

**President/****Président**

Timothy Ramsay
tramsay@ohri.ca

Vice-President/**Vice-président**

John Nash
(613) 562-5800x4796(work/travail)
email: nashjc@uottawa.ca

President-Elect/**Président-désigné****Past-President/****Président-sortant**

Eric Rancourt
(613) 951-5046 (work/travail)
(613) 951-1462 (FAX/télécopieur)
email: Eric.Rancourt@statcan.ca

Program Coordinator/**Coordonnateur de programme**

Dena Schanzer
(613) 946-0461 (work/travail)
(613) 954-5414 (FAX/télécopieur)
email : Dena_Schanzer@phac-aspc.gc.ca

Treasurer/**Trésorier**

Manchun Fang
(613) 737-7600 x4107 (work/travail)
(613) 738-4800 (FAX/télécopieur)
email: mfang@cheo.on.ca

Secretary/**Secrétaire**

Cynthia Bocci
(613) 951-4885 (work/travail)
(613) 951-1462 (FAX/télécopieur)
email: cynthia.bocci@statcan.ca

Symposium 2006 de la SSO**Le 20 avril 2006****Nouvelles Frontières en Statistique**

Statistique Canada
120 Parkdale

Pré Tunney, immeuble R.-H. -Coats
Salle Simon Goldberg (au sous-sol)
Ottawa, Ontario

Le thème du symposium se concentre sur les nouvelles frontières en statistique. De nos jours, une grande partie du travail statistique le plus excitant ce trait aux problèmes de bases de données énormes et complexes pour lesquelles les modèles linéaires habituels ne s'appliquent pas tout à fait. Habituellement, l'évolution de ce genre de problème implique l'application de techniques statistiques ou problème courant tout en développant des méthodes d'analyse novatrices qui, à prime abord, peuvent ne pas sembler être des modèles statistiques. Souvent, il y a un aspect laborieux et quantitatif à l'analyse avec une idée d'apprentissage informatique ou de « data-mining ». Le symposium regroupe des présentateurs provenant de plusieurs domaines.

L'inscription à l'avance est fortement encouragée (voir formulaire).

Présentateurs confirmés (les résumés suivent):

Giles Hooker, Université McGill, Département de psychologie

Les équations différentielles en statistique: L'analyse des données pour les systèmes dynamiques

Alan Karr, National Institute of Statistical Sciences (North Carolina)

L'analyse statistique de haute sécurité pour les bases de données distribuées

David Martell, Université de Toronto, Faculté de sylviculture

L'analyse statistique sur l'incidence de feux de forêt

Steven Wang, Université York, Département de mathématique et statistique

Le regroupement des données catégoriques dans les grandes bases de données

Douglas B. Woolford, Université de Western Ontario, Département de statistique et des sciences actuarielles

Amélioration convergente des données appliquées aux données de foudre

Mu Zhu, Université de Waterloo, Département de statistique et des sciences actuarielles

La détection des cibles rares avec LAGO

Programme :

9h00 – 9h20	Mot de bienvenue
9h20 -10h00	Présentation de Alan F. Karr
10h00-10h30	Pause-café
10h30-11h10	Présentation de Mu Zhu
11h10-11h50	Présentation de Steven Wang
11h50-13h00	Dîner
13h00-13h40	Présentation de David Martell
13h40-14h20	Présentation de Douglas B. Woolford
14h20-14h50	Pause-café

Les équations différentielles en statistique: L'analyse des données pour les systèmes dynamiques

Giles Hooker, Université McGill

Resumé

Les équations différentielles ordinaires (EDO) ont une longue histoire dans la modélisation des systèmes au fil du temps. Elles sont largement utilisées dans les sciences physiques et en génie et suscitent de plus en plus d'intérêt en sciences biologique; pour la modélisation des maladies dynamiques (à la fois pour les populations et les individus), les processus de tirs neuronaux, la cinétique humaine et les systèmes écologiques. Ces nouvelles applications correspondent aux systèmes pour lesquelles les mesures sont moins précises que celles qui existent en sciences physiques et donc les modèles proposés ne sont que des approximations : ils ne sont plus dérivés des premiers principes. Ainsi, ces modèles requièrent une nouvelle méthodologie statistique.

Dans cette présentation, on souligne la puissance des EDOs pour décrire intuitivement une grande variété de comportements complexes et on discute quelques difficultés associées à leur modélisation et les méthodes pour ce faire. On présente aussi le développement des techniques diagnostiques pour la compréhension de modèles mal spécifiés et on montre la correspondance entre les techniques d'analyse des données en régression linéaire et celles nécessaires pour comprendre des systèmes dynamiques. Finalement, on décrit des problèmes ouverts en suspens qui touchent à une grande variété de préoccupations statistiques traditionnelles.

L'analyse statistique de haute sécurité pour les bases de données distribuées

Alan F. Karr, National Institute of Statistical Sciences (North Carolina)

Résumé

Dans les contextes de statistiques nationales, en passant par la sécurité de la patrie jusqu'au commerce, il y a un besoin continu pour des analyses statistiques qui intègrent les données entreposées dans les bases de données multiples qui sont distribuées. Pourtant, les barrières à l'intégration des données, incluant la confidentialité des données, droits de propriété et l'envergure, sont nombreuses et difficiles à surmonter.

Toutefois, pour plusieurs analyses, il n'est pas nécessaire d'intégrer les données. À la place, les détenteurs des données, utilisant une méthode basée sur des techniques informatiques connues sous le nom « calculs multi-partis sécuritaire », peuvent partager anonymement les statistiques suffisantes spécifiques à l'analyse qui permettent de faire cette analyse statistique de façon valide. On présente quatre analyses : la régression pour les données divisées horizontalement, l'intégration des données sans risque, les tableaux de contingence sécuritaire et la vraisemblance maximum sécuritaire pour les familles de modèles exponentiels.

On introduit aussi l'idée de « Tierces partis partiellement fiable (TPPF) » qui disposent d'informations non disponibles aux détenteurs des données. Dans l'intérêt commun, ils servent à enlever ou à aider les détenteurs des données à résister à la tentation de tricher en déclarant des données fausses ou statistiques suffisantes.

L'analyse statistique sur l'incidence de feux de forêt

David Martell, Université de Toronto

Résumé

Les feux de forêt sont fréquents dans plusieurs régions boisées du Canada. Les gestionnaires de forêt et des incendies doivent tenir compte de leur nature prévisible et de l'impact potentiel sur les gens, les biens immobiliers et les processus naturels de l'écosystème. On présente un bref survol de la gestion des feux de forêt en mettant l'accent sur les processus décisionnels et sur l'importance de comprendre et de prédire l'incidence des feux à travers les vastes dimensions spatiales et temporelles.

Ensuite, on décrit des analyses statistiques complétées sur l'incidence des feux et les défis importants qui restent.

Le regroupement des données catégoriques dans les grandes bases de données

Steven Wang, Université York

Résumé

On présente deux algorithmes pour regrouper les données catégoriques. Le premier algorithme de regroupement, basé sur la distance vectorielle de Hamming, est conçu pour les données catégoriques nominales. La méthode proposée est conceptuellement simple et directe puisqu'elle n'utilise aucun modèle statistique. Elle est aussi capable de détecter le nombre de grappes automatiquement sans aucune entrée de la part de l'utilisateur. Un test de Pearson Chi-deux modifié détermine si une grappe potentielle est significative. Le deuxième algorithme traite les données catégoriques ordinales avec les structures dépendantes. Au début, il était fondé sur la distribution de probabilité empirique. Bien qu'il soit plus général que le premier algorithme, il est plus intensif sur l'ordinateur et il ne s'applique pas aux grandes bases de données. On propose donc un algorithme moins rigoureux mais plus efficace basé sur la distribution de probabilité empirique. On montre que l'algorithme proposé a des meilleurs résultats par rapport à des autres algorithmes de regroupement bien connus comme K-modes et Autoclass avec des bases de données réelles. On discute aussi la complexité des calculs informatiques et le travail futur.

Amélioration convergente des données appliquées aux données de foudre

Douglas B. Woolford and W. John Braun, Université de Western Ontario

Résumé

On relie les démarrages de feu de forêt aux occurrences de foudre en analysant les données provenant du Ministère de ressources naturelles d'Ontario sur les feux de forêt et la foudre en Ontario. Toutefois, il est nécessaire de modifier les données avant de faire cette analyse à cause de la quantité de données de foudre et des problèmes reliés à la précision et les valeurs manquantes. Des examens préliminaires des données suggèrent que la formation de grappes par rapport aux occurrences de foudre dans les dimensions spatiales et temporelles sera utile. On propose un algorithme de grappe qui cherche le mode basé sur une forme de méthode convergente pour rendre les données plus nettes. Cette amélioration des données, comme algorithme, amène les observations près de leur mode le plus proche à chaque itération. Nous proposons d'appliquer l'itération jusqu'à la convergence des données qui est atteint à un mode local ou global. L'utilité de l'algorithme dans le contexte de données de foudre est triple : Premièrement, les données de foudre peuvent être réduites à des modes locaux respectifs dans les dimensions spatiales et temporelles; Deuxièmement, des petites modifications mènent à une méthode de réduction de bruit qui peut être appliquée pour estimer la trajectoire spatiale à court terme du système de foudre; Troisièmement, les données améliorées permettent une simulation de tendances de trajectoires de foudre basée sur le Bootstrap.

La détection des cibles rares avec LAGO

Mu Zhu, Université de Waterloo

Résumé

LAGO est un outil informatique efficace utilisé pour identifier les cibles rares dans une base de données. Pour retrouver des cibles rares dans une base de données, LAGO donne un score à chaque item dans la base de données avec un réseau spécialisé de fonctions radiales de base (RBFnet) formé des données d'apprentissage. Soit p_1 la fonction de densité pour la classe rare et p_0 la fonction de densité pour la classe d'origine. Le RBFnet construit par LAGO est un estimateur de p_1 de noyau densité avec une bande adaptée localement ajustée par un facteur qui se rapproche de p_0 au premier degré. La fonction de score résultant, $f(x)$, est donc à peu près une transformation monotone de la probabilité postérieure que l'item x appartienne à une classe rare.

La version originale (maintenant appelé eLAGO) se sert des fonctions radiales de base elliptiques qui peuvent s'adapter de façon plus souple aux données d'apprentissage. Une variation plus simple (mais peut-être plus utile en général) qu'utilise des fonctions radiales de base sphérique est maintenant disponible.



Formulaire d'inscription

SYMPOSIUM ANNUEL DE LA SOCIÉTÉ STATISTIQUE D'OTTAWA

Nouvelles frontières en Statistique

Judi, le 20 avril 2006

à

Statistique Canada
Pré Tunney , immeuble R.-H. -Coats
Salle Simon Goldberg (au sous-sol)
Ottawa, Ontario

S'il vous plaît utiliser l'entrée de la rue Holland et noter qu'une carte d'identité est requise pour accéder à l'édifice.

Frais d'inscription (comptant ou chèque seulement)

75\$ pour les membres avant le 7 avril 2006, 85\$ frais d'inscription en retard

75\$ +10\$ frais d'adhésion pour les non-membres avant le 7 avril 2006, 95\$ frais d'inscription en retard

50\$ pour les étudiants à temps plein avant le 7 avril 2006, 60\$ frais d'inscription en retard

Les frais d'inscription comprennent le dîner et les pauses-café

Nom: _____

Affiliation: _____

Téléphone: _____

Je suis membre de ASA (), SSC (), SSO ()

Prière de faire le chèque à l'ordre de la « Société statistique d'Ottawa ». Veuillez envoyer le formulaire d'inscription complété avec le chèque à

Carole Jean-Marie
Statistique Canada
Division des méthodes d'enquêtes auprès des entreprises
Pré Tunney, 11e étage, immeuble R.-H. -Coats
Ottawa, Ontario K1A 0T6

téléphone (613) 951-0827
télécopieur (613) 951-1462
courriel: Carole.Jean-Marie@statcan.ca