borrowed from biology, is a Markov process on a lattice and can be controlled through two parameters. These parameters give some insight on the long term behavior of the process. We will discuss various methods for estimating these parameters. The methods are based on differential equations associated with the biased voter model. It is hoped that these methods will be helpful in analyzing multi-temporal spatial data and to make inference on global climate change.

Beaucoup de chercheurs conviennent maintenant que le changement du climat est une vraie menace à notre survie et à notre écosystème. Le rôle du gel saisonnier des mers aux pôles est complexe et étroitement lié au climat de la terre. On pense que la quantité de glace peut avoir un effet significatif sur les énergies transférées entre l'atmosphère et l'océan. La compréhension du processus saisonnier de gel aux pôles est donc d'un grand intérêt pour les scientifiques.

Les jeu de données sur la concentration en glace dérivés des satellites orbitaux de la terre sont aisément disponibles et contiennent des observations qui couvrent plusieurs décennies. Il peut être difficile d'analyser ces données car elles sont à la fois de nature spatiale et temporelle. Les méthodes d'analyse pour ce type de données peuvent demander des calculs intensifs.

Nous présentons le modèle d'électeur biaisé comme candidat pour décrire le processus de gel. Le modèle, qui est emprunté à la biologie, est un processus de Markov sur un treillis et peut être contrôlé par deux paramètres. Ces paramètres donnent de l'information sur le comportement à long terme du processus. Nous discuterons de diverses méthodes pour estimer ces paramètres. Les méthodes sont basées sur des équations différentielles associées au modèle d'électeur biaisé. Nous pouvons espérer que ces méthodes seront utiles dans l'analyse multitemporelle et spatiale des données et pour faire de l'inférence sur le changement global du climat.

Tuesday, May 28th/Mardi 28 mai, 11:00 **TSH B106**

Estimation of the order of a hidden Markov model
L'estimation de l'ordre d'une chaîne de Markov cachée

Rachel MacKay, University of British Columbia

While the estimation of the parameters of a hidden Markov model (HMM) has been studied extensively, the consistent estimation of the number of hidden states has received far less attention. The AIC and BIC are used most commonly, but their use in this context has not been justified theoretically. Capitalizing on the relationship between HMMs and finite mixture models, we show that, under mild conditions, the method of penalized minimum-distance yields a consistent estimate of the number of hidden states in a stationary HMM. Identifiability issues are also addressed. The method is applied to a multiple sclerosis data set, and its performance in finite samples assessed via simulation.

Alors que l'estimation des paramètres d'une chaîne de Markov cachée (CMC) a été étudiée de manière extensive, l'estimation cohérente du nombre d'états cachés a reçu beaucoup moins d'attention. Les critères AIC et BIC sont davantage utilisés, mais leur utilisation dans ce contexte n'a pas été justifiée théoriquement. En profitant de la relation entre les CMC et les modèles de mélange finis, nous montrons que, sous des conditions convenables, la méthode de distance minimum pénalisée donne une estimation cohérente du nombre d'états cachés dans une CMC stationnaire. Nous adressons aussi le problème d'identifiabilité. Nous appliquons la méthode à des données sur la sclérose en plaques et, par le biais d'une simulation, nous discutons son efficacité pour des échantillons de tailles finies.

Tuesday, May 28th/Mardi 28 mai, 11:15 **TSH B106**

Bayesian modelling for social networks data
Modélisation bayésienne de données de réseaux sociaux

Paramjit Gill, Okanagan University College

Sociologists, anthropologists and social psychologists study relations amongst a set of "nodes" with the aim of quantitative modelling of properties of a social network. Each node in the network plays a dual role of an "actor" and a "partner" which produces dyadic data. When the relation is of binary nature (yes/no), a social network can be described as a directed graph (digraph). For example, in a study of friendship patterns, a directed edge from node A to node B means that individual A says that individual B is a friend. There exists a rich history of deterministic and statistical methods used for the analysis of social networks (Wasserman and Faust, 1994). In a seminal paper, Holland

& Leinhardt (1981) proposed an exponential model, called the p1 model, for the analysis of dgraphs. This presentation considers a fully Bayesian analysis of the p1 and related models. The vehicle for doing so is modern Bayesian computation made accessible in the software package WinBUGS. We demonstrate how the basic Bayesian model can be modified easily to incorporate covariates and how the stochastic block models (Wang & Wong, 1987) can be fitted. We use well known data examples for illustration and for comparisons with the non-Bayesian analyses.

References:

Holland, P.W. & Leinhardt, S. (1981). An exponential family of probability distributions for directed graphs. JASA, 76, 33-65.

Wang, Y.J. & Wong, G.Y. (1987). Stochastic blockmodels for directed graphs. JASA, 82, 8-19.

Wasserman, S. & Faust, K. (1994). Social Networks Analysis: Methods and Applications. Cambridge University Press.

Les sociologues, les anthropologues et les psychologues sociaux étudient les relations parmi un ensemble de "regroupements de facteurs" dans le but de modéliser quantitativement les propriétés d'un réseau social. Chaque regroupements de facteurs dans le réseau joue un rôle double: celui d'acteur et celui de partenaire qui produit des données dyadiques. Quand la relation est de nature dichotomique (oui/non), un réseau social peut être décrit comme un graphique orienté (digraphe).

Par exemple, dans une étude sur les relations d'amitié, une relation orienté du regroupement de facteurs A au regroupement de facteurs B signifie que l'individu A dit que l'individu B est un ami. Il existe une histoire riche sur les méthodes déterministes et statistiques employées pour l'analyse des réseaux sociaux (Wasserman et Faust, 1994).

Dans un article, Holland et Leinhardt (1981) ont proposé un modèle exponentiel, appelé modèle p1, pour l'analyse des graphes. Cette présentation considère une analyse entièrement bayésienne du modèle p1 et des modèles reliés. La motivation pour faire ainsi est que le calcul bayésien moderne est rendu accessible dans le progiciel WinBUGS. Nous démontrons comment le modèle bayésien de base peut être facilement modifié pour incorporer des covariables et comment les modèles stochastiques par blocs (Wang et Wong, 1987) peuvent être adaptés. Nous utilisons des exemples de données bien connus pour l'illustration et pour des comparaisons avec des analyses de non bayésiennes.

Références:

Holland, P.W. & Leinhardt, S. (1981). An exponential family of probability distributions for directed graphs. JASA, 76, 33–65.

Wang, Y.J. & Wong, G.Y. (1987). Stochastic blockmodels for directed graphs. JASA, 82, 8–19.

Wasserman, S. & Faust, K. (1994). Social Networks Analysis: Methods and Applications. Cambridge University Press.

Tuesday, May 28th/Mardi 28 mai, 11:30 **TSH B106**

Structural analysis of stationary germ grain models
Analyse de la structure de modèles stationnaires de germes de grain

Jeffrey Picka, University of Maryland

If monosized discs are centered at points in the realization of a stationary point process, then the union of those discs is a realization of a spatial process known as a random set, which is useful for modeling disordered materials. To select models of this type, statistics are required that summarize important aspects of a coverage pattern in an interpretable fashion.

A decomposition of the moments of a random set has been found which is ideal for model identification. Unlike the covariogram in random field theory, its features are interpretable, and reveal fine details of structure that are easily overlooked when naive estimators are used. Further, if the notion

of a random set moment is suitably extended, this decomposition can be extended to provide multiple points of view not available through point process analyses of the same information. This extended definition of moments also provides a clear explanation of why balanced estimators of cumulants are best in single-sample inference in random sets.

Si des disques unidimensionnels sont centrés aux points dans la réalisation d'un processus stationnaire ponctuel, alors l'union de ces disques est la réalisation d'un processus spatial connu sous le nom de positionnement aléatoire, et est utile pour modéliser du matériel désordonné. Pour choisir des modèles de ce type, des statistiques sont requises pour résumer les aspects importants d'une configuration de recouvrement dans un mode interprétable.

On a trouvé une décomposition des moments pour un ensemble aléatoire qui est idéal pour l'identification du modèle. à la différence du covariogramme dans la théorie aléatoire des zones, ses dispositifs sont interprétables, et indiquent les détails précis de structure qui sont facilement obtenus quand les estimateurs naïfs sont utilisés. De plus, si la notion de moment aléatoire pour un ensemble est généralisée convenablement, cette décomposition peut être généralisée pour fournir plusieurs points de vue non disponibles par une analyse du processus ponctuel avec la même information. Ceci a généralisé la définition des moments et fournit également une explication claire de la raison pourquoi les estimateurs des cumulants balancés sont les meilleurs pour l'inférence faite pour un seul échantillon aléatoire d'ensembles.

Tuesday, May 28th/Mardi 28 mai, 11:45 **TSH B106**

Effective estimation in CLT for the Markov chains with Doeblin condition
Estimation efficace du théorème de la limite centrale pour les chaînes de Markov sous la condition de Doeblin
Janusz Kawczak et/and Stanislav Molchanov, UNCC

The weak convergence of the general Markov chain is studied under Doeblin condition with applications to the tests for quality of random number generators (RNG). Also, some results of Berry-Esseen type for general MC are obtained.

La convergence faible des chaînes de Markov généralisées est étudiée sous la condition de Doeblin avec des applications aux tests sur la qualité des générateurs de nombres aléatoires (GNA). Quelques résultats du type Berry-Esseen pour les chaînes de Markov généralisées sont obtenus.

Session 22: Statistics and Public Policy/Statistique et ordre public

Tuesday, May 28th/Mardi 28 mai, 10:30 **TSH B128**

Information for public policies
Informations d'ordre public
Miron Straf, National Academy of Sciences

Policy information is taken to mean information used in the formulation, design, and selection of public policies. It comprises both data and analysis. Often it is the result of policy analysis, the objective evaluation of the social and economic implications of changes in public policy. Key ingredients of policy analysis are data; analytical models, including their documentation and validation; and effective communication of the results to policy makers, in particular of the assumptions in models and uncertainty in estimates. Many models are employed in policy analysis. Microsimulation models, for example, simulate how records on individual people or units, such as taxpayers or health care providers,

change over time and as a result of a proposed policy. Characterizing uncertainty through external as well as internal validation is important but more difficult for policy analysis models, because they involve conditional forecasts.

The nature of information as a public good provides an economic rationale for a government to collect statistics and provide information from policy analysis, but there are many other reasons for it to do so. One is to assure accurate, timely, and credible information. The most common problem for policy analysis in the United States is the lack of integrated data. As a result, analysts must often patch together a variety of data and research results of varying quality and make assumptions that may be unsupported. Governments must take the lead in data collection if they are to have high quality, comprehensive databases for effective policy analysis.

A government agency managing research for public policies faces a number of fundamental questions: When should the agency conduct research or procure it from others? What balance should be struck between policy analysis and descriptive information? How should its research services and the performance of individual researchers be evaluated? How can the performance of the agency be improved? How can the agency reinforce credibility in its research? Principles for the management of public policy research in government are articulated to address these questions.

L'information d'ordre public est utilisée comme information moyenne dans la formulation, la planification, et la sélection d'information d'ordre public. Elle comporte des données et des analyses. Souvent, c'est le résultat de l'analyse de politiques, l'évaluation objective des implications sociales et économiques de certains changements de politiques publics. Les principaux ingrédients de l'analyse de politiques sont des données; des modèles analytiques, y compris leur documentation et validation; et la transmission pertinente des résultats aux décideurs, en particulier des hypothèses des modèles et l'incertitude dans les estimations. Plusieurs modèles sont utilisés dans l'analyse de politiques. Par exemple, des modèles de micro-simulations simulent comment des enregistrements sur différentes personnes ou unités, telles que des contribuables ou des participants au régime de santé, changent avec le temps et ceci mis en relation avec une politique proposée. Il est important de caractériser l'incertitude autant à l'externe par de la validation qu'à l'interne. Toutefois, c'est une tâche plus difficile pour des modèles d'analyse de politiques, parce qu'ils impliquent des prévisions conditionnelles.

La nature de l'information en tant que bien public fournit un raisonnement économique pour un gouvernement dans le but de recueillir des statistiques et de fournir les informations de l'analyse de politiques, mais il existe plusieurs autres raisons pour justifier cette approche. On doit s'assurer que l'information soit précise, opportune, et crédible. Le problème le plus commun dans l'analyse de politiques aux États-Unis est le manque de données intégrées. En conséquence, les analystes doivent souvent raccorder ensemble une variété de données et de résultats de recherches de qualité variable et faire des hypothèses qui peuvent être peu réalistes. Les gouvernements doivent prendre en charge la collecte de données s'ils veulent obtenir des analyses de bonne qualité et des bases de données complètes pour ces analyses de politiques.

Un organisme gouvernemental effectuant de la recherche sur des politiques publics fait face à un certain nombre de questions fondamentales : Quand l'agence devrait-elle conduire la recherche ou l'obtenir d'autres? Quel équilibre devrait-il y avoir entre l'analyse de politiques et l'information descriptive? Comment ses services de recherches et les performances des différents chercheurs devraient-ils être évalués? Comment les performances de l'agence peut-elle être améliorée? Comment l'agence peut-elle renforcer la crédibilité de ses recherches? Les principes de gestion de recherches d'ordre public dans le gouvernement sont articulés pour répondre à ces questions.

Session 23: Statistics and Brain Mapping/La statistique et le mappage du cerveau

Tuesday, May 28th/Mardi 28 mai, 10:30 **KTH B135**

Image fusion for concurrently recorded spontaneous EEG and fMRI
Fusion d'images spontanées obtenues de façon concurrentielle par EEG et fMRI

Pedro Valdes, Eduardo Martinez and Nelson Trujillo, Cuban Neuroscience Center

A current challenging problem is how to obtain brain images with very high temporal and spatial resolution by the fusion of data obtained from different neuroimaging modalities. The electroencephalogram (EEG) has a millisecond accuracy which the highest currently obtainable. By means of a hierarchical Bayesian model it is possible to solve the EEG inverse problem to obtain electrophysiological images with the same high temporal resolution. This is known as EEG Tomography (ET). Time domain, frequency domain and time/frequency domain estimators are presented for ET via the (penalized)EM algorithm. The method is applied to the identification of the neural sources of the alpha rhythm. Two drawbacks of ET are the very low spatial resolution obtainable as well as an "depth bias" in which source strength is overestimated for positions near the electrodes. Functional MRI (fMRI), on the other hand, has none of these spatial problems but offers a very low temporal resolution. It is shown how the hierarchical Bayesian model for ET may be extended to provide EEG/fMRI image fusion. The resulting EEG/fMRI images have both high spatial and temporal resolution. A particular, very informative, case is when the EEG and fMRI are recorded concurrently. Advantage may then be taken of the empirically demonstrated linear relation between the fMRI and components of the time varying EEG spectrum (Goldman et. al 2002,Martinez et. al 2002) to obtain an image fusion model with a minimum of hyper parameters. It is illustrated how this fusion technique can identify spatially extended neural subsystems involved in alpha rhythm generation. It is discussed how EEG/fMRI poses new inferential problems both due to the multimodal nature of the data and to the increased dimensionality of the problem. In particular, the issue of extending the concept of Granger Causality to these spatially extended systems is touched upon.

Un des défis actuels est de trouver un moyen d'obtenir des images du cerveau avec une résolution très élevée dans le temps et l'espace par la fusion des données obtenues à partir de différentes modalités d'imagerie du cerveau. L'électroencéphalogramme (EEG) a une exactitude à la milliseconde, ce qui est le plus haut degré de précision actuel. Au moyen d'un modèle bayésien hiérarchique, il est possible de résoudre le problème de l'inverse du EEG pour obtenir des images électrophysiologiques avec cette même résolution temporelle élevée. Ceci est connu sous le nom de tomographie EEG (ET). Les estimateurs basés sur un domaine de temps, de fréquence et de fréquence/temps sont présentés pour ET par l'intermédiaire de l'algorithme EM (avec pénalité). La méthode est appliquée à l'identification des sources neurales du rythme alpha. Deux faiblesses de ET sont la résolution spatiale très lente ainsi qu'un "biais de profondeur" dans lequel la force de la source est surestimée pour des positions près des électrodes. D'autre part, la fonctionnelle MRI (fMRI), n'a aucune de ces problèmes spatiaux mais offre une résolution temporelle très basse. Nous avons démontré la façon dont le modèle bayésien hiérarchique pour ET peut être généralisé afin de fournir une fusion des images EEG et fMRI. Les images résultantes EEG/fMRI ont des résolutions spatiale et temporelle élevées. En particulier, un cas informatif survient quand les EEG et fMRI sont enregistrés concurremment. Nous pouvons alors tirer avantage de la relation linéaire empiriquement démontrée entre le fMRI et les composants du temps changeant sur le spectre de l'EEG (Goldman et al 2002, Martinez et al 2002) pour obtenir un modèle de fusion d'images avec un minimum d'hyper-paramètres. Nous illustrons comment cette technique de fusion peut identifier des sous-ensembles de neurones dans l'espace impliqués dans la génération

du rythme alpha. Nous discutons aussi de nouveaux problèmes d'inférence relatifs à EEG et au fMRI lesquels sont dus à l'allure multimodale des données et à la dimension plus élevée du problème. Nous discuterons également de la généralisation du concept de causalité de Granger à ces systèmes spatiaux.

Tuesday, May 28th/Mardi 28 mai, 11:30 **KTH B135**

How to smooth surface data ?
Comment lisser des données de surface?
Moo Chung, University of Wisconsin-Madison

Gaussian kernel smoothing has been widely used in 3D medical images such as magnetic resonance images (MRIs) and positron emission tomography to increase the signal-to-noise ratio. However, it does not work on curved surfaces such as the cortical surfaces extracted from 3D brain MRIs.

By reformulating the Gaussian kernel smoothing as a solution to a diffusion equation on a Riemannian manifold, we can overcome the inherent limitations of the Gaussian kernel smoothing. This generalization is called diffusion smoothing. For surfaces, the counterpart of the Euclidean Laplacian is the Laplace-Beltrami operator. Based on the finite element method, we present an explicit method of estimating the Laplace-Beltrami operator. Afterwards, the diffusion equation is numerically solved via the finite difference method.

As an illustration, we show how to smooth data on the triangulated brain surfaces consisting of 81920 triangles and characterize the brain shape changes.

Le lissage par le noyau gaussien a été largement répandu dans les images médicales tridimensionnelles telles que les images de résonance magnétique (MRIs) et la tomographie d'émission de protons pour augmenter le taux signal/bruit. Cependant, cela ne fonctionne pas sur les surfaces courbes telles que les surfaces corticales extraites à partir de MRI tridimensionnelle du cerveau.

En reformulant le lissage par le noyau gaussien comme solution à une équation de diffusion multiple de Riemann, nous pouvons surmonter les limitations inhérentes du lissage par le noyau gaussien. Cette généralisation s'appelle le lissage de diffusion. Pour des surfaces, l'équivalent du laplacien euclidien est l'opérateur de Laplace-Beltrami. Basé sur la méthode par élément fini, nous présentons une méthode explicite pour estimer l'opérateur de Laplace-Beltrami.

Ensuite, l'équation de diffusion est numériquement résolue par l'intermédiaire de la méthode des différences finies. Nous démontrons comment lisser des données sur des surfaces triangulées du cerveau se composant de 81920 triangles et caractérisant le changement dans la forme du cerveau.

Session 24: Theoretical Aspects of Monte Carlo/Aspects théoriques de la méthode de Monte Carlo

Tuesday, May 28th/Mardi 28 mai, 13:30 **TSH B105**

Perfect sampling algorithms: connections
Algorithmes d'échantillonage parfait : les liens
Duncan Murdoch, University of Western Ontario

The arrival of Propp and Wilson's (1996, Random Structures and Algorithms) coupling from the past (CFTP) algorithm caused a big stir in the Markov chain Monte Carlo community, because it did the seemingly impossible task of using a finite run of a Markov chain to obtain samples from the steady-state distribution of the chain. At around the same time Fill (1998, Annals of Applied

Probability) developed quite a different algorithm that accomplished the same thing using rejection sampling.

In this talk (which is loosely based on Fill, Machida, Murdoch and Rosenthal, 2000, Random Structures and Algorithms) I will summarize both algorithms, and show how Wilson's (1999, Random Structures and Algorithms) "read-once" variation on CFTP is in some sense also a variation on Fill's algorithm.

L'arrivée de l'algorithme de couplage à partir du passé (CFTP) de Propp et de Wilson (1996, Random Structures and Algorithms) a causé un grand remous dans le domaine des chaînes de Markov Monte-Carlo, parce qu'il fait la tâche apparemment impossible d'employer un nombre fini de passages d'une chaîne de Markov pour obtenir des échantillons provenant de distribution d'équilibre de la chaîne. Environ à la même période, Fill (1998, Annals of Applied Probability) a développé un algorithme tout à fait différent qui a accompli la même chose en utilisant l'échantillon de rejet.

Dans cette présentation (qui est largement basé sur Fill, Machida, Murdoch et Rosenthal, 2000, Random Structures and Algorithms), je présenterai les deux algorithmes, et j'exposerai comment la variante "lecture unique" de CFTP proposée par Wilson (1999, Random Structures and Algorithms) est dans un certain sens également une variante de l'algorithme de Fill.

Tuesday, May 28th/Mardi 28 mai, 14:00 **TSH B105**

Optimal scaling of random walk Metropolis algorithms
Échelle optimale pour l'algorithme de marche aléatoire de Métropolis

John Yuen, York University et/and G. Roberts, Lancaster University

Scaling the proposal distribution of a multi-dimensional random walk Metropolis algorithm in order to maximize the efficiency of the algorithm has been studied for some collections of target distributions. In particular, Roberts, Gelman and Gilks (1997) proved a weak convergence result as the dimension n, of a sequence of target densities converges to infinity, if the target density has a symmetric product form and is 'smooth' enough. In this talk, I will discuss some recent work on extending this kind of results to algorithms with other target distribution, such as piece-wise continuous densities.

Mesurer la distribution proposée d'un algorithme de Metropolis multidimensionnel de promenade aléatoire afin de maximiser l'efficacité de l'algorithme a été étudié pour quelques regroupements de distributions cibles. En particulier, Roberts, Gelman et Gilks (1997) ont prouvé un résultat de convergence faible en dimension n, d'une suite de densités cibles qui converge à l'infini, si la densité cible a une forme du produit symétrique et est assez lisse. Dans cette présentation, je discuterai de quelques travaux récents sur l'étendu de ce genre de résultats aux algorithmes avec d'autres distributions cibles, telle que des densités continues par morceaux.

Tuesday, May 28th/Mardi 28 mai, 14:30 **TSH B105**

Getting perfect more quickly: speed-up methods for perfect sampling algorithms
Comment devenir parfait au plus vite : des méthodes d'accélération pour les algorithmes d'échantillonage parfait

Radu Craiu, University of Toronto

Perfect sampling algorithms are a novel addition to the MCMC literature. For this class of algorithms the challenging issue of assessing convergence completely vanishes. This feature makes them very attractive for many computational problems. Despite recent efforts, the usage of perfect sampling for Bayesian inference is quite limited due to the large memory and/or computer time required to perform a simulation experiment. In this talk we explore two alternative designs of a perfect sampling

algorithm which may result in significant time savings. The first method uses antithetic variates in the context of backward coupling and relies heavily on the concept of negative association. As a bonus, the study of antithetic coupling brings on interesting issues regarding the stationarity of a coupled chain. The perfect sampling algorithm's counterpart in sample survey is the simple random sampling. The second method, called multi-stage coupling, is inspired by the fact that, in survey studies, multi-stage sampling is more cost efficient than simple random sampling. Both techniques are illustrated with simulations for Bayesian estimation problems.

Les algorithmes d'échantillonnage parfait représentent une contribution récente à la littérature du MCMC. Pour ces algorithmes la détermination de la convergence n'est plus nécessaire. Cet aspect les rend très attrayants pour plusieurs situations où le MCMC est nécessaire. Malgré des récents efforts, l'utilisation des algorithmes d'échantillonnage parfait pour les problèmes d'estimation bayésienne est tout à fait limitée. Dans cette présentation, on explore deux possibilités alternatives de l'algorithme d'échantillonnage parfait qui peuvent conduire à des réductions considérables du temps nécessaire à compléter la simulation. La première méthode utilise des variables antithétiques dans le contexte du couplage à rebours. En bonus, le couplage antithétique soulève quelques questions intéressantes sur la stationnarité d'une chaîne couplée. La contrepartie des algorithmes d'échantillonnage parfait est, en théorie des sondages, le sondage aléatoire simple. La seconde méthode qu'on appelle couplage en multistages est inspirée du fait que les échantillonnages en plusieurs phases sont moins coûteux que les échantillonnages aléatoires simples. Les deux techniques sont illustrées avec des simulations pour des problèmes d'estimation bayésienne.

Session 25: Environmetrics I/Mésométrie I

Tuesday, May 28th/Mardi 28 mai, 13:30 **KTH B135**

Statistical framework for a waterborne infectious outbreak
Cadre statistique pour la détection, la description et la prédiction des épidémies de maladies transmises par l'eau

Elena Naumova, Tufts University

Investigation of a waterborne infectious outbreak can be substantially improved by a better understanding of the mechanism behind an outbreak that, in part, can be achieved by describing and modeling documented outbreaks. Every outbreak has a unique signature, which depends on the type of exposure and the population characteristics. Every outbreak is initiated by the introduction of a pathogen to a susceptible population. Every outbreak represents a unique sequence of elementary events in time and space and can be characterized by duration, magnitude and shape of an epidemic curve. In our previous study, we demonstrated the Poisson regression modeling of a single outbreak with a single source of exposure using simulated data and actual data from the Milwaukee waterborne cryptosporidiosis outbreak of 1993. Then, we added some complexity to the model by assuming a secondary spread and demonstrated this method using data from the suspected outbreak of cryptosporidiosis in Worcester, MA in the late summer of 1995. This work is an overview of statistical approaches for modeling, detecting and describing waterborne infectious outbreaks. General considerations for the modeling and potential use of new and well-established techniques for characterizing an outbreak, such as a change-point detection approach, approximations by a mixture of distributions, and fractal description will be discussed.

La recherche sur une épidémie de maladie infectieuse transmise par l'eau peut être sensiblement améliorée par une meilleure compréhension du mécanisme derrière cette épidémie qui peut être réalisée en partie par la description et la modélisation des épidémies documentées. Chaque épidémie a une seule

signature, qui dépend du type d'exposition et des caractéristiques de la population. Chaque épidémie débute par l'introduction d'un microbe pathogène dans une population sensible. Chaque épidémie représente une suite unique d'événements simples dans le temps et l'espace et peut se caractériser par la durée, l'importance et la forme d'une courbe épidémique. Dans notre étude précédente, nous avons fait la modélisation par une régression de Poisson d'une épidémie simple avec une source d'exposition unique en utilisant des données simulées et des données réelles de l'épidémie d'une maladie transmise par l'eau du cryptosporidiosis à Milwaukee en 1993. Puis, nous avons ajouté une certaine complexité au modèle en assumant une diffusion secondaire et avons démontré cette méthode en utilisant des données sur l'épidémie probable du cryptosporidiosis à Worcester, MA vers la fin de l'été 1995. Cette présentation se veut une vue d'ensemble des approches statistiques pour modéliser, détecter et décrire des épidémies de maladies infectieuses transmises par l'eau. Des propositions générales pour la modélisation seront faites. L'usage potentiel de nouvelles techniques bien établies pour caractériser une épidémie, telle qu'une approche de détection par le point de rupture, approximations par un mélange des distributions, et la description fractale sera discuté.

Tuesday, May 28th/Mardi 28 mai, 14:00 **KTH B135**

Visualization of time and geographical distribution of cancer mortality in Japan
Visualisation de la distribution du cancer au Japon dans le temps et géographiquement

Megu Ohtaki, Hiromi Kawasaki, et/and Kenichi Satoh, Hiroshima University

We developed a statistical method for visualizing time and geographical distribution of standardized mortality rario in Japan. An empirical Bayes method with Poisson-gamma model and a nonparametric smoothing with respect to calendar time were used. The method, which is based on an empirical Bayes method with Poisson-gamma model and nonparametric smoothing with respect to calendar time, works for sparse data as well as dense ones without extensive calculations. Using the method with year-specific demographic data for 3342 municipalites, we made animated disease maps for major cancers in Japan 1975-1994. It can be seen from the resulting animated map that lung, colon and breast cancers have nationwide rapid increasing trends while stomach, uterus and skin cancers have decreasing one. It is also remarkable that liver, brain cancers and leukemia have their own peculiar features for the time-geographical distribution.

Nous avons développé une méthode statistique pour visualiser la répartition géographique dans le temps du ratio de mortalité normalisé au Japon. La méthode consiste en une approche bayésienne empirique avec le modèle de Poisson-gamma et un lissage non paramétrique par rapport au temps. Elle est fonctionnelle pour les données éparses ou denses sans aucun calculs intensifs. En utilisant la méthode avec des données démographiques pour une année donnée pour 3342 municipalités, nous avons fait les cartes animées de la maladie pour les principaux types de cancers au Japon entre 1975-1994. On peut voir sur la carte animée résultante que les cancers de poumon, du colon et du sein ont une tendance d'augmentation rapides dans tout le pays tandis que les cancers de l'estomac, de l'utérus et de la peau diminuent. Il est également remarquable que le cancer du foie, du cerveau et la leucémie aient leurs propres dispositifs particuliers pour la répartition géographique dans le temps.

Tuesday, May 28th/Mardi 28 mai, 14:30 **KTH B135**

Linear regression with correlated residuals. An application to monthly temperature measurements in Lisbon.

Régression linéaire avec des erreurs corrélés. Une application à une série de températures mensuelles à Lisbonne.

Teresa Alpuim, University of Lisbon, et/and Abdel El-Shaarawi, National Water Research Institute

A common procedure in environmental statistics is to estimate and test for trends in time series data. Usually, this is done with the help of a regression model including time as one of the independent variables. However, very often, the residuals present an autocorrelation structure which may cause a distortion in the variance of Least Squares estimators and, therefore, in the tests results. Detection of climatic changes or increase in concentration of pollutants provide good examples of such series, where the need for the development of appropriate and rigorous tools is of the utmost importance. In this talk we will consider, first, the linear regression model with any set of independent variables and errors following an autoregressive, AR(p), process. In this general case, the Maximum Likelihood estimators are difficult to obtain and not much is known about their statistical properties. On the other hand, the least squares estimators, although loosing some of their optimal properties, are easy to evaluate and keep some important properties, namely, they are unbiased, consistent, their theoretical variances are known and may be obtained from the sample through consistent estimators. Consequently, under the assumption of normality, it is possible to derive asymptotic tests and confidence intervals for the regression parameters. However, under some typical cases of the design matrix X, we will show that the Maximum Likelihood and Least Squares estimators are asymptotically equivalent and, in such cases, it is possible to prove optimality properties. An application of these methods will be made to a series of monthly average temperature measurements in Lisbon, from January 1856 to December 1999, through the use of a model that includes trend, seasonality and a covariate correponding to the North Athlantic Oscillation index.

Une procédure utilisée dans le domaine des statistiques de l'environnement doit estimer et tester des tendances pour des données de séries chronologiques. Habituellement, ceci est fait avec l'aide d'un modèle de régression comprenant le temps comme une des variables indépendantes. Très souvent, les résidus présentent une structure d'autocorrélation qui peut causer une distorsion dans la variance des estimateurs obtenus par la méthode des moindres carrés et, en conséquence, dans les résultats des tests. La détection des changements climatiques ou l'augmentation de la concentration des polluants fournissent de bons exemples de telles séries, où le besoin du développement d'outils appropriés et rigoureux est primordial.

Dans cette conférence, nous considérerons d'abord le modèle de régression linéaire avec tous les sous-ensembles de variables indépendantes et d'erreurs suivant un processus autorégressif AR(p). Dans ce cas général, il est difficile d'obtenir les estimateurs de maximum de vraisemblance et peu de propriétés statistiques sont connues. D'autre part, il est facile d'évaluer les estimateurs par les moindres carrés qui, bien que perdant certaines de leurs propriétés d'optimalité, gardent quelques propriétés importantes, à savoir qu'ils sont non biaisés, convergent, leurs variances théoriques sont connues et peuvent être obtenues de l'échantillon à partir d'estimateurs convergents. En conséquence, sous l'hypothèse de normalité, il est possible de déduire les tests asymptotiques et les intervalles de confiance pour les paramètres de régression. Cependant, dans quelques cas typiques de la matrice X du plan d'expérience, nous prouverons que le maximum de vraisemblance et les estimateurs des moindres carrés sont asymptotiquement équivalent et, dans ces cas-ci, il est possible de prouver des propriétés d'optimalité.

On appliquera cette méthode à un série mensuelle de mesures de température moyenne à Lisbonne,

de janvier 1856 à décembre 1999, par l'utilisation d'un modèle qui inclut une tendance, une saisonnalité et une covariable correspondant à l'indice d'oscillation de l'Atlantique Nord.

Session 26: Longitudinal Data - Theory and Applications/Données longitudinales - theorie et applications

Tuesday, May 28th/Mardi 28 mai, 13:30 **TSH B128**

A nonparametric procedure for the analysis of balanced crossover designs
Tests non paramétriques pour l'analyse de données issues de plans croisés équilibrés

François Bellavance, l'École des Hautes Études Commerciales, Serge Tardif, Université de Montréal, et/and Constance van Eeden, University of British Columbia

Nonparametric tests will be presented for the hypotheses of no direct treatment and carryover effects for balanced M-period M-treatment crossover designs. The design consists of n replicates of balanced crossover designs. The tests are permutation tests based on the n vectors of least squares estimators of the parameter of interest obtained from the n replications of the experiment. The exact as well as the asymptotic distribution of the test statistics are obtained and it is shown that the tests have, asymptotically, the same power as the F-ratio test.

Des tests non paramétriques seront présentés pour vérifier les hypothèses d'absence d'un effet de traitement et d'un effet rémanent pour les données issues de plans croisés équilibrés avec M traitements et M périodes. Le plan d'expérience est formé de n répliques de plans croisés équilibrés. Les tests proposés sont des tests de permutations basés sur les n vecteurs des estimateurs des moindres carrés des paramètres d'intérêts obtenus pour chacune des n répliques du plan d'expérience. La distribution exacte ainsi que la distribution asymptotique des tests statistiques seront présentées et nous montrerons que ces tests ont, asymptotiquement, la même puissance que le test F de l'analyse de variance.

Tuesday, May 28th/Mardi 28 mai, 13:45 **TSH B128**

A likelihood-based method for analyzing longitudinal binary responses with informative drop-outs
Une méthode basée sur la vraisemblance pour l'analyse longitudinale de réponses dichotomiques avec abandons informatifs

Grace Yi et/and Mary Thompson, University of Waterloo

In this talk, we discuss a likelihood-based approach for analyzing longitudinal binary data with informative drop-outs. The joint distribution for the underlying responses is modeled in terms of the Bahadur representation (Bahadur, 1961), while the drop-out process is characterized by a logistic regression model. Such a model is appealing because of the tractable form of the resulting likelihood function. Inferences for the interest parameters may be carried out by applying standard maximum likelihood theory. Simulation studies are conducted to assess the performance of the proposed model.

Dans cette présentation, nous discutons une approche basée sur la vraisemblance pour l'analyse de données binaires longitudinales avec des abandons informatifs. La distribution conjointe pour les réponses sous-jacentes est modélisée en termes de la représentation de Bahadur (Bahadur, 1961), alors que le processus d'abandon est caractérisé par un modèle de régression logistique. Un tel modèle est intéressant en raison de la forme particulière de la fonction de vraisemblance résultante. Des inférences pour les paramètres d'intérêt peuvent être effectués en appliquant la théorie du maximum de vraisemblance. Des études de simulations sont entreprises pour évaluer la performance du modèle proposé.

Tuesday, May 28th/Mardi 28 mai, 14:00 **TSH B128**

Performance of nonparametric estimates in Poisson time series with small counts
Performance des estimateurs non paramétriques pour des séries chronologiques de Poisson avec faibles fréquences

John Holt et/and O. Brian Allen, University of Guelph

Large scale investigations of disease occurrence in longitudinal studies of putative harmful effects of pollutants often involve generalized additive models. Monte Carlo studies suggest caution in using standard large sample methods with large time series with small daily counts.

Les recherches à grande échelle sur l'occurrence d'une maladie dans des études longitudinales avec des effets nocifs des polluants impliquent souvent les modèles additifs généralisés. Les études de Monte-Carlo suggèrent une prudence quant à l'utilisation des méthodes standards pour de grands échantillons avec de grandes séries chronologiques comportant de petites fréquences quotidiennes.

Tuesday, May 28th/Mardi 28 mai, 14:15 **TSH B128**

Consistent estimation in age-period-cohort analysis
Estimateurs convergents dans une analyse de cohorte âge-période

Wenjiang Fu, Michigan State University

Age-period-cohort (APC) model has been popular in studying chronic disease rates (cancer, stroke, etc.) (see Kupper et al. 1985, Clayton and Schifflers 1987), and social event rates (crime, suicide, depression, etc.) (see Mason and Fienberg 1985). It analyzes rates in a table of a rows of age group and p columns of period (calendar year) group by estimating fixed effects of age groups, period groups and birth cohorts (diagonals). However, the linear relationship Period + Age = Cohort leads to a singular design in regression and yields multiple estimations. Consequently, true trend in age, period and cohort has not been correctly determined with current methods so far. This is the well known identifiability problem in APC analysis, and has been deemed unsolvable in the literature. In this presentation, I will present a novel approach, and correct a mistake in the literature upon which previous conclusions were made. The new estimation method is justifiable from both epidemiological/clinical/sociological and statistical points of view. To address the identifiability problem, I will provide results on asymptotic consistency and normality of the new estimator through profile likelihood as the number of period groups p goes to infinity. Even though the maximum likelihood estimator is well known to be inconsistent, the maximum profile likelihood estimator (MaPLE) is. The method will be demonstrated through two data sets, including cervical cancer rates and homicide rates. A home-made computer software will be presented to help visualize the new approach. Future studies will be discussed at the end.

Le modèle avec cohorte âge-période (RPA) a été populaire dans l'étude des taux de maladies chroniques (cancer, crise cardiaque, etc.) (voir Kupper et al. 1985, Clayton et Schifflers 1987), et dans taux d'événements sociaux (crime, suicide, dépression, etc.) (voir Mason et Fienberg 1985). Ce modèle analyse des taux dans une table où les a lignes sont des catégories d'âge et les p colonnes sont des périodes de temps (année civile) en estimant des effets fixes pour les catégories d'âge, la période et les cohortes de naissance (diagonales). Cependant, la relation linéaire âge-cohorte mène à une seule planification d'expérience pour la régression et engendre de multiples estimations. Par conséquent, les tendances dans l'âge, la période et la cohorte n'ont pas été correctement déterminée jusqu'ici avec les méthodes actuelles.

C'est le problème bien connu d'identifiabilité dans l'analyse en composantes principales (ACP), et il a été considéré comme non résoluble dans la littérature. Dans cette présentation, je donnerai

une nouvelle approche, et je corrigerai une erreur dans la littérature sur laquelle des conclusions précédentes ont été faites. La nouvelle méthode d'estimation est justifiable autant d'un point de vue épidémiologique/clinique/sociologique que statistique. En ce qui concerne le problème d'identifiabilité, je fournirai des résultats sur l'uniformité et la normalité asymptotique du nouvel estimateur par la méthode du profil de vraisemblance quand le nombre de périodes tend vers l'infini. Même si l'estimateur du maximum de vraisemblance est bien connu pour être non convergent, l'estimateur de la méthode du maximum du profil de vraisemblance (MaPLE) l'est. La méthode sera utilisée sur deux jeux de données, incluant des taux de cancer cervical et des taux d'homicide. Un logiciel maison sera présenté pour aider à visualiser la nouvelle approche. De futures études seront aussi discutées.

Tuesday, May 28th/Mardi 28 mai, 14:30 **TSH B128**

Finite sample properties in using GEE
Propriétés des échantillons finis par l'utilisation des équations d'estimation généralisées

Shenghai Zhang et/and Mary E. Thompson , Unversity of Waterloo

The generalized estimating equations (GEE) methodology is considered the most popular approach for estimating both the regression parameters and correlations in marginal models for repeated responses. Based on the method by Liang and Zeger(1986), the software solving GEE based on Liang and Zeger's method is widely used, even though the authors pointed out that the finite sample performance of the estimator required further study, but the bias in estimating regression parameters for finite samples has not been studied except in some simulation studies. In this talk, we will give approximations for the bias, MSE and variances of the estimator.

La méthodologie des équations d'estimation généralisées (GEE) est considérée comme l'approche la plus populaire pour estimer les deux paramètres et la corrélation de la régression dans les modèles marginaux pour les mesures répétées. Basé sur la méthode de Liang et Zeger(1986), les logiciels résolvant les GEE basés sur la méthode de Liang et Zeger sont largement répandus, même si les auteurs ont mentionnés que la performance des estimateurs pour un échantillon fini nécessite plus d'étude, mais le biais des paramètres de régression pour les échantillons finis n'a pas été étudié excepté dans certaines études de simulations. Dans cette présentation, nous donnerons des approximations pour le biais, l'erreur quadratique moyenne et variances de l'estimateur.

Tuesday, May 28th/Mardi 28 mai, 14:45 **TSH B128**

Searching for thresholds in a simulation study
Recherche de seuils dans une étude de simulations

Andrea Benedetti et/and Michal Abrahamowicz, McGill University

In a variety of research settings, investigators may wish to detect and estimate a threshold in the association of interest. A spectrum of methods appears in the recent literature; however their statistical properties remain, for the most part, unevaluated. In any method, there are two steps to identify thresholds. The first is determining potential threshold locations, while the second is deciding if a threshold association indeed exists and where. We compare four methods and three criteria (totalling 12 distinct procedures) to decide if a threshold exists, in a simulation study. One of the criteria was based on an F test of our own design. Type 1 error is estimated for each method. We also investigate the impact of the position of the true threshold on power, and precision and bias of the estimated threshold. In one of the more novel methods, the use of GAM smoothing spline models is incorporated.[1] We search for potential thresholds in a neighbourhood of suspicious points located

by taking all points whose mean numerical second derivative (a measure of local curvature) is more than 1 standard deviation away from the mean of all second derivatives. A threshold association was declared if our F-test indicated that a threshold model fitted significantly better than a linear model. This method had acceptable type 1 error and had, on average, better power than alternative methods. We applied these methods and tests to determine if a threshold exists in the association between systolic blood pressure (SBP) and body mass index (BMI) in several real data sets.

1.Hastie T, Tibshirani R. Generalized Additive Models. New York: Chapman and Hall, 1990.

Dans une variété de configurations de recherches, les investigateurs peuvent souhaiter détecter et estimer un seuil dans une association d'intérêt. Un éventail les méthodes apparaît dans la littérature récente; toutefois leurs propriétés statistiques demeurent, pour la plupart, non évaluées. Dans n'importe quelle méthode, il y a deux étapes pour identifier des seuils. La première détermine les emplacements potentiels du seuil, alors que la seconde décide si une association de ce seuil existe et où elle se situe.

Nous comparons quatre méthodes et trois critères (pour un total de 12 procédures distinctes) pour décider si un seuil existe, dans une étude de simulation. Un des critères a été basé sur un test F de notre propre conception. L'erreur de type 1 est estimée pour chaque méthode. Nous étudions également l'impact de la position du vrai seuil sur la puissance, la précision et la polarisation du seuil estimé.

Dans une des méthodes les plus récentes, l'utilisation des splines de lissage GAM est incorporée [1]. Nous recherchons pour des seuils potentiels dans un voisinage d'un point critique en prenant tout point dont la moyenne numérique de la deuxième dérivée (une mesure locale de courbure) est supérieur à un écart type de la moyenne de toutes les dérivées secondes. Une association pour le seuil a été déclarée si notre test F indiquait qu'un modèle avec seuil s'ajustait significativement mieux à un modèle linéaire.

Cette méthode avait une erreur de type 1 acceptable et avait, en moyenne, une meilleure puissance que les méthodes alternatives. Nous avons appliqué ces méthodes et les tests pour déterminer si un seuil existe pour l'association entre la tension artérielle systolique (TAS) et l'indice de masse corporel (IMC) dans plusieurs vrais jeux de données.

1. Hastie T, Tibshirani R. Generalized Additive Models. New York: Chapman et Hall, 1990.

Session 27: Theoretical Survey Methods/Méthodes d'enquête - théorie

Tuesday, May 28th/Mardi 28 mai, 13:30 **TSH B106**

Re-calibration of higher-order calibration weights
Recalibration des poids de calibration d'ordre élevé

Patrick Farrell et/and Sarjinder Singh, Carleton University

A new technique for re-calibrating the higher-order calibrated estimators of the variances of various estimators of the population total is proposed. Recent advances in programming techniques and computational speed make the approach appealing for practical use. Estimators of the variances of the sample mean and the ratio and regression estimators under different sampling schemes are shown to be special cases of the proposed technique. A new system of predictors for the population variance is shown to be a special case of the approach as well. The results of an empirical study designed to investigate the properties of the proposed methodology under simple sampling designs are also reported.

On propose une nouvelle technique pour recalibrer les estimateurs de grand ordre calibrés des variances de divers estimateurs du total d'une population. Les progrès récents dans les techniques de

programmation et la vitesse de calcul font de cette approche une solution intéressante en pratique. Les estimateurs des variances de la moyenne échantillonnale et les estimateurs de taux et de régression sous différentes méthodes d'échantilonnage s'avèrent des cas spéciaux de la technique proposée. Un nouveau système des prédicteurs pour la variance dans la population s'avère également un cas spécial de cette approche. Les résultats d'une étude empirique conçue pour étudier les propriétés de la méthodologie proposée sous des plans d'échantilonnage simples sont également donnés.

Tuesday, May 28th/Mardi 28 mai, 13:45 **TSH B106**

Benchmarking hierarchical Bayes small area estimators with application in census undercoverage estimation

Standardisation des estimateurs de Bayes hiérarchiques pour de petits domaines avec des applications à la sous-estimation dans un recensement

Yong You, Statistics Canada/Statistique Canada et/ and J.N.K. Rao, Carleton University

Linear mixed effects models such as the Fay-Herriot model (1979) and non-linear mixed effects models such as the unmatched area level models proposed by You and Rao (2002) have been used in small area estimation to obtain efficient model-based small area estimators. It is often desirable to benchmark the model-based estimates so that they add up to the direct survey estimates for large areas to protect against possible model mis-specification and possible overshrinkage. In this paper, hierarchical Bayes (HB) unmatched area level models are considered. Posterior means and posterior variances of parameters of interest are first obtained using the Gibbs sampling method. Then we benchmark the HB estimators (posterior means) to obtain the benchmarked HB (BHB) estimators. Posterior mean squared error (PMSE) is then used as a measure of uncertainty for the BHB estimators. The PMSE can be represented as the sum of the usual posterior variance and a bias correction term. We evaluate the HB and the BHB estimators in the application of Canadian census undercoverage estimation. The sum of the provincial BHB census undercount estimates is equal to the direct survey estimate of the census undercount for the whole nation.

Les modèles linéaires à effets mixtes tels le modèle de Fay-Herriot(1979) et les modèles non linéaires à effets mixtes tels que les modèles à niveaux de zones non appariées proposés par You et Rao (2002) ont été utilisés pour des estimations sur de petits domaines pour obtenir des estimateurs efficaces basés sur un modèle. Il est souvent souhaitable de standardiser les estimateurs obtenus, toujours basés sur un modèle, pour qu'ils puissent s'additionner afin de fournir l'estimateur direct donné par le sondage sur un plus grand domaine et pour qu'ils fournissent une protection contre la mauvaise spécification et la trop grande réduction du modèle.

Dans cette présentation, les modèles à niveaux de domaines non appariés avec une approche de Bayes hiérarchique(HB) sont considérés. Les moyenne et variance a posteriori des paramètres d'intérêt sont d'abord obtenus en utilisant la méthode d'échantillonnage de Gibbs. Ensuite, nous standardisons les estimateurs HB (moyenne a posteriori) pour obtenir des estimateurs standardisés de HB (BHB). L'erreur quadratique moyenne a posteriori(PMSE) est alors utilisée comme mesure d'incertitude pour les estimateurs BHB. Le PMSE peut être représenté comme la somme de la variance a posteriori habituelle et d'un facteur de correction pour le biais. Nous évaluons le HB et les estimateurs BHB dans une application de l'estimation de la sous représentation dans le recensement canadien. La somme des estimations provinciales BHB du sous-dénombrement dans le recensement est égale à l'estimation directe de l'étude du sous-dénombrement de recensement pour toute la nation.

Tuesday, May 28th/Mardi 28 mai, 14:00 **TSH B106**

Linearization variance estimators for survey data
Estimateurs de la variance par linéarisation applicables aux données d'enquêtes
Abdellatif Demnati, Statistics Canada/Statistique Canada, et/and J.N.K. Rao, Carleton University

In survey sampling, Taylor linearization is often used to estimate nonlinear finite population parameters such as ratios, regression and correlation coefficients which can be expressed as smooth functions of totals. It is generally applicable to any sampling design, but it can lead to more than one variance estimator that are asymptotically equivalent under repeated sampling. The choice among the variance estimators requires other considerations such as conditional properties of the variance estimators. A new approach to deriving Taylor linearization variance estimators is proposed. This method is based on representing Taylor linearization in terms of partial derivatives with respect to design weights. It leads to variance estimators with good conditional properties and agrees with a jackknife linearization variance estimator when the latter is applicable. We apply the method to a variety of problems, covering general calibration estimators of a total as well as other estimators defined either explicitly or implicitly as solutions of estimating equations. In particular, estimators of logistic regression parameters with calibration weights are studied. Extensions to two phase sampling are also presented. Our approach leads to a new variance estimator for a general class of calibration estimators that includes generalized raking ratio and generalized regression estimators.

Dans le domaine de l'échantillonnage d'enquête, on utilise souvent la linéarisation par série de Taylor pour obtenir les estimateurs de la variance de paramètres non linéaires d'une population finie, comme les quotients ou les coefficients de régression et de corrélation qui peuvent être exprimés sous forme de fonction continue des totaux. La linéarisation par série de Taylor s'applique généralement à tout plan de sondage, mais elle peut produire plusieurs estimateurs de la variance asymptotiquement non biaisés par rapport au plan de sondage dans des conditions d'échantillonnage répété. Le choix de l'estimateur approprié de la variance doit se fonder sur d'autres critères, tels que i) l'absence approximative de biais dans la variance du modèle de l'estimateur dans les conditions du modèle considéré, ii) la validité en cas d'échantillonnage conditionnel répété. Nous proposons une nouvelle méthode de calcul des estimateurs de la variance par linéarisation de Taylor qui mène directement à un estimateur unique de la variance satisfaisant aux critères mentionnés ci-haut. Cet estimateur concorde avec l'estimateur de la variance par linéarisation selon la méthode du jackknife lorsque celui-ci est applicable. Nous appliquons notre méthode à la résolution de divers problèmes, allant des estimateurs d'un total à d'autres estimateurs définis explicitement ou implicitement comme solution d'équations d'estimation. Nous étudions notamment les estimateurs des paramètres de régression logistique avec poids de calibration. Notre méthode produit un nouvel estimateur de la variance pour une classe générale d'estimateurs par calage qui englobe les estimateurs généralisés par la méthode itérative du quotient et les estimateurs généralisés par régression. Nous étendons la méthode proposée à l'échantillonnage à deux phases pour obtenir un estimateur de la variance qui utilise plus complètement les données du premier degré d'échantillonnage que les estimateurs de la variance par linéarisation classique.

Tuesday, May 28th/Mardi 28 mai, 14:15 **TSH B106**

Unrounding procedures for systematically rounded survey income data
Procédures exactes pour des données sur le revenus arrondies systématiquement
Emile Allie, Statistics Canada/Statistique Canada

In most surveys on income, rounding from respondents occurs. It is well known that respondent rounding can be problematic to statistical measures. To insure confidentiality of published data, some statistical agencies are systematically rounding income data, which significantly increase the rounding problems. This paper evaluates various unrounding procedures like the inversion of the rounding procedure, the use of fiscal information applied to rounded and original data. The evaluation will be based on two continuous and two upper truncated distributions of income variables from the Survey on Labour and Income Dynamics (SLID),Statistics Canada, 1998. Fiscal information are from administrative files.

Dans la plupart des études sur le revenu, les données founies par les répondants sont fréquemment arrondies. Il est bien connu que cet arrondissement par les répondants peut être problématique dans le calcul des mesures statistiques. Pour assurer la confidentialité des données publiées, quelques agences statistiques arrondissent systématiquement les données sur le revenu, ce qui augmente de manière significative les problèmes d'arrondissements. Cette présentation évalue diverses procédures de non arrondissements comme l'inversion du procédé d'arrondissement, l'utilisation d'information fiscale appliquée aux données arrondies et initiales. L'estimation sera basée sur deux distributions continues ainsi que deux ditribution continues tronquées à droite pour les variables du revenu dans l'étude sur la dynamique du travail et du revenu (SLID) réalisée par Statistiques Canada, 1998. L'information fiscale sont obtenues à partir des dossiers administratifs.

Tuesday, May 28th/Mardi 28 mai, 14:30 **TSH B106**

Estimation of variance from missing data
Estimation de la variance avec des données manquantes

Sarjinder Singh, Carleton University et/and Raghunath Arnab, University of Durban-Westville

Any large scale survey may be prone to nonresponse problems. No exact formulation of the nature of nonresponse in survey is available. So, several methods of handling nonresponse problems are proposed by survey statisticians. In this paper, the problems of estimation of population total and its variance have been studied in the presence of nonresponse. The proposed methodology is based on the assumption that the set of respondents (comprises response sample) is a Poisson sample elected by nature. The proposed method is more general than the method proposed by Sarndal (1992) since one can use auxiliary information in estimating the unknown response probabilities. The proposed estimators are compared with Sarndal's (1992) estimator using similar simulation studies as prescribed by Sarndal (1992). The simulation results reveals that the proposed method performs better than the existing competitors.

N'importe quelle enquête à grande échelle peut être sujette aux problèmes de non réponse. Aucune formulation exacte de la nature de la non réponse pour l'étude n'est disponible. Ainsi, des statisticiens ont proposé plusieurs méthodes pour traiter ces problèmes de non réponse dans une étude. Dans cette présentation, les problèmes d'estimation de la taille de la population et de sa variance ont été étudiés en présence de non réponse. La méthodologie proposée est fondée sur l'hypothèse que l'ensemble de répondants (composé de l'échantillon ayant répondu) est un échantillon aléatoire de Poisson. La méthode proposée est plus générale que la méthode proposée par Sarndal (1992) puisqu'on peut utiliser de l'information auxiliaire en estimant les probabilités de réponse inconnues. Les estimateurs proposés sont comparés aux estimateurs semblables de Sarndal en utilisant des études de simulations similaires à celles décrites par Sarndal (1992). Les résultats de simulations indiquent que la méthode proposée performe mieux que les concurrents existants.

Session 28: Multivariate Methods/Méthodes multidimensionels

Tuesday, May 28th/Mardi 28 mai, 15:30 **TSH B106**

Sensitivity to the criterion in generalized canonical correlation analysis. Influence du critère sur les analyses canoniques généralisées

Victor Nzobounsana et/and Dhorne Thierry, Université Rennes 2 - Haute Bretagne

Canonical Correlation Analysis measures and represents the linear relationship between two subsets of variables in an unambiguous way. Many procedures have been proposed to generalize Canonical Correlation Analysis to three or more sets of variables (cf Steel, R.G.D. (1951), Horst, P. (1961 a and b), Carroll, J.D. (1968), Kettenring, J.R. (1971), Lafosse (1989)). All of these procedures are based on the principle of optimizing some functions of the correlation matrix of linear combinations.

The solution of such analysis depends therefore on a criterion used explicitely or implicitely by the method. In this paper, is proposed a general class of criteria for Generalized Correlation Canonical Analysis. This general class includes, as specific cases, all the criteria presented in the literature. This gives the possibility of studying the sensitivity of the analysis with respect to criteria in order to assess empirical remarks made by Kettenring, J.R. (1971) on this subject. A numerical procedure to solve the optimized problem is presented. To illustrate the method, the real and some artificial data sets are used. The real data are given by Thurstone and Thurstone (1941). These real data have also been used by Horst, P. (1961 a,b) and Kettenring, J.R. (1971) in their papers.

L'analyse canonique des corrélations mesure et représente le rapport linéaire entre deux sous-ensembles de variables d'une façon non ambiguë. Plusieurs procédures ont été proposé pour généraliser l'analyse canonique des corrélations à trois ensembles ou plus de variables (voir Steel, R.G.D. (1951), Horst, P. (1961 a et b), Carroll, J.D. (1968), Kettenring, J.R. (1971), Lafosse (1989)).

Toutes ces procédures sont basées sur le principe d'optimiser quelques fonctions de la matrice de corrélation des combinaisons linéaires.

La solution d'une telle analyse dépend donc d'un critère employé explicitement ou implicitement par la méthode. Dans cet article, une classe générale de critères est proposé pour l'analyse canonique généralisée des corrélations. Cette classe générale inclut, en tant que cas spécifiques, tous les critères présentés dans la littérature. Ceci donne la possibilité d'étudier la sensibilité de l'analyse en ce qui concerne les critères afin d'évaluer certaines remarques empiriques faites par Kettenring, J.R. (1971). Une procédure numérique pour résoudre le problème optimisé est présentée. Pour illustrer la méthode, de vrais et quelques jeux de données artificiels sont employés. Les vraies données sont tirées de Thurstone et Thurstone (1941). Ces vraies données ont été également été employées par Horst, P. (1961 a et b) et Kettenring, J.R. (1971) dans leur article.

Tuesday, May 28th/Mardi 28 mai, 15:45 **TSH B106**

Hypothesis testing and power calculations under the generalized Bessel-type model Tests d'hypothèses et calcul de puissance sous un modèle du type Bessel généralisé

Lehana Thabane, McMaster University et/and Steve Drekic, University of Waterloo

In this paper, we consider hypothesis testing problems in which the involved samples are drawn from generalized multivariate modified Bessel populations. This is a much more general distribution that includes both the multivariate normal and multivariate-t distributions as special cases. We derive the distribution of the Hotelling's T^2-statistic for both the one-sample and two-sample problems, as well as the distribution of the Scheffe's T^2-statistic for the Behrens-Fisher problem. In all cases, the

non-null distribution of the corresponding F-statistic follows a new distribution which we introduce as the non-central F-Bessel distribution. Some statistical properties of the distribution are studied. Further, the distribution was utilized to perform some power calculations for tests of means for different models which are special cases of the generalized multivariate modified Bessel distribution, and the results compared with those obtained under the multivariate normal case. Under the null hypothesis, however, the non-central F-Bessel distribution reduces to the central F-distribution obtained under the classical normal model.

Dans cette conférence, nous considérons les problèmes reliés aux tests d'hypothèses pour lesquels les échantillons impliqués sont tirés de populations suivant une densité de Bessel modifiée multidimensionnelle généralisée. C'est une distribution très générale dont les distributions normales multidimensionnelles et les lois de Student multidimensionnelles sont cas spéciaux. Nous dérivons la distribution de la statistique T^2 de Hotelling pour les problèmes avec un seul échantillon et deux échantillons, aussi bien que la distribution de la statistique T^2 de Scheffe pour le problème de Behrens-Fisher. Dans tous ces cas, la distribution de la statistique F sous l'hypothèse alternative correspondant suit une nouvelle distribution que nous présentons comme une distribution F-Bessel non centrée. Quelques propriétés statistiques de cette distribution sont étudiées. De plus, la distribution a été utilisée pour exécuter quelques calculs de puissance pour des tests pour la moyenne pour différents modèles qui sont des cas spéciaux de la distribution de Bessel modifiée multidimensionnelle généralisée, et les résultats ont rivalisé avec ceux obtenus sous le cas normal multidimensionnel. Cependant, sous l'hypothèse nulle, la distribution F-Bessel non centrée se réduit à la distribution F centrée obtenue sous le modèle normal classique.

Tuesday, May 28th/Mardi 28 mai, 16:00 **TSH B106**

Improving on the MLE of a bounded mean for spherical distributions
Sur l'estimation d'un paramètre de position borné pour des distributions à symétrie sphérique

François Perron, Universite de Montrél et/and Eric Marchand, University of New Brunswick

For the problem of estimating under squared error loss the mean of a p-variate spherically symmetric distribution where the mean lies in a ball of radius m, a sufficient condition for an estimator to dominate the maximum likelihood estimator is obtained. We use this condition to show that the Bayes estimator with respect to a uniform prior on the boundary of the parameter space dominates the maximum likelihood estimator whenever $m \leq \sqrt{p}$ in the case of a multivariate student distribution with d degrees of freedom, $d \geq p$. The sufficient condition $m \leq \sqrt{p}$ matches the one obtained by Marchand and Perron (2001) in the normal case with identity matrix. Furthermore, we derive a class of estimators which, for $m < \sqrt{p}$, dominates the maximum likelihood estimator simultaneously for the normal distribution with identity matrix and for all multivariate student distributions with d degrees of freedom, $d \geq p$. The family of distributions where dominance occurs includes the normal case; and includes all student distributions with d degrees of freedom, $d \geq 1$, for the case $p = 1$.

On considère le problème de l'estimation du paramètre de position d'un vecteur aléatoire en dimension p dont la distribution est à symétrie sphérique. On assume que ce paramètre est contenu dans une boule de rayon m centrée à l'origine. Dans ce contexte, on obtient une condition suffisante qui permet à un estimateur symétrique de dominer l'estimateur par la méthode du maximum de vraisemblance sous une perte quadratique. On vérifie ensuite que cette condition s'applique sur l'estimateur de Bayes par rapport à la loi a priori uniforme sur la frontière de la boule (c'est-à-dire: la sphère de rayon m centrée à l'origine) pour une distribution de Student à d degrés de liberté en autant que $m^2 \leq p \leq d$.

La condition $m \leq \sqrt{p}$ reproduit le même type de résultat que celui obtenu dans Marchand et Perron (2001) pour le cas particulier de la loi multinormale dont le paramètre de dispersion est la matrice identité. Nous proposons également toute une classe d'estimateurs qui dominent l'estimateur par la méthode du maximum de vraisemblance lorsque $m < \sqrt{p}$ et ce, simultanément, pour toute les lois de student à plusieurs dimensions qu'importe d où d est le nombre de degrés de liberté, en autant que $d \geq p$. Cette domination se vérifie églament pour la loi multinormale dont le paramètre de dispersion est la matrice identité.

Tuesday, May 28th/Mardi 28 mai, 16:15 **TSH B106**

A generalization of the Mahalanobis distance to mixed quantitative and qualitative multivariate data

Une généralisation de la distance de Mahalanobis pour jumeler des données multidimensionnelles quantitatives et qualitatives

Alexander de Leon et/and K.C. Carriere, University of Alberta

The analysis of multivariate data from several populations or groups frequently entails the estimation of a distance measure. For quantitative data, the standard distance measure used in practice is the so-called Mahalanobis distance. Bar-Hen and Daudin (1995) and Bedrick, Lapidus, and Powell (2000) recently proposed generalizations of the Mahalanobis distance to mixed data, the former for the case of mixed nominal and quantitative data and the latter for mixed ordinal and quantitative data. In this talk, the Mahalanobis distance is further generalized to data with mixtures of nominal, ordinal and quantitative variables. This is accomplished by specifying a model for the joint distribution of nominal, ordinal and quantitative variables which generalizes models previously studied in the literature including the general location model (Schafer, 1997), the grouped continuous model (Anderson and Pemberton, 1995; Poon and Lee, 1987), and the model of Bedrick, Lapidus and Powell (2000). From this model, a distance measure based on the Kullback-Leibler divergence is derived. The resulting generalized distance measure includes as special cases the Mahalanobis distance and its extensions mentioned previously. Maximum likelihood estimation is outlined and the asymptotic distribution of the resulting estimate is obtained. Finally, several related problems and issues for future research are identified.

L'analyse des données multidimensionnelles de plusieurs populations ou groupes nécessite fréquemment l'estimation d'une mesure de distance. Pour des données quantitatives, la mesure standard de distance utilisée dans la pratique est la distance de Mahalanobis. Bar-Hen et Daudin (1995) et Bedrick, Lapidus, et Powell (2000) ont proposé récemment une généralisation de la distance de Mahalanobis pour des mélanges de données, anciennement pour le cas de mélanges de données nominales et quantitatives et dernièrement pour un mélange de données ordinales et quantitatives. Dans cette présentation, la distance de Mahalanobis est encore généralisée aux données avec des mélanges des variables nominales, ordinales et quantitatives. Ceci est accompli par la spécification d'un modèle pour la densité conjointe des variables nominale, ordinale et quantitative qui généralise le modèle précédent étudié dans littérature comprenant le modèle général de position (Schafer, 1997), le modèle continu regroupé (Anderson et Pemberton, 1995; Poon et Lee, 1987), et le modèle de Bedrick, Lapidus et Powell (2000). De ce dernier modèle, une mesure de distance basée sur la divergence de Kullback-Leibler est déduite. La mesure généralisée résultante de distance inclue comme cas spécial la distance de Mahalanobis et ses généralisations sont mentionné précédemment. L'estimation du maximum de vraisemblance est donnée et la distribution asymptotique de l'estimation résultante est obtenue. Finalement, plusieurs problèmes et idées relativement aux recherches futures sont identifiés.

Tuesday, May 28th/Mardi 28 mai, 16:30 **TSH B106**

Sequential comparison of two treatments by means of parametric tests
Comparaisons séquentielles de deux traitements par des tests paramétriques basés sur les moyennes
Abdulkadir Hussein and/et Edit Gombay, University of Alberta

A class of statistics based on Rao's efficient score is proposed for testing composite hypothesis concerning the sequential comparison of two treatments (populations). Large sample approximations of the statistics and approximate critical values of the tests are given. The tests are then illustrated by normal and binomial population examples. Monte Carlo simulations are carried out in order to asses empirically the power and the average sample numbers of the tests. Finally, the tests are compared with some commonly used group sequential procedures.

On propose une classe de statistiques basées sur les scores efficaces de Rao pour tester une hypothèse composée concernant la comparaison séquentielle de deux traitements (populations). Des approximations de ces statistiques sont fournies par de grands échantillon et des valeurs critiques approximatives des tests sont données. Les tests sont illustrés par des exemples avec des populations normales et binomiales. Des simulations de Monte-Carlo sont faites pour estimer la puissance empirique et le nombre moyen d'échantillons nécessaires pour les tests. Enfin, les tests sont comparés à quelques procédures séquentielles généralement utilisées.

Tuesday, May 28th/Mardi 28 mai, 16:45 **TSH B106**

On a class of multivariate generalized exponential distributions.
Sur une classe de distributions exponentielles multidimensionnelles généralisées
Veeresh Gadag et/and K. Jauakumar, Memorial University of Newfoundland

We introduce a class of multivariate generalized exponential distributions and study some of its properties. Freund(1961), Moran(1967), Marshall and Olkin (1967), Gumbel(1961) among others studied bivariate and multivariate exptensions of exponential distribution. As a particular case of our class of multivariate generalized exponential distributions we introduce a new multivariate exponential distribution. We obtain explicit expressions for the joint probability density of some members of the new class.

Nous introduisons une classe de distributions exponentielles généralisées multidimensionnelles et étudions quelques unes de ses propriétés. Freund(1961), Moran(1967), Marshall and Olkin (1967), Gumbel(1961) sont parmi ceux qui ont étudié les généralisations bidimensionnelles et multidimensionnelles de la distribution exponentielle. En tant que cas spécial de notre distribution exponentielle généralisée multidimensionnelle, nous introduisons une nouvelle distribution exponentielle multidimensionnelle. Nous obtenons des expressions analytiques pour la densité conjointe de quelques membres de cette nouvelle classe.

Session 29: Stochastic Modeling/Modélisation stochastique

Tuesday, May 28th/Mardi 28 mai, 15:30 **KTH B105**

Associated random variables in the testing of Poisson-Voronoï tesselations
Variables aléatoires associées dans certains tests sur les mosaïques de Poisson-Voronoï
Jean Vaillancourt, Université du Québec à Hull et/and Susie Fortier, Statistique Canada/Statistics Canada

Associated random variables were introduced in 1967 by Esary, Proschan and Walkup. After reviewing some of the first and second order asymptotics emanating from this concept of association, we shall show how they can be used to get valuable statistical information on the complex Archambault-Moore (1995) test statistic for the blackened proportion of a Poisson-Voronoï tessalation randomly painted in black and white. Most competitors to this statistic can be shown to be asymptotically unbiased but CLT's and LIL's are usually harder to come by. This is joint work with Susie Fortier (Statistics Canada).

L'introduction par Esary, Proschan et Walkup des variables aléatoires associées remonte à 1967. Après avoir passé en revue quelques uns des résultats asymptotiques de premier et de second ordre rattachés à ce concept d'association, nous montrerons comment ils peuvent être exploités afin d'extraire des informations statistiques fort utiles sur le complexe test statistique d'Archambault et Moore (1995) pour la proportion noircie d'une mosaïque de Poisson-Voronoï peinturée au hasard en noir et blanc. On peut montrer que la plupart des compétiteurs de cette statistique sont asymptotiquement sans biais, mais les théorèmes limites centraux et lois du logarithme itéré correspondants sont souvent plus difficiles à obtenir. Ce travail est conjoint avec Susie Fortier (Statistique Canada).

Tuesday, May 28th/Mardi 28 mai, 16:00 **KTH B105**

Randomized quasi-Monte Carlo methods for multivariate integration
Méthodes quasi-Monte Carlo randomisées pour l'intégration multidimensionnelle

Christiane Lemieux, University of Calgary

In this talk, we review commonly used quasi-Monte Carlo methods for high-dimensional numerical integration. We present different randomization schemes and the variance of the associated estimators. The combination of these methods with other variance reduction techniques is discussed, as well as their connection with functional ANOVA decompositions. Numerical results are presented to illustrate the efficiency gains that can be provided by randomized quasi-Monte Carlo methods.

Dans cette présentation, nous présentons les méthodes quasi-Monte-Carlo qui sont les plus souvent utilisées pour l'intégration numérique en haute dimension. Nous revoyons différentes techniques de randomisation pouvant être utilisées avec ces méthodes et examinons la variance des estimateurs associés. La combinaison des méthodes quasi-Monte-Carlo randomisées avec d'autres techniques de réduction de la variance est discutée, ainsi que leur lien avec une décomposition ANOVA de l'intégrant. Nous présentons également des résultats numériques qui illustrent comment ces méthodes peuvent fournir des estimateurs plus efficaces pour différents problèmes.

Tuesday, May 28th/Mardi 28 mai, 16:30 **KTH B105**

Invariance principles for Studentized partial sum processes
Principes d'invariance pour des processus studentisés de sommes partielles

Miklós Csörgő, Barbara Szyszkowicz et/and Qiying Wang, Carleton University

Let $X, X_1, X_2, ...$, be i.i.d. non-degenerate random variables, $S_n = \sum_{j=1}^n X_j$ and $V_n^2 = \sum_{j=1}^n X_j^2$. In [1] we investigate the asymptotic behavior in distribution of the maximum of self-normalized sums, $\max_{1\le k\le n} S_k/V_k$, and the law of the iterated logarithm for self-normalized sums, S_n/V_n, when X belongs to the domain of attraction of the normal law. In this context, we establish a Darling-Erdős type theorem as well as an Erdős-Feller-Kolmogorov-Petrovski type test for self-normalized sums. In [2] we show that a self-normalized version of Donsker's theorem holds only under the assumption that X belongs to the domain of attraction of the normal law. Weighted approximations for the sequence

of self-normalized partial sum processes $\{S_{[nt]}/V_n,\ 0 \le t \le 1\}$, with applications to arc sine law and changepoint problems, are also established. This presentation will review some of the results of [1] and [2], as well as some earlier works on laws of the iterated logarithm for self-normalized partial sums and their increments.

References:

1. M. Csörgő, B. Szyszkowicz, Q. Wang, Darling-Erdős Theorems for Self-normalized Sums, *Technical Report Series of the Laboratory for Research in Statistics and Probability,* No. 354-July 2001, Carleton University - University of Ottawa.

2. M. Csörgő, B. Szyszkowicz, Q. Wang, Donsker's Theorem and Weighted Approximations for Self-normalized partial sums processes, *Technical Report Series of the Laboratory for Research in Statistics and Probability,* No. 360-October 2001, Carleton University - University of Ottawa.

Soient $X, X_1, X_2, ...,$ n variables aléatoires i.i.d. non dégénérées et $S_n = \sum_{j=1}^n X_j$ et $V_n^2 = \sum_{j=1}^n X_j^2$. En [1], on explore le comportement asymptotique de la distribution du maximum de la somme normalisée, $\max_{1\le k\le n} S_k/V_k$, et la loi du logarithme itéré pour la somme normalisée, S_n/V_n, quand X appartient au domaine d'attraction de la loi normale. Dans ce contexte, nous établissons un théorème du type Darling-Erdős de même qu'un test pour la somme normalisée du type Erdős-Feller-Kolmogorov-Petrovski. En [2], nous démontrons que la version normalisée du théorème de Donsker est valide uniquement sous l'hypothèse que X appartient au domaine d'attraction de la loi normale. Des approximations pondérées pour la suite des sommes partielles normalisées $\{S_{[nt]}/V_n,\ 0 \le t \le 1\}$ avec des applications à la loi arc sinus et au problème des points de rupture sont aussi établies. Cette présentation inclura des résultats de [1] et [2] de même que certains résultats obtenus précédemment sur les loi du logarithme itéré pour les sommes partielles normalisées et leurs incréments.

Références:

1. M. Csörgő, B. Szyszkowicz, Q. Wang, Darling-Erdős Theorems for Self-normalized Sums, Technical Report Series of the Laboratory for Research in Statistics and Probability, No. 354-July 2001, Carleton University - University of Ottawa.

2. M. Csörgő, B. Szyszkowicz, Q. Wang, Donsker's Theorem and Weighted Approximations for Self-normalized partial sums processes, Technical Report Series of the Laboratory for Research in Statistics and Probability, No. 360-October 2001, Carleton University - University of Ottawa.

Session 30: Spatial Sampling/L'échantillonnage spatial

Tuesday, May 28th/Mardi 28 mai, 15:30 **TSH B105**

Forest structure and forest birds: a balanced, model-based design for multivariate glms. Structure de la forêt et oiseaux : un plan d'expérience balancé basé sur un modèle pour des modèles linéaires généralisés multidimensionnels

Steve Cumming, Boreal Ecosystems Research Ltd, et/and Subash Lele, University of Alberta

Regional biomes, such as the boreal mixedwood forest, are heterogeneous at multiple spatial scales. For example, our 100,000 km^2 study region in northeast Alberta can be considered as a grid of 100 km^2 landscapes, each of which is a mosaic of different vegetation types, or habitats. Within landscapes, habitat structure is characterized by habitat abundance and configuration, or the spatial arrangement of habitat patches. The regional between-landscape variation in these two attributes is naturally very high. Increasingly, however, natural landscapes are being modified by industrial development, resulting in both habitat loss and the changes in configuration known as fragmentation. Unfortunately, the consequences for populations of habitat specialists cannot be predicted, because the relations between

landscape structure and species distributions and abundances are poorly understood. To elucidate some of these relations, a comprehensive, multi-year field survey was initiated in 2001. The sample units are landscapes, and the observations are Bernoulli or count data. Thus, data analysis involves model selection and inference for multivariate generalized linear models. Here, we describe the design of this study. We adopted a prediction-based approach to sample design. To achieve a form of balance, samples are constructed by a greedy iterative algorithm which minimizes the distances between the marginal distributions of the sample and the population. A new information-based measure, based on the sum of estimation and intrinsic errors, is being used as an additional sample design criteria for the second and subsequent years of data collection. We conclude by considering how our methodology could be adapted for optimal prediction given an expected future distribution of landscape structures.

Les biomes régionaux, tel que la forêt boréale, sont hétérogènes à différentes échelles spatiales. Par exemple, nos 100 000 km^2 de région d'étude au nord-est de l'Alberta peuvent être considérés comme des grilles horizontales de $100km^2$, dont chacune est une mosaïque de différents types de végétation ou d'habitats. Dans chacune des grilles, la structure de l'habitat est caractérisé par l'abondance d'habitats et leur configuration, ou l'agencement spatial des liens d'habitat. D'une grille à l'autre, la variation régionale de ces deux attributs est naturellement très élevée. De plus en plus, ces grilles naturelles normales sont modifiées par le développement industriel, ayant pour résultat la perte d'habitats et des changements dans la configuration connue sous le nom de fragmentation.

Malheureusement, les conséquences pour les populations avec des habitats spécifiques ne peuvent pas être prévues, parce que les relations entre la structure de ces grilles et les distributions des espèces et leur abondance sont mal comprises. Pour élucider certaines de ces relations, un sondage sur le terrain échelonné sur plusieurs années a débuté en 2001. Les unités échantillonnales sont les grilles, et les observations sont des données de compte de Bernoulli. Ainsi, l'analyse des données implique une sélection de modèle et de l'inférence pour des modèles linéaires multidimensionnels généralisés. Ici, nous décrivons la planification de cette étude. Nous avons adopté une approche basée sur la prévision lors de la conception de l'échantillon. Pour obtenir une sorte d'équilibre, les échantillons sont construits selon un algorithme itératif qui réduit au minimum les distances entre les distributions marginales de l'échantillon et de la population. Une nouvelle mesure basée sur l'information comme fonction de la somme des erreurs intrinsèques et des erreurs d'estimation est utilisée en tant que critère supplémentaire d'échantillonnage pour la deuxième année et les années subséquentes de la collecte de données. Nous concluons en considérant comment notre méthodologie peut être adapté pour la prévision optimale sachant une distribution moyenne future de la structure des grilles.

Tuesday, May 28th/Mardi 28 mai, 16:00 **TSH B105**

On optimal spatial designs
Sur les plans d'expérience spatiaux optimaux

Steve Thompson, Pennsylvania State University

Many real populations are characterized by spatial covariance functions that are positive and decreasing with distance. Another commonly encountered trait of spatially uneven populations is that variability is higher where values are higher. More specifically, given the observed value of the variable of interest at a sampling location, the conditional variance at nearby sites is an increasing function of the value observed.

The decreasing covariance function (or increasing variogram) motivates the use of spatial designs with spread the sample sites apart, such as systematic designs and designs partitioning the region into many strata. The increasing conditional variance function, on the other hand, suggests the use of adaptive designs. Optimal conventional and adaptive designs in a spatial setting will be discussed in this talk.

Plusieurs vraies populations se caractérisent par des fonctions spatiales de covariance positives diminuant avec la distance. Un autre trait généralement rencontré des populations dans des espaces différents est que la variabilité est plus grande où les valeurs sont plus grandes. Plus spécifiquement, si la valeur observée de la variable d'intérêt à un emplacement de prélèvement est donnée, la variance conditionnelle aux sites voisins est une fonction croissante de cette valeur observée.

La fonction décroissante de covariance (ou le variogramme croissant) motive l'utilisation de plans d'expérience dans l'espace avec des sites d'échantillonnage séparés, comme des plans systématiques et des plans divisant la région en plusieurs strates. La fonction conditionnelle croissante de variance, d'autre part, suggère l'utilisation de plans adaptatifs. Des plans conventionnels et adaptatifs optimaux dans une configuration spatiale seront discutés dans cette présentation.

Tuesday, May 28th/Mardi 28 mai, 16:30 **TSH B105**

Designs for predicting the extremes of spatial processes
Plan d'expérience pour prédire les valeurs extrêmes dans un processus dans l'espace

James Zidek, University of British Columbia, Nhu D. Le, BC Cancer Agency, et/and Li Sun, Ericsson Berkeley Research Center

In this paper, we describe methods for adding sites to an environmental monitoring networks. One such approach relies on maximizing entropy and thus side-steps the need to specify specific design objectives. (Generally the designer will face a multiplicity of competing objectives and even some that have not yet been foreseen.) We demonstrate how that approach can be used to extend a monitoring network in Vancouver, that measures hourly small particulate airborne concentrations. This approach requires a spatial predictive distribution of the concentrations for potential sites, given the data measured by the existing network. Finding that distribution is itself a challenging problem since, besides temporal correlation, the observed data also have a staircase pattern due to different operational initiations of the stations in the existing network. We will describe a Bayesian framework for obtaining this spatial predictive distribution. Here, the multivariate response is assumed to have a joint Gaussian distribution conditional on its mean and covariance function. A conjugate prior is used for these parameters with its hyperparameters being fitted empirically. We will also examine whether such a design could be a satisfactory solution when different metrics, eg. daily 1-hour maximum, are considered. Since the method described above yields a predictive distribution, a second design approach becomes worthy of consideration. The latter, more in line with the objectives of regulating air pollution levels, seeks to locate the monitoring stations so that the probability of violating a regulatory standard at unmonitoring sites is minimized. This leads to some interesting design issues that we discuss.

Dans cet article, nous décrivons des méthodes pour ajouter des sites aux réseaux de contrôle de l'environnement. Une telle approche se fonde sur la maximisation de l'entropie et évite ainsi la nécessité d'indiquer des objectifs spécifiques pour le plan d'expérience. (Généralement, le concepteur fera face à une multitude d'objectifs dont certains n'ayant pas été originalement prévus.) Nous démontrons comment cette approche peut être employée pour étendre le réseau de surveillance à Vancouver mesurant à chaque heure la concentrations de petites particules dans l'air. Cette approche exige une distribution prédictive spatiale des concentrations pour les sites potentiels, étant donné les quantités mesurées par le réseau déjà existant. Trouver cette distribution est un problème intéressant puisque, outre la corrélation temporelle, les données observées ont également une configuration en escalier due à différents déclenchements opérationnels des stations dans le réseau existant. Nous décrirons un cadre bayésien pour obtenir cette distribution prédictive spatiale. Ici, on suppose que la réponse multidimensionnelle a une distribution conjointe de Gauss conditionnelle à sa moyenne et sa covariance. Une densité a priori conjuguée est utilisée pour ces paramètres avec des hyperparamètres

trouvés empiriquement. Nous examinerons également si une telle expérience pourrait être une solution satisfaisante quand la mesure est différente, par exemple le maximum quotidien de données horaires, sont considérés.

Puisque la méthode décrite ci-dessus produit une distribution prédictive, une deuxième approche mérite considération. Cette dernière, plus en conformité avec les objectifs de régulation des niveaux de pollution atmosphérique, tente de localiser les stations de surveillance de sorte que la probabilité de violer une norme aux sites non surveillés soit minimisée. Ceci mène à quelques plans d'expérience intéressants que nous discutons.

Session 31: Statistical Methods for Occupational Risk Assessment/ Les méthodes statistiques pour mesurer le risque professionnel

Tuesday, May 28th/Mardi 28 mai, 15:30 **KTH B135**

Exposure assessment for studying relationships between air pollution and chronic diseases: A case-control study of lung cancer patients in British Columbia

Estimation de l'exposition pour l'étude des relations entre les polluants atmosphérique et les maladies chroniques : une étude cas-témoin pour les patients atteint d'un cancer des poumons en Colombie-Britannique

Nhu Le, BC Cancer Agency

In this talk, statistical problems related to the estimation of the cumulative air pollution exposure for individuals, in a case-control setting, whose locations of residence might have changed several times, will be discussed. Some recent advances in statistical theory for dealing with such problems will be presented. Specifically, a hierarchical Bayesian approach using the lognormal distribution, in conjunction with a conjugate generalized inverted Wishart distribution, will be used to obtain the joint predictive distribution of concentration levels at locations of interest. The predictive distirbution will be conditional on historical data collected by monitoring networks where stations have been added over time, creating a monotone staircase pattern with the highest step corresponding to the oldest station. The prior distribution will allow different degrees of freedom to be fitted for individual steps, taking into account the differential amounts of information available from stations at the different steps in the staircase. The approach will be demonstrated with a case-control study of lung cancer patients in British Columbia. Preliminary results on the lung cancer impact of air pollution will be presented.

Dans cette présentation, les problèmes statistiques reliés à l'estimation de l'exposition répétitive à des polluants atmosphériques pour des individus, dans une étude cas-témoin, dont les emplacements de la résidence pourraient avoir changé plusieurs fois, seront discutés. Quelques progrès récents dans la théorie statistique pour traiter de tels problèmes seront présentés. Spécifiquement, une approche bayésienne hiérarchique utilisant la distribution log-normale, avec une distribution inverse-Wishart généralisée conjuguée, sera employée pour obtenir la distribution conjointe prédictive des niveaux de concentration aux emplacements d'intérêt. Le distribution prédictive sera fonction des données historiques rassemblées par la surveillance des réseaux où des stations ont été ajoutées avec le temps, créant une densité monotone en escalier, où le pallier le plus élevé correspond à la plus vieille station. La distribution a priori permettra l'ajustement de différents degrés de liberté pour chacun des palliers, et tenant compte de la qualité d'information différente fournie par les stations aux différents palliers de l'escalier. L'approche sera illustrée avec une étude cas-témoin sur des patients ayant le cancer du poumon en Colombie-Britannique. Les résultats préliminaires sur l'impact de la pollution atmosphérique sur le cancer du poumon seront présentés.

Tuesday, May 28th/Mardi 28 mai, 16:00 **KTH B135**

Application of statistical principles to the determination of occupational health risk
Application des principes statistiques à la détermination du risque professionel pour la santé

Mik Bickis, University of Saskatchewan, Ugis Bickis and/et Tom Beardall, Phoenix OHC

Population health risks are normally expressed in probabilistic terms; for example, the "acceptable" incremental lifetime cancer risk (ILCR) is usually regarded as 10^-6. On the other hand, such risks are typically managed in the context of deterministic exposure limits. Although sometimes described as "safe" levels, these limits need not represent a complete lack of risk, but rather an acceptable level of risk. In the occupational setting, the population may be very small, and the intent is to protect the health of all workers by sampling a few exposures. Often, attempts are made to classify the workplace population into "homogeneous exposure groups" (HEGs), but this approach may be akin to a self-fulfilling prophesy ... in order to determine empirically what the exposures are, individuals are classified into exposure groups on the basis of expert judgement. Furthermore, the exposure may be determined over less than one work shift, with the implicit assumption that this is representative of a working lifetime. For some situations / substances (e.g. that have exposure limits expressed as a "ceiling" or other short-term limits), so-called time-weighted average (TWA) exposures may not be as relevant as excursions lasting only a few minutes; again, professional judgement is used to select an appropriate sampling time frame. What statistical methods are applied, and are they truly applicable, in this context?

Les risques concernant la santé d'une population sont normalement exprimés en termes probabilistes. Par exemple, l'accroissement "acceptable" du risque de cancer dans une vie (ILCR) est habituellement considéré comme 10^{-6}*. D'autre part, de tels risques sont typiquement contrôlés dans un contexte de limites d'exposition déterministes. Bien que parfois décrit comme des niveaux "sûrs", ces limites ne représentent pas un manque de risque complet, mais plutôt un niveau de risque acceptable.*

Dans un contexte professionnel, la population peut être très petite, et l'intention est de protéger la santé de tous les ouvriers en prélevant quelques mesures. Souvent, il existe une volonté de classifier la population reliée au lieu de travail dans des groupes "d'exposition homogène" (HEGs). Cependant cette approche peut être apparentée à une prophétie subjective... car, afin de déterminer empiriquement ce que sont les niveaux d'exposition, les individus sont classifiés dans des groupes d'exposition sur la base du jugement d'un expert. De plus, l'exposition peut être déterminé dans une période de temps d'un seul quart de travail, avec l'hypothèse implicite que ce temps est représentatif de l'exposition sur une période aussi longue qu'une vie entière.

Pour quelques situations/substances (par exemple qui ont des limites d'exposition exprimées comme un "plafond" ou d'autres limites à court terme), les moyennes d'exposition avec temps pondérés (TWA) peuvent ne pas être aussi appropriées que des excursions durant seulement quelques minutes. De plus, le jugement d'un professionnel est employé pour déterminer un intervalle de temps approprié pour le prélèvement de la mesure d'exposition. Quelles méthodes statistiques sont appliquées, et sont-elles vraiment applicables dans ce contexte?

Tuesday, May 28th/Mardi 28 mai, 16:30 **KTH B135**

Issues in exposure-reponse models for occupational risk assessment
Les modèles dose-réponse pour mesurer le risque professionnel

Kyle Steenland, National Institute for Occupational Safety and Health, James A. Deddens et/and Siva Sivaganesan, University of Cincinnati

Occupational risk assessment typically must answer the question: "how much excess risk exists at any given level of exposure for an exposed individual vs. a nonexposed individual." An exposure-response model is required to answer this question. Here we focus on issues in statistical modeling of exposure-response trends in mortality studies. Categorical analyses are useful for detecting the shape of the exposure-response curve, but are dependent on choice of cutpoints and cannot be used directly in risk assessment which requires a smooth parametric curve. Restricted cubic splines and penalized splines are useful intermediates between categorical and simpler parametric curves, and may help to choose a simpler parametric curve. However, their shapes will depend on the degree of "smoothing". Exposure-response curves in occupational epidemiology often flatten out at high exposures, for a number of reasons (eg., greater misclassification at high exposure, saturation of metabolic pathways, etc). A log transformation of exposure often provides a good parametric fit to such curves, but has the disadvantage of a very high slope at low exposures, which may be the relevant exposures. The model with the best statistical fit may not be the "best" model for risk assessment. Bayesian restrictions on exposure-response curves may prove useful in some settings in increasing precision. These points are illustrated using data from published studies.

L'estimation du risque professionnel doit typiquement répondre à la question: "Dans quel mesure le risque augmente-t-il selon le niveau d'exposition pour un individu exposé versus un individu non exposé." Un modèle de réponse selon l'exposition est exigé pour répondre à cette question. Ici nous nous concentrons sur des questions sur les tendances de la modélisation statistique dans des telles études concernant la mortalité. Les analyses catégorielles sont utiles pour déduire la forme de la courbe dose-réponse, mais dépendent du choix des points de rupture et ne peuvent pas être utilisées directement dans l'estimation des risques qui exige une courbe paramétrique lisse. Les splines cubiques sous contraintes et les splines avec pénalité sont les intermédiaires utiles entre les courbes paramétriques simples et les courbes catégorielles, et peuvent aider à choisir une courbe paramétrique encore plus simple. Cependant, leur forme dépendra du niveau de "lissage". Les courbes dose-réponse en épidémiologie professionnelle sont souvent applaties lorsque le niveau d'exposition est élevé et ce, pour un certain nombre de raisons (par exemple, un taux plus élevé de fausses classifications, une saturation des voies métaboliques, etc.). Une transformation logarithmique du niveau d'exposition fournit souvent un bon ajustement paramétrique à de telles courbes, mais a l'inconvénient de créer une pente très élevée à des niveaux de basses expositions, qui peut toutefois être appropriée. Le modèle avec le meilleur ajustement statistique peut ne pas être le "meilleur" modèle pour l'estimation des risques. Les restrictions bayésiennes aux courbes dose-réponse peuvent s'avérer utiles dans quelques plan d'expérience par l'augmentation de la précision. Ces points sont illustrés en utilisant des données d'études déjà publiées.

Session 33: Business and Industry Section Special Invited Address/Groupe de statistique industrielle et de gestion : Allocution sur invitation spéciale

Wednesday, May 29th/Mercredi 29 mai, 8:30 **TSH B105**

The changing nature of data and its implications for applied statistics
Le changement dans la nature des données et ses implications en statistique appliquée

John MacGregor, McMaster University

The nature of much of the data collected routinely in industrial, financial and research settings has changed significantly in the past decade with the presence of on-line computer systems and with the collection of large data banks. These data sets are not only large, but typically are non-full rank and

non-causal in nature. Traditional statistical analysis and design approaches are very ill-suited to the investigation of these problems. In this presentation, we look at the latent variable model and latent variable estimation methods such as Principal Component Analysis (PCA) and Partial Least Squares (PLS) as a natural way of treating many of these problems. The success of these methods in three areas will be discussed and illustrated through industrial examples. The first involves the analysis of industrial databases for trouble- shooting process problems, and for establishing multivariate SPC schemes. The second looks at the problem of designing experiments (DOE) in such high dimensional systems for purposes such as drug discovery. The third considers the problem of extracting information from data intensive sensors such as multispectral digital images.

La nature d'une grande partie des données recueillies continuellement dans les domaines industriel, financier et de recherches a changé sensiblement dans la dernière décennie avec la présence des systèmes informatiques en ligne et avec la cueillette de grandes banques de données. Ces jeux de données sont non seulement grands, mais ne sont typiquement pas de plein rang et non causal par nature. Les approches traditionnelles d'analyse statistique et de planification d'expérience sont très mal assorties à la recherche sur ces problèmes. Dans cette présentation, nous regardons le modèle avec une variable latente et les méthodes d'estimation rattachées telles que l'analyse en composantes principales (ACP) et la méthode des moindres carrés partiels(PLS) comme des méthodes naturelles pour traiter plusieurs de ces problèmes. Le succès de ces méthodes dans trois domaines sera discuté et illustré par des exemples industriels. Le premier implique l'analyse de bases de données industrielles pour des problèmes de procédé de dépannage, et pour établir des modèles SPC multidimensionnels. Le second regarde le problème de la conception des expériences dans de tels systèmes à dimension élevée pour des buts tels que la découverte de médicament. Le dernier considère le problème d'extraire de l'information à partir des capteurs de données puissants tels que pour des images digitales multispectrales.

Session 34: Statistical Inference for Mixture Models/ Inférence statistique pour les modèles de mélange

Wednesday, May 29th/Mercredi 29 mai, 8:30 **TSH B128**

A mixture model approach for finding informative genes in microarray studies
Une approche de mélange de lois pour trouver les gènes informatifs dans des études de microréseaux

Jing Qin et/and Glenn Heller, Memorial Sloan-Kettering Cancer Center

The statistical analysis of microarray data focuses on the association between genetic expression and an outcome variable. For each gene, a test of non-association generates a p-value. The multiple testing associated with the tens of thousands of genes typically incorporated into a microarray data analysis poses an interesting statistical challenge. Our approach is based on a two-stage procedure; an initial test is performed for the global null hypothesis that gene expression is not associated with an outcome variable. If this hypothesis is rejected, a parametric mixture model is proposed for the set of p-values to determine which genes are associated with outcome. The p-values are modeled as a mixture of uniform $(0, 1)$ random variables, representing the genes with no association to the outcome variable, and Beta (ξ, θ) random variables representing those related to outcome. The parameters of the Beta distribution are constrained to form a density decreasing in p. Likelihood analysis is used to estimate the mixture proportion and the Beta parameters. These estimates, along with the Bayesian false discovery rate, are used to determine a threshold p-value, whereby there is a high level of confidence that all genes with p-values less than the threshold, are associated with the outcome variable.

L'analyse statistique des données de microtableau se concentre sur l'association entre une expression génétique et une variable résultat.

Pour chaque gène, un test de non-association produit une valeur-p. Les tests multiples associés aux dizaines de milliers de gènes typiquement incorporés à une analyse de données de microtableau pose un défi statistique intéressant. Notre approche est basée sur un procédé à deux étapes; un premier test est réalisé pour l'hypothèse nulle globale que l'expression du gène n'est pas associée à une variable résultat. Si cette hypothèse est rejetée, on propose un modèle paramétrique de mélange pour l'ensemble des valeurs-p afin de déterminer quels gènes sont associés aux résultats. Les valeurs-p sont modélisées comme un mélange de variables aléatoires uniforme sur l'intervalle (0,1), représentant les gènes sans association à la variable résultat, et de variables aléatoires bêta, représentant ceux liés aux résultats. On pose comme contrainte sur les paramètres de la distribution bêta qu'ils engendrent une densité décroissante en p. L'analyse de la vraisemblance est utilisée pour estimer la proportion du mélange et les paramètres de la distribution bêta. Ces estimateurs, avec le taux bayésien de fausses découvertes, sont employés pour déterminer un seuil pour la valeur-p, pour lequel il y ait un niveau de confiance élevé que tous les gènes ayant des valeurs-p inférieurs à ce seuil, sont associés à la variable de résultats.

Wednesday, May 29th/Mercredi 29 mai, 9:00 **TSH B128**

On computing information in semiparametric mixture models
Sur le calcul de l'information contenue dans des mélanges semi-paramétriques

Mary Lesperance, University of Victoria et/and Bruce Lindsay, Penn State University

Semiparametric mixture models pose many interesting inferential problems. When the mixing distribution is unspecified, the parameter space is infinite-dimensional, and hence standard asymptotic results do not apply in general. Fisher's information arises naturally in standard models as the inverse of the asymptotic variance of maximum likelihood estimators. In this talk, we define what is meant by information in a semiparametric mixture model, and consider a conjugate gradient method for calculating it.

Les modèles de mélanges de semi-paramétriques posent beaucoup de problèmes intéressants d'inférence. Quand la distribution de mélange est non spécifiée, l'espace des paramètres est de dimension infinie, et par conséquent les résultats asymptotiques standards ne s'appliquent pas en général. L'information de Fisher survient naturellement dans les modèles standards comme l'inverse de la variance asymptotique des estimateurs du maximum de vraisemblance. Dans cette présentation, nous définissons ce qu'on entend par information dans un modèle de mélanges semi-paramétriques, et nous considérons une méthode de gradient conjugué pour la calculer.

Wednesday, May 29th/Mercredi 29 mai, 9:30 **TSH B128**

A modified likelihood ratio test for a mixed treatment effect
Un test du maximum de vraisemblance modifié pour l'effet d'un traitement mixte

Cindy Fu, Jiahua Chen, University of Waterloo et/and Jack Kalbfleisch, University of Michigan

We consider a two-sample problem, comparing treatment and control groups, in which only a portion of individuals in the treatment group are expected to respond to the treatment applied. In this situation, the observations in the treatment group arise from a mixture distribution with two components, one corresponding to the subgroup that does not. The problem is motivated by certain applications in genetics.

Of particular interest is the statistical problem of testing for no treatment effect versus this mixture alternative. Even in simple parametric cases, this problem is nonregular, and usual likelihood ratio and related tests have relatively complicated asymptotic distributions.

We consider semi-parametric and parametric models for this problem and discuss a modified likelihood ratio test. In the parametric case, it is found that the statistic has a simple chi-square limiting distribution. Simulations indicate that this test performs as well or better than alternative procedures that have been suggested in the literature. Areas for additional investigation are also identified.

Nous considérons un problème à deux échantillons, comparant le groupe traitement et le groupe contrôle, dans lesquels on s'attend seulement à ce qu'une fraction d'individus dans le groupe traitement réagisse au traitement appliqué. Dans cette situation, les observations dans le groupe traitement résultent d'une distribution de mélange avec deux composantes, une correspondant au sous groupe qui ne réagit pas. Le problème est motivé par certaines applications en génétique.

Le problème statistique de tester "aucun effet du traitement" contre cette alternative de mélange est d'un intérêt particulier. Même dans des cas paramétriques simples, ce problème est non régulier, et le rapport de vraisemblances et les tests relatifs ont des distributions asymptotiques relativement compliquées.

Nous considérons les modèles semi-paramétriques et paramétriques pour ce problème et discutons d'un test du rapport de vraisemblances modifié. Dans le cas paramétrique, on constate que la statistique a une simple distribution limite de khi-deux. Les simulations indiquent que ce test performe aussi bien ou mieux que les procédures alternatives qui sont suggérées dans la littérature. Des champs de recherche supplémentaire sont également identifiés.

Session 35: Survival Analysis/Analyse de survie

Wednesday, May 29th/Mercredi 29 mai, 8:30 **KTH B135**

Cure models for survival data with a nonsusceptible fraction
Modèles de traitement pour des analyses de survie avec une fraction non susceptible

Yingwei Peng, Memorial University of Newfoundland

Cure models are useful to model survival data with a nonsusceptible fraction. In this work, we review recent cure models proposed independently in statistical literature. Several important relationships among the cure models are revealed. We also discuss the pros and cons of the cure models. The results in this paper are useful for practitioners to understand and to determine appropriate cure models for survival data. Finally we illustrate the models with a leukemia data set.

Les modèles de traitement sont utiles pour modéliser des données de survie avec une fraction non susceptible. Dans cette présentation, nous passons en revue les modèles de traitement récents proposés de façon indépendante dans la littérature statistique. Plusieurs relations importantes parmi les modèles de traitement sont révélées. Nous discutons également du pour et du contre de ces modèles de traitement. Les résultats de cette recherche sont utiles pour les praticiens afin de comprendre et de déterminer les modèles de traitement appropriés pour des données de survie. Enfin nous illustrons les modèles avec un jeu de données sur la leucémie.

Wednesday, May 29th/Mercredi 29 mai, 8:45 **KTH B135**

Flexible regression models for three state progressive processes
Modèles de régression flexibles pour les processus progressifs à trois états

Karen Kopciuk, Samuel Lunenfeld Research Institute and/et David E. Matthews, University of Waterloo

Casting survival data within a three-state framework is an effective way to incorporate intermediate events into an analysis, especially when the right censoring is heavy. By exploiting the unidirectional nature of these processes, other types of incomplete data, such as interval-censored observations, can also be accommodated in a model. In this work, we develop a flexible regression model for a three-state progressive process where the terminal (third) state represents the event of interest and the intermediate state represents an auxiliary variable which must occur for all subjects prior to the terminal event. Transition times to a higher state can be known exactly, known to occur in some interval of time, or known to be greater than the study follow-up time. The effects of both fixed and time-varying explanatory variates can be assessed in our model. Different underlying time scales can be adopted by assuming the process has Markov, semi-Markov, modulated Markov, or modulated semi-Markov structure. Our regression model is quite general and combines features of the models proposed by Frydman (1995, Biometrics 51, 502-511) and Kim et al. (1993, Biometrics 49, 13-22). A version of the EM algorithm is developed and used to find the self-consistent estimates. An AIDS data set analyzed by these authors will be used to illustrate our regression approach.

Insérer des données de survie dans une structure à trois états est un moyen efficace d'incorporer des évènements intermédiaires dans une analyse, surtout lorsque la censure à droite est importante. En exploitant la nature unidirectionnel de ces processus, d'autres types de données incomplètes telles que les observations censurées par intervalle, peuvent aussi être prises en compte dans un modèle. Dans cette étude, nous développons un modèle flexible de régression pour un processus à trois états ou l'état final (le troisième) représente l'évènement d'intérêt et l'état intermédiaire représente une variable auxiliaire qui doit survenir pour tous les sujets avant l'état final. Les durées de transition jusqu'à un état plus avancé peuvent être connues exactement, connues comme survenant dans un intervalle de temps, ou connus comme étant plus longues que la durée de suivi de l'étude. Les effets de variables exploratoires fixes ou variables avec le temps peuvent tous les deux être évalués dans notre modèle. Différentes échelles de temps sous-jacentes peuvent être adoptées sous l'hypothèse d'un processus de Markov, semi-Markov, Markov modulé, ou d'une structure semi-Markov modulée. Notre modèle de régression est très général et combine les caractéristiques des modèles proposés par Frydman (1995, Biometrics 51, 502–511) et Kim et al. (1993, Biometrics 49, 13–22). Une variante de l'algorithme EM est développée et utilisée pour trouver les estimateurs autoconvergents. Un ensemble de données sur le SIDA analysée par ces auteurs, sera utilisé pour illustrer notre approche régressive.

Wednesday, May 29th/Mercredi 29 mai, 9:00 **KTH B135**

Bivariate location-scale models for log lifetimes
Modèles de position-échelle bidimensionnel pour les logarithmes des durées de vie

Wenqing He, Samuel Lunenfeld Research Institute Mt. Sinai Hospital et/and Jerry F. Lawless, University of Waterloo

Multivariate failure time data arise frequently in scientific investigations. Typically interest lies in the cases when failure times within the same cluster are correlated. In many models the logarithms of lifetimes have location-scale marginal distributions. We investigate the estimation of regression coefficients for bivariate location-scale models and show that the maximum likelihood estimators for the regression coefficients may be consistent even when the joint distribution is misspecified. Efficiency and robustness issues will be considered through a simulation study and illustrated using the Diabetic Retinopathy Study (Huster et al., 1989).

Les données de temps de panne multidimensionnelles surgissent fréquemment dans expériences scientifiques. L'intérêt se situe principalement dans les cas où les temps de panne dans un même

regroupement sont corrélées. Dans plusieurs modèles, les logarithmes des durées de vie ont des distributions marginales position-échelle. Nous étudions l'estimation des coefficients de régression pour des modèles bidimensionnelles de position-échelle et nous montrons que les estimateurs du maximum de vraisemblance pour les coefficients de régression peuvent être convergents même lorsque la distribution conjointe est mal spécifiée. L'efficacité et la robustesse seront considérées par des études de simulations et illustrées en utilisant l'étude sur la rétinophatie causée par le diabète (Huster et al., 1989).

Wednesday, May 29th/Mercredi 29 mai, 9:15 **KTH B135**

A Class of Estimators for the Parameters in the Location-Scale Model with Censored Data

Une classe d'estimateurs pour les paramètres dans un modèle de position-échelle des données censurées

Xuewen Lu, Agriculture and Agri-Food Canada et/and Radhey Singh, University of Guelph

This paper considers a class of estimators of the location and scale parameters in the location-scale model based on unbiased data transformation when the observations are randomly censored on the right. The asymptotic normality of the estimators is established using counting process and martingale techniques when the censoring distribution is known and unknown respectively. In the case that the censoring distribution is unknown, we show that it can be estimated from the Product-Limit estimator and this class of estimators has the same asymptotic variance, which is the lower bound of variances of this class of estimators when the censoring distribution is given. Therefore, the proposed estimators, compared with the corresponding maximum likelihood estimators, provide a simpler procedure while maintaining high asymptotic relative efficiencies. An application is presented to illustrate the proposed method.

Cette présentation considère une classe d'estimateurs de position et d'échelle dans un modèle basé sur la transformation non biaisée des données quand les observations sont censurées aléatoirement à droite. La normalité asymptotique des estimateurs est établie par l'utilisation des processus de comptage quand la distribution de la censure est connue et des techniques impliquant des martingales quand cette distribution est inconnue. Dans le cas où la distribution de la censure est inconnue, nous montrons que nous pouvons l'estimer par l'estimateur du produit-limite. Cette classe d'estimateurs a la même variance asymptotique, qui est une borne inférieure des variances des estimateurs de cette classe d'estimateurs quand la distribution de la censure est donnée. Par conséquent, les estimateurs proposés, comparés aux estimateurs correspondants obtenus par la méthode du maximum de vraisemblance, fournissent un procédé simple tout en maintenant un bon niveau de convergence asymptotique. Une application est présentée pour illustrer la méthode proposée.

Wednesday, May 29th/Mercredi 29 mai, 9:30 **KTH B135**

Smooth and aggregated incidence rate estimation for interval-censored HIV data

Estimation lisse et agrégée du taux d'incidence du VIH pour des données censurées par intervalle

Thierry Duchesne, James E. Stafford, et/and Fasil Woldegeorgis, University of Toronto

We consider the problem of incidence rate estimation when data are interval-censored. We compare an aggregated incidence rate estimator based on a uniform distribution assumption over censoring intervals that is widely used in HIV epidemiology to a natural extension of the kernel intensity function estimator of Ramlau-Hansen (Ann. Stat., 1983) so that it can handle interval censoring. We show

how the former estimator relates to the latter and we investigate the properties of both estimators analytically and through simulation.

Nous considérons l'estimation du taux d'incidence lorsque les données sont censurées par intervalle. Nous comparons un estimateur agrégé de l'incidence, qui suppose une distribution uniforme des données dans leurs intervalles de censure et qui est fréquemment utilisé dans les études épidémiologiques sur le SIDA, à un estimateur basé sur une extension naturelle de l'estimateur du noyau de la fonction d'intensité de Ramlau-Hansen (Ann. Stat., 1983) afin de permettre l'estimation avec données censurées par intervalle. Nous démontrons comment le premier estimateur est lié au second et nous étudions les propriétés des deux estimateurs de façon analytique et par simulations.

Wednesday, May 29th/Mercredi 29 mai, 9:45 **KTH B135**

Some over-dispersed life time models and associated tests
Quelques modèles de temps de vie surdispersés et tests correspondants
Sudhir Paul, University of Windsor et/and Uditha Balasoorya, Nayang Technological University

The one parameter exponential distribution is useful for modeling lifetime data. This distribution belongs to the exponential family of distributions. The one parameter exponential distribution is a special case of a more richer family of distributions, such as the Pareto distribution or the generalized Pareto distribution, the gamma distribution and the Weibull distribution all of which are two parameter distributions and can be expressed as a family of over-disperesed exponential models. The purpose of this paper is to develop tests of goodness of fit of the exponential model against the over-dispersion family of distributions.

La distribution exponentielle à un seul paramètre est utile pour modéliser des données reliées à des temps de vie. Cette distribution appartient à la famille des distributions exponentielles. La distribution exponentielle à un paramètre est un cas spécial d'une famille plus riche de distributions, telles que la distribution de Pareto, la distribution de Pareto généralisée, la distribution gamma et la distribution de Weibull qui sont deux distributions à deux paramètres et peuvent être exprimées comme une famille de modèles exponentiels surdispersés. Le but de cette présentation est de développer des tests d'ajustement pour le modèle exponentiel contre la famille des distributions surdispersées.

Session 36: Robust Methods, Outlier Detection and Bootstrapping/Méthodes robustes, détection de valeurs aberrantes et rééchantillonnage.

Wednesday, May 29th/Mercredi 29 mai, 8:30 **TSH B106**

Robust tests for independence of two time series
Tests robustes d'indépendance entre deux séries chronologiques
Pierre Duchesne et/and Roch Roy, HEC-Montréal

This paper aims at developing a robust and omnibus procedure for checking the independence of two time series. Li and Hui (1994) proposed a robustified version of Haugh's (1976) classic portmanteau statistic which is based on a fixed number of lagged residual cross-correlations. In order to obtain a consistent test for independence against an alternative of serial cross-correlation of an arbitrary form between the two series, Hong (1996) introduced a class of statistics that take into account all possible lags. The test statistic is a weighted sum of residual cross-correlations and the weighting is

determined by a kernel function. With the truncated uniform kernel, we retrieve a normalized version of Haugh's statistic. However, several kernels lead to a greater power. Here, we introduce a robustified version of Hong's statistic. We suppose that for each series, the true ARMA model is estimated by a consistent robust method and the robust cross-correlation is so obtained. Under the null hypothesis of independence, we show that the robust statistic asymptotically follows a N(0,1) distribution. Using a result of Li and Hui, we also propose a robust procedure for checking independence at individual lags and a descriptive causality analysis in the Granger's sense is discussed. The level and power of the robust version of Hong's statistic are studied by simulation in finite samples. Finally, the proposed robust procedures are applied to a set of financial data.

L'article développe une procédure robuste et omnibus for vérifier l'indépendance de deux séries chronologiques. Li et Hui (1994) ont proposé une version robustifiée de la statistique classique de Haugh (1976), qui est basée sur un nombre fixé de corrélations croisées résiduelles. Afin d'obtenir un test convergent pour tester l'indépendance contre une alternative de dépendance croisée de forme arbitraire entre deux séries, Hong (1996) a introduit une classe de statistiques qui peuvent prendre en compte l'ensemble des délais. La statistique de test est une somme pondérée de corrélations croisées résiduelles et la pondération est déterminée à l'aide d'un noyau. Avec le noyau uniforme tronqué, nous retrouvons une version normalisée de la statistique de Haugh. Cependant, plusieurs noyaux permettent d'obtenir une meilleure puissance. Ici, nous introduisons une version robustifiée de la statistique de Hong. Nous supposons que pour chaque série chronologique le bon modèle ARM! A est estimé par une méthode robuste et convergente et la fonction de corrélation croisée est obtenue. Sous l'hypothèse nulle d'indépendance, nous montrons que la statistique robuste proposée converge en loi vers une loi normale centrée réduite. Utilisant un résultat de Li et Hui, nous proposons également une procédure robuste pour vérifier l'indépendance à des délais individuels et une analyse descriptive de causalité dans le sens de Granger est discutée. Le niveau et la puissance de la version robuste du test de Hong sont étudiés par simulations. Finalement, les méthodes robustes proposées sont appliquées à des données financières.

Wednesday, May 29th/Mercredi 29 mai, 8:45 **TSH B106**

Bootstrap adaptive least trimmed squares
Moindres carrés tronqués adaptatifs par rééchantillonnage

Jean-François Boudreau et/and Christian Léger, Université de Montréal

The least trimmed squares family of estimators is indexed by trimming proportion α. A large value of α (maximum 50% will give a very robust estimator that could resist to a large percentage of outliers). But smaller value of α might be advantageous by decreasing the variance of the estimator provided that bias is not induced by including an outlier.

We propose a technique based on resampling to adaptively estimate the appropriate trimming proportion for a data set. Performance of the technique is illustrated on simple linear regression models. Data are generated from a straight line for the correct data, and a ball far from the line for the outliers.

In addition to simulation results, different aspects of the proposed technique will be discussed. In particular, we show why it is preferable to use pairs resampling instead of residuals resampling, how to adapt the method whether the number of outliers is fixed or random, and will discuss the criterion used to compare estimators.

Les moindres carrés tronqués forment une famille d'estimateurs de régression indicée par le pourcentage de troncature α. Une valeur de α élevée (au maximum 50% donnera un estimateur très robuste pouvant résister à une grande proportion de données aberrantes. Mais il y a avantage à utiliser une valeur de α plus petite, ce qui diminue la variance de l'estimateur, à condition de ne pas créer un biais en obligeant l'estimateur à inclure une donnée aberrante.

Nous proposons une technique basée sur le rééchantillonnage pour estimer de façon adaptative le pourcentage de troncature approprié pour le jeu de données à l'étude. La performance de la technique est illustrée pour des modèles de régression linéaire simple. Les données sont générées à partir d'une droite, et d'une boule éloignée de la droite pour les données aberrantes.

En plus des résultats des simulations, différents aspects de la technique retenue seront discutés. Nous verrons en particulier pourquoi il est préférable d'utiliser le rééchantillonnage par paires plutôt que par résidus, comment adapter la méthode selon que l'on considère le nombre de valeurs aberrantes comme fixe ou aléatoire, et les critères utilisés pour comparer les estimateurs entre eux.

Wednesday, May 29th/Mercredi 29 mai, 9:00 **TSH B106**

Outliers in large data sets
Valeurs aberrantes dans de grands jeux de données

Sharmila Banerjee et/and Boris Iglewicz, Temple University

Procedures for identifying outliers are among the oldest investigated statistical tools. A great variety of statistical outlier identification techniques have been proposed and several outlier-generating models were introduced. These methods were designed for relatively small data sets and required knowledge as to the number of outliers and where they were located. The majority of these methods were designed for observations coming from the normal distribution, with entirely different methodology typically used for other distributions. Approaches for handling outliers include accommodation, testing for discordant outliers, outlier labeling, and construction of outlier identification intervals. This extensive work is nicely summarized in Barnett and Lewis (1994). Tukey (1977) introduced a graphical procedure for summarizing batches of univariate data. That procedure contained a popular rule for flagging outliers. Hoaglin and Iglewicz (1987) converted this rule into a proper outlier identification procedure for normally distributed observations. A large sample version of this boxplot outlier identification procedure is studied here. The modified rule can be easily used for a variety of distributions, but still using the same basic formulas, one for symmetric distributions and another for skewed distributions. This large sample outlier detection rule is timely as large data sets are presently commonly encountered in practice, sometimes involving millions of observations. The popular normal, t, gamma, and Weibull distributions are studied in some detail. Additionally, this modified boxplot procedure does not require prior knowledge as to the number nor the location of outliers. Simple adjustments are used for smaller data sets. The probability properties and critical values are derived mathematically and also obtained through simulation. This studied boxplot rule is compared with several alternative procedures. The investigated methodology is illustrated with the identification of outliers in two relatively large data sets, one consisting of about 5,000 patients in a health study, while the second dealing with approximately 15,000 oceanographic observations.

Les procédures pour identifier les données aberrantes sont parmi les plus anciens outils statistiques étudiés. On a proposé une grande variété de techniques statistiques pour l'identification des données aberrantes et plusieurs modèles pour générer des données aberrantes ont été présentés. Ces méthodes ont été conçues pour des jeux de données relativement petits et exigent la connaissance du nombre de données aberrantes et où elles étaient localisées. La majorité de ces méthodes ont été conçues pour des observations venant de la distribution normale, et une méthodologie totalement différente est utilisée pour d'autres distributions. Les approches pour manipuler les données aberrantes incluent l'accomodation, la détermination des données aberrantes discordantes, les étiquettes des données aberrantes, et la construction d'intervalles d'identification des données aberrantes. Ce travail est bien résumé dans Barnett et Lewis (1994). Tukey (1977) a présenté une procédure graphique pour résumer des séries de données univariés. Ce procédé contient une règle populaire pour l'indentification

des données aberrantes. Hoaglin et Iglewicz (1987) ont converti cette règle en un procédé approprié d'identification des données aberrantes pour des observations normalement distribuées. Une version de ce procédé d'identification basé sur un diagramme en boîte pour les grands échantillons est étudiée ici. La règle modifiée peut être facilement utilisée pour une variété de distributions, mais utilisant toujours les mêmes formules de base, une pour des distributions symétriques et une autres pour des distributions asymétriques. Cette règle de détection des données aberrantes pour de grands échantillons est opportune car de grands jeux de données sont produits généralement en pratique, impliquant parfois des millions d'observations. La loi normale, la loi t, la gamma, et les distributions de Weibull sont étudiés de manière assez détaillée. En plus, ce procédé modifié pour le diagramme en boîte n'exige pas de connaissance antérieure quant au nombre ni à l'emplacement des données aberrantes. Des ajustements simples sont utilisés pour de plus petits échantillons. Les propriétés des probabilités et les valeurs critiques sont dérivées mathématiquement et également obtenues par la simulation. Cette règle basée sur le diagramme en boîte est comparée à plusieurs variantes. La méthodologie étudiée est illustrée avec l'identification des données aberrantes dans deux jeux de données relativement grands, un premier se composant d'environ 5 000 patients dans une étude médicale, alors que le second traite approximativement 15 000 observations océanographiques.

Wednesday, May 29th/Mercredi 29 mai, 9:15 **TSH B106**

Bootstrapping simultaneous prediction intervals for autoregressive time series models
Bootstrap simultané d'intervalles de prédiction pour des modèles de séries chronologiques autorégressives

Ka Ho Wu et/and Siu Hung Cheung, Chinese University of Hong Kong

Multiple forecasting is extremely useful in areas such as business and economics. In many circumstances, instead of a single forecast, simultaneous prediction intervals for multiple forecasts are necessary. For example, based on past monthly sales records, a production manager is interested to obtain monthly sales forecasts for the coming year. These forecasts are important for inventory and production planning. For Gaussian autoregressive processes, several procedures for obtaining simultaneous prediction intervals have been proposed in the literature. These methods assume a normal error distribution and can be adversely affected by departures from normality. In this paper, we propose bootstrap methods for constructing simultaneous prediction intervals for autoregressive processes. To understand the mechanisms and characteristics of the proposed bootstrap procedures, several time series are selected for illustrative purposes. The major ideas discussed in this paper with autoregressive processes can be generalized to other more complicated time series models.

Les prévisions multiples sont extrêmement utiles dans les domaines tels que les affaires et les sciences économiques. En plusieurs circonstances, plutôt que d'effectuer une prévision simple, des intervalles de prévision simultanés pour des prévisions multiples sont nécessaires. Par exemple, basé sur des données mensuels de ventes antérieures, un gestionnaire de production est intéressé à obtenir des prévisions mensuelles de ventes pour la prochaine année. Ces prévisions sont importantes pour l'inventaire et la planification de la production. Pour des processus autorégressifs gaussiens, on a proposé dans la littérature plusieurs procédures pour obtenir des intervalles de prévision simultanés. Ces méthodes assument une distribution normale des erreurs et peuvent être compromises par des écarts à l'hypothèse de normalité. Dans cette présentation, nous proposons des méthodes bootstrap pour construire des intervalles simultanés de prévision pour des processus autorégressifs. Pour comprendre les mécanismes et les caractéristiques des procédures de bootstrap proposées, plusieurs séries chronologiques sont choisies pour l'illustrer. Les idées principales discutées dans cette présentation avec des processus autorégressifs peuvent être généralisées à d'autres modèles plus compliqués de séries chronologiques.

Wednesday, May 29th/Mercredi 29 mai, 9:30 **TSH B106**

Bootstrap confidence intervals for periodic replacement policies
Intervalles de confiance bootstrap pour les remplacements préventifs périodiques
Pascal Croteau, Robert Cléroux et/and Christian Léger, Université de Montréal

Preventive replacement policies are important in reliability. Their ability to reduce costs make them popular to industrial companies. A piece of equipment is either replaced preventively according to the policy or at failure, depending of which occurs first. When the lifetime distribution F is known, we can compute the cost function for a preventive replacement at time T. The optimal replacement time is therefore the argument which minimizes this cost function. When F is unknown, several estimators of this function exist and one is now able to obtain point estimators of the optimal replacement time and of the minimal cost. We now propose confidence intervals based on four estimators. We will use bootstrap methods to obtain such intervals, more specifically basic and percentile intervals and will check via simulations whether they have the right asymptotic coverage probabilities. Léger & Cléroux (1992) have shown theoretically and via simulations that in the case of age replacement policy, bootstrap intervals have asymptotically the right coverage probabilities only in the case of the minimal cost. We consider the case of periodic replacement policy for both the optimal cost and the optimal replacement time.

Le principe du remplacement préventif est important en fiabilité et trouve couramment application en industrie puisqu'il permet d'instaurer une politique de remplacement optimale réduisant les coûts du système. Une pièce est donc remplacée soit selon la politique, soit à la panne, selon ce qui arrive en premier. Lorsque la distribution F du temps de vie d'une pièce est connue, on peut calculer la fonction de coût pour un remplacement préventif au temps T. Le temps optimal de remplacement est celui qui minimise cette fonction de coût. Lorsque F est inconnue, il existe des estimateurs de la fonction de coût et on peut donc calculer des estimateurs ponctuels du temps optimal et du coût minimal. Nous proposons maintenant de calculer des intervalles de confiance basés sur quatre de ces estimateurs. Nous allons utiliser le bootstrap et construire des intervalles de type percentile et de base et vérifier via simulations si les probabilités de couverture sont asymptotiquement les bonnes. Léger & Cléroux (1992) ont obtenu des résultats sur le remplacement préventif de type âge et ont démontré théoriquement et via simulations que le bootstrap donnait des intervalles asymptotiquement de bon niveau pour le coût minimal uniquement. Pour notre part, nous considérerons le cas du remplacement préventif périodique et construirons des intervalles de confiance pour le coût minimal et le temps optimal.

Session 37: Financial Modeling/Modélisation financière

Wednesday, May 29th/Mercredi 29 mai 10:30 **TSH B105**

Maximum likelihood estimation of credit risk models
L'estimation par la méthode du maximum de vraisemblance pour des modèles de risque de crédit
Genevieve Gauthier, l'École des Hautes Études Commerciales, Jin-Chuan Duan, University of Toronto, Jean-Guy Simonato et/and Sophia Zaanoun, l'École des Hautes Études Commerciales

One of the limitations associated with the use of Merton's (1974) credit risk model is the necessity to obtain estimates for the asset values and the variance of their returns. This paper examines a solution to this problem with a maximum likelihood approach. The estimator is developped in a

multivariate framework and allows the assessment of credit risk in a portfolio context. Besides the usual benefits associated with maximum likelihood, it is shown here that an advantage of the approach is the possibility to obtain an estimate of the default probability under the data generating measure. Such an estimation is usually not possible within the context of the implicit estimation approaches often adopted by academics and market participants. A Monte Carlo study is conducted to examine the performance of the implicit and maximum likelihood estimators in finite samples. The results show that the implicit parameter estimates tend to be biased for large debt to asset value ratio while the usual properties of the maximum likelihood estimator hold well in finite samples of moderate sizes. A small set of Canadian firms is used to illustrate the methodology. We extend the results to other credit risk model such as the Longstaff-Schwartz model.

Une des limitations associées à l'utilisation des modèles de risque de crédit de Merton (1974) est la nécessité d'obtenir des estimations pour les valeurs des actifs et la variance de leurs retours. Cette présentation examine une solution à ce problème avec une approche par le maximum de vraisemblance. L'estimateur est développé dans un cadre multivarié et permet l'estimation du risque de crédit dans un contexte de portfolio. Mis à part les avantages habituels associés au maximum de vraisemblance, on montre ici qu'un avantage de l'approche est la possibilité d'obtenir une estimation de la probabilité de défaut sous la mesure qui a générée les données.

Une telle estimation n'est habituellement pas possible dans le contexte des approches implicites d'estimation souvent adoptées par des chercheurs autant du cadre académique et de léntreprise. Une étude de Monte-Carlo est entreprise pour examiner la performance des estimateurs implicite et du maximum de vraisemblance dans des échantillons finis. Les résultats prouvent que les estimations implicites des paramètres tendent à être biaisées pour de grandes dettes au taux de valeur des actifs tandis que les propriétés habituelles de l'estimateur de maximum de vraisemblance se comportent bien dans des échantillons finis de tailles modérées. Un petit ensemble de sociétés canadiennes est employé pour illustrer la méthodologie. Nous étendons les résultats à d'autre modèle de risque de crédit tel que le modèle de Longstaff-Schwartz.

Wednesday, May 29th/Mercredi 29 mai, 11:15 **TSH B105**

Pricing derivatives in incomplete markets
Les prix dérivés dans des marchés incomplets

Tom Hurd, McMaster University

Arbitrage pricing theory becomes more demanding and interesting when the utopian assumptions of the Black–Scholes model and its cousins are relaxed. In particular, throwing away the idea of replicating portfolios leads to incomplete market models, in which real world nitty-gritty again becomes critical. In this talk I will discuss the general picture of incomplete markets with attention to how key concepts can be generalized. Some specific models will illustrate these points. Then I will discuss several complementary points of view about pricing contingent claims in incomplete markets: the risk-neutral measure, implied estimation (statistical inference of parameters from current option data), pricing via optimal portfolio theory and dynamics of the state price density.

Le modèle d'évaluation des prix par arbitrage devient plus exigeant et intéressant lorsque les hypothèses utopiques des modèles de Black-Scholes et ses dérivés sont plus ou moins respectées. En particulier, ne pas considérer l'idée de répéter les portefeuilles mène aux modèles de marché incomplet, dans lesquels la réalité devient critique. Dans cette présentation, je présenterai un tableau d'ensemble des marchés incomplets en portant une attention particulière à la façon dont des concepts clés peuvent être généralisés. Quelques modèles spécifiques illustreront ces points. Je discuterai ensuite de plusieurs points de vue complémentaires au sujet des réclamations contingentes relatives au prix sur les marchés

incomplets : la mesure risque neutre, l'estimation implicite (inférence statistique des paramètres à partir des données actuelles), estimation du prix par l'intermédiaire de la théorie optimale du portefeuille et la dynamique de la densité des prix fixés par l'état.

Session 38: Environmetrics II/Mésométrie II

Wednesday, May 29th/Mercredi 29 mai, 10:30 **KTH B135**

Nearly nonparametric multivariate density estimates that incorporate marginal parametric density information for the environment
Estimations de densités multivariées non paramétriques approximatives qui incluent de l'information sur la densité marginale paramétrique dans l'environnement

Cliff Spiegelman et/and Eun Sug Park, Texas A&M University

When data analysts have multivariate data often they have partial knowledge about the form of the marginal densities, but frequently they have little information about the bivariate and higher dimensional densities. This talk provides nonparametric estimators that nearly equal the MLE estimates for the marginal densities while being close to the kernel nonparametric density estimates for the joint density estimates. The motivation for this talk came from recollections of a 15 year old conversation with Ingram Olkin where the problem at hand was how to model multivariate data with fixed marginals yet having a flexible and rich multivariate structure. We apply our procedure to environmental data sets and show that while the marginal distributions of many VOC's are well modeled by lognormal distributions their joint distributions are not easily parametrically modeled.

Quand les analystes statistiques ont des données multidimensionnelles, souvent ils ont une connaissance partielle de la forme des densités marginales, mais fréquemment ils ont peu d'information sur les densités bidimensionnelles et de dimensions plus élevées. Cette présentation fournit les estimateurs non paramétriques qui égalent presque les estimateurs du maximum de vraisemblance des densités marginales tout en étant près des estimateurs par la méthode du noyau non paramétrique de la densité conjointe. La motivation de ce travail est venue des souvenirs d'une conversation tenue il y a 15 ans avec Ingram Olkin où le problème en vogue était de modéliser des données multidimensionnelles avec des marginales fixes tout en ayant une structure multidimensionnelle flexible et riche. Nous appliquons notre procédé à un jeu de données sur l'environnement et prouvons que tandis que les distributions marginales de plusieurs VOC sont bien modélisées par des distributions lognormales, leur distribution conjointe n'est pas facilement modélisée de façon paramétrique.

Wednesday, May 29th/Mercredi 29 mai, 11:00 **KTH B135**

Nonlinear spatio-temporal statistics via Monte Carlo methods implemented in a Javaspaces distributed computer
Statistiques spatio-temporels non linéaires via des méthodes de Monte Carlo implémentées dans des ordinateurs avec JavaSpaces

Timothy Haas, University of Wisconsin at Milwaukee

Coarse grained parallel computing has the potential for allowing a variety of computationally intensive spatio-temporal statistical calculations to be performed by anyone with access to a network of 50 or more PCs. These calculations include robust estimation of nonlinear spatio-temporal trend models, Monte Carlo assessment of model goodness-of-fit and parameter estimate reliability, optimal prediction of a spatio-temporal random field at many spatio-temporal locations under asymmetric loss,

and the construction of and high speed access to a distributed spatio-temporal statistical database. A distributed computer that performs these calculations can be constructed through a transaction space protocol using either JavaSpaces from Sun Microsystems or TSpaces from IBM. As an example, a nonlinear spatio-temporal trend model is estimated with Minimum Distance (a robust statistical parameter estimator) followed by a Monte Carlo computation of parameter estimate standard errors with a JavaSpaces program running on the 50 PCs contained in an instructional computer laboratory during hours that the laboratory is closed.

Le calcul parallèle à grains bruts offre des possibilités intéressantes pour permettre une variété de calculs statistiques spatio-temporels intensifs par ordinateur pour être exécuté par n'importe qui avec l'accès à un réseau de 50 PC ou plus. Ces calculs incluent l'estimation robuste des modèles spatio-temporels non linéaires avec une tendance, l'évaluation du modèle d'ajustement par Monte-Carlo et la fiabilité de l'estimation des paramètres, la prévision optimale d'une zone aléatoire spatio-temporelle à plusieurs emplacements spatio-temporels sous un perte asymétrique, et la construction d'un accès à grande vitesse à une base de données statistique spatio-temporelle. Un ordinateur qui exécute ces calculs peut être construit par un protocole de l'espace de transaction en utilisant soit JavaSpaces sur un système Sun ou TSpaces sur IBM.

Comme exemple, un modèle spatio-temporel non linéaire avec tendance est estimé par la méthode de la distance minimale (un estimateur statistique robuste pour les paramètres) suivi d'un calcul de Monte-Carlo des écarts types des estimations des paramètres avec JavaSpaces fonctionnant sur les 50 PC contenus dans un laboratoire informatique durant les heures où le laboratoire est fermé.

Wednesday, May 29th/Mercredi 29 mai, 11:30 **KTH B135**

Bayesian analysis of nonstationary spatial covariance
Analyse bayésienne de la covariance spatiale non stationnaire
Peter Guttorp, University of Washington, Doris Damian, Worcester Polytechnic Institute et/and Paul D. Sampson, University of Washington

A fully Bayesian analysis of spatial air quality data, allowing for nonhomogeneity in both variance and covariance structure, is introduced. The method is based on the deformation approach by Sampson and Guttorp. We illustrate the level of uncertainty in the deformation as well as in the estimation of covariance using data on precipitation and air quality.

Une analyse complètement bayésienne est présentée pour des données sur la qualité de l'air dans l'espace, tenant compte de la non-homogénéité de la variance et de la structure de covariance. La méthode est basée sur l'approche par la déformation introduite par Sampson et Guttorp. Nous illustrons le niveau d'incertitude dans la déformation aussi bien que dans l'estimation de la covariance en utilisant des données sur la qualité de l'air et sur les précipitations.

Session 40: New research findings in analysis methods for survey data/Nouveaux résultats de recherche dans les méthodes d'analyse pour les donnés d'enquête

Wednesday, May 29th/Mercredi 29 mai, 10:30 **TSH B106**

Proportional hazards models and ignorable sampling
Modèles à risques proportionnels et plans d'échantillonnage non informatifs
Christian Boudreau et/and Jerry Lawless, University of Waterloo

A considerable amount of event history data is collected through longitudinal surveys. The goal is to understand different events that individuals experience over time; for example, marriage, divorce, unemployment, etc. Data collected through longitudinal surveys involve the use of complex survey designs with clustering and stratification. Statistical analysis of such data must account for intra-cluster dependence. When the goal is analytical inference for duration or survival time variables and the sampling design is ignorable, marginal proportional hazards methodology can account for stratification and for intra-cluster correlation. Variance estimation is achieved by using the theory of estimating equations.

In this talk, we propose variance estimation methods for the estimator of regression coefficients in the Cox model and for the Breslow-Aalen estimator that account for intra-cluster dependence. We prove that, under mild conditions, this estimator of regression coefficients is consistent and converges weakly to a mean-zero Gaussian process. The asymptotic properties of the Breslow-Aalen estimator for the baseline cumulative hazard function are also studied in the presence of intra-cluster correlation and it is shown that it converges weakly to a zero-mean Gaussian process. The proposed methods are illustrated using the Survey of Labour and Income Dynamics (SLID).

Les enquêtes longitudinales sont une source importante de données pour l'analyse historique des événements. Le but est de comprendre les différents événements que les individus vivent au fil du temps ; par exemple, mariage, divorce, chômage, etc. Les enquêtes longitudinales ont généralement recourt à des plans d'échantillonnage complexes, tels que l'échantillonnage en grappes et la stratification. Cependant, la dépendance intra-grappes, c'est-à-dire entre les observations provenant de la même grappe, doit être prise en considération lors de l'inférence statistique. Ces enquêtes ont plusieurs buts, certains sont analytiques, d'autres descriptifs. Nous portons notre attention sur celles dont le but est analytique et où le plan d'échantillonnage est non informatif. Dans ce cas, les modèles marginaux à risques proportionnels peuvent accommoder la stratification et la corrélation intra-grappes. De plus, la théorie des équations d'estimation permet d'obtenir des estimateurs de la variance prenant en compte cette corrélation.

Dans cette présentation, nous proposons des méthodes pour estimer la variance de l'estimateur des coefficients de régression du modèle de Cox et de l'estimateur de Breslow-Aalen. Ces méthodes tiennent compte de la corrélation intra-grappes. Nous démontrons, sous des hypothèses peu contraignantes, que l'estimateur des coefficients de régression est un estimateur convergent et qu'il converge en loi vers un processus gaussien de moyenne zéro. Les propriétés asymptotiques de l'estimateur de Breslow-Aalen pour la fonction cumulative du taux de base sont également étudiées en présence de la corrélation intra-grappes. Il est démontré que cet estimateur converge en loi vers un processus gaussien de moyenne zéro. Les méthodes proposées sont illustrées par le biais de l'Enquête sur la dynamique du travail et du revenu (EDTR) de Statistique Canada.

Wednesday, May 29th/Mercredi 29 mai, 10:55 **TSH B106**

Analysis of longitudinal survey data in the presence of missing outcomes
Analyse de données de sondage longitudinal en présence de réponses manquantes

Brajendra Sutradhar, Memorial University et/and Milorad Kovacevic, Statistics Canada/Statistique Canada

In longitudinal survey data, outcomes that are repeatedly measured over time may be correlated and some may be missing. One of the main goals of the longitudinal survey is to describe the marginal expectation of the response variable as a function of the associated covariates while accounting for the longitudinal correlations and the missingness nature of the data. In this talk, we discuss a generalized estimating equations (GEE) approach for the analysis of the longitudinal survey data under both cases

whether the missing data are imputed or not. The regression estimators are shown to be consistent as well as efficient. The GEE approach will be illustrated by using the Survey of Labour and Income Dynamics (SLID) data from Statistics Canada.

Dans des données d'études longitudinales, les résultats qui sont mesurés à plusieurs reprises dans le temps peuvent être corrélés et certains peuvent être manquants. Un des buts principaux de l'étude longitudinale est de décrire l'espérance marginale de la variable réponse en fonction des covariables associés en tenant compte des corrélations longitudinales et de la nature des données manquantes. Dans cette présentation, nous discutons de l'approche par les équations d'estimation généralisées(GEE) pour l'analyse des données d'étude longitudinale sous deux hypothèses, que les données manquantes soient imputées ou non. Les estimateurs de régression s'avèrent convergents et efficaces. L'approche par les GEE sera illustrée en utilisant l'étude des données sur les dynamiques du travail et du revenu (SLID) de Statistiques Canada.

Wednesday, May 29th/Mercredi 29 mai, 11:20 **TSH B106**

Applying item response theory methods to complex survey data
Application de la théorie des éléments de réponse aux données d'enquête complexe

Roland Thomas, Carleton University et/and Andre Cyr, Statistics Canada/Statistique Canada

Item response theory (IRT) offers many advantages to researchers who need to quantify children's reading and writing abilities, and for this reason, IRT methods have been adopted in Statistics Canada's National Longitudinal Survey for Children and Youth. IRT methods have a long history in the field of psychometrics, and provide a model based method for characterizing both test items and subject abilities, and for generating predictions of individual abilities. For the most part, IRT methods implicitly assume i.i.d. observations, so that the application of these methods to complex surveys raises a number of issues. Questions arise as to the use or otherwise of the sample weights, appropriate methods for predicting individual abilities (for inclusion on public use datasets), and appropriate methods for estimating parameters and their variances.

This talk will provide an overview of IRT theory and its application in the NLSCY, together with a brief review of the alternative IRT implementations used in some other large surveys. Specific empirical results based on NLSCY data will be provided, including: (1) the potential for biases due to ignoring survey weights; (2) biases in the distribution of ability predictors, and the dependence of this bias on test length. Issues requiring further research will be identified.

La théorie des éléments de réponse (IRT) offre beaucoup d'avantages aux chercheurs qui doivent mesurer les capacités de lecture et d'écriture des enfants, et pour cette raison, les méthodes relatives à IRT ont été adoptées dans l'Enquête longitudinale sur les enfants et la jeunesse de Statistique Canada (NLSCY). Les méthodes d'IRT ont une longue histoire dans le domaine de la psychométrie, et fournissent une méthode basée sur un modèle pour caractériser des éléments de test et des capacités du sujet à l'étude, et pour produire des prévisions concernant les capacités individuelles. La plupart du temps, les méthodes d'IRT assument implicitement des observations i.i.d., de sorte que l'application de ces méthodes aux enquêtes complexes soulève un certain nombre de questions. Des questions se posent quant à l'utilisation de poids échantillonnaux, quant aux méthodes appropriées pour prévoir différentes capacités individuelles (pour l'inclusion dans l'ensemble des données publiques), et des méthodes appropriées pour estimer des paramètres et leurs variances.

Cette présentation fournira une vue d'ensemble de la théorie IRT et de son application dans le NLSCY, ainsi qu'un bref examen des réalisations des méthodes IRT alternatives utilisées dans quelques grandes études. Des résultats empiriques spécifiques basés sur les données du NLSCY seront fournis,

incluant: (1) le potentiel de biais dû à la non-utilisation des poids échantillonnaux; (2) le biais dans la distribution des prédicteurs de capacités, et la dépendance de ce biais à l'égard de la durée du test. Des questions exigeant davantage de recherche seront identifiées.

Session 41: Statistical Indexes/Index statistiques

Wednesday, May 29th/Mercredi 29 mai, 13:30 **TSH B106**

Development of the Labour Cost Index at Statistics Canada
Développement de l'Indice des coûts de main-d'oeuvre à Statistique Canada

Lenka Mach et/and Abdelnasser Saïdi, Statistics Canada/Statistique Canada

Statistics Canada is currently working on development of a new index, the Labour Cost Index (LCI), to respond to the need for a measure of change in the total labour cost in Canada. LCI is supposed to measure the rate of change in the cost of one unit of labour, equivalent to one hour of labour, where the cost of labour should include both wages and salaries as well as non-wage benefits. This measure should not be affected by shifts in either the occupational or the industrial mix of employment and therefore it should be a Laspeyres type of index. A similar index is being produced in a few other countries, for example in the USA and New Zealand.

In the development of the LCI methodology, we can borrow a lot from the methodology of price indices and learn from the experiences with CPI and other price indices. However, there are some fundamental differences between the LCI and a price index that will lead to differences in the methodology. For example, while good expenditure data for the different commodity groups usually exist for the calculation of the economic weights, the total cost data for occupational groups are not readily available. Another major challenge is to measure the total labour cost for an hour of labour for each employee in the sample since there are many components in the total cost of labour.

In this presentation, we will describe the similarities and differences between the LCI and the price indices as well as the proposed LCI methodology. We will discuss the construction of the fixed basket of occupations as well as the derivation of the total cost and cost relatives for a unit of labour. The LCI estimation, including variance estimation, will also be discussed. As Statistics Canada must minimise the respondent burden, the employees will be rotated out of the LCI sample every two or three years. This certainly represents yet another challenge since indices are usually calculated only using the units common to the two subsequent samples. Some plans to overcome this difficulty will be discussed.

Statistique Canada développe actuellement un nouvel indice, l'indice des coûts de main-d'œuvre (ICM) pour répondre au besoin de mesure de la variation du coût total du travail au Canada. L'ICM est supposé mesurer le taux de variation du coût total d'une unité de main-d'œuvre, égale à une heure de travail. Le coût d'une heure de travail devrait inclure aussi bien les gains et les salaires que les avantages non salariaux. Cette mesure devrait être non affectée par des changements d'emploi tant au niveau de l'industrie ou de la profession et doit donc être un indice de type Laspeyres. Un indice similaire est déjà calculé dans quelques pays comme les États-Unis et la Nouvelle-Zélande.

Dans le développement de la méthodologie de L'ICM, nous pouvons nous inspirer beaucoup de la méthodologie des indices des prix et bénéficier des expériences du calcul de l'indice IPC et d'autres indices de prix. Cependant, il existe des différences fondamentales entre l'indice ICM et les indices de prix qui entraînent des différences dans la méthodologie de calcul de l'ICM. Par exemple, alors que les données de dépenses de consommation pour différents groupes de biens existent habituellement pour le calcul des poids économiques, les données de coût total pour différents groupes d'occupations ne sont pas facilement disponibles. Un autre élément essentiel de ce défi est de mesurer le coût total

d'une heure de main-d'œuvre pour chaque employé de l'échantillon car il existe plusieurs composantes du coût total.

Dans cette présentation, nous décrirons les similarités et différences entre l'ICM et les indices de prix ainsi que la méthodologie de calcul de l'ICM. Nous discuterons des problèmes posés par la construction d'un panier fixe d'occupations et du calcul des coûts totaux et relatifs d'une unité de travail. L'estimation de l'indice et de sa variance seront aussi discutés. Comme Statistique Canada doit minimiser le fardeau de réponse, les employés devront sortir de l'échantillon ICM tous les deux ou trois ans. Cette pratique constitue aussi un autre défi puisque les indices sont habituellement calculés en n'utilisant que les unités communes de deux échantillons successifs. Des stratégies pour résoudre cette difficulté seront discutées.

Wednesday, May 29th/Mercredi 29 mai, 14:00 **TSH B106**

Using a weighted average of the Jevons and Laspeyres indexes to approximate a superlative index

Utilisation des indices podérés de Jevons and Laspeyres pour approximer un indice superlatif

Janice Lent, U.S. Bureau of Transportation Statistics et/and Alan Dorfman, U.S. Bureau of Labor Statistics

Shapiro and Wilcox (1997) advocated the Lloyd-Moulton price index (Lloyd 1975, Moulton 1996), or constant elasticity of substitution (CES) index, as a timely approximation to a Fisher or Trnqvist index. Using Taylor series expansions, we show that a weighted average of the Jevons and Laspeyres indexes can serve as a simple, robust approximation to the Lloyd-Moulton. This approximation also suggests a simple method of estimating the elasticity of substitution parameter on which the Lloyd-Moulton depends. An empirical study, based on CPI data from the U.S. Bureau of Labor Statistics, indicates that we may compute timely approximations to a superlative index using a weighted Jevons-Laspeyres mean; with weights estimated from past data. We also apply the elasticity estimator to experimental airfare index data from the U.S. Bureau of Transportation Statistics.

Shapiro et Wilcox (1997) ont recommandé l'indice des prix de Lloyd-Moulton (Lloyd 1975, Moulton 1996), ou l'indice de l'élasticité constante de la substitution (CES), comme approximation à l'indice de Fisher ou de Trnqvist. En utilisant des développements en séries de Taylor, nous prouvons qu'une moyenne pondérée des indices de Jevons et de Laspeyres peut servir d'approximation facile et robuste à l'indice de Lloyd-Moulton. Cette approximation suggère également une méthode simple d'estimer le paramètre "d'élasticité de la substitution" impliqué dans la méthode de Lloyd-Moulton. Une étude empirique, basée sur des données de CPI du U.S. Bureau of Labor Statistics, indique que nous pouvons calculer des approximations appropriées à un indice superlatif en utilisant une moyenne pondérée par la méthode de Jevons-Laspeyres; avec des poids estimés à partir des données antérieures. Nous appliquons également l'estimateur d'élasticité aux données de l'indice expérimental des tarifs aériens produit par le U.S. Bureau of Transportation Statistics.

Wednesday, May 29th/Mercredi 29 mai, 14:30 **TSH B106**

IT, Hedonic Price Indexes, and Productivity: International Comparability Issues

Jack E. Triplett, The Brookings Institution

U.S. labor productivity (LP) increased from 1.4 percent per year before 1995 to about 1.8 percent per year after 1995. Around a quarter to a third of the acceleration in U.S. LP came from increased

growth in capital services per worker (capital deepening), mainly in IT (information technology) capital. IT capital is also a major factor in boosting productivity in the services industries. Similar but not identical trends apply to Canada.

Economists want to determine the contribution of IT in other OECD economies, but the basic analytic statistics on IT are not always available, mainly because of inadequate IT deflators. National accounts deflators for computers and IT equipment are constructed by non-comparable methodologies and show variations that are far too large to be caused by differences in national distribution systems and market conditions.

This paper reports on a research project that is just getting underway to explore cost-effective solutions to providing internationally comparable quality-adjusted price indexes for IT products, which would then permit international comparisons of productivity. It involves an international collaboration between the OECD, the European Hedonic Center at Eurostat, Statistics Canada and the Australian Bureau of Statistics, and it will therefore cover a substantial number of OECD countries.

La productivité des travailleurs américains (PTA) a augmenté de 1,4 pour cent par année avant 1995 à d'environ 1,8 pour cent par année après. Près du quart à un tiers de l'accélération du PTA provient de la croissance accrue des services capitaux par ouvrier (rationalisation d'équipements), principalement dans le capital associé aux technologies de l'information (TI). Ce capital est également un facteur important dans le gain de productivité dans le secteur tertiaire. Les tendances semblables mais non identiques, s'appliquent au Canada.

Les économistes veulent déterminer la contribution des TI dans les économies des autres pays membres de l'OCDE, mais les statistiques analytiques de base sur ce sujet ne sont pas toujours disponibles, principalement en raison de l'absence de coefficients correcteurs adéquat pour les TI. Des coefficients correcteurs des comptes nationaux pour les ordinateurs et les équipements associés aux TI sont construits à partir de méthodologies non comparables et ils montrent de trop grandes variations causées par les différences dans les systèmes de distribution nationaux et les conditions du marché.

Cet article rend compte d'un projet de recherche venant de commencer et visant à explorer des solutions rentables afin de fournir des indices des prix comparables, tenant compte de la qualité, des équipements associés au TI. Ceci permettra alors des comparaisons, au niveau international, de la productivité. Il implique une collaboration internationale entre l'OCDE, le centre hédonistique européen à l'Eurostat, Statistiques Canada et le Bureau australien des statistiques, couvrant ainsi un bon nombre de pays de l'OCDE.

Session 42: Stochastic Operations Research/Recherche opérationnelle stochastique

Wednesday, May 29th/Mercredi 29 mai, 13:30 **TSH B105**

Exact asymptotics for polling models
Asymptotique exacte pour des modèles de regroupements

Doug Down, McMaster University

Polling models arise in the analysis of single server systems with multiple classes of customers, such as in communications switches. We will provide an overview of a body of work concerned with finding exact asymptotics for the tails of queue length distributions for polling models operating under various service disciplines. This involves methodology developed by David McDonald, which allows one to convert the problem of computing exact asymptotics into one of finding an appropriate change of measure for the underlying stochastic process and then solving a series of stability problems. The

talk will conclude with implications of these results for designing scheduling policies for systems with "quality of service" constraints, which typically involve tail behaviour.

Les modèles d'intégration surgissent dans l'analyse des systèmes à serveur unique avec plusieurs classes de clients, comme dans les transferts de communication. Nous fournirons une vue d'ensemble du cadre de travail entourant la recherche sur l'asymptotique exacte des queues des distributions de la longueur des files d'attente pour des modèles d'intégration pour diverses politiques de service. Ceci implique la méthodologie développée par David McDonald, qui permet de convertir le problème du calcul de l'asymptotique exacte en un problème consistant plutôt à trouver un changement de mesure approprié pour le processus stochastique sous-jacent et ensuite à résoudre une série de problèmes reliés à la stabilité. La conférence va se conclure sur des conséquences de ces résultats sur la conception des politiques d'établissement de programme pour des systèmes avec des contraintes sur la "qualité de service", qui impliquent typiquement le comportement de queue.

Wednesday, May 29th/Mercredi 29 mai, 14:00 **TSH B105**

Using simulation to evaluate robustness of forest operations plans
Utilisation de simulations pour évaluer la robustesse des plans sur les opérations en forêts

Evelyn Richards et/and John A. Kershaw, University of New Brunswick

Multi-year forest management operational plans schedule harvesting, silviculture, and distribution of timber products over a 3-5 year planning horizon. In this production - distribution problem, operations are constrained by system availability, site factors, seasons and regulatory constraints. Demands for timber products must be met. This system is highly stochastic, with fluctuations in actual versus expected product yields, operating condition changes due to weather, and delivery expectations which depend on dynamic markets. In this situation, plans which are robust with respect to change in conditions are desirable. This paper describes a simulation approach to evaluating operational plans for robustness under variation in standing timber inventories.

Les plans opérationnels sur plusieurs années pour la gestion des forêts planifient la moisson, la sylviculture, et la distribution des produits de bois de construction sur un intervalle de 3 à 5 ans. Dans ce problème de production distribution, les opérations sont contraintes par la disponibilité des systèmes, les facteurs du site, les saisons et les contraintes de réglementation. La demande des produits de bois de construction doit être satisfaite. Ce système est fortement stochastique, avec des fluctuations entre le réel et les rendements prévus dans les produits, les changements de condition de fonctionnement dus à la température, et les livraisons espérées qui dépendent de la dynamique des marchés. Dans cette situation, des plans robustes sont souhaitables tout en respectant le changement des conditions. Cet article décrit une approche de simulations pour évaluer des plans opérationnels quant à la robustesse sous la variation des inventaires disponibles de bois de construction.

Wednesday, May 29th/Mercredi 29 mai, 14:30 **TSH B105**

Modelling forest fires stochastically
Modélisation stochastique des feux de forêt

Reg Kulperger, W.J. Braun et/and D.A. Stanford, The University Of Western Ontario

What is the optimal or near-optimal day-to-day allocation of resources (planes, fire crews, etc.) for fighting forest fires in Ontario? The operations research solution of this problem requires an appropriate model of forest fire growth. The research project described in this paper presents a

flexible statistical model based on an interacting particle system (previously used by, among others, Braun (1992) to model cancer growth) for the evolution of forest. This model will be used in the future as a basis for simulation on a large-scale cluster of computers and may used to determine the best means of controlling of forest fires, and to gain insight into fire "spotting".

Qu'elle est la répartition optimale et quasi-optimale des ressources (avions, équipages de feu, etc...) de jour en jour pour le combat des feux de forêt en Ontario? La solution proposée par la recherche opérationnelle à ce problème exige un modèle approprié de croissance des feux de forêt. Le projet de recherche décrit dans cette conférence présente un modèle statistique flexible basé sur un système de particules agissant l'une sur l'autre (précédemment employé par, entre d'autres, Braun (1992) pour modéliser la croissance du cancer) dans l'évolution de la forêt. Ce modèle sera utilisé à l'avenir comme base pour des simulations sur de très grands nombres de regroupements d'ordinateurs et peut être utilisé pour déterminer les meilleurs moyens pour le contrôle des feux de forêt, et pour gagner du discernement dans le repérage des feux.

Session 43: Survival Analysis of Case-Control and Case-Cohort Data/Analyse de survie des données d'études cas-témoins et d'études cas-cohorte

Wednesday, May 29th/Mercredi 29 mai, 13:30 **KTH B135**

The nature of and the methodological challenges in epidemiological case-control studies.
La nature et les défis méthodologiques dans les études épidémiologique cas-contrôles.
Jack Siemiatycki, Université de Montréal

Wednesday, May 29th/Mercredi 29 mai, 14:00 **KTH B135**

Some recent developments of case-cohort analysis with Cox's regression model
Développements récents dans l'analyse de cas-cohorte avec un modèle de régression de Cox
Shaw-Hwa Lo, Columbia University et/and Kani Chen , Hong Kong University of Science and Technology

Prentice (1986) proposed the case-cohort design and studied a pseudolikehood estimator of regression parameters in Cox's model. The method is useful in both epidemiological studies and clinical trials. The original estimators suggested by Prentice have been well studied. We discuss in this talk several possible improvements based on simple estimating equations and properly weighted functions. The issue of extending the methods to other sampling schemes will be discussed.

Prentice (1986) a proposé une expérience de cas-cohorte et a étudié un estimateur de pseudo-vraisemblance des paramètres de régression dans le modèle Cox. La méthode est utile dans des études épidémiologiques et des essais cliniques. Les estimateurs initiaux suggérés par Prentice ont été abondamment étudiés. Dans cette présentation, nous discutons de plusieurs améliorations possibles basées sur des équations d'estimation simples et des fonctions de poids adéquates. La possibilité d'appliquer ces méthodes à d'autres plans d'échantillonnage sera discutée.

Wednesday, May 29th/Mercredi 29 mai, 14:30 **KTH B135**

Comparison of Cox's model versus logistic regression for case-control data with time-varying exposure: a simulation study.

Comparaison des modèles de Cox et de régression logistique pour des données cas-témoins dont l'exposition varie au cours du temps : une étude de simulation.

Karen Leffondré, Michal Abrahamowicz, McGill University, et/and Jack Siemiatycki, Université de Montréal

Many environmental and occupational exposures investigated in case-control studies may vary over time. These exposures are ascertained retrospectively and their effects are typically assessed using the conventional logistic model. However, logistic regression does not directly account for changes in the covariate values over time. On the other hand, some adaptations of the Cox model to case-control data have been proposed (Prentice and Breslow, Biometrika 1978; Chen and Lo, Biometrika 1999), but their performance in the case of time-varying exposure has not been systematically evaluated. Through a comprehensive simulation study, we propose to investigate (1) the advantages of using time-dependent variables in an adaptation of the Cox model for case-control data, compared to the conventional logistic regression model with fixed covariates; (2) the dependence of the bias and precision of the Cox model estimates on the definition of the risk sets, especially in studies where cases are identified over a long time period. The lifetime experience of a hypothetical population is first generated and a series of case-control studies is then simulated from this population. We control the frequency, intensity, interruptions, and duration of exposure; the type (proportional hazards or not) and strength of the association between exposure and outcome, and different aspects of study design (e.g. the length of the period of cases' identification). Different ways of representing exposure history, as well as different models and risk set definitions, are considered for the analysis of these data.

De nombreuses expositions environnementales et professionnelles étudiées dans les enquêtes cas-témoins peuvent varier au cours du temps. Les données concernant ces expositions sont recueillies rétrospectivement et leurs effets sont habituellement étudiés à l'aide du modèle logistique conventionnel. Cependant, la régression logistique ne tient pas compte directement des changements au cours du temps des valeurs de la covariable. D'autre part, quelques modifications du modèle de Cox aux données cas-témoins ont déjà été proposées (Prentice et Breslow, Biometrika 1978; Chen et Lo, Biometrika 1999), mais leur performance dans le cas d'une exposition variant au cours du temps n'a jamais été systématiquement évaluée. À l'aide d'une étude de simulation, nous proposons d'étudier : (1) les avantages d'utiliser des variables dépendantes du temps dans une adaptation du modèle de Cox pour des données cas-témoins, en comparaison avec le modèle de régression logistique avec des variables fixes dans le temps; (2) la dépendance du biais et de la précision des estimations découlant du modèle de Cox, par rapport à la définition des ensembles à risque, en particulier dans les études où les cas sont identifiés sur une longue période de temps. L'expérience de vie d'une population fictive est d'abord générée et une série d'études cas-témoins est ensuite simulée à partir de cette population. Nous contrôlons la fréquence, l'intensité, les interruptions, et la durée de l'exposition; le type (risques proportionnels ou non) et la force d'association entre l'exposition et l'événement d'intérêt; ainsi que différents aspects de la conception de l'étude (par exemple, l'étendue de la période d'identification des cas). Différentes façons de représenter l'histoire de l'exposition, ainsi que différents modèles et définitions des ensembles à risque, sont considérés pour analyser ces données.

Session 44: Probability and Statistical Inference
Probabilité et inférence statistique

Wednesday, May 29th/Mercredi 29 mai, 13:30 **TSH B128**

True odds with a biased coin
Vrais probabilités avec une pièce de monnaie biaisée

Adrienne Kemp, University of St Andrews Scotland

Many games begin with a question. Who goes first? Which team starts at which end of the pitch? How can four people randomly form two pairs? During play, how do we simulate a dreidel (equiprobable four-sided top) or a true six-sided die?

Many people know the 'trick' of deciding between two alternatives by tossing a (possibly biased) coin twice, assuming no autocorrelation. Heads-then-tails implies A; tails-then-heads implies B. If the outcome is two heads or two tails, then two more tosses are needed.

The talk explores other, more efficient ways of obtaining true odds when choosing between n=2 alternatives. Methods for n=3, 4, 5, 6, . . . are then constructed. A surprising use of Fermat's Little Theorem is revealed.

Plusieurs jeux débutent avec une question. Qui commence? Quelle équipe débute à quelle extrémité du terrain? De combien de façon quatre personnes peuvent-elles aléatoirement former deux pairs?

Durant le jeu, comment peut-on simuler un dreidel (dé à quatre faces équiprobables) ou un véritable dé à six faces?

Plusieurs personnes connaissent la solution pour décider entre deux choix simplement en lançant une pièce de monnaie (possiblement biaisée), assumant aucune corrélation. La combinaison face-pile implique A tandis que pile-face implique B. Si le résultat est deux piles ou deux faces, alors deux autres lancers sont requis.

Cette présentation explore entres autres la façon la plus efficace d'obtenir les vrais résultats lorsque nous devons choisir parmi n=2 alternatives. Des méthodes pour n=3,4,5,6,... sont alors construites. Une utilisation surprenante du petit théorème de Fermat est révélée.

Wednesday, May 29th/Mercredi 29 mai, 13:45 **TSH B128**

Poisson limits for U-statistics
Limites poissoniennes pour les U-statistiques

Andre Dabrowski, Université d'Ottawa, H.G. Dehling, Bochum, T. Mikosch, Copenhagen et/and O. Sharipov, Uzbek Academy of Sciences

We study Poisson limits for U-statistics with nonegative kernels and lacking a second moment. The limit theory is derived from the Poisson convergence of suitable point processes of U-statistic structure. We obtain infinite variance stable limits for U-statistics with regularly varying kernel. We also apply these results to explore the asymptotic behaviour of some standard estimators of correlation dimension.

Nous étudions des limites poissoniennes pour des U-statistiques avec les noyaux non négatifs et avec le deuxième moment manquant. Les théorie sur les limites est dérivée de la convergence de Poisson d'un processus ponctuel convenable ayant une structure de U-statistique. Nous obtenons des limites stables de variance infinie pour des U-statistiques avec le noyau variable. Nous appliquons également ces résultats à l'étude du comportement asymptotique de quelques estimateurs standards de dimension de corrélation.

Wednesday, May 29th/Mercredi 29 mai, 14:00 **TSH B128**

A note on maximum autoregressive processes of order one
Une note sur le processus autorégressif maximum d'ordre un

Mahmoud Zarepour, Université d'Ottawa et/and Dragan Banjevic, University of Toronto

We consider estimating the coefficient of a maximum autoregressive process of order one. Under a parametric assumption for innovations the exact distribution of this estimate is calculated using a recursion method, while under the assumption that the distribution for the innovations has a regularly varying tail at infinity we derive its limiting distribution.

Nous considérons l'estimation du coefficient d'un processus maximum autorégressif d'ordre un. Sous l'hypothèse paramétrique, la distribution exact de cette estimation est calculée en utilisant une méthode récursive, tandis que sous l'hypothèse que la distribution a une queue à variations régulières à l'infini, nous dérivons sa distribution limite.

Wednesday, May 29th/Mercredi 29 mai, 14:15 **TSH B128**

Evaluation of the asymptotic variance-covariance matrix for finite mixture distributions
L'estimation de la matrice des variances-covariances asymptotiques dans des modèles de mélanges finis

Murray Jorgensen, University of Waikato and Roger Littlejohn, Invermay Agricultural Research Centre, New Zealand

An analytical calculation of the missing-data information matrix for the general finite mixture model is given, from which the asymptotic variance-covariance matrix follows. This leads to an adapted Newton-Raphson algorithm which is compared to the EM and ECME algorithms.

Un calcul analytique de la matrice d'information des données manquantes pour le modèle de mélanges finis est donné, duquel la matrice asymptotique de variance-covariance suit. Ceci mène à un algorithme adapté de Newton-Raphson qui est comparé aux algorithmes EM et ECME.

Wednesday, May 29th/Mercredi 29 mai, 14:30 **TSH B128**

A formula for the density of solutions to estimating equations
Une formule pour la densité des solutions d'une équation d'estimation

Tony Almudevar, Acadia University

In Almudevar, Field and Robinson (Annals of Statistics, 2000) an exact formula for the density of the solution to an estimating equation $\Psi(X, \theta) = 0$ is derived. This density is expressed directly in terms of the density of the estimating function Ψ and its derivative, and so may simplify considerably the calculation of the density of maximum likelihood estimates, or of any other type of estimator calculable as the solution to an estimating equation.

A new derivation of this formula is presented which permits the calculation of conditional densities, in a matter suitable, for example, for conditioning on ancillary statistics. In addition, the original formula was given as a double limit, whereas the new derivation eliminates the need for this.

Some applications of the formula will be given. In particular, in the exponential family case the formula permits the density of an MLE to be expressed directly in terms of the density of the sufficient statistic. This work will be compared to similar results given in Pazman (1984), Skovgaard (1990) and Jensen & Wood (1998).

Almudevar, Field and Robinson (Annals of Statistics, 2000) propose une formule exacte pour la densité des solutions d'une équation d'estimation de la forme $\Psi(X, \theta) = 0$. Cette densité est exprimée directement en terme de la densité de la fonction estimée Ψ et sa dérivée, et peut ainsi simplifier considérablement les calculs de la densité du maximum de vraisemblance, ou de tout autre type d'estimateur qui est aussi solution de l'équation d'estimation.

On présente une nouvelle expression pour cette formule qui permet de calculer des densités conditionnelles, de façon appropriée, comme par exemple, pour conditionner sur des statistiques auxiliaires. De plus, la formule initiale a été donnée comme une limite double, tandis que la nouvelle expression élimine ce besoin.

Quelques applications de la formule seront données. En particulier, dans le cas de familles exponentielles, la formule permet à la densité d'un EMV d'être exprimée directement en termes de la densité de la statistique exhaustive.

Ce travail sera comparé aux résultats similaires obtenus par Pazman (1984), Skovgaard (1990) et Jensen et Wood (1998).

Session 45: Point Processes and Applications/Processus ponctuels et applications

Wednesday, May 29th/Mercredi 29 mai, 15:30 **KTH B135**

Stress release and transfer models for earthquakes
Stabilité du relâchement de tension et modèles de transferts pour des tremblements de terre.

Mark Bebbington, Massey University et/and Kostya Borovkov, University of Melbourne

The stress release process is a Markov model for the long-term build up of stress by tectonic forces and its release in the form of earthquakes. This can be generalised to multiple regions, which creates the additional possibility of transfer (positive or negative) between the regions. By maximising the point-process likelihood, the model can be fitted to earthquake catalogs and, through simulation, create probabilistic forecasts of earthquake risk. We will outline procedures for using the model and consider some conditions for the process to be stable.

Le processus de relâchement du stress est un modèle de Markov pour l'accumulation à long terme du stress causé par les forces tectoniques et son relâchement sous forme de tremblements de terre. Ceci peut être généralisé aux régions multiples, qui crée la possibilité supplémentaire de transfert (positif ou négatif) entre les régions. En maximisant la vraisemblance du processus ponctuel, le modèle peut être adapté aux phénomènes de tremblement de terre et, par la simulation, créer des prévisions probabilistes de risque de tremblement de terre. Nous tracerons les grandes lignes des conditions pour que le processus soit stable, et à la lumière de ceci, nous considérerons quelques résultats d'adaptation de modèle aux données de tremblement de terre.

Wednesday, May 29th/Mercredi 29 mai, 16:00 **KTH B135**

Point process models and probability forecasts for earthquakes
Modèles basés sur les processus ponctuels et prévision de la probabilité d'un tremblement de terre

David Vere-Jones, Victoria University of Wellington

Probability forecasts require good models, good data, and rather sophisticated users. While none of these requirements is fully met for earthquakes, the last decade has seen major improvements in the quality and scope of earthquake data, which has lead in turn to improved models and the beginnings of probability forecasting.

Space-time point processes form a natural modelling framework for earthquake probability forecasts. Important statistical questions relate to the extraction of predictive information from complex, spatially distributed data, the processing of such data to produce synoptic (ie regional) forecasts, and assessment of the forecasts so produced.

This work will review some recent developments in this field, based largely on collaborative work with scientists in New Zealand, China and Japan. A particular concern will be the development of point-process models which capture some part of the evolution of the local stress field towards a critical regime.

Les prévisions de probabilités exigent de bons modèles, de bonnes données, et des usagers plutôt sophistiqués. Malgré le fait qu'aucune de ces exigences n'est entièrement disponible pour les tremblements de terre, la dernière décennie a connu des améliorations notables quant à la qualité et à la précision des données sur les tremblements de terre, qui ont conduit à des modèles améliorés et aux débuts des prévisions de probabilités.

Les processus ponctuels espace-temps forment un cadre de modélisation naturel pour la prévision de la probabilité d'un tremblement de terre. Des questions statistiques importantes associent à l'extraction d'information prévisionnelle de données complexes distribuées dans l'espace, le traitement de telles données pour produire des prévisions synoptiques (c'est-à-dire régional), et l'évaluation des prévisions ainsi produites.

Ce travail passera en revue quelques développements récents dans ce domaine, basé en grande partie sur le travail d'une équipe de scientifiques de Nouvelle-Zélande, de Chine et du Japon. Un intérêt particulier sera le développement des modèles de processus ponctuels qui capturent une certaine partie de l'évolution du champ de contraintes locales vers un régime critique.

Wednesday, May 29th/Mercredi 29 mai, 16:30 **KTH B135**

Analyses of bivariate time series in which the components are sampled at different instants

Analyses de séries chronologiques bidimensionnelles dans lesquelles les composantes sont échantillonnées à des temps différents

David Brillinger, University of California Berkeley

We consider a bivariate time series $(X(t), Y(t))$, where X is sampled at instants s_j and Y at instants t_k. We formulate the following questions: i) are the components X and Y associated? ii) Can they be related via a linear system? iii) Are they related in an instantaneous non-linear way? We develop estimates and an analysis of error. The methods are illustrated via discharge data on the River Solimoes and its difluent, the Parana do Coreiro, in Amazonia, Brazil.

Nous considérons une série chronologique bidimensionnelle $X(t), Y(t)$, où X est échantillonné à l'instant s_j et Y à l'instant t_k. Nous formulons les questions suivantes : i) Est-ce que les composantes de X et Y sont associées? ii) Peuvent-elles être reliées via un système linéaire? iii) Sont-elles reliées d'une façon instantanée non linéaire?

Nous développons des estimateurs et une analyse des erreurs. La méthode est illustrée par des données sur les décharges dans la rivière Solimoes et son affluent, le Parana do Coreiro en Amazonie au Brésil.

Session 46: Linear Models and Design/Modèles linéaires et plan d'expérience

Wednesday, May 29th/Mercredi 29 mai, 15:30 **TSH B128**

Aligned rank test for the bivariate randomized block model
Test de rangs alignés pour le plan de blocs aléatoires bidimensionnels

Denis Larocque, École des Hautes Études Commerciales et/and Isabelle Bussières, Université du Québec à Trois-Rivières

An aligned rank test for treatment effects in the bivariate randomized block model is proposed. The test is easy to implement and its validity requires only minimal assumptions. Furthermore, the test statistic is affine-invariant and has a limiting chi-square distribution under the null hypothesis when the number of blocks goes to infinity. If the number of blocks is not large enough, we show how to perform a permutation test and illustrate this method with an example. Finally, a simulation study indicates that the new test performs well compared to the likelihood ratio test, to a coordinate-wise aligned rank test and to a sign test based on the Oja measure of scatter.

On propose un test d'effets de traitements basé sur les rangs alignés dans le plan de blocs aléatoires bidimensionnels. Le test est facile à utiliser et est valide sous des hypothèse très peu restrictives. De plus, la statistique de test est affine-invariante et converge vers la loi khi-deux, sous l'hypothèse nulle, lorsque le nombre de blocs tend vers l'infini. Nous montrons comment effectuer un test de permutation lorsque le nombre de blocs est petit et illustrons cette méthode par un exemple. Finalement, une étude par simulation démontre que le nouveau test est très performant comparativement au test du rapport de vraisemblance, à un test basé sur les rangs alignés coordonnées par coordonnées et à un test du signe basé sur la mesure de dispersion d'Oja.

Wednesday, May 29th/Mercredi 29 mai, 15:45 **TSH B128**

Effect of W, LR, and LM tests on the performance of preliminary test ridge regression estimators
Effet des tests de W, LR, et LM sur la performance des estimateurs de régression ridge pour des tests préliminaires

B. M. Golam Kibria, Florida International University et/and A.Saleh, Carleton University

This paper combines the idea of preliminary test and ridge regression methodology, when it is suspected that the regression coefficients may be restricted to a subspace. The preliminary test ridge regression estimators (PTRRE) based on the Wald (W), Likelihood Ratio (LR) and Lagrangian Multiplier (LM) tests are considered. The bias and the mean square errors (MSE) of the proposed estimators are derived under both null and alternative hypotheses. By studying the MSE criterion, the regions of optimality of the estimators are determined. Under the null hypothesis, the PTRRE based on LM test has the smallest risk followed by the estimators based on LR and W tests. However, the PTRRE based on W test performs the best followed by the LR and LM based estimators when the parameter moves away from the subspace of the restrictions. The conditions of superiority of the proposed estimator for both ridge parameter k and departure parameter Delta are provided. Some graphical representations have been presented which support the findings of the paper. Some tables for maximum and minimum guaranteed relative efficiency of the proposed estimators have been provided. These tables allow us to determine the optimum level of significance corresponding to the optimum

estimators among proposed estimators. Finally, weconcluded that the optimum choice of the level of significance becomes the traditional choice by using the W test for all non-negative ridge parameter, k.

Cet article combine l'idée de tests préliminaires et la méthodologie reliée à la régression "ridge", quand on suspecte que les coefficients de régression puissent être limité à un sous-espace. Les estimateurs de régression "ridge" pour tests préliminaires (PTRRE) basés sur les tests de Wald (W), du rapport de vraisemblance (LR) et des multiplicateurs de Lagrange (LM) sont considérés.

Le biais et les erreurs quadratiques moyennes (MSE) des estimateurs proposés sont déduits sous pour les hypothèses nulle et alternative. En étudiant le critère MSE, les régions qui optimisent les estimateurs sont déterminées. Sous l'hypothèse nulle, les PTRRE basés sur le test LM ont le plus petit risque suivi des estimateurs basés sur LR et sur le test W.

Cependant, les PTRRE basés sur le test W performent le mieux suivi de ceux basés sur les tests du LR et du LM quand les paramètres s'écartent du sous-espace des restrictions. Les conditions de supériorité de l'estimateur "ridge" proposé pour le paramètre k et le paramètre de d'écart Delta sont fournis. On présente quelques représentations graphiques qui illustreront les résultats de l'exposé. Quelques tables pour le maximum et le minimum de l'efficacité relative garantie par les estimateurs proposés sont fournies. Ces tables nous permettent de déterminer le niveau de confiance optimal correspondant aux estimateurs optimaux parmi les estimateurs proposés. Finalement, nous conclurons que le choix optimal du niveau de confiance devient le choix traditionnel en utilisant le test W pour tout les paramètres "ridge" non négatifs, k.

Wednesday, May 29th/Mercredi 29 mai, 16:00 **TSH B128**

Multivariate outlier detection
Détection de valeurs aberrantes dans un contexte multidimensionnel

Jiaqiong Xu, Bovas Abraham and Stefan Steiner, University of Waterloo

A Generalized Cook statistic for detecting multiple outliers in multivariate linear regression models is proposed. An approximate distribution of the proposed statistic is also obtained to get a suitable cutoff point for a test of hypothesis of no outliers. In addition, a simulation study has been conducted to examine the performance of the approximate distribution.

On propose une statistique généralisée de Cook pour détecter plusieurs données aberrantes dans des modèles de régression linéaire multidimensionnelle. Une distribution approximative de la statistique proposée est également obtenue pour obtenir un point de rupture approprié pour un test d'hypothèses sans données aberrantes. En plus, une étude de simulations a été entreprise pour examiner la performance de la distribution estimée.

Wednesday, May 29th/Mercredi 29 mai, 16:15 **TSH B128**

Optimal designs for model discrimination and fixed efficiency
Plans d'expérience optimaux pour la discrimination des modèles avec convergence fixe

Saumendranath Mandal et/and K.C. Carriere, University of Alberta

In optimal designs we generally assume that the regression model is known at the design stage. But in many situations this is not the case. Our goal is to implement a design that is efficient for two or more models that might fit the experiment, to discriminate between them, and select the best model. We consider different approaches to evaluate a criterion or a mixture of various criteria, and also optimize one objective subject to achieving a given efficiency of a parameter. In the latter case,

we solve the constrained problem by Lagrangian approach which eventually transforms the problem to a simultaneous optimization problem. To find optimal designs, we use a class of multiplicative algorithms. As an example, we consider discrimination between polynomial models. We also consider a practical example arising in Chemistry. In conclusion, some algorithmic results will be reported and discussed.

Dans des plans d'expérience optimaux, nous supposons généralement que le modèle de régression est connu à l'étape de la conception. Toutefois, dans beaucoup de situations, ce n'est pas le cas. Notre but est de mettre en application un plan d'expérience qui est efficace pour deux modèles ou plus et qui pourraient modéliser l'expérience, pour distinguer entre eux, et aussi pour choisir le meilleur modèle. Nous considérons différentes approches pour évaluer un critère ou un mélange de divers critères, et optimalisons également une fonction objective sous la contrainte d'observer une certaine efficacité pour un paramètre. Dans le dernier cas, nous résolvons le problème sous contraintes par l'approche du lagrangien qui transforme par la suite le problème à un problème d'optimisation simultanée. Pour trouver des plans d'expérience optimaux, nous employons une classe d'algorithmes multiplicatifs. Nous considérons par exemple la discrimination entre les modèles polynômiaux. Nous considérons également un exemple pratique surgissant en chimie. En conclusion, quelques résultats algorithmiques seront donnés et discutés.

Session 47: Statistical Inference/Inférence statistique

Wednesday, May 29th/Mercredi 29 mai, 15:30 **TSH B106**

On weighted least squares and related estimators in linear functional error-in-variables models

Sur les moindres carrés pondérés et les estimateurs reliés dans les modèles linéaires fonctionnels avec erreur dans les variables linéaires

Yuliya Martsynyuk, Carleton University

Consider the linear functional error-in-variables model

$y_i = \xi_i\beta_0 + \alpha_0 + \delta_i$, $x_i = \xi_i + \epsilon_i$,

where the scalar variables y_i and x_i are observed, and $\delta_i, i = 1, 2, ..., n$ and $\epsilon_i, i = 1, 2, ..., n$ are independent vectors of independent mean zero and finite variance random variables, $n >= 1$. We assume $E\delta_i^2 = \Gamma_{11} > 0$ and $E\epsilon_i^2 = \Gamma_{22} > 0, i = 1, 2, ..., n, n >= 1$. However, we do not assume that the error terms are normally and/or identically distributed. The variables ξ_i are nonrandom nuisance parameters, satisfying some mild regularity conditions, while β_0 and α_0 are the true values of parameters that are to be estimated.

The weighted least squares estimators (WLSE's) $\hat{\beta}_n$ of β_0 and thus also $\hat{\alpha}_n$ of α_0 depend on the ratio σ of the variances of error terms δ_i and ϵ_i. When the latter ratio σ is partially or completely unknown, one can replace it with its consistent estimator and retain also the consistency of the thus modified WLSE's of β_0 and α_0. The above WLSE's are studied for their asymptotic normality along with a few other known estimators of β_0 and α_0. The following cases are treated: the ratio σ is known; either the variance of δ_i or that of ϵ_i is unknown; both of the variances of error terms are unknown. In each of the above cases we compare asymptotic variances of already known estimators of β_0 and α_0 to those obtained in this exposition.

References.

1. Cheng, C.-L. and Van Ness, J.W. (1999). Statistical Regression with Measurement Error. Arnold, London.

2. Yu. Martsynyuk. Asymptotic Behaviour of Weighted Least Squares Estimator in Linear Functional Error-in-Variables Models, Technical Report Series of Laboratory for Research in Statistics and Probability, No.353-July 2001, Carleton University of Ottawa.

Considérons le modèle avec erreur dans les variables sur une fonctionnelle linéaire suivant $y_i = \xi_i\beta_0 + \alpha_0 + \delta_i$, $x_i = \xi_i + \epsilon_i$, *où les variables scalaires* y_i *et* x_i *sont observées, et* $\delta_i, i = 1, 2, ..., n$ *et* $\epsilon_i, i = 1, 2, ..., n$ *sont des des vecteurs indépendants de variables aléatoires indépendantes de moyenne 0 et de variance finie,* $n >= 1$. *Nous assumons* $E\delta_i^2 = \Gamma_{11} > 0$ *et* $E\epsilon_i^2 = \Gamma_{22} > 0, i = 1, 2, ..., n, n >= 1$. *Cependant, nous ne supposons pas que les termes d'erreur sont normalement et/ou identiquement distribuées. Les variables* ξ_i *sont des paramètres de nuisance non aléatoires, satisfaisant quelques conditions faibles de régularité, alors que* β_0 *et* α_0 *sont les vraies valeurs des paramètres qui doivent être estimés.*

Les estimateurs des moindres carrés pondérés (WLSE's) $\hat{\beta}_n$ *de* β_0 *et également* $\hat{\alpha}_n$ *de* α_0 *dépendent aussi du rapport* σ *des variances des termes d'erreur* δ_i *et* ϵ_i. *Quand le dernier rapport* σ *est partiellement ou complètement inconnu, nous pouvons le remplacer par son estimateur convergent et maintenir également la convergence du WLSE's modifié de* β_0 *et de* α_0. *Les WLSE ci-dessus sont étudiés pour leur normalité asymptotique avec quelques autres estimateurs connus de* β_0 *et de* α_0. *Les cas suivants sont traités : le rapport* σ *est connu; la variance de* δ_i *ou de* ϵ_i *est inconnue; les deux variances des termes d'erreur sont inconnues. Dans chacun des cas, nous comparons les variances asymptotiques des estimateurs déjà connus de* β_0 *et de* α_0 *à celles obtenues.*

Références :

1. Cheng, C.-L. and Van Ness, J.W. (1999). Statistical Regression with Measurement Error. Arnold, London.

2. Yu. Martsynyuk. Asymptotic Behaviour of Weighted Least Squares Estimator in Linear Functional Error-in-Variables Models, Technical Report Series of Laboratory for Research in Statistics and Probability, No.353-July 2001, Carleton University of Ottawa.

Wednesday, May 29th/Mercredi 29 mai, 15:50 **TSH B106**

Incorporating inter-item correlations for item response data analysis
Incorporation de corrélations inter-items pour l'analyse de données de réponses par item

Xiaoming Sheng, A. Biswas and/et K.C. Carriere, University of Alberta

This work concerns with item response data, which are usually measured on a rating scale and therefore ordinal. These study items are inter-correlated quite highly. Rasch models are widely used in ordinal data analysis, which convert ordinal categorical scales into linear measurements. However, the correlation between the study items, known as the polychoric correlation, has not been taken into consideration in the methodology development for the Rasch models. In this paper, we improve the current methodology to incorporate the inter-item correlations. Ignoring the presence of significant correlation can lead to serious bias in the study conclusions and loss in efficiency. We took a latent variable approach for this purpose, in combination with the generalized estimating equations to expand the estimation method for the Rasch model parameters. Simulation study clearly shows the relative efficiency of the estimates when the inter-item correlation is incorporated. As expected, the efficiency loss of the estimates increases as the level of the polychoric correlation gets high.

Ce travail étudie les données de réponse par item, qui sont habituellement mesurées sur une échelle d'évaluation et, par conséquent, ordinal. Ces éléments d'étude sont intercorrélés plutôt fortement. Les modèles de Recht sont largement répandus dans l'analyse de données ordinales, qui convertissent les échelles catégoriques ordinales en mesures linéaires. Cependant, la corrélation entre les éléments

de l'étude, connus sous le nom de corrélation polychorique, n'a pas été prise en compte dans le développement de la méthodologie pour les modèles de Recht.

Dans cette présentation, nous améliorons cette méthodologie en incorporant la corrélation interélément. Ignorer la présence d'une corrélation significative peut mener à un sérieux biais dans les conclusions de l'étude et une perte dans la convergence. Nous avons adopté une approche avec une variable latente, en combinaison avec les équations d'estimation généralisées pour augmenter la méthode d'évaluation des paramètres dans les modèles de Recht. L'étude de simulations montre clairement la convergence relative des estimateurs quand la corrélation interélément est incorporée. Tel que prévu, la perte de convergence des estimateur augmente pendant que le niveau de la corrélation polychorique devient élevée.

Wednesday, May 29th/Mercredi 29 mai, 16:10 **TSH B106**

Hierarchical quasi-likelihoods and their applications to hierarchical generalized linear models (hglms) and survival models with frailty

Quasi-vraisemblance hiérarchique et ses applications aux modèles linéaires généralisés et modèles de survie avec effets aléatoires

Changchun Xie, Anthony Desmond and Radhey Singh, University of Guelph

When we consider a random effect in our model, we usually suppose its distribution is normal as in generalized mixed model and integrate out the random effect from a generalization of Henderson's joint likelihood to get the marginal likelihood. However, since the random effect (or frailty) is not directly observed, it is hard to know its distribution without estimating the random effect. Also, the integration above is difficult to deal with. We need a method to estimate the random effect without assuming the distribution of the random effect. In this talk, we will introduce hierarchical models with quasi-likelihood for the random effect (or frailty) in which we make assumption only about its mean and variance function. We also claim the two classical frailty models: gamma frailty models, lognormal frailty models are only special cases of our frailty models. Examples and simulation are also given in the talk.

Lorsque nous considérons des effets aléatoires dans notre modèle, nous supposons généralement que sa distribution est normale comme dans les modèles mixtes généralisés et nous intégrons l'effet aléatoire par une généralisation de la distribution conjointe de Henderson pour obtenir la vraisemblance de la marginale. Toutefois, comme l'effet aléatoire n'est pas directement observé, il est difficile de connaître sa distribution sans estimer l'effet aléatoire. De plus, il est difficile de résoudre l'intégration mentionnée précédemment. Nous avons besoin d'une méthode pour estimer l'effet aléatoire sans supposer sa distribution. Dans cette conférence, nous présenterons les modèles hiérarchiques avec une quasi-vraisemblance pour l'effet aléatoire dans lesquels nous faisons seulement des hypothèses au sujet des densités de la moyenne et de la variance. Nous montrons également que les deux modèles classiques (les modèles gamma et lognormal) sont des cas spéciaux de notre modèle. Des exemples et des simulations sont également donnés dans la présentation.

Wednesday, May 29th/Mercredi 29 mai, 16:30 **TSH B106**

Hierarchical Bayesian estimation in conditional autoregressive disease mapping models

Estimation bayésienne hiérarchique dans des modèles autorégressif conditionnels pour le marquage d'une maladie

Ying MacNab, University of British Columbia

The we present a hybrid Monte Carlo algorithm for calculating Bayesian marginal posterior and illustrate its implementation in full Bayesian conditional autoregressive disease mapping model inference. The method can be seen as an elaborate form of the Metropolis algorithm applicable to Bayesian statistical problems when the derivatives of the log posterior are available. The key feature of this algorithm is that it utilises posterior gradient information to search for directions in which candidate moves have a high probability of being accepted and the Markov chain moves throughout the support of its target distribution via a suppressed random walk. The hybrid Monte Carlo algorithm is hitherto relatively unknown in Bayesian disease mapping literature. Here, we show that it can be an efficient algorithm that produces good risk prediction and permits adequate assessment of the prediction uncertainty. The algorithm is illustrated by analysing infant mortality in the province of British Columbia in Canada.

Cette présentation présente un algorithme hybride de Monte-Carlo pour calculer la distribution marginale bayésienne a posteriori et illustre son rôle au niveau de l'inférence dans un modèle bayésien autorégressif conditionnel du marquage d'une maladie. La méthode peut être vue comme une forme élaborée de l'algorithme Metropolis appliqué à un problème bayésien où les dérivées du logarithme de la densité a posteriori sont disponibles. La principale caractéristique de cet algorithme est l'utilisation de l'information fournit par le gradient a posteriori pour rechercher la direction vers laquelle le candidat se dirige ayant la plus forte probabilité d'être acceptée, et la chaîne de Markov se promène dans tout le support de la distribution de sa cible par l'intermédiaire d'une marche aléatoire restreinte. L'algorithme de Monte-Carlo hybride est jusqu'ici relativement inconnu dans la littérature bayésienne sur le marquage de maladie. Ici, nous prouvons qu'il peut être efficace et produire une bonne prévision du risque et permet l'estimation adéquate de l'incertitude de la prévision. L'algorithme est illustré par une analyse sur la mortalité infantile dans la province de Colombie-Britannique au Canada.

Session 48: Statistics in Finance and Marketing/La statistique en finance et en marketing

Wednesday, May 29th/Mercredi 29 mai, 15:30 **TSH B105**

Persuading people to trust a model
Persuader les gens à faire confiance à un modèle

Daymond Ling, CIBC

You now work for a financial institution, and you've been commissioned to build a predictive model for mortgage acquisition. Excited, you gather data across the whole bank, attack the problem with all of the statistical wizardry known to man and then some, and came out with a highly predictive model. Proud of the fruits of your labor, you rush into the VP's office and declare Victory. He listens to you excitedly sprout off every known statistical diagnosis known to man supporting the validity of the model, and asks you "What is it telling me" and "How does it work". If you can't answer these two questions in plain language within 5 minutes, all would be lost. What do you say? Here are some tips that may help and make you a winner at the end of the day.

Vous travaillez maintenant pour une institution financière, et vous avez été mandaté pour établir un modèle prédictif pour l'acquisition des hypothèques. Passionné, vous recueillez des données à travers toute la banque, attaquez le problème avec tous les concepts statistiques connus de l'homme à ce jour et puis certain, vous obtenez un modèle fortement prédictif. Fier des fruits de votre travail, vous vous précipitez dans le bureau du vice-président et déclarez victoire. Il écoute votre discours quant à la validation du modèle et vous demande : "qu'est-ce que ce modèle m'indique" et "comment fonctionne-t-il". Si vous ne pouvez pas répondre à ces deux questions en langage ordinaire dans un

délai de cinq minutes, tout serait détruit. Que dites-vous? Voici quelques trucs qui peuvent vous aider et faire de vous un gagnant à la fin de la journée.

Wednesday, May 29th/Mercredi 29 mai, 16:00 **TSH B105**

AVM: how to value over 3 million properties . . . every month!
AVM : Comment évaluer 3 millions de propriétés ...tous les mois!
Anthony Percaccio, Municipal Property Assessment Corporation

Have you ever wondered how your property assessment value is generated? Would you have thought that statistical inference would have been used in the development of this value? In fact over the past five years the Municipal Property Assessment Corporation (MPAC) has assessed more than three million residential properties using the sales comparison approach to value utilizing multiple regression analysis (MRA) representing a total value of over 500 billion. In today's active real estate market, the June 30 th 1999 assessed value may not represent current market value necessary for financing such as the collateral value of property required for loan underwriting. To address this gap, MPAC has embarked on producing Automated Valuation Models (AVM) in which continuous real-time value estimates are achieved as a result of extensive statistical analysis of market activity at provincial, regional, city and neighbourhood levels. This could only be achieved through maintaining the most comprehensive and dynamic database available in Ontario, containing more than two billion data elements. The talk will discuss the AVM Product and how it provides a reliable and accurate real-time estimate of market value for all residential properties in Ontario and conceivably beyond.

Vous êtes-vous déjà demandé comment l'évaluation de la valeur de votre propriété est obtenue? Auriez-vous pensé que l'inférence statistique ait été utilisée dans le calcul de cette valeur? En fait au cours des cinq dernières années, Municipal Property Assessment Corporation (MPAC) a évalué plus de trois millions de propriétés résidentielles en utilisant l'approche de comparaison des ventes selon la valeur utilisant l'analyse de régression multiple (MRA) représentant une valeur totale de plus de $500 milliards.

Dans un marché immobiliers actif tel qu'actuellement, la valeur évaluée au 30 juin 1999 peut ne pas représenter la valeur marchande actuelle nécessaire comme garantie exigée pour obtenir un prêt. Pour corriger cette imprécision, MPAC a voulu produire des modèles automatisés d'évaluation (AVM) dans lesquels des évaluations de la valeur (en temps réel continu) sont réalisées à l'aide d'analyses statistiques multiples des activités du marché aux niveaux provinciaux, régionaux, des villes et même du voisinage. Ceci a seulement pu être réalisé en mettant à jour la base de données la plus complète et la plus dynamique disponible en Ontario, contenant plus de deux milliards d'informations. La conférence discutera du produit de AVM et comment il fournit une estimation fiable et précise de la valeur marchande pour toutes les propriétés résidentielles en Ontario et peut-être même ailleurs.

Wednesday, May 29th/Mercredi 29 mai, 16:30 **TSH B105**

Credit scoring and the use of statistics in the world of consumer lending
Score de crédit et utilisation des statistiques dans le monde du prêt aux consommateurs
Mark Merritt, TransUnion of Canada

The world of consumer lending changed forever in the 1950's when credit scoring was introduced as a way to statistically evaluate potential customers. In today's competitive world, the use of new and more predictive data sources, sophisticated strategy design, and technology has forced lenders and the credit scoring industry to rethink the way their business must be managed.

To help you better understand this evolving industry, this session will explore several different issues: We will review the history of credit scoring and discuss some of the revolutionary advancements the industry experienced. We will examine some of the statistical techniques used and challenges faced in the process of application scorecard development. We will explore what benefits empirically derived scorecards bring to the consumer lending environment. We will discuss the typical results that consumer lenders achieve using credit scoring in an automated decision making process.

Le monde du prêt au consommateur a changé pour toujours dans les années 50 où le score de crédit a été présenté comme une méthode pour évaluer statistiquement les clients potentiels. Dans un monde concurrentiel comme aujourd'hui, l'utilisation de nouvelles sources de données plus prédictives, de plans d'expérience sophistiqués, et de la technologie a forcé les prêteurs et l'industrie du score de crédit à repenser la gestion de leurs entreprises.

Pour vous aider mieux à comprendre cette industrie en pleine évolution, cette présentation explorera plusieurs problèmes : nous passerons en revue l'histoire du score de crédit et discuterons de certains des avancements révolutionnaires de l'industrie; nous examinerons certaines des techniques statistiques utilisées et les défis auxquels ils ont fait face au cours du développement de l'application de cartes de score; nous explorerons quels avantages ont été empiriquement dérivés des cartes de score dans l'environnement du prêt au consommateur; nous discuterons des résultats typiques que les prêteurs au consommateur réalisent en utilisant le score de crédit dans un processus décisionnel automatisé.

Index