



SSO 2006 Symposium April 20, 2006

New Frontiers in Statistics

Statistics Canada
120 Parkdale Ave.

R.H. Coats Building, Tunney's Pasture
Simon Goldberg Room (basement level)
Ottawa, Ontario

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

The theme of this symposium is new frontiers of statistics. Nowadays, much of the most exciting statistical work revolves around problems with large, complex databases for which the standard linear models are not particularly useful. Progress with these types of problems usually comes from applying so-called 'statistical thinking' to the issue at hand while developing novel methods of analysis that may superficially not even seem to qualify as statistical models. Usually, there is a heavy computational aspect to the analysis with a machine learning or data mining flavor. This symposium features speakers from many areas of statistics.

Pre-registration is HIGHLY RECOMMENDED (see form).

The confirmed speakers (abstracts follow):

Giles Hooker, McGill University, Department of Psychology
An ODE to Statistics: Data Analysis for System Dynamics

Alan F. Karr, National Institute of Statistical Sciences (North Carolina)
Secure Statistical Analysis of Distributed Databases

David Martell, University of Toronto, Faculty of Forestry
Statistical Analysis of Forest Fire Activity

Steven Wang, York University, Department of Mathematics and Statistics
Clustering Categorical Data in Large Databases

Douglas B. Woolford, University of Western Ontario, Department of Statistics and Actuarial Sciences
Convergent Data Sharpening Applied to Lightning

Mu Zhu, University of Waterloo, Department of Statistics and Actuarial Sciences
Rare Target Detection with LAGO

Agenda :

9:00 - 9:20	Welcoming remarks
9:20 -10:00	Alan F. Karr's presentation
10:00-10:30	Coffee Break
10:30-11:10	Mu Zhu's presentation
11:10-11:50	Steven Wang's presentation
11:50-13:00	Lunch
13:00-13:40	David Martell's presentation
13:40-14:20	Douglas B. Woolford's presentation
14:20-14:50	Coffee Break
14:50-15:30	Gilles Hooker's presentation

An ODE to Statistics: Data Analysis for System Dynamics

Giles Hooker, McGill University

Abstract

Ordinary differential equations (ODEs) have a long history in modelling the evolution of systems over time. They have been widely used in the physical sciences and engineering and are the object of increasing interest in biological sciences; modelling disease dynamics -- both in populations and individuals, neural firing processes, human kinematics and ecological systems. These new applications correspond to systems with less accurate measurements than are found in the physical sciences and the proposed models can only be described as approximations: no longer being derived from first principles. As such, there is a new need for statistical methodology for such models.

In this talk, I will showcase the power of ODEs to intuitively describe a wide range of complex behavior and discuss some of the difficulties associated with fitting them to data and methods for doing so. I will present the development of diagnostic techniques for understanding poorly fit models and show that there is a correspondence between the techniques of data analysis in linear regression and those needed to understand system dynamics. Finally, I will list some important open problems that touch on a wide range of traditional statistical concerns.

Secure Statistical Analysis of Distributed Databases

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

Abstract

A continuing need in contexts from national statistics to homeland security to business is for statistical analyses that "integrate" data stored in multiple, distributed databases. However, barriers to actually integrating the databases, which include data confidentiality, proprietary data and scale, are numerous and not easy to overcome.

For many analyses, however, it is not necessary actually to integrate the data. Instead, using methods based on techniques from computer science known as secure multi-party computation, the database holders can share analysis-specific sufficient statistics anonymously, but in a way that the desired analysis can be performed in a statistically valid manner. Four illustrative analyses will be presented: regression for horizontally partitioned data, secure data integration, secure contingency tables and secure maximum likelihood for exponential family models.

Partially trusted third parties (PTTPs) will also be introduced. PTTPs hold some information not available to the database holders, but to their mutual benefit, removing or at least attenuating unilateral incentives for database holders to "cheat" by reporting false data or sufficient statistics.

Statistical Analysis of Forest Fire Activity

David Martell, University of Toronto

Abstract

Forest fires are common in many of the forest regions of Canada and fire and forest managers must account for their uncertain occurrence and their potential impact on people, property and natural ecosystem processes. I will present a brief overview of forest fire management with emphasis on decision-making and the importance of understanding and predicting fire activity across broad spatial and temporal scales. I will then describe some of the statistical analyses of fire activity that have been completed and important challenges that remain.

Clustering categorical data in large databases

Steven Wang, York University

Abstract

We introduce two algorithms to cluster categorical data. The first clustering algorithm is designed to handle nominal categorical data based on Hamming distance vectors. The proposed method is conceptually simple and straightforward as it does not require any statistical model. It also can detect the number of clusters automatically without any user input. The significance of a possible cluster is determined by a modified Pearson Chi-square test. The second algorithm tries to handle ordinal categorical data sets with dependent structures. It was initially based on the empirical probability distribution. Although it is more general than the first algorithm, it is computationally intensive and not applicable to large data sets. We then propose a less rigorous but more efficient algorithm based on the empirical probability distribution. Comparisons with well known clustering algorithms such as K-modes and AutoClass show that the proposed algorithms outperform their competitors for some well known real data sets. The computational complexity and future works will also be discussed.

Convergent Data Sharpening Applied to Lightning

Douglas B. Woolford and W. John Braun, University of Western Ontario

Abstract

We wish to relate forest fire ignitions to lightning strike occurrences through the analysis of Ontario lightning and fire data, supplied by the Ontario Ministry of Natural Resources. However, due to the sheer volume of the lightning data, as well as accuracy and missing data issues, changes to the data are required prior to any such investigation. Initial explorations of the data indicate that it may be useful to cluster the lightning strokes in space-time. We propose a mode-seeking clustering algorithm based on a convergent form of data sharpening methods. Data sharpening, as an algorithm, nudges observations closer to their nearest local mode(s) at each iteration. We propose to iterate the algorithm until convergence, showing that the data will converge to either local or global modes. The usefulness of the algorithm in the lightning context is threefold: First, the lightning data can be reduced to corresponding local spatial-temporal modes; second, slight modifications result in a noise-reduction method that can be applied to estimate short-term spatial track(s) of lightning storm system(s); third, the sharpened data provides a means for a bootstrap based simulation of spatial lightning strike patterns.

Rare target detection with LAGO

Mu Zhu, University of Waterloo

Abstract

LAGO is a computationally efficient tool for finding rare targets in a database. To do so, LAGO scores every item in the database with a specialized radial basis function network (RBFnet), trained with some learning data. Suppose p_1 is the density function of the rare class and p_0 , the density function of the background class. The RBFnet constructed by LAGO is an adaptive-bandwidth kernel density estimator of p_1 adjusted locally by a factor that approximates p_0 to the first-order. The resulting scoring function $f(x)$ is thus approximately a monotonic transformation of the posterior probability that item x belongs to the rare class.

The original LAGO (now called eLAGO) uses elliptical radial basis functions which can adapt more flexibly to the training data. A simpler (but perhaps more generally useful) variation that uses spherical radial basis functions (sLAGO) is now available.

STATISTICAL SOCIETY OF OTTAWA

SOCIÉTÉ STATISTIQUE D'OTTAWA

CHAPTER - ASA
REGIONAL ASSOCIATION - SSC



SECTION - ASA
ASSOCIATION RÉGIONALE - SSC

Registration Form

ANNUAL SYMPOSIUM OF THE STATISTICAL SOCIETY OF OTTAWA

New Frontiers in Statistics

Thursday, April 20, 2006

8:30 a.m. - 4:00 p.m.

Statistics Canada

R.H. Coats Building, Tunney's Pasture
Simon Goldberg Room (basement level)
Ottawa, Ontario

Please use the Holland Street entrance and note that ID is required to access the building.

Registration Fees- Cash or cheque only

\$75 for members if received by April 7, 2006, \$85 late-registration fee

\$75 + \$10 membership for non-members if received by April 7, 2006, \$95 late-registration fee

\$50 for full-time students, if received by April 7, 2006, \$60 late-registration fee

Registration fee includes coffee breaks and a catered lunch.

Please make cheques payable to "The Statistical Society of Ottawa" and provide the following information:

Name: _____

Affiliation: _____

Telephone: _____

I am a member of ASA (), SSC (), SSO ()

Please send completed registration form along with cheque to:

Carole Jean-Marie

Statistics Canada

Business Survey Methods Division

11th Floor, R.H. Coats Building, Tunney's Pasture

Ottawa, Ontario K1A 0T6

Tel. (613) 951-0827

Fax. (613) 951-1462

Email: Carole.Jean-Marie@statcan.ca