NZSA 2011 Conference

28 – 31 August 2011

Owen G Glenn Building, University of Auckland

62nd Annual Meeting of the New Zealand Statistical Association

Organisers

ORGANISING COMMITTEE

James Curran (Chair): j.curran@auckland.ac.nz

Harold Henderson (Sponsorship): harold.henderson@agresearch.co.nz

Thomas Lumley: t.lumley@auckland.ac.nz

Alexandra Miliotis: a.miliotis@auckland.ac.nz

Kathy Ruggiero (Treasurer): k.ruggiero@auckland.ac.nz

Chris Triggs: cm.triggs@auckland.ac.nz

Alain Vandal: alain.vandal@aut.ac.nz

Thomas Yee (Scientific Programme Chair): t.yee@auckland.ac.nz

Sponsors

NZSA 2011 gratefully acknowledges the financial support from the following organisations:

sas

THE POWER TO KNOW

SAS Institute NZ Limited

Travelex worldwide money

Global Business Payments

Travelex Financial Services New Zealand Limited

THE UNIVERSITY OF AUCKLAND

FACULTY OF SCIENCE

Department of Statistics

Department of Statistics, University of Auckland

Contents

General Information

Name Tags

Please wear your name badge at all times during the conference and at social events.

Mobile Phones

As a courtesy to presenters and colleagues, please ensure that your mobile phone is switched off during the conference sessions.

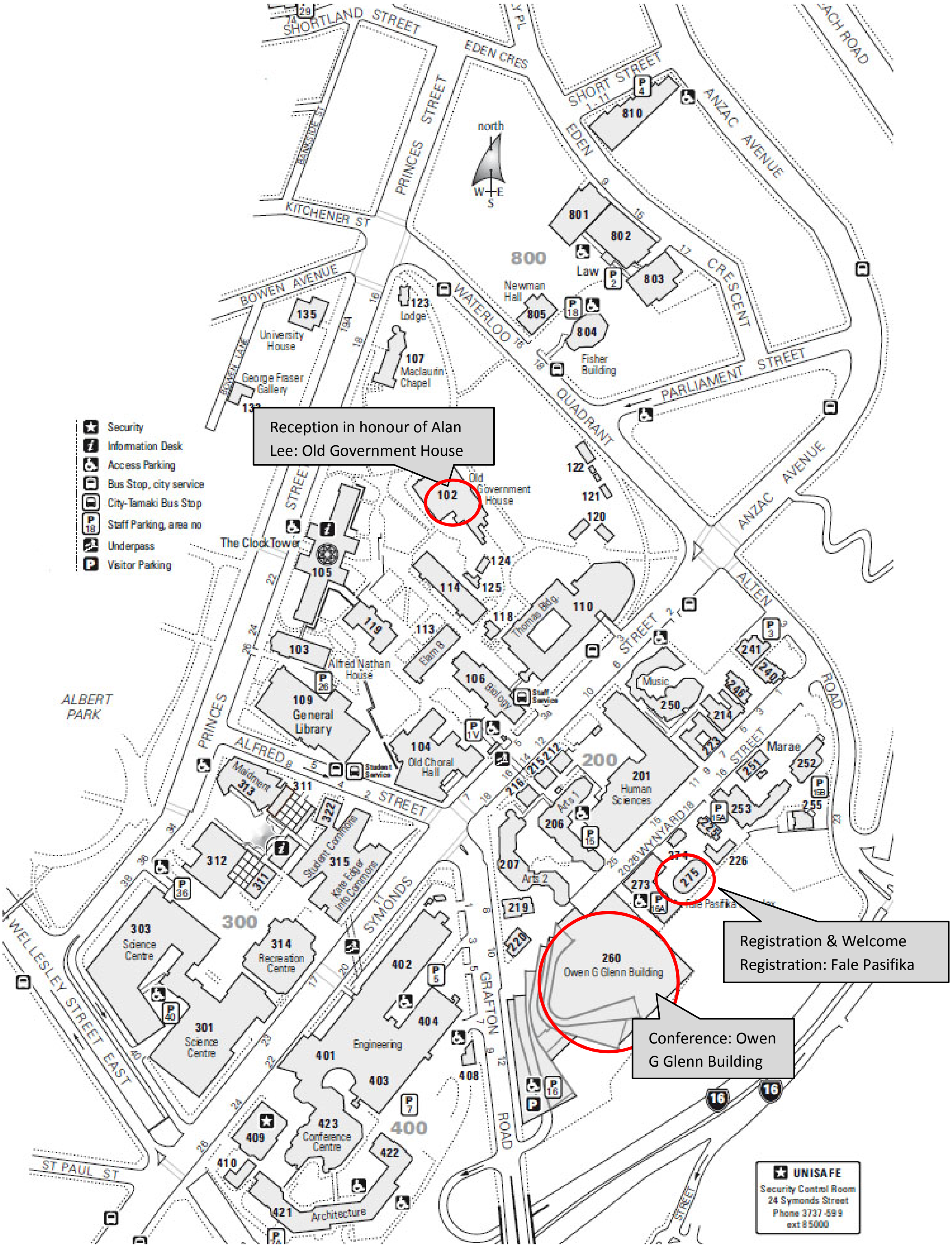
Social Events

Conference Registration and Welcome Reception
Sunday 28 Aug 6 – 8pm
Fale Pasifika, Building 275, 22 Wynyard Street, University of Auckland

Reception in honour of Professor Alan Lee
Monday 29 Aug 6 – 8pm
Old Government House, University of Auckland

Conference Dinner
Tuesday 30 Aug 6pm onwards
The Bluestone Room, 9 – 11 Durham Lane, Auckland CBD

Maps

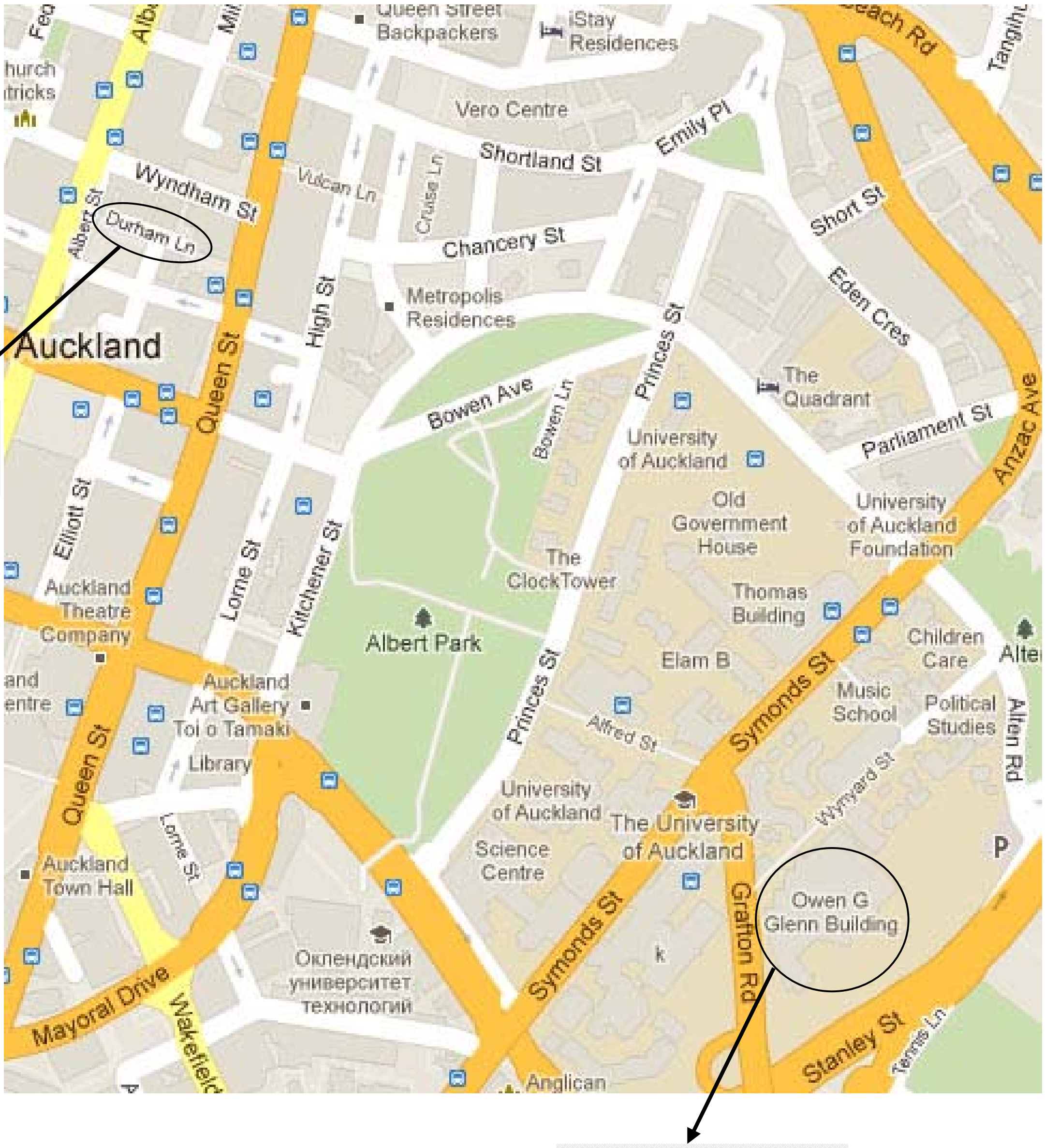


Reception in honour of Alan Lee: Old Government House

Registration & Welcome Registration: Fale Pasifika

Conference: Owen G Glenn Building

Dinner venue: The Bluestone Room, 9 – 11 Durham Lane



Dinner venue:
Durham Lane

Conference venue:
Owen G Glenn Building

Keynote & Invited Speakers

Nick Fisher

Dr Nick Fisher is the founder of ValueMetrics Australia, an organisation that carries out research and consulting primarily in the area of Performance Measurement. After 30 years as a statistical consultant and researcher in CSIRO, he left his position as a Chief Research Scientist in May 2001. Whilst in CSIRO, he led the development of CSIRO's Organisational Performance Measurement (OPM®) system, which has been applied successfully in a number of private and public enterprises, and has been part of graduate programs at a number of institutions.

Nick carries out research and consulting in Performance Measurement, with particular focus on improving quantitative reports to Boards and top management, and the associated business improvement processes.

He holds honorary appointments as Visiting Professor of Statistics at the University of Sydney, and Visiting Professor of Quality Management at Macquarie University. He is also professionally accredited by the Statistical Society of Australia.

Robert Gentleman

Dr Robert Gentleman is a Senior Director of Genentech specializing in Bioinformatics and Computational Biology. Prior to his move to Genentech in 2009 Robert has held a variety of full and adjunct academic appointments including the University of Washington, the University of Ghent, Harvard University, the University of Auckland, and the University of Waterloo.

Robert is extremely well known for his foundational work on the R project (www.r-project.org), being one of the two original authors along with Ross Ihaka, as well as his work with the associated Bioconductor project (http://www.bioconductor.org), which provides a large set of public domain tools for bioinformatics academics and professionals.

His current fields of interest relate to the use of high throughput sequencing to advance our knowledge of many basic biological mechanisms. Of particular interest is developing methods for understanding transcriptional regulation through the use of careful experimentation and ChIP-seq data. His group is also interested in helping to develop a better understanding of the role that transposable elements play in human disease.

Trevor Hastie

Prof Trevor Hastie joined Stanford University in Palo Alto, CA. as Professor of Statistics and Biostatistics in 1994. Trevor is famous worldwide for his work in data mining and machine learning. His main research contributions have been in the field of applied nonparametric regression and classification, and he has written two books in this area: "Generalized Additive Models" (with R. Tibshirani, Chapman and Hall, 1991), and "Elements of Statistical Learning" (with R. Tibshirani and J. Friedman, Springer 2009, Second Edition). He has also made contributions in statistical computing, co-editing (with J. Chambers) a large software library on modelling tools in the S language ("Statistical Models in S", Wadsworth, 1992), which form the basis for much of the statistical modelling in R and S-plus. His current research focuses on applied problems in biology and genomics, medicine and industry, in particular data mining, prediction and classification problems.

Xihong Lin

Prof Xihong Lin is a Professor of Biostatistics at the Harvard School of Public Health. She also currently serves as the coordinating director of the Program of Quantitative Genomics (http://www.hsph.harvard.edu/pqg).

Prof Lin's major statistical research interests lie in developing statistical methods for high-dimensional and correlated data. Examples of high-dimensional data include genomic and proteomic data in basic, population and clinical sciences. Examples of correlated data include longitudinal data, clustered data, hierarchical data and spatial data. She is particularly interested in developing statistical and computational methods for "omics" data in population-based studies, such as genetic epidemiology, genetic environmental sciences and clinical studies.

Prof Lin's specific areas of statistical research include statistical learning methods for high-dimensional data, dimension reduction, variable selection, nonparametric and semiparametric regression models, measurement error, mixed (frailty) models, estimating equations, missing data.

Prof Lin's areas of applications include cancer, genetic epidemiology, gene and environment, genome-wide association studies, genomics in population science, biomarkers and proteomics.

Alan Welsh

Prof Alan Welsh is the EJ Hannan Professor of Statistics and the head of the Centre for Mathematics and its Applications at the Australian National University. His current interests include statistical inference, modelling, robustness, nonparametric methods, adaptive estimation, and the analysis of survey data.

Alan obtained a BSc from the University of Sydney in 1982 and a PhD from the ANU in 1985. He was an Assistant Professor at the University of Chicago from 1984 to 1987 before he became a lecturer at the ANU. He held the Chair of Statistics at the University of Southampton in the UK from 2001 to 2003 before returning to the ANU as EJ Hannan Professor of Statistics. He was awarded the Moran Medal in 1990 and is a Fellow of the Institute for Mathematical Statistics and the American Statistical Association.

Conference Timetable

	SUNDAY, 28 AUGUST
1800–2000	**Conference Registration Opens/Social Mixer** Fale Pasifika Complex, Building 275, 20 Wynyard Street

	MONDAY, 29 AUGUST		
0850	**Opening [OGGB4]: Chris Triggs, Stuart McCutcheon**		
	Plenary session [OGGB4] *Chair: Alain Vandal*		
0900	*Genome variant calling: A statistical perspective* **Robert Gentleman** **Genentech**		
	Invited session	**Contributed sessions**	
	Methods 1 **[OGGB4]** *Chair: Thomas Yee*	**Applied statistics 1** **[OGGB5]** *Chair: John Pearson*	**Theory** **[OGGB3]** *Chair: Alan Wan*
0950	*Statistical disclosure from the margins of very high-dimensional genomic data* Thomas Lumley University of Auckland	*Hypothesis testing for navigation cues of homing pigeons* Chieh-Hsi Wu University of Auckland	*Coping in the absence of likelihoods* Steven Miller Waikato University
1010	*Prediction and confidence regions for ecological ordination plots* Matt Pawley Massey University	*Multilevel models for team sports* Denny Meyer Swinburne University of Technology	*Hierarchical covariance selection models* Insha Ullah Massey University
1030	**Morning Tea (30 minutes)**		
	Contributed sessions		
	Smoothing and applications **[OGGB4]** *Chair: Martin Upsdell*	**Applied bioinformatics** **[OGGB5]** *Chair: Stéphane Guindon*	**Fisheries** **[OGGB3]** *Chair: Marti Anderson*
1100	*Testing for log-concavity of densities* Martin Hazelton Massey University	*Estimating the relative roles of recombination & point mutation to the generation of single locus variants in* **C. jejuni** & **C. coli** Shoukai Yu Massey University	*Estimating the number of salmon returning to spawn* Russell Millar University of Auckland
1120	*Non-parametric estimation of covariate effects for spatial point processes* Rolf Turner University of Auckland	*Extracting knowledge from graphical models of microarray data* Beatrix Jones Massey University	*Assessment of lunar and indigenous fishing calendar predictions using recreational catch data of snapper* **Pagrus auratus** Ben Stevenson University of Auckland
1140	*Boundary kernels for adaptive kernel density estimators* Jonathan Marshall Massey University	*Reproducibility assessment & statistical quantification in mass spectral data from clinical proteomic study* Irene Zeng University of Auckland	*Fisheries in Ngati Kahungunu Rohe* Kylie Reiri Victoria University
1200	*On monotone regression* Berwin Turlach University of Western Australia	*Expression array analysis and interpretation* John Pearson University of Otago	*The potential of life-history models of Coho salmon dynamics* Sam McKechnie University of Auckland
1220	**Lunch (1 hour 10 minutes)**		

	Plenary session [OGGB4] *Chair: Alastair Scott*		
1330	*Alan's Statistical Saga: from Chapel Hill to Highland Park, with some Unexpected Directions, an Optimal Bet, and Trees that grow from People* **Nick Fisher** **University of Sydney**		
	Contributed sessions		
	Methods 2 **[OGGB4]** *Chair: Kate Lee*	**Sampling and surveys** **[OGGB5]** *Chair: Alan Welsh*	**Distributions** **[OGGB3]** *Chair: Robin Hankin*
1410	*Correlated winds and the risk of extreme power fluctuations* Barry McDonald Massey University	*Surveying in a time of earthquakes* Richard Penny Statistics New Zealand	*On a new subclass of the generalized inverse Gaussian distribution* Thomas Tran University of Auckland
1430	*Functional data classification via subspace projection* Pai-Ling Li Tamkang University	*Discarding the glass half full: An investigation into how households are discarded in the Household Economic Survey* Jessica Adams Statistics New Zealand	*Robust linear modeling using the hyperbolic distribution* Xinxing Li University of Auckland
1450	*Iterative methods in model fitting and diagnostics* Murray Jorgensen Waikato University	*LR tests with survival data from a sample survey* Alastair Scott University of Auckland	*A Poisson-Weibull model for comparing two independent populations* M. E. Ghitany Kuwait University
1510	**Afternoon tea (30 minutes)**		
	Contributed sessions		
	Multivariate statistics **[OGGB4]** *Chair: David Scott*	**Environment and ecology** **[OGGB5]** *Chair: Ian Westbrooke*	**Model selection and averaging** **[OGGB3]** *Chair: Beatrix Jones*
1540	*Generalizations of Ward's method* Alan Lee University of Auckland	*Where did you get that rat? Using genetics to study the origins and swimming patterns of invasive pests* Rachel Fewster University of Auckland	*Bootstrapped model-averaged confidence interval* Jiaxu (Jimmy) Zeng University of Otago
1600	*Modelling longitudinal functional response data* Steve Lane University of Melbourne	*Estimating species richness and similarity under different treatment conditions* Austina Clark Otago University	*Focused Information Criteria, model selection and model averaging in a tobit model with a non-zero threshold* Alan Wan City University of Hong Kong
1620	*Microdata for the masses: investigating the safety and utility of synthetic microdata* Mike Camden Statistics New Zealand	*The influence of near-pith material on the common climate signal* Maryann Pirie University of Auckland	*Threshold selection for modeling exceedances over a high threshold using a Bayesian measure of surprise* Kate Lee AUT University
1640	*Effects of heterogeneity of dispersions on multivariate distance-based permutation tests* Daniel Walsh Massey University	*Heterogeneous capture-recapture models with covariates: A partial likelihood approach for closed populations* Jakub Stoklosa University of Melbourne	*Graphical models: penalized likelihood or decomposable Bayesian?* Marie Fitch University of Auckland
1700	*Factor analysis for Hektner's Future Emotion Scale* Thewaporn Anwar SHSS - College Humanities	*Linking zeros to abundance in zero-inflated models of species count data* Adam Smith Massey University	*Model selection for three-mode three-way data* Lynette Hunt Waikato University
1720-1740	*Multivariate Gaussian processes* Robin Hankin AUT University	*A Bayesian state-space capture-recapture model for insular rat population dynamics* James Russell University of Auckland	
1800	**Reception in honour of Alan Lee** **Old Government House**		

	Tuesday 30 August		
0850	Housekeeping [OGGB4]		
	Plenary session [OGGB4] *Chair: Chris Triggs*		
0900	*Learning with sparsity* **Trevor Hastie** **Stanford University**		
	Plenary session on SAS data mining [OGGB4] *Chair: Thomas Yee*		
0950	*Feedback from the field: Two years later, are we any better off?* **Evan Stubbs** **SAS Australia & New Zealand**		
1030	Morning Tea (30 minutes)		
	Contributed sessions		
	Machine learning and data mining **[OGGB4]** *Chair: Trevor Hastie*	**Applied probability 1** **[OGGB5]** *Chair: Ilze Ziedins*	**Biostatistics** **[OGGB3]** *Chair: John Koolaard*
1100	*Tree-structured models for difference and change detection in a complex environment* Yong Wang University of Auckland	*Queueing up for enzymatic processing: correlations through coupled degradation* Ruth Williams University of California at San Diego	*Fitting regression models for response-biased problems* Gustavo Amorim University of Auckland
1120	*A learning experience* William Grant Massey University		*Case-cohort designs for the time failure data* Patricia Metcalf University of Auckland
1140	*Complexity measurement: A systematic approach to oversampling in imbalance data sets* Nafees Anwar Massey University	*Heavy-traffic control and pricing for systems with leadtime sensitive customers* Tava Olsen University of Auckland	*The effect of early growth and development on life-long health. A case study of the Helsinki Birth Cohort* Elena Moltchanova University of Canterbury
1200	*Classifying digital ink* Beryl Plimmer University of Auckland	*Mechanism design for wholesale market clearing under uncertainty* Golbon Zakeri University of Auckland	*Genome-wide association analysis with PLINK* Dug Yeo Han University of Auckland
1220	Lunch + Poster session (1 hour 10 minutes)		
	Plenary session [OGGB4] ***Chair: Andrew Balemi***		
1330	*Bayesian statistics: The Second Coming* **Wayne Stewart** **University of Auckland**		

	Contributed sessions		
	Statistics education: general [OGGB4] *Chair: Andrew Balemi*	**Applied probability 2 [OGGB5]** *Chair: Ilze Ziedins*	**Quantile regression [OGGB3]** *Chair: Alain Vandal*
1410	*Using media reports to promote statistical literacy for non-quantitative majors* Stephanie Budgett University of Auckland	*Models and measurements for cognitive radio systems* Peter Smith University of Canterbury	*On quantile regression* Arash Ardalan University of Auckland
1430	*Merging visions for statistics and mathematics education* Mike Camden Statistics New Zealand	*Unsteady volcanic modelling* Mark Bebbington Massey University	*An exploratory approach of modeling nonstationarity for spatial quantile-based data analysis* Vivian Yi-Ju Chen Tamkang University
1450	*The introductory statistics course and inference* Maxine Pfannkuch University of Auckland	*A regionalization method based on a cluster probability model* Paul Cowpertwait AUT University	*On testing convex transform ordering* Muhyiddin Izadi University of Auckland
1510	**Afternoon tea (30 minutes)**		
	Contributed sessions		
	Statistics education: Bayesian [OGGB4] *Chair: Andrew Balemi*	**Applied probability 3 [OGGB5]** *Chair: David Scott*	**Mixture models and methods 1 [OGGB3]** *Chair: Yong Wang*
1540	*Bayesian statistics in NZ universities undergraduate curriculum* Bill Bolstad Waikato University	*User optimal policies for a stochastic transportation network* Heti Afimeimounga University of Auckland	*Mixture survival models for identifying infant and senescent mortality* Rebecca Green Massey University
1600	*Teaching MCMC in Bayesian statistics: What goes on behind the algorithm* Wayne Stewart University of Auckland	*Sequential analysis of the Moran Process* Peter Green Otago University	*Application of a non-linear mixed model to stress-strain relationship of flax fibres* Chikako van Koten AgResearch
1620-1720	**AGM [OGGB5]** **Official start: 16:30** **Please arrive early**		
1800	**Conference Dinner** **The Bluestone Room** **9—11 Durham Lane**		

	Wednesday 31 August		
0850	**Housekeeping [OGGB4]**		
	Plenary session [OGGB4] *Chair: Thomas Lumley*		
0900	*Statistical issues and challenges in analyzing high-throughput 'omics data in population-based studies* **Xihong Lin** **Harvard University**		
	Contributed sessions		
0950	**Genetics** **[OGGB4]** *Chair: Thomas Lumley*	**Categorical data analysis 1** **[OGGB5]** *Chair: Ivy Liu*	**Copulas** **[OGGB3]** *Chair: David Scott*
1010	*iDArTs: thinking differently about genetic markers to unlock new resources* Emma Huang CSIRO	*Using algebraic methods to test the independence of ethnicity and resolution for drug-related crimes in NZ* Irene van Woerden University of Canterbury	*Finite mixtures of Archimedean copulas* Renate Meyer University of Auckland
1030	*Diffusion approximation and maximum entropy* Jing Liu University of Auckland	*Row-column association models* Thomas Yee University of Auckland	*Sharp bounds on a class of copulas and quasi-copulas* Heydar Ali Mardani-Fard Yasouj University
	Morning Tea (30 minutes)		
	Contributed sessions		
	Design **[OGGB4]** *Chair: Xihong Lin*	**Categorical data analysis 2** **[OGGB5]** *Chair: Beatrix Jones*	**Bayesian statistics 2** **[OGGB3]** *Chair: Kate Lee*
1100	*MDS-optimal supersaturated designs* Arden Miller University of Auckland	*Biclustering and pattern detection for binary and count data* Shirley Pledger Victoria University	*Benchmarking WinBUGS and OpenBUGS to independent Metropolis-Hastings with heavy-tailed candidate for GLMs : Part 1* Toufiq Al Gheilani Waikato University
1120	*Sensitivity of EWMA control charts* Saddam Akber Abbasi University of Auckland	*Biclustering models for ordinal data* Eleni Matechou Victoria University	*Benchmarking WinBUGS and OpenBUGS to independent Metropolis-Hastings with heavy-tailed candidate for GLMs : Part 2* Bill Bolstad Waikato University
1140	*Designing a two-phase experiment for many treatments and few replicates in blocks of size two* Kathy Ruggiero University of Auckland	*Goodness-of-fit tests for logistic regression models using stochastic processes* Ivy Liu Victoria University	*Effects of incorporating GPS routing information into current traffic models* Katharina Parry Massey University
1200	*Complete allocation sampling: An efficient and easily implemented adaptive sampling design* Jennifer Brown University of Canterbury	*Modelling strategies for repeated multiple response data* Thomas Suesse University of Wollongong	*A principle for quality control in Bayesian analyses* Robin Willink
1220	**Lunch (1 hour 10 minutes)**		
	Plenary session [OGGB4] *Chair: Alastair Scott*		
1330	*Handling nonresponse when fitting models to survey data* **Alan Welsh** **Australian National University**		

	Contributed sessions		
	Applied statistics 2 [OGGB4] *Chair: Alastair Scott*	**Statistical computing [OGGB5]** *Chair: Robert Gentleman*	**Mixture models and methods 2 [OGGB3]** *Chair: Yong Wang*
1410	*Attrition in the Longitudinal Immigration Survey: New Zealand (Wave1 to Wave3)* Maoxin Luo Statistics New Zealand	*What's in a name?* Paul Murrell University of Auckland	*Fitting mixture models made easy* Murray Jorgensen Waikato University
1430	*An assumption-free small-sample procedure for the difference in medians* Robin Willink	*InfoDecompuTE: an R package for information decomposition in two-phase experiments* Kevin Chang University of Auckland	*GARCH model with scale normal mixture errors* Michael Kao University of Auckland
1450	*Measuring the price movements of used cars and residential rents in the New Zealand Consumers Price Index* Frances Krsinich Statistics New Zealand	*Software for distributions* David Scott University of Auckland	*Mode-based clustering using nonparametric mixture models* Xuxu Wang University of Auckland
1510	**Afternoon tea (30 minutes)**		
1540- 1600	**Closing and student prizes announcement [OGGB4]**		

Oral Presentation Abstracts

MONDAY, 29 AUGUST	
	Plenary session [OGGB4] *Chair: Alain Vandal*
0900- 0920	**Robert Gentleman** **Genentech**

GENOME VARIANT CALLING: A STATISTICAL PERSPECTIVE

Robert Gentleman*
Genetech
E-mail: gentleman.robert@gene.com

The advent of low cost sequencing has given rise to an enormous amount of sequence data. It is possible to sequence many hundreds of genomes at substantial depth. These data provide us with a basis for estimating the genome (usually diploid) of the individual. However, the technology still has limitations and statistical methods are essential for ensuring that the estimates are reasonable and comparable. Varying error rates and coverage both between and within individual samples make this estimation problem challenging. We will discuss some of them and present our current results and algorithms.

*****

0950- 1030	**MONDAY, 29 AUGUST** **Contributed Session: Methods 1 [OGGB4]** *Chair: Thomas Yee*

STATISTICAL DISCLOSURE FROM THE MARGINS OF VERY HIGH-DIMENSIONAL GENOMIC DATA

Thomas Lumley*
University of Auckland
E-mail: t.lumley@auckland.ac.nz

Research on identifying DNA from mixtures has recently shown that marginal allele frequencies at a very large number of loci provide sufficient information to determine whether an individual is part of a sample. The same approach can be used to determine case status from genotype in a case-control study if the marginal allele frequencies for cases and controls are known. In this paper we explain this counterintuitive finding in terms of overfitting and show how it applies much more generally to regression models and obtain accurate closed-form approximations for the accuracy of prediction. We contrast this form of disclosure, which provides only probabilistic bounds, with previous research on regression and loglinear models, which exploit non-negativity and discreteness constraints.

*****

PREDICTION AND CONFIDENCE REGIONS FOR ECOLOGICAL ORDINATION PLOTS

Matthew Pawley* and Marti Anderson
Massey University
E-mail: m.pawley@massey.ac.nz

The high-dimensional nature of ecological data requires ordination techniques to reduce the data cloud down to just two or three dimensions that can be plotted in order to visualise patterns. Traditional techniques of inference for classical ordination methods (PCA, CVA) draw circles or ellipses (ellipsoids if the ordination is plotted in 3-D) to show confidence regions, which rely on the assumption of multivariate normality, which is unrealistic in ecology and inappropriate for non-parametric MDS ordination. As a consequence, researchers typically just present the sample units as a configuration of points - but in what region(s) of the plot might new samples from a population be expected to lie? Using ecological examples , we will present (using a simple intuitive framework) the development of data-driven (non-parametric) methods for drawing empirical regions for valid statistical inference directly in rdination plots such as PCA, PCO or MDS.

*****

0950- 1030	**MONDAY, 29 AUGUST** **Contributed Session: Applied statistics 1 [OGGB5]** *Chair: John Pearson*

HYPOTHESIS TESTING FOR NAVIGATION CUES OF HOMING PIGEONS

Chieh-Hsi M. Wu*, Megan M. Marcotte, Joshua M. Guilbert and Rachel M. Fewster
*University of Auckland**
E-mail: cwu080@aucklanduni.ac.nz

For hundreds of years, navigation has been studied in homing pigeons, *Columbia livia*, due to their ability to home rapidly upon demand. The propensity of the birds to use a variety of stimuli while navigating requires an analysis that can determine if the birds use more than one navigational stimulus and/or are affected by behavioral barriers. Here we present a framework which aims to test the influence of multiple geographical factors (e.g. the Earth's magnetic field intensity, roads, topography, and forest cover) on the pigeons' flight trajectories. The method is applied to a dataset of flight trajectories of homing pigeons recorded by GPS-based tracking devices. We use time series models to simulate a large number of pigeon flight trajectories to determine whether the correlations between the real pigeon tracks and the potential navigational stimuli are stronger than expected by random chance.

*****

MULTILEVEL MODELS FOR TEAM SPORTS

D. Meyer*, K. Jackson, J. Cook, P. Gastin, E. Huntsman
*Swinburne University of Technology**
E-mail: lssfinance@swin.edu.au

There are at least two reasons why multi-level models are important for the analysis of individual performance in team sports. The first relates to missing information. It is common for players to miss matches in the case of team sports. Multi-level models can handle repeated measures data with missing values. The second reason for using multi-level models is the nested nature of the data. Firstly there is the nesting of match performances for individual players over a season and secondly there is the nesting of individual performances for each team within a season. Multi-level models allow the variation contribution of each level to be assessed and they allow independent variables for all levels to be incorporated in the model as main effects. In addition the contribution of moderation effects between levels and heterogeneity effects can be tested under a variety of response distributions. In this paper individual player performance and the risk of injury are explored in the context of Australian Rules Football using multi-level models.

*****

0950- 1030	**MONDAY, 29 AUGUST** **Contributed Session: Theory [OGGB3]** *Chair: Alan Wan*

COPING IN THE ABSENCE OF LIKELIHOODS

Steven Miller*
University of Waikato
E-mail: s.miller@waikato.ac.nz

Many modern problems investigate complex systems where it is infeasible to construct likelihoods for inference on the many parameters involved. This situation often arises in applications using detailed scientific models to try to recreate data from systems via simulation. One approach to cope with this complexity is "indirect inference". This approach is useful where an analytic likelihood is not available, but it is possible to simulate data according to a scientific model based on information that can be specified in terms of the parameters of interest. Prior information regarding these parameters can be incorporated into the simulation. Simulation can be expensive in terms of time or resources. We are interested in incorporating a statistical model for simulation output that can act as a descriptive approximation of the underlying system so that we can express our uncertainty concerning the estimated parameter values that result from a potentially small number of simulations.

*****

HIERARCHICAL CAVARIANCE SELECTION MODELS

Insha Ullah*, Beatrix Jones
Massey University
E-mail: i.ullah@massey.ac.nz

The off-diagonal zero elements in the inverse covariance matrix are of particular interest in order to know the conditional independence structure in a multivariate normal distribution. A zero off-diagonal element will mean that the corresponding variables are independent given all the other variables in the model. The maximum likelihood estimate of the inverse covariance matrix never appears with an off-diagonal element exactly equal to zero. A model selection procedure is therefore required. The penalized likelihood has been effectively used to force some of the small coefficients exactly equal to zero. This use of penalized likelihood is so far restricted to the situation where variables are measured on the same group. In this paper the method of penalized likelihood is extended to the two-level normal hierarchical models, where together with an estimate for within-group inverse covariance matrix, a separate estimate is needed for between-group inverse covariance matrix. For example, consider the study of patients at different hospitals or patients who had their data processed at different laboratories. We wish to examine the relationship between variables (e.g., gene expression measurements) while controlling for differences between hospitals/laboratories. On the other hand, if the variation within a group denotes the measurement error i.e. each group contains repeated measurements on a single subject then the interest will be in developing a model for between-group variation while controlling for within-group variation. If there exist any grouping then the idea is to infer both within-group and between-group inverse covariance matrices. This is achieved via EM-algorithm. The performance of the method is illustrated by a number of simulated examples and a real gene expression data set.

*****

1100- 1220	**MONDAY, 29 AUGUST** **Contributed Session: Smoothing and applications [OGGB4]** *Chair: Martin Upsdell*

TESTING FOR LOG-CONCAVITY OF DENSITIES

Martin Hazelton*
Massey University
E-mail: m.hazelton@massey.ac.nz

Many commonly encountered densities are log-concave, including normal, Laplace, Gumbel and logistic densities. The class of all log-concave densities (whether univariate of multivariate) allows considerable flexiblility for modelling purposes, but has sufficient structure to support of workable theory. For example, a random sample from a continuous distribution will provide a unique log-concave (nonparametric) maximum likelihood estimator of the density with probability 1. It is therefore of interest to test for log-concavity of a density. This problem has received significant attention for univariate data, but limited progress has been made in the multivariate case. We describe a multivariate methodology using kernel density estimation, where the test statistic is the smallest (isotropic) bandwidth for which the estimate is log-concave. We illustrate its use on measurements taken on digitized images of fine needle aspirates of breast masses, intended to help distinguish benign cases from malignant.

*****

NON-PARAMETRIC ESTIMATION OF COVARIATE EFFECTS FOR SPATIAL POINT PROCESSES

Rolf Turner*
University of Auckland
E-mail: r.tuner@auckland.ac.nz

Frequently it is obvious that the intensity of a spatial point process depends upon one or more covariate effects. Almost equally frequently it is unclear just what the actual nature of the dependence is. In this talk I will describe one method for obtaining a non-parametric estimate of a single covariate effect. This may, with a bit of luck, suggest a parametric model. The ideas are intriguingly simple. A nice feature is that an estimate of the variance of the effect estimate is also readily available.

*****

BOUNDARY KERNELS FOR ADAPTIVE KERNEL DENSITY ESTIMATORS

J C Marshall* and M L Hazelton
*Massey University**
E-mail: J.C.Marshall@massey.ac.nz

In many applications of kernel density estimation, the data have a highly non-uniform distribution and are confined to a compact region. In addition, many applications have a significant portion of the data located near the region boundary. Standard fixed-bandwidth density estimates typically struggle to cope with the spatially variable smoothing requirements, and in addition are subject to excessive bias at the boundary of the region. Adaptive kernel estimators can address the variable smoothing requirement, and with the use of a boundary kernel can account for the boundary bias. A simple linear boundary kernel which reduces the asymptotic order of the bias at the boundary is introduced, and theoretical and numerical results presented.

*****

ON MONOTONE REGRESSION

Berwin Turlach*
University of Western Australia
E-mail: Berwin.Turlach@gmail.com

Many practical application require a monotone regression function to be fitted to data. Several smoothing methods for monotone regression have been proposed and a selective overview of these will be given. Time permitting, the talk will also compare some of these non-parametric approaches to parametric approaches, e.g. via monotone polynomials.

*****

1100-1220	**MONDAY, 29 AUGUST** **Contributed Session: Applied bioinformatics [OGGB5]** *Chair: Stéphane Guindon*

ESTIMATING THE RELATIVE ROLES OF RECOMBINATION AND POINT MUTATION TO THE GENERATION OF SINGLE LOCUS VARIANTS IN *CAMPYLOBACTER JEJUNI* AND *CAMPYLOBACTER COLI*

Shoukai Yu*; Paul Fearnhead; Barbara Holland; Patrick Biggs; Martin Maiden and Nigel French
IVABS, *Massey University**
E-mail: s.yu1@massey.ac.nz

Single locus variants (SLVs) are bacterial sequence types that differ at only one of the seven canonical MLST loci. Estimating the relative roles of recombination and point mutation in the generation of new alleles that lead to SLVs is helpful in understanding how organisms evolve. An expectation-maximization (EM) algorithm was applied to estimate the relative rates of recombination and mutation for *Campylobacter jejuni* and *Campylobacter coli* for seven different housekeeping loci from publically available multilocus sequence typing (MLST) data. The probability of recombination generating a new allele that leads to an SLV is estimated to be roughly three times more than that of mutation. This estimate is much larger than estimates of the relative rate of recombination to mutation calculated from more distantly related isolates using MLST data. One explanation for this is that purifying selection plays an important role in the evolution of *Campylobacter*. A simulation study was performed to test performance of our method under a range of biologically realistic parameters.

*****

EXTRACTING KNOWLEDGE FROM GRAPHICAL MODELS OF MICROARRAY DATA

Beatrix Jones* and Cristin Print
*Massey University**
E-mail: m.b.jones@massey.ac.nz

This talk considers network models of high dimensional microarray data and considers the problem of extracting biological information from them. There has been extensive work on inferring high dimensional network structures, but less on what to do with these hard-won structures once they have been obtained. The talk is illustrated with a data set of expression measurements from breast cancer tumors; we look at the 'fingerprint' that some well understood pathways leave in a Gaussian graphical model inferred with the adaptive lasso. This fingerprint is less straightforward than one might hope, and verifying hypotheses about individual gene function is difficult. However, the network can be usefully combined with information from the TRANSFAC database, a set of DNA sequence-based predictions about transcription factor targets. We demonstrate a technique for identifying TRANSFAC sets that co-vary across breast cancer tumors-i.e., are strongly connected in the Gausian graphical model. We also consider refining the TRANSFAC sets to include only genes whose subgraph is more densely connected than randomly selected genes.

*****

REPRODUCIBILITY ASSESSMENT AND STATISTICAL QUANTIFICATION IN MASS SPECTRAL DATA FROM CLINICAL PROTEOMIC STUDY

Irene Zeng*, Sharon Browning and Ralph Stewart
*University of Melbourne**
E-mail: zeng@stat.auckland.ac.nz

Use of mass spectrometry to investigate disease-associated proteins among thousands of candidates simultaneously creates challenges with the evaluation of operational and biological variation. Traditional statistical methods, which evaluate reproducibility of a single feature, are likely to provide an inadequate assessment of reproducibility. This presentation proposes a systematic approach for evaluation of the global reproducibility of multi-dimensional mass spectral data at the post- identification stage. It will also discuss the quantification strategy for protein in the reproducibility assessment. The proposed systematic approach combines dimensional reduction and permutation to test and summarize the reproducibility. First, principal component analysis is applied to the mean quantities from identified features of paired replicated samples. An eigenvalue test is used to identify the number of significant principal components which reflect the underlying correlation pattern of the multiple features. Second, a simulation-based permutation test is applied to the derived paired principal components. Third, a modified form of Bland Altman or MA Plot is produced to visualize agreement between the replicates. Last, a discordance index is used to summarize the agreement. Application of this method to data from both a cardiac LC-MS/MS experiment with iTRAQ labelling and simulation experiments derived from an ovarian cancer SELDI-MS experiment demonstrate that the proposed global reproducibility test is sensitive to the simulated systematic bias when the sample size is above 15. The two proposed test statistics (max t statistics and a sign score statistic) for the permutation tests are shown to be reliable.

*****

EXPRESSION ARRAY ANALYSIS AND INTERPRETATION

John F Pearson*, Anna P Pilbrow, Les McNoe, Wendy E Sweet, WH Wilson Tang, Richard W Troughton, A Mark Richards, Christine S Moravec and Vicky A Cameron
*University of Otago, Christchurch**
E-mail: john.pearson@otago.ac.nz

Expression arrays are an established component of the molecular biology toolkit and a wide range of statistical techniques are available to process, analyse and interpret the data. An emerging problem is integrating these datasets with data from other platforms. Polymorphisms, particularly those identified by genome wide association studies, can be investigated for association with differential expression levels for individual genes and sets of genes. Quality control to remove technical effects are critical when comparing cross platform data and the use of appropriate genetic models in particular the log-additive or per-allele model can aid identification of individual genes and genesets. Techniques will be illustrated with data from healthy heart donor tissue courtesy of the Cleveland Heart Study.

*****

1100- 1220	**MONDAY, 29 AUGUST** **Contributed Session: Fisheries [OGGB3]** *Chair: Marti Anderson*

ESTIMATING THE NUMBER OF SALMON RETURNING TO SPAWN

Russell Millar*, Sam McKechnie and Chris Jordan
*University of Auckland**
E-mail: r.millar@auckland.ac.nz

This talk presents a new estimator of the number of salmon spawners in a stream, derived from weekly counts of the number of spawners observed in the stream. The estimator that is currently most widely used does not provide an estimate of its precision. Other estimators do, but are too complex to be widely used. The new estimator is derived from a generalized linear model and so it is easy to implement, and to obtain its estimate of precision. It is also shown to have good relative performance.

*****

ASSESSMENT OF LUNAR AND INDIGENOUS FISHING CALENDAR PREDICTIONS USING RECREATIONAL CATCH DATA OF SNAPPER *PAGRUS AURATUS*

Ben Stevenson* and Russell Millar
University of Auckland
E-mail: bste085@aucklanduni.ac.nz

Recreational fishers in New Zealand often make use of lunar and indigenous Maori fishing calendars in order to predict fishing success on specific days. Little is known as to the performance of such predictions and whether or not they hold any practical use to the everyday angler. Here, generalised nonlinear mixed effects models are fitted in AD Model Builder (ADMB) using recreational fishing data provided by the Ministry of Fisheries. These data are based on catches of snapper *Pagrus auratus*, New Zealand's most popular recreational species, and take the form of diary and boatramp surveys. Some evidence is found for the effect of lunar phase on fishing success, and also to support some aspects of the performance of the Maori fishing calendar predictions. The magnitudes of these effects are small, however, casting doubt on the practicality of lunar based fishing predictions. These analyses also allow for the assessment of ADMB in its capabilities of explaining large datasets using complex non-linear mixed effects models.

*****

FISHERIES IN NGATI KAHUNGUNU ROHE

Kylie Reiri*
Victoria University of Wellington
E-mail: kyliereiri@gmail.com

Background: There are three (legal) mechanisms through which kai moana are taken from the coastal waters of Aotearoa New Zealand: Customary, Recreational and Commercial. Data of high quality on the commercial catch are reported to the Ministry of Fisheries, whereas data on the Recreational catch are very patchy, being based on voluntary reporting. Customary catch data are collected through kaitiaki, the iwi representatives who issue customary permits, and summary data are reported to the Ministry of Fisheries. Currently these data are published only at very high aggregate level (Fisheries or Quota Management Areas), which do not allow for analysis or monitoring at finer spatial scales. Te Ohu Kai Moana (TOKM), who oversee Maori interests in fisheries at a national level, are developing new methods for collecting customary and recreational data from iwi, and are piloting these in the Ngāti Kahungunu Rohe in 2010.

Research Goals: The overall aim of this project is to provide an improved view of the fisheries data that are currently being collected in the Ngāti Kahungunu rohe, including reporting results at a finer geographical scale. Specific goals are: 1. Assess the quality of existing customary data (from Kaitiaki) and (if possible) recreational data (from voluntary reports by individuals and boat clubs); 2. Combine these data with the commercial data from the Ministry of Fisheries to provide total catch estimates, as an estimate of stocks of key species (e.g. paua, kina, snapper, kahawai), including temporal and geographical variation.

*****

THE POTENTIAL OF LIFE-HISTORY MODELS OF COHO SALMON DYNAMICS

Sam McKechnie* and Russell Millar
University of Auckland
E-mail: mckechniesam@yahoo.com.au

Models of salmon population dynamics are frequently used to guide conservation management decisions. Traditionally models have been simple, ignoring several sources of uncertainty, and were typically fitted to datasets from only one part of the life-cycle (spawning adults). Recently, government agencies have attempted to expand monitoring programmes to include several important stages of the life-cycle. Coupled with more sophisticated modelling tools there is now the potential to significantly improve inferences from salmon models, though whether this potential will be met has yet to adequately assessed. Herein we fit state-space models to stage-structured coho salmon data from Lobster Creek, Oregon, with the intention of identifying the value of these more complex models. This includes comparing the breadth of questions that can be answered and also the precision and accuracy of predictions of abundance.

*****

	MONDAY, 29 AUGUST
	Plenary Session [OGGB4]
	Chair: Alastair Scott
1330-	**Nick Fisher**
1410	***University of Sydney***

ALAN'S STATISTICAL SAGA: FROM CHAPEL HILL TO HIGHLAND PARK, WITH SOME UNEXPECTED DIRECTIONS, AN OPTIMAL BET, AND TREES THAT GROW FROM PEOPLE

Nicholas Fisher*
University of Sydney
E-mail: nickf@maths.usyd.edu.au

Alan Lee has collaborated on research projects with many people over the years. This paper looks at some of the projects that he has been involved in, and attempts to answer the question: 'Is this really the person who started out his professional life practising Abstract Harmonic Analysis in the dark?'.

*****

1410-1510

MONDAY, 29 AUGUST
Contributed Session: Methods 2 [OGGB4]
Chair: Kate Lee

CORRELATED WINDS AND THE RISK OF EXTREME POWER FLUCTUATIONS

Barry McDonald*
*Massey University**
E-mail: b.mcdonald@massey.ac.nz

Wind power is set to generate an increasing percentage of New Zealand's electricity supply. The minute-by-minute variability of power output from a wind farm is high, due to natural variation in wind speed and a highly nonlinear relationship between wind and the power output. The national power grid is currently protected from the effects of sudden extreme power fluctuations (surges or brownouts) by control measures, but the control measures may prove inadequate as more wind farms come online. Winds affecting different wind farms are correlated, with time lags, and hence so is power output but in a highly nonlinear way. I discuss a method of incorporating these correlations into the task of quantifying the risk of extreme power events.

*****

FUNCTIONAL DATA CALSSIFICATION VIA SUBSPACE PROJECTION

Pai-Ling Li* and Che-Chiu Wang
Tamkang University
E-mail: plli@stat.tku.edu.tw

A subspace projected functional classification method is proposed for classifying curves or functional data. We will present a framework of subspace projected functional data classification based on the functional random-effects model. Curves are embedded in the cluster subspace spanned by a mean function and eigenfunctions of the covariance kernel. The cluster membership prediction for each curve attempts to minimize the distance between the observed and predicted curves via subspace projection among all the clusters. The proposed method accounts for both the means and the modes of variation differentials between clusters while the other classical approaches mainly consider the differences in mean functions. The performance of proposed approach will be demonstrated through simulation studies and a real data example.

*****

ITERATIVE METHODS IN MODEL FITTING AND DIAGNOSTICS

Murray Jorgensen*
Waikato University
E-mail: maj@waikato.ac.nz

I will delve into some of the mathematics behind iterative fitting algorithms and discuss a matrix with applications to both the fitting process and constructing leverage diagnostics for nonlinear statistical models.

*****

1410- 1510	**MONDAY, 29 AUGUST** **Contributed Session: Sampling and surveys [OGGB5]** *Chair: Alan Welsh*

SURVEYING IN A TIME OF EARTHQUAKES

Richard Penny*
Statistics New Zealand
E-mail: Rchard.Penny@stats.govt.nz

As you are all aware Christchurch has experienced a series of earthquakes since September 4 last year. These earthquakes have seriously affected many of the Christchurch respondents to Statistics New Zealand's surveys. Also a large amount of the work producing outputs from the surveys of New Zealand businesses is done in Statistics New Zealand's Christchurch office. Continuing to produce reliable, consistent, and timely outputs is an ongoing challenge for both Statistics New Zealand and its respondents. The work required to meet this challenge has provided interesting insights into what can be important in producing survey outputs to a suitable quality standard. Some of these insights will be of interest to both those producing and using survey outputs.

*****

DISCARDING THE GLASS HALF FULL: AN INVESTIGATION INTO HOW HOUSEHOLDS ARE DISCARDED IN THE HOUSEHOLD ECONOMIC SURVEY

Jessica Adams*
Statistics New Zealand
E-mail: Jessica.adams@stats.govt.nz

The Household Economic Survey (HES), run every three years, collects comprehensive information relating to household expenditure and income, as well as a range of demographic information on individuals and households. In the two years between the three-yearly HES, a shortened version of the survey (HES (Income)) is conducted to collect information on income and housing expenditure. The core use of HES is to provide detailed expenditure data for the revision of the Commodity Price Index (CPI). It is also used in tax modelling and development, and monitoring and evaluating policies. In HES the discard process removes households who are ineligible or who have not responded sufficiently from the results. A study has been undertaken to determine whether the discard process is working efficiently. It was found that improvements could be made to reduce the amount of information lost due to discarding households. This presentation covers the results of this study and the key improvements.

*****

LR TESTS WITH SURVIVAL DATA FROM A SAMPLE SURVEY

Alastair Scott* and Thomas Lumley
University of Auckland
E-mail: a.scott@auckland.ac.nz

With the advent of large-scale health surveys such as NHANES, methods for the analysis of survey data have became increasingly important ? PubMed lists almost 8000 papers published in the last 5 years with NHANES in the abstract, for example. We review the work of Rao and Scott (1981, 1984) on likelihood-ratio tests for contingency tables and show that the key results are valid in a much wider setting. We illustrate with tests for the Cox proportional hazards model fitted to survey survival data.

*****

1410- 1510	**MONDAY, 29 AUGUST** **Contributed Session: Distributions [OGGB3]** *Chair: Robin Hankin*

ON A NEW SUBCLASS OF THE GENERALIZED INVERSE GAUSSIAN DISTRIBUTION

Thomas Tran*
University of Auckland
E-mail: tt.tran@auckland.ac.nz

On a new subclass of the GIG The generalized inverse Gaussian (GIG) is a very flexible family of continuous distributions because its density function contains the modified Bessel function of the second kind $K_\lambda(z)$, which is also known as the modified Bessel function of the third kind, $\lambda \in \mathbb{R}$ and $z > \mathbb{R}^+$. Various subclasses of the GIG are obtainable by changing the value of λ. However, research effort has only been focused on the inverse Gaussian (IG) distribution $\left(\lambda = -\frac{1}{2}\right)$ because of its tractability, applicability and close relation with the Gaussian distribution. The IG is the only subclass whose cumulative distribution function has been derived explicitly in the literature. This talk is to propose a new subclass of the GIG with $\lambda = \frac{1}{2}$ named the complimentary inverse Gaussian (CIG). Explicit formulae for the cumulative and complimentary cumulative distribution function of the IG and CIG are introduced together with some appealing statistical features of the latter.

*****

ROBUST LINEAR MODELLING USING THE HYPERBOLIC DISTRIBUTION

Xinxing Li*
University of Auckland
E-mail: xli053@aucklanduni.ac.nz

The Generalized Hyperbolic distribution family possesses the non-Gaussian characters which typically are present in financial asset returns data which often are time series data. As a special case of the family, the hyperbolic distribution features semi-heavy tails and exhibits skewness for certain parameter values. Linear modeling with hyperbolic error provides an approach for analyzing data with heavy-tailed and skewed errors. This talk describes the implementation of R routines for implementing linear models with hyperbolic errors, and gives examples of the use of these routines.

*****

A POISSON-WEIBULL MODEL FOR COMPARING TWO INDEPENDENT POPULATIONS

M.E. Ghitany*
Kuwait University
E-mail: meghitany@yahoo.com

We propose a Poisson-Weibull model for comparing two independent populations. Specifically, we are interested in estimating the reliability parameter $R = P(X > Y)$ where X and Y are independent Poisson-Weibull random variables. This parameter R provides a general measure of difference between two populations and has applications in different areas. The confidence interval of R, based on the maximum likelihood method, is developed. The performance of this confidence interval is studied through extensive simulations. A numerical example, based on real data representing failure stresses of single carbon fibers of two different lengths is presented to illustrate the implementation of the proposed procedure.

*****

1540- 1700	MONDAY, 29 AUGUST Contributed Session: Multivariate statistics [OGGB4] *Chair: David Scott*

GENERALIZATIONS OF WARD'S METHOD

Alan Lee* and Bobby Wilcox
University of Auckland
E-mail: lee@stat.auckland.ac.nz

In this talk, we consider several generalizations of the popular Ward's method for agglomerative hierarchical clustering. Our work was motivated by clustering software, such as the R function hclust, which accepts a distance matrix as input and applies Ward's definition of inter-cluster distance to produce a clustering. The standard version of Ward's method uses squared Euclidean distance to form the distance matrix. We explore the properties and effect on the clustering of using other definitions of distance, such as the Minkowski distance and powers of the Minkowski distance. We explore the effect of these on several examples and find that using powers of the Manhattan metric is particularly effective.

*****

MODELLING LONGITUDINAL FUNCTIONAL RESPONSE DATA

Steve Lane*
University of Melbourne
E-mail: s.lane@ms.unimelb.edu.au

Functional response data appear in many biological studies. Motivated by the prediction of tree diameter distributions, we present a nonparametric method that allows prediction of longitudinal functional responses, accounting for (possible) covariate information.

*****

MICRODATA FOR THE MASSES: INVESTIGATING THE SAFETY AND UTILITY OF SYNTHETIC MICRODATA

Mike.Camden*, Clare Bycroft and Crystal Symes
Statistics New Zealand
E-mail: mike.camden@stats.govt.nz

We describe an innovative data product that aims to benefit statistical educators at universities and elsewhere: a set of 100 Synthetic Unit Record Files (SURFs), where each SURF contains 11,000 records and 6 variables (categorical and numerical). Each SURF retains many of the properties of the parent dataset, which is a subset of the NZ Income Survey 2003 dataset. Taken together, the set also contains inferential information about the parent dataset: a user can replicate any statistic 100 times and create its sampling distribution. The process to create this set of SURFs results from work done by Patrick Graham under the Official Statistics Research programme. It takes sets of draft SURFs, and measures two properties: safety for public release, and utility for users seeking information about income and some of the variables that it varies with. Safety here means that the SURF sets can be released publically, while complying with the confidentiality requirement of the Statistics Act (1975). We describe the methods we used to test for safety, the unexpected issues that arose with these SURFs, the further data modifications that we made, and the final 'hacking' tests. Utility requires that the SURF sets contain much of the structure of the parent dataset: properties that appear in summary statistics and parameter estimates, and the unique features that are important to the Income Survey context. We use graphical methods to describe the information disturbance.

*****

EFFECTS OF HETEROGENEITY OF DISPERSIONS ON MULTIVARIATE DISTANCE-BASED PERMUTATION TESTS

Daniel Walsh*[1] and Marti Anderson*[2]
[1]Institute of Information and Mathematical Sciences (IIMS) and [2]New Zealand Institute for Advanced Study (NZIAS),
Massey University, Albany Campus, Auckland, New Zealand
E-mail: d.c.walsh@massey.ac.nz and m.j.anderson@massey.ac.nz

In recent years, a number of robust tests to compare groups of multivariate sample units have become widely used in the biological and ecological sciences. These tests, including ANOSIM and PERMANOVA, construct ANOVA-like test statistics from the matrix of distances (or dissimilarities) among samples and obtain p-values to test the general null hypothesis of 'no differences among groups' with random permutations, assuming only exchangeability for the one-way case. We did a simulation study to investigate the effect of heterogeneity of multivariate dispersions on the rejection rates (size) of these tests (for $\alpha = 0.05$), and compared their robustness to that of a classical MANOVA test (Pillai's trace). Pillai's trace and PERMANOVA were very robust to heterogeneity for balanced sample sizes ($n_1 = n_2$), while ANOSIM was not. For the unbalanced case ($n_1 < n_2$), however, all tests had inflated type I error when the group with the large variance had the smaller sample size. Tests were overly conservative when the group with the large variance had the larger sample size. Counter to intuition under the central limit theorem, increasing the total sample size did not help, as exchangeability was still violated; type I error became worse for ANOSIM for larger total numbers of samples, and error rates for PERMANOVA and Pillai's trace remained constant for a given ratio of (n_1/n_2). Increasing correlation structure and increasing numbers of variables can also exacerbate these problems. We propose some potential solutions to isolate differences in centroids vs differences in dispersions for multivariate hypothesis-testing.

[1]Institute of Information and Mathematical Sciences (IIMS) and [2]New Zealand Institute for Advanced Study (NZIAS), Massey University, Albany Campus, Auckland, New Zealand.

*****

FACTOR ANALAYSIS FOR HEKTNER'S FUTURE EMOTION SCALE

Thewaporn Thimasarn Anwar*, Jackie Sanders and Robyn Munford
SHSS – College Humanities
E-mail: dev-null@stat.auckland.ac.nz

This paper presents the findings of reliability of Hektner's future emotion scale, a questionnaire administered to 1500 young people from all over New Zealand who are involved in Pathways to Resilience study in 2009. Our first concern of was to establish the scale's internal consistency. The results showed a low Cronbach's alpha value for the whole scale meaning that our scale had very low internal consistency. The Hektner's future emotion scale clearly split into two different areas which are; positive and negative, we can also see that our samples tended to give very high rating to the positive question and give very low rating for negative question.

*****

MULTIVARIATE GAUSSIAN PROCESSES

Robin Hankin*
AUT University
E-mail: robin.hankin@aut.ac.nz

The "emulator" technique allows one to predict the output of complex computer codes without actually running them. Based on the statistical theory of Gaussian Processes, emulators have been used in climate science, econometrics, and oceanography.

In the standard univariate case, a finite number of observations of a single type is made; here, observations of different types are considered. A multivariate generalization of the emulator technique is presented in which an arbitrary multivariate function may be assessed. The representative example chosen is temperature and rainfall which are predicted over the surface of the Earth using a climate model.

The technique has the property that marginal analysis (that is, considering only a single observation type) reduces exactly to the univariate theory. The associated software is used to analyze datasets from the fields of economics and climate change. This work is currently under review at the Journal of Statistical Software.

*****

1540- 1720	**MONDAY, 29 AUGUST** **Contributed Session: Environment and ecology [OGGB5]** *Chair: Ian Westbrooke*

WHERE DID YOU GET THAT RAT? USING GENETICS TO STUDY THE ORIGINS AND SWIMMING PATTERNS OF INVASIVE PESTS

Rachel Fewster*
University of Auckland
E-mail: r.fewster@auckland.ac.nz

Every week, islands around New Zealand are subject to a barrage of invasions by four-legged creatures with sharp teeth and big appetites. These invaders are mammal pests, including rats, stoats, and mice, and they have plenty of tricks up their furry sleeves for reintroducing themselves to conservation sanctuaries. They are excellent and eager swimmers, hitch rides on boats, abound in resourcefulness, and can cost tens of thousands of dollars - each - to track down and remove when discovered on a sanctuary island. Understanding where mammal invaders are coming from is pivotal to the long-term protection of sanctuaries. I will describe genetic assignment methods to estimate the origin of individuals, and give examples from island conservation sanctuaries around the country.

*****

ESTIMATING SPECIES RICHNESS AND SIMILARITY UNDER DIFFERENT TREATMENT CONDITIONS

Austina Clark*
University of Otago
E-mail: aclark@maths.otago.ac.nz

To conserve regional biodiversity effectively, ecologists need to know how diversity is distributed geographically within the region under different conditions. In particular, they need to know whether the regions consist of distinct communities or they are homogeneous. It is difficult to compare the shared species using finite resources, particularly with communities with large species richness and a large portion of rare species. (Colwell & Coddington 1994; Colwell et al 2004; Magurran 2004). I will illustrate the above using a large data set collected by DoC over 4 years from 1990 to 1993 in the South Island, New Zealand. A total of 1200 samples (via quadrats) were collected. There were 10 sites involved. Three treatments were allocated to each site randomly. An ecologist regrouped the 10 sites into three areas according to the plants growing there: A1 is land covered with tall tussocks, A2 covered with hieracium, and A3 is bare land. Earlier study (Clark, 2009) showed that there is no obvious difference of species richness for all treatments in the tall tussock and hieracium areas. However in the bare land area the estimated species was much higher for treatment 2 than treatment 1. Here we further investigate the species richness for each of the 10 sites respectively under the three treatments. Next we use the multiple-community diversity measure to compare the similarity (Chao et al, 2008) among the sites as well as the treatments.

*****

THE INFLUENCE OF NEAR-PITH MATERIAL ON THE COMMON CLIMATE SIGNAL

Maryann Pirie*
University of Auckland
E-mail: m.pirie@auckland.ac.nz

Kauri (Agathis australis (D. Don) Lindley) ring widths sequences have been used to reconstruct the past activeness of ENSO (El Niño/La Niña-Southern Oscillation) events within Northern New Zealand. However, there are concerns that the observed trends are influenced by differences in the common signal produced by kauri during its different life stages ('juvenile' and mature). This issue was investigated by comparing the common signal produced by two subsets of the data; near-pith (biological centre of the tree) material and material further from the pith. This talk will introduce a method for determining the similarity of two ragged arrays of time series, where there is both correlation between and within the time series. Then this method will be applied to tree-ring data to investigate any difference in the common signal produced by near and far-pith material for Kauri trees. Finally, correlation functions are used to determine the effect of differences in the common signal on the climate reconstruction.

*****

HETEROGENEOUS CAPTURE-RECAPTURE MODELS WITH COVARIATES: A PARTIAL LIKELIHOOD APPROACH FOR CLOSED POPULATIONS

Jakub Stoklosa*
University of Melbourne
E-mail: jhs@ms.unimelb.edu.au

In practice, when analyzing data from a capture-recapture experiment it is tempting to apply modern advanced statistical methods to the observed capture histories. However, unless the analysis takes into account that the data have only been collected from individuals who have been captured at least once, the results may be biased. Without the development of new software packages, methods such as generalized additive models, generalized linear mixed models, and simulation-extrapolation cannot be readily implemented. In contrast, the partial likelihood approach allows the analysis of a capture-recapture experiment to be conducted using commonly available software. Here we examine the efficiency of this approach and apply it to a data set.

*****

LINKING ZEROS TO ABUNDANCE IN ZERO-INFLATED MODELS OF SPECIES COUNT DATA

Adam N. H. Smith* and Marti J. Anderson
Massey University
E-mail: anhsmith@gmail.com

Many ecological studies collect data in order to assess the abundance of organisms. The resulting data are often counts that are over-dispersed and contain more zeros than expected by standard statistical distributions. To model such data, distributions such as the Poisson or negative binomial are used in conjunction with zero-inflation to cope with the extra zeros. Zero-inflated distributions have at least two parameters: the mean (mu) and the probability of an additional zero (p). Normally, the mean is of primary interest and some sort of linear model is used to explain variation in the mean with a set of covariates. But what of p, the probability of an extra zero? The current list of options for modelling p is fairly limiting. One can estimate a constant p across the whole dataset, an option with a fairly strong assumption but adding only a single parameter. Alternatively, one can model p with a separate linear model, which provides more flexibility but can double the number of parameters and complicate interpretation of the effects of covariates. An appealing alternative to these two approaches will be presented, where p is functionally linked to the mean of the count process. This model was applied here to estimate the effect of a marine reserve on the relative abundance of snapper (Pagrus auratus). This approach proved to have strong predictive power while requiring few extra parameters, and had the advantage of providing simple and sensible ecological interpretation.

*****

A BAYESIAN STATE-SPACE CAPTURE-RECAPTURE MODEL FOR INSULAR RAT POPULATION DYNAMICS

James C. Russell* and L. Ruffino
*University of Auckland**
E-mail: j.russell@auckland.ac.nz

The effects of spatio-temporal resource fluctuations on animal survival may be mediated by individual movement among habitat patches, but simultaneously analysing survival, resource availability and habitat selection requires sophisticated analytical methods. We use a Bayesian multi-state capture-recapture model to estimate survival and movement probabilities of introduced black rats across three habitats seasonally varying in resource availability. We find that survival varied most strongly with temporal rainfall patterns, overwhelming minor spatial variation among habitats. Climate is likely the main driver of rodent population dynamics on islands, and even substantial habitat and seasonal spatial subsidies are overwhelmed by predictable annual patterns in resource pulses. Temporal results on survival and capture probability provide important information for the timing of pest control operations such as eradication.

*****

1540- 1720	**MONDAY, 29 AUGUST** **Contributed Session: Model selection and averaging [OGGB3]** *Chair: Beatrix Jones*

BOOTSTRAPPED MODEL-AVERAGED CONFIDENCE INTERVAL

Jiaxu Zeng* and David Fletcher
*University of Otago**
E-mail: jzeng@maths.otago.ac.nz

Model averaging is commonly used to make allowance for model uncertainty in parameter estimation. In the frequentist setting, a model-averaged estimate of a parameter is a weighted mean of the estimates from individual models, the weights being based on an information criterion or on bootstrapping. In this talk, I will review current methods for calculating model-averaging confidence intervals and propose a new bootstrap-based method that appears to provide better coverage rates.

*****

FOCUSED INFORMATION CRITERIA, MODEL SELECTION AND MODEL AVERAGING IN A TOBIT MODEL WITH A NON-ZERO THRESHOLD

Alan Wan*
City University of Hong Kong
E-mail: msawan@cityu.edu.hk

A well-cited paper by Claeskens and Hjort (2003) developed a Focused Information Criterion (FIC) for model selection that selects different models based on different focused functions. In another paper, Hjort and Claeskens (2003) presented model averaging as an alternative to model selection, and suggested a local mis-specification framework for studying the limiting distributions and asymptotic risk properties of post model selection and model average estimators in parametric models. Despite the bourgeoning literature on Tobit models, little work has been undertaken with respect to model selection explicitly in the Tobit context. In this paper, we propose FICs for variable selection allowing for such measures as the MAD, MSE, and expected LINEX errors in a Type I Tobit model with an unknown threshold. We also develop a model average Tobit estimator using values of a smoothed version of the FIC as weights. The finite sample performance of model selection and model average estimators resulting from various FICs is studied via a Monte Carlo experiment, where the possibility of using a model screening procedure prior to combining the models is also demonstrated. Finally, we present an example from a well-known study on married women’s working hours to illustrate the estimation methods discussed.

*****

THRESHOLD SELECTION FOR MODELLING EXCEEDANCES OVER A HIGH THRESHOLD USING A BAYESIAN MEASURE OF SURPRISE

Jeong Eun (Kate) Lee*, Scott Sisson and Yannan Fan
*AUT University**
E-mail: jelee@aut.ac.nz

The Generalized Pareto distribution (GPD) is commonly used to model exceedances over high thresholds. As such, threshold selection is a key aspect in extreme value analysis. Ideally the threshold is high enough so that the asymptotic theory underlying the GPD holds, but also low enough that enough data is available for inference. In this poster, we adopt a Bayesian measure of surprise (MS) for the threshold selection problem. The MS is perceived as a quantification of the degree of incompatibility of data to some hypothesized model, which is useful in guiding the development of models at preliminary stage. The MS is estimated by the posterior predictive p- value and only threshold exceedances are used for inference on the parameters. The MS approach is applied to simulated data and to the Danish fire loss data set. We also compare to standard threshold estimation approaches including graphical approaches, nonparametric goodness-of-fit tests (Cramer-von Mises and Anderson-Darling statistics), and mixture models (MacDonald et al 2011). Based on the simulation studies, we show that the MS approach is a credible approach for threshold estimation in extreme value analyses.

*****

GRAPHICAL MODELS: PENALIZED LIKELIHOOD OR DECOMPOSABLE BAYESIAN

Marie Fitch* and Beatrix Jones
*University of Auckland**
E-mail: m.fitch@auckland.ac.nz

Gaussian graphical models are a popular and useful tool for describing patterns of conditional independence. We compare the costs and benefits of decomposable Bayesian and penalized likelihood approaches to model selection and parameter estimation.

*****

MODEL SELECTION FOR THREE-MODE THREE-WAY DATA

Lynette A. Hunt* and Kaye E. Basford
*University of Waikato**
E-mail: lah@waikato.ac.nz

The mixture likelihood approach to clustering is a model based approach that requires the specification of both the form of the density functions of each of the underlying groups and the number of groups that are fitted to the model. There has been extensive use of mixtures where the component distributions are multivariate normal and where the data would be described as two mode two way data. This talk investigates the behaviour of some commonly used model selection criteria when using the finite mixture model to cluster three way data containing mixed categorical and continuous attributes. We illustrate the performance of these criteria in selecting both the number of components in the model and the form of the correlation structure amongst the attributes when fitting a mixture model to classical three way data sets.

*****

	TUESDAY, 30 AUGUST
0900- 0920	**Plenary session [OGGB4]** *Chair: Chris Triggs* **Trevor Hastie** **Stanford University**

LEARNING WITH SPARSITY

Trevor Hastie*, Jerome Friedman and Rob Tibshirani
Stanford University
E-mail: hastie@stanford.edu

Many problems in machine learning have to deal with wide data – many more features than observations. Most of the features are of no use, and even the useful ones are often too sparse. For these problems L1 regularization and its variants have proven to be useful for both feature selection and complexity control. This talk is a review of a number of topics in this area, with a focus on computational aspects.

*****

0950- 1030	**TUESDAY, 30 AUGUST** **Plenary Session on SAS data mining [OGGB4]** *Chair: Thomas Yee* **Evan Stubbs** **SAS Australia & New Zealand**

FEEDBACK FROM THE FIELD: TWO YEARS LATER, ARE WE ANY BETTER OFF?

Evan Stubbs
Stanford University
E-mail: evan.stubbs@sas.com

The rise in the prominence of business analytics has been inexorable. However, in the two years since our last conference, how much progress has been made in helping bridge the gap between the skills required and the skills possessed by our recent graduates? In this presentation we look into whether or not we've made any progress and, based on specific feedback from regional and global organisations and educators, highlight some promising trends.

*****

1100- 1220	**TUESDAY, 30 AUGUST** **Contributed Session: Machine Learning and data mining [OGGB4]** *Chair: Trevor Hastie*

TREE-STRUCTURED MODELS FOR DIFFERENCE AND CHANGE DETECTION IN A COMPLETE ENVIRONMENT

Yong Wang*, Ilze Ziedins, Mark Holmes and Neil Challands
*University of Auckland**
E-mail: yongwang@auckland.ac.nz

A new family of tree-structured models is described, which we call “differential trees.” A differential tree model is constructed from multiple data sets and aims to detect distributional differences between them. The new methodology differs from the existing difference and change detection techniques, in its nonparametric nature, model construction from multiple data sets, and applicability to high-dimensional data. Through a case study, we illustrate how these models can help detect changes in the frequencies of event occurrences and uncover unusual clusters of events in a complex environment.

*****

A LEARNING EXPERIENCE

William Grant* and Siva Ganesh
*Massey University**
E-mail: billgrant1001@gmail.com

What happens through the course of a real life data mining exercise? We discuss what actually happened, contrasted with what we thought would happen, the data issues dealt with and the extent to which the final outcome was driven by the (not always objective) decisions of the analyst. The exercise aimed to identify clustering within a 3rd party data set used for a secondary analysis. Techniques utilised include Self Organising Maps, Support Vector Machines (focussing on its novelty detection properties), randomForests and multi-dimensional scaling for visualisation, k-means, application in a non-bio-statistics environment of the R ‘clValid’ cluster validation package and, finally, tree classification routines. An emphasis was placed on graphical approaches to present and analyse the data. Variable selection, data pre-processing and sampling to cope with processing constraints all had critical impacts upon the project and required frequent decision-making by the analyst. A heavy reliance was placed upon small sample test data to validate the usefulness and applicability of the R packages used and shape the final approach adopted for the project. Ultimately data hotspots or clusters were successfully identified. A summary, including the lessons learnt, is presented which we hope will aid other analysts wrestling with data mining in ‘real life’.

*****

COMPLEXITY MEASUREMENT: A SYSTEMATIC APPROACH TO OVERSAMPLING IN IMBALANCE DATA SETS

Nafees Anwar*, Siva Ganesh, Geoff Jones and Ganes Ganesalingam
*Massey University**
E-mail: m.n.anwar@massey.ac.nz

Traditional classification algorithms can be inadequate in their performance on extremely Imbalance data. Researchers working on different methods in different domains have yet to agree on standard benchmark datasets or on a systematic approach to solving the problem. Given the suggestion that the skewed class distribution may not be the only problem, we proposes a complexity measure for a systematic way of investigating and explaining what intrinsic features of the data are affecting the degraded learning performance of an imbalanced data set. The research objective of this presentation is to devise a structured study for learning about class imbalance problems. In this presentation we are using a complexity measure for oversampling of minority class examples, which do not need any optimization (iterative over sampling of minority class as proposed in Synthetic minority oversampling technique (SMOTE)). For the minority class, experiments show that our approach achieves better TP (true positive) and G-mean than SMOTE, Border Line SMOTE and random over-sampling (RO) method. An overview of the proposed methodology as well as SMOTE and RO will be presented followed some results and discussions.

*****

CLASSIFYING DIGITAL INK

Beryl Plimmer*
University of Auckland
E-mail: beryl@cs.auckland.ac.nz

The screens of many new computing devices, from phones to walls, have the hardware to sense touch or pen interactions with the surface. These interactions generate what we refer to as digital ink gestures. Sometimes the gestures are intended to fire commands, for example tapping or dragging on an iPhone, other times they may be artistic drawings. Yet another possibility is that they are a diagram or a semi structure document (such as a ‘to do’ list). In each case the computer software could offer intelligent support if it ‘understood’ the content – thus the need for classifiers.

The possible contexts for using digital ink are vast and thus the classification complexity is also vast. Much of the existing research has been on building a classifier for a particular context; this is time consuming! While the problem space is not yet well enough understood for us to produce general purpose classifiers, we are moving in this direction by automatically generating classifiers for a particular context. In this talk I will describe our layered approach to the problem and how we employ data mining techniques to automatically generate classifiers.

*****

1100- 1220	TUESDAY, 30 AUGUST Contributed Session: Applied probability 1 [OGGB5] *Chair: Ilze Ziedins*

QUEUEING UP FOR ENZYMATIC PROCESSING: CORRELATIONS THROUGH COUPLED DEGRADATION

Ruth Williams*, Natalie Cookson, Tal Danino, Jeff Hasty, Will Mather, Octavio-Mondragon-Palomino and Lev Tsimring
University of California, San Diego
E-mail: rjwilliams@ucsd.edu

A major challenge for systems biology is to deduce the molecular interactions that underlie correlations observed between concentrations of different intracellular molecules. Although direct explanations such as coupled transcription or direct protein-protein interactions are often considered, potential indirect sources of coupling have received much less attention. Here we show how correlations can arise generically from a posttranslational coupling mechanism involving the processing of multiple protein species by a limited number of copies of a common enzyme. By observing a connection between a stochastic model and a multiclass queue, we obtain a closed form expression for the steady-state distribution of the numbers of molecules of each protein species. From analytic expressions for the moments and correlations associated with this distribution, we observe a striking phenomenon that we call correlation resonance: for small dilution rate, correlations peak near the balance-point where the total rate of influx of proteins into the system is equal to the maximum processing capacity of the enzymes. The talk will describe the theoretical developments and the results of related experiments.

*****

HEAVY-TRAFFIC CONTROL AND PRICING FOR SYSTEMS WITH LEADTIME SENSITIVE CUSTOMERS

Tava Olsen*, Baris Ata
*University of Auckland**
E-mail: t.olsen@auckland.ac.nz

This paper studies a queueing model where two customer classes compete for a given resource and each customer is dynamically quoted a menu of price and leadtime pairs upon arrival. Customers select their preferred pairs from the menu and the server is obligated to meet the quoted leadtime. Customers have convex-concave delay costs. The firm does not have information on a given customer's type, so the offered menus must be incentive compatible. A menu quotation policy is given and proven to be asymptotically optimal under traditional large-capacity heavy-traffic scaling.

*****

MECHANISM DESIGN FOR WHOLESALE MARKET CLEARING UNDER UNCERTAINTY

Golbon Zakeri*, Javad Khazaei and Shmuel Oren
*University of Auckland**
E-mail: g.zakeri@auckland.ac.nz

We will discuss difficulties arising from integrating volatile renewable resources such as wind into a deregulated electricity market. We will also examine proposed market clearing mechanisms to harness wind effectively into an electricity market. Finally we will present some examples of what may happen should these mechanisms be adopted when there is exercise of market power.

*****

1100- 1220	**TUESDAY, 30 AUGUST** **Contributed Session: Biostatistics[OGGB3]** *Chair: John Koolard*

MECHANISM DESIGN FOR WHOLESALE MARKET CLEARING UNDER UNCERTAINTY

Gustavo Amorim* and Chris Wild
University of Auckland
E-mail: gustavodecastro@hotmail.com

Biased sampling can be found in most data collection processes and is quite common in medical studies. It may arise by happenstance, for example when some selected units fail to respond, or by design, for example if potential covariates are observed only for a subsample of the finite population. We are interested in a multiphase sampling scheme, where the probability of the ith individual being selected for the jht-phase is denoted by π_{ji}. Although these probabilities are controlled by the researcher and so can be treated as constants while estimating the regression coefficients β, gains in efficiency can be obtained by considering them unknown and modelled by a parametric function. In this talk we will discuss some simulation results for the case where continuous information is available for the entire population and a logistic function is assumed for the selection probabilities.

*****

CASE-COHORT DESIGNS FOR THE TIME FAILURE DATA

Patricia Metcalf*
University of Auckland
E-mail: p.metcalf@auckland.ac.nz

Prentice (1986) proposed a case-cohort design for large survey studies where the populations size makes it infeasible to collect data on all of the cases. A case-cohort study for failure time data consists of a random sample from the full cohort, the subcohort, and any additional cases not in the subcohort. Using a case-cohort design can very cost-efficient in that a sample much smaller than the full cohort results in only a small loss of statistical efficiency. A further advantage of this design is that the same subcohort may be used as a control group for multiple outcomes. I will discuss some of the computational methods that can be implemented using the standard Cox proportional hazards model software.

*****

THE EFFECT OF EARLY GROWTH AND DEVELOPMENT ON LIFE-LONG HEALTH. A CASE STUDY OF THE HELSINKI BIRTH COHORT

Elena Moltchanova* and Eero Kajantie
University of Canterbury
E-mail: elena.moltchanova@canterbury.ac.nz

In epidemiology, the Barker hypothesis suggests that prenatal and early childhood development has a significant effect on health later in life. Many studies, that have analysed the association between low birth weight and risks of hypertension, heart disease, diabetes, as well as cognitive ability at different stages of life, appear to confirm it. The Helsinki Birth Cohort Study is based on extensive data, collected for the 13345 people born in Helsinki, Finland during the period 1934-1944. Birth and early growth measurements (between 0 and 11 years of age) and socio-economic factors were recorded by the childhood welfare clinics and collected for the study. A unique personal identification number, allocated to each Finnish resident has permitted to link these data to the nationwide Hospital Discharge Registry and the Death Registry in order to obtain information on later diagnoses. In this case study, we aim to examine the association between the childhood growth (i.e. height and weight trajectories) and the later risk of diabetes, heart disease and hypertension. In particular, we look for ways to summarize the growth development and to distinguish aspects, if any, which are likely to influence on the later health.

*****

GENOME-WIDE ASSOCIATION ANALYSIS WITH PLINK

Dug Yeo Han*
University of Auckland
E-mail: dy.han@auckland.ac.nz

A GWA study undertaken in the New Zealand Caucasian (excluded non- Caucasian ancestry), which has examined 348 individuals with Crohn's disease and 488 controls. All 878 individuals were genotyped using ImmunoChip (Illumina Infinium genotyping chip), which comprise 196524 SNPs. A total of 129217 SNPs (65.8%) passed our control filters (MAF, genotyping call rate, and HWE). Standard 1-df test of case-control association analysis was performed using PLINK. Thirty-four SNPs from six chromosomes show evidence for association on allelic test.

*****

1330- 1410	**TUESDAY, 30 AUGUST** **Plenary Session[OGGB4]** *Chair: Andrew Balemi* **Wayne Stewart,** **University of Auckland**

BAYESIAN STATISTICS: THE SECOND COMING

Wayne Stewart*
*University of Auckland**
E-mail: w.stewart@auckland.ac.nz

The once commonly used Bayesian paradigm is making its way back and has the potential to re define modern statistics. The credibility, applicability and the richness of the archetype are self evident. Most of the agreed disadvantages of Bayesianism namely, priors and MCMC are two edged. The priors are a backdoor for information previously found and relevant to the study. MCMC, a mathematical tool, has the added value that functions of random variables can be easily summarized even when they are analytically intractable. Although the theory is well developed and reasonably straightforward for statisticians, the teaching of it still remains a challenge specially to undergraduate students with less mathematical knowledge and skills. How can we teach Bayesian statistics in ways that will actively facilitate the use of this incredibly powerful paradigm rather than procrastinating and watching the unexploited opportunities float by?

In this talk I will illustrate the effectiveness of Bayesian statistics and how it differs from classical statistics. I will also show some fascinating examples of the paradigm by a meaningful comparison of confidence intervals with Bayesian credible intervals to point out its interpretational simplicity and advantages. Bayesian estimates can be biased but will often have better frequentist mean squared errors. Hierarchical modeling which is easily accomplished in a Bayesian framework and difficult to perform within the classical paradigm is a natural way of pooling information to produce smaller interval estimates for parameters.

*****

1410- 1510	**TUESDAY, 30 AUGUST** **Contributed Session: Statistics education: general [OGGB4]** *Chair: Andrew Balemi*

USING MEDIA REPORTS TO PROMOTE STATISTICAL LITERACY FOR NON-QUANTITATIVE MAJORS

Stephanie Budgett*
*University of Auckland**
E-mail: s.budgett@auckland.ac.nz

At The University of Auckland we teach an undergraduate course entitled Lies, Damned Lies and Statistics, the purpose of which is to facilitate students to “think statistically” when confronted with evidence-based arguments. In this paper we first describe how we use media reports in teaching to enhance students’ ability to understand and evaluate statistically based information. Second, we report our observations on interviews we conducted with three non-quantitative majors and three quantitative majors seven months after they completed the course. A comparison of their responses to two media reports suggested there was little difference between the two groups. Possible reasons for these observations are discussed.

*****

MERGING VISIONS FOR STATISTICS AND MATHEMATICS EDUCATION

Mike Camden*
*Statistics NZ**
E-mail: mike.camden@stats.govt.nz

The Senior Secondary and Undergraduate Mathematical Science project has developed a vision for education in the mathematical sciences, for the last two secondary levels and the undergraduate tertiary levels. For some years the NZ Statistical Association's Education Committee has been promoting a vision for statistics education, and this vision is built in to the current statistics strand of Mathematics and Statistics in the NZ school Curriculum. This presentation compares the two visions, and finds ways in which they can strengthen each other. For example, the statistics vision includes the statistical enquiry cycle, which adds the strength and motivation of practical context. Both visions include some new methodologies from their fields.

*****

THE INTRODUCTORY STATISTICS COURSE AND INFERENCE

Maxine Pfannkuch*
University of Auckland
E-mail: m.pfannkuch@auckland.ac.nz

This presentation sets out some of the rationale and arguments for making major changes to the teaching and learning of statistical inference in introductory courses at our university by changing from a norm-based, mathematical approach to more conceptually accessible computer-based approaches. The core problem of the inferential argument with its hypothetical probabilistic reasoning process is examined in some depth. We argue that the revolution in the teaching of inference must begin. We also discuss some problematic areas associated with introducing the logic of inference through randomisation.

*****

1410- 1510	**TUESDAY, 30 AUGUST** **Contributed Session: Applied probability 2 [OGGB5]** *Chair: Ilze Ziedins*

MODELS AND MEASUREMENTS FOR COGNITIVE RADIO SYSTEMS

Peter Smith*
University of Canterbury
E-mail: p.smith@elec.canterbury.ac.nz

Cognitive radio is an emerging technology which aims to increase the efficiency in which the radio spectrum is accessed. The fundamental idea is that licensed users (for example TV broadcasts in the TV channels or mobile phone calls in the cellular band) may not use all their allocated spectrum all the time in all locations. Hence, there are times or places where other people could use the spectrum without causing interference. In this talk I will outline the fundamental ideas of cognitive radio and present measured data on a real cellular system. The data is then used to develop very simple queueing type models for the spectral occupancy. Using these models we evaluate the possibilities for cognitive radio and also derive some very simple blocking and dropping probabilities for the cognitive radio users.

*****

UNSTEADY VOLCANIC MODELLING

Mark Bebbington*
Massey University
E-mail: m.bebbington@massey.ac.nz

Standard models for medium-long term volcanic hazard treat the magmatic system as being statistically stationary. Examples include renewal processes of various types and, surprisingly often, even the simple Poisson process. However, it has been suggested that such stationarity is a phenomenon measured in decades, not centuries. Although a nonhomogeneous Poisson process can be constructed to allow for nonstationarity, the potential degrees of are difficult to reconcile with the stationarity/quiescence dichotomy supposed for most volcanoes. The trend renewal process allows us to use a 'building-block' approach capable of dealing with cases of constant activity level, increasing trend, wax and wane of activity, and cyclic (plus trend) behaviour. A consequence is that the observed tendency of volcanic eruptions to cluster in time can best be explained by trends in the activity. In the case of the Auckland Volcanic Field, approximately two thirds of the eruptions have occurred during less than a tenth of the life of the field. Such clustering can only be modelled via a self exciting process.

*****

A REGIONALIZATION METHOD BASED ON A CLUSTER PROBABILITY MODEL

Paul Cowpertwait*
AUT University
E-mail: paul.cowpertwait@aut.ac.nz

A regionalization method based on a cluster probability model (a mixed multivariate Gaussian model) is proposed for grouping points $(x, y) \in \mathbb{R}^2$ into non-overlapping contiguous homogeneous regions defined by a Voronoi tessellation. The cluster probability model is applied to second-order standardized annual sample properties (mean, coefficient of variation, and autocorrelation) evaluated at the daily level of aggregation taken from each of 234 daily rainfall records with positions (x_i, y_i)located in the Basque Country, Spain (i = 1, . . . , 234). Using the bayesian information criterion (BIC), four clusters are identified, which are seen to overlap in $\mathbb{R}^2$ when the site coordinates (x_i, y_i) are plotted. For each point (x_i, y_i), the probability $p_i(k)$ that the point belongs to the k-th cluster is extracted from the fitted Gaussian model. Edges $E_{ij}$are found for the Delaunay complete planar graph $\{V_i, E_{ij}\}$ of the points $(x_i, y_i) = V_i \in \mathbb{R}^2$, and the Euclidean distance d_{ij} corresponding to each edge E_{ij} is found. For each V_i a probability-distance score is calculated for each cluster k by summing $p_i(k)p_{ij}(k)/d_{ij}^2$ over all edges corresponding to V_i, where $p_{ij}(k)$ is the probability that the j-th point linked to V_i is in the k-th cluster. Regions from the Voronoi tessellation of the points are classed based on this score and according to whether they are spatially isolated from other regions of the same class. Points that have the least influence on the variance of residual errors of the three-class model are found using a criteria based on Wilk's Lambda for multivariate analysis of variance, and the least influential regions adjusted to ensure the overall regions are contiguous.

*****

1540- 1620	**TUESDAY, 30 AUGUST** **Contributed Session: Quantile regression [OGGB4]** *Chair: Alain Vandal*

ON QUANTILE REGRESSION

Arash Ardalan* and Thomas W. Yee
University of Auckland
E-mail: arash@stat.auckland.ac.nz

This talk consists of two sections, at the first we will present a new approach to quantile regression, in the second section we will present a Bayesian approach to quantile regression. The classical quantile regression approach has been given by Koenker and Bassett (1978) which estimates quantiles by specialized linear programming techniques, expectile/percentile regression has been proposed by Newey and Powell (1987) and Efron (1991) which is very much related to the classical quantile regression. It is known that the quantiles can be coincided with the maximum likelihood solution of the location parameters of a class of two-piece distribution. In this regard, we present a new class of two-piece distribution which is useful in quantile regression, the location parameter is the pth quantile of it, we investigate the properties and asymptotic behaviour of the maximum likelihood estimators of the parameters. In the second section, we will present the quantile regression using the idea of Bayesian semi-parametric regression. Key words and phrases: Tow-piece distribution; Quantile regression; Maximum likelihood estimation; Asymptotic normality; Markov chain Monte Carlo; Mixed models; Smoothing splines.

*****

AN EXPLORATORY APPROACH OF MODELING NONSTATIONARITY FOR SPATIAL QUANTILE-BASED DATA ANALYSIS

Vivian Yi-Ju Chen*
Tamkang University
E-mail: viviyjchen@stat.tku.edu.tw

There has been a trend in spatial data analysis to analyze spatial nonstationarity by fitting a regression model that allows for geographically varying (local) coefficients. An emergent advanced exploratory spatial analytic tool that is widely used to implement this task in the field of geostatistics is the geographically weighted regression (GWR) proposed by Fotheringham, Brunsdon, and Charlton. However, the current GWR is only capable of computing the parameter estimates to the mean function of the conditional distribution of the dependent variable. Quantile regression (QR) has been known as a statistical technique that moves beyond traditional mean modeling. While recent years have witnessed developments in both GWR and QR, there is lack of spatial analysis technique that generalizes GWR to allow for estimating various conditional quantile functions. The intent of this study is to develop a new geostatistical methodology, geographically weighted quantile regression (GWQR), which bridges the concept of QR with the GWR framework. The proposed method is applied to an empirical data set and consequently proven as an innovative technique for exploratory spatial quantile-dased data analysis.

*****

ON TESTING COVEX TRANSFORM ORDERING

Muhyiddin Izadi*
Massey University
E-mail: Izadi_552@yahoo.com

Suppose $F$and G are two life distribution functions. It is said that F is less than G in the convex transform order (written by $F \leq_c G$) if $G^{-1}F(x)$ is convex on $(0, \infty)$. In this paper the problem of testing $H_0: F =_c G$ against $H_1: F \leq_c G$ and $F \neq_c G$ is considered in both cases when G is known and when G is unknown. We obtain the asymptotic distribution of the test statistics that are based on U-statistics. To establish our test, we compare the test with Ahmad and Kochar's test [1990, Testing whether F is more IFR than G. *Metrika*, **37**. 45-58] by the Pitman's asymptotic relative efficiency.

*****

1540- 1620	**TUESDAY, 30 AUGUST** **Contributed Session: Statistics education: Bayesian [OGGB4]** *Chair: Andrew Balemi*

BAYESIAN STATISTICS IN NZ UNIVERSITIES UNDERGRADUATE CURRICULUM

Bill Bolstad*
Waikato University
E-mail: bolstad@waikato.ac.nz

A survey the current state of teaching Bayesian statistics to undergraduates in NZ universities

*****

TEACHING MCMC IN BAYESIAN STATISTICS: WHAT GOES ON BEHIND THE ALGORITHM

Wayne Stewart*
University of Auckland
E-mail: w.stewart@auckland.ac.nz

Markov Chain Monte Carlo (MCMC) simulation, a mathematical tool, is an important and necessary component of a Bayesian statistics course. The simulation is often taught by presenting an algorithm and translating it into an appropriate computer program. Consequently, undergraduate students with no prior background of the algorithm, are diverted from Bayesian concepts and confronted by formal mathematical ideas. In this talk I show how to teach MCMC simulation conceptually in the context of a Bayesian paradigm without revealing the formal algorithm first. This is achieved when a two state discrete parameter is used within a tactile simulation where a coin supplies the proposal values and given the acceptance sets, the die value determines whether to accept the proposal or not.

*****

1540- 1620	**TUESDAY, 30 AUGUST** **Contributed Session: Applied probability 3 [OGGB5]** *Chair: David Scott*

USER OPTIMAL POLICIES FOR A STOCHASTIC TRANSPORTATION NETWORK

Heti Afimeimounga*
University of Auckland
E-mail: h.afimeimounga@auckland.ac.nz

Consider a queueing network where two routes are available for users wishing to travel from a source to a destination. On one route (private transport) service slows as traffic increases. On the other (public transport) the service frequency increases with demand. We study the properties of the user optimal policy for this network in the setting of state-dependent and probabilistic routing.

*****

SEQUENTIAL ANALYSIS OF THE MORAN PROCESS

Peter Green*, Travis Monk, Mike Paulin
*University of Otago**
E-mail: pgreen@maths.otago.ac.nz

Sequential analysis is the theory of cumulative sums of random variables. The central result in sequential analysis, Wald’s Fundamental Identity, can be used to calculate absorption probabilities in random walks with barriers, when the moment generating functions of the random increments have the same roots. The Moran process from theoretical biology is a birth-death process used to model the spread on mutant genes in a population. This process can be used to calculate the probability that a beneficial mutation will spread to an entire population. The Moran process is the cumulative sum of random changes in the population state, and is therefore amenable to sequential analysis. The Moran process can also be extended to populations on graphs. Star graphs – graphs with a single central vertex joined to a number of point vertices – are known to amplify the advantage of a beneficial mutation, at least in the limit of a large number of points. Sequential analysis can be used to calculate exact fixation probabilities for the Moran process on a finite star graph.

*****

1540- 1620	**TUESDAY, 30 AUGUST** **Contributed Session: Mixture models and methods 1 [OGGB3]** *Chair: Yong Wang*

MIXTURE SURVIVAL MODELS FOR IDENTIFYING INFANT AND SENESCENT MORTALITY

Rebecca Green* and M. S. Bebbington
Massey University Palmerston North
E-mail: bex.green@hotmail.com

It is suggested that life can be separated into stages such as development, ageing and late life. Late life mortality is difficult to estimate since only small amounts of the data are representative. Aging and mortality is widely modelled by the Gompertz distribution, which assumes that mortality rate increases exponentially. However, this distribution can provide a poor fit for infant mortality and mortality of humans at very advanced ages. The aim of this presentation is, to propose an alternative model which can be used to provide an accurate fit for all ages of life. In this talk a mixture distribution including infant, exogenous and both Gompertzian and non-Gompertzian senescent mortality will be proposed. Each component of the mixture represents a separate phase of life, which corresponds to a subpopulation with its own life distribution. Using data of mortality in Swedish females from 1751, it will be shown that this model outperforms models without these features. The differences between using period data rather than cohort data to fit a parametric model of aging will also be discussed. The proposed mixture model will be used to compare the trends in these two different ways of analyzing mortality over time.

*****

APPLICATION OF A NON-LINEAR MIXED MODEL TO STRESS-STRAIN RELATIONSHIP OF FLAX FIBRES

Chikako Van Koten*
AgResearch, Lincoln
E-mail: chikako.vankoten@agresearch.co.nz

In agricultural, environmental, and biomedical applications, it is common to observe a continuous response evolving over time, or other condition, within same samples/individuals, where the response profile is not necessarily linear. A nonlinear mixed model is one way to analyse such data.

The composite flax-polymer fibres (composites of themoplastic polymers with natural flax fibres as reinforcing agents) have in recent years attracted increasing attention in light of the growing environmental awareness, as well as due to their attractive price/performance ratio in many manufacturing products. This growing flax-polymer fibre market is a very promising opportunity for New Zealand. However, studies to date have reported that the composite flax-polymer fibres exhibit a unique nonlinear stress-strain relationship due to their rather intricate structures, and hence, their ultimate performance cannot be predicted in the same way as for glass fiber reinforced composites.

In this presentation, experience in analysing the stress-strain relationship of composite flax-polymer fibres using the following Michaelis-Menten nonlinear regression model with a random coefficient is discussed:

$$Stress_{i,j} = \frac{\left(A_j + u_i\right) \times Strain_{i,j}}{B_j + Strain_{i,j}}$$

where A and B are parameters that define treatment-group-specific fixed effects, and u is a sample-specific random effect. The fibres are grouped into four treatment groups, thus j = 1; ….; 4. The suffix i indicates i-th sample. Since many different algorithms are available for estimating parameters of this nonlinear mixed model, this presentation reports and compares the results from the NLMIXED procedure in SAS and the nlme() function in the nlme package in R. This nonlinear mixed model seems to fit well to the observed relationships, allowing us to compare the treatments.

*****

WEDNESDAY 31 AUGUST

	Plenary session [OGGB4]
	Chair: Thomas Lumley
0900- 0920	**Xihong Lin** **Harvard University**

STATISTICAL ISSUES AND CHALLENGES IN ANALYZING HIGH-THROUGHPUT ‘OMICS DATA IN PUPULATION-BASED STUDIES

Xihong Lin*
Havard School of Public Health
E-mail: xlin@hsph.harvard.edu

With the advance of biotechnology, massive "omics" data, such as genomic and proteomic data, become rapidly available in population based studies to study interplay of genes and environment in causing human diseases. An increasing challenge is how to design such studies, managing the data, analyze such high-throughput "omics" data, interpret the results, make the findings reproducible. We discuss several statistical issues in analysis of high-dimensional "omics" data in population based "omics" studies. We present statistical methods for analysis of several types of "omics" data, including incorporation of biological structures in analysis of data from genome-wide association studies, next generation sequencing data for rare variants. Data examples are presented to illustrate the methods. Strategies for interdisciplinary training in statistical genetics, computational biology and genetic epidemiology will also be discussed.

*****

0950- 1030	**WEDNESDAY, 31 AUGUST** **Contributed Session: Genetics [OGGB4]** *Chair: Thomas Lumley*

iDArTs: THINKING DIFFERENTLY ABOUT GENETIC MARKERS TO UNLOCK NEW RESOURCES

Emma Huang*, Colin Cavanagh, Andrzej Kilian and Andrew George
*CSIRO**
E-mail: emma.huang@csiro.au

For many years, genetic markers have been the building blocks in assembling genomic knowledge. Improved technology and methods for collecting marker data have increased accuracy, increased throughput, and reduced cost. However, this improvement has not been uniform across organisms. There are still far fewer markers available in many plant species than in animals and humans. We propose a new type of genetic marker based on the Diversity Arrays Technology (DArT) genotyping system for organisms lacking reference genetic sequence. These markers are based directly on microarray probe intensity profiles and hence are called iDArTs. They require no additional genotyping beyond screening with a DArT array. Since standard methods of genetic analysis cannot be used with these continuous markers, we develop novel methods for common bi-parental experimental designs, including doubled haploids, recombinant inbred lines, and backcrosses. These enable the augmentation of genetic maps with iDArTs and permit QTL mapping with both discrete and continuous markers. To demonstrate the value of this approach, I will present results for genetic map construction and QTL mapping from simulation and a doubled haploid wheat cross. These methods allow access to a previously untapped genetic resource by extracting additional information from the raw data. With no additional genotyping cost, we are able to double the number of markers mapped and thereby increase genome coverage.

*****

DIFFUSION APPROXIMATION AND MAXIMUM ENTROPY

Jing Liu*
University of Auckland
E-mail: jliu070@aucklanduni.ac.nz

The diffusion approximation and maximum entropy principle together provide a powerful procedure for obtaining stationary distributions for certain stochastic systems. I will talk about the use of the diffusion approximation in population genetics, and give intuitive explanations of the maximum entropy principle and current usage. The strengths and weaknesses of combining the two methods will be discussed, and I will demonstrate performance using some models in population genetics.

*****

0950- 1030	**WEDNESDAY, 31 AUGUST** **Contributed Session: Categorical data analysis 1 [OGGB5]** *Chair: Ivy Liu*

USING ALGEBRAIC METHODS TO TEST THE INDEPENDENCE OF ETHNICITY AND RESOLUTION FOR DRUG-RELATED CRIMES IN NZ

Irene van Woerden* and Raazesh Sainudiin
*University of Christchurch**
E-mail: icv10@UClive.ac.nz

Using Markov bases the basic assumption of the cell values in contingency tables being sufficiently high no longer needs to be met. This algebraic statistical method allows us to conduct exact tests for independence in two-way contingency tables. In this study we obtain Monte Carlo estimates of the exact P-value to test the null hypothesis of independence between ethnicity and resolution of various drug-related crimes in NZ. The exact P-value is given by the fraction of tables with the same row and column marginals as the observed table but with at least as extreme a Chi-squared deviation as the observed table from the expected table. This is estimated by an irreducible random walk on the graph of such marginal-preserving tables. We conducted nine tests of independence between ethnicity and resolutions of various drug-related offences in 2009 in NZ using publicly available crime statistics and concluded that there is strong evidence against independence at least in the case of one drug.

*****

ROW-COLUMN ASSOCIATION MODELS

Thomas W. Yee* and Alfian F. Hadi
*University of Auckland**
E-mail: t.yee@auckland.ac.nz

This paper describes a statistical framework and software for fitting row-column association models (RCAMs) to two-way table responses. We consider some link function applied to the mean (say) of a cell equalling a row effect plus a column effect plus an interaction term. The interaction term is modelled as a reduced-rank regression (with complexities ranging from rank-1 and upwards), while the row and column (main) effects are handled using simple indicator variables. What sets apart this work from others is that our framework incorporates a very wide range of statistical models. For example, (i) log-link with Poisson counts is Goodman's RC model, (ii) zero-inflated Poisson distribution may be suitable with a two-way table with lots of zeros, (iii) identity-link with a double exponential (Laplace) distribution is akin to median polish, (iv) identity-link with normal errors is similar to two-way ANOVA with one observation per cell, (v) log-link with negative binomial counts may help handle overdispersion relative to the Poisson model. New software within the first author's VGAM R package makes it very easy to fit a wide range of RCAMs to data. Altogether, the main result of this work is that RCAMs facilitates the analysis of two-way tables of many data types, therefore is potentially very useful in many areas of applied statistics.

*****

0950- 1030	**WEDNESDAY, 31 AUGUST** **Contributed Session: Copulas [OGGB3]** *Chair: Alan Wan*

FINITE MIXTURES OF ARCHIMEDEAN COPULAS

Renate Meyer* and G. Kauermann
*University of Auckland**
E-mail: meyer@stat.auckland.ac.nz

Copulas allow for stochastic modelling of multivariate distributions with a flexibility well beyond that of the classical normal distribution. To further increase the versatility we propose the use of mixtures of different Archimedean copula families like Clayton, Frank, Gumbel, etc. Using a Bayesian approach, each family-specific parameter is modelled by imposing a prior distribution on the parameter. The mixture model itself is fitted in two ways. We first present a fully Bayesian approach with MCMC-based posterior computation. Then, a computationally much faster marginal likelihood estimate is proposed using a penalized version of a classical quadrature which approximates the integrals. The performance of the new approach is evaluated on simulations and an example in the context of modelling the dependence structure of the log-returns of exchange rates.

*****

SHARP BOUNDS ON A CLASS OF COPULAS AND QUASI-COPULAS

Heydar Ali Mardani-Fard*, S. M. Sadooghi-Alvandi and Z. Shishebor
*Yasouj University**
E-mail: h_mardanifard@yahoo.com

The diagonal section of a copula (or quasi-copula) has several probabilistic interpretations and is useful in studying the tail dependence of pairs of random variables which has applications in insurance and finance. In this article, we consider the class of copulas with known values at several diagonal points, and derive best-possible bounds on such copulas. We then use these bounds to establish best-possible bounds on measures of association, Kendall's and Spearman's, for such copulas. We also use our results to establish best-possible bounds on the distribution function of the sum of two random variables with known marginal distributions when the values of the joint distribution function are known at several diagonal points.

References:
[1] Mardani-Fard, H.A., Sadooghi-Alvandi, S.M., and Shishebor, Z., (2010). Bounds on Bivariate Distribution Functions with Given Margins and Known Values at Several Points, Communications in Statistics: Theory and Methods, 39: 20, 3596-3621.
[2] Sadooghi-Alvandi, S.M., Shishebor, Z., and Mardani-Fard, H.A., Sharp Bounds on a Class of Copulas with Known Values at Several Points, Communications in Statistics: Theory and Methods, Accepted for Publication.

*****

1100- **1220**	**WEDNESDAY, 31 AUGUST** **Contributed Session: Design [OGGB4]** *Chair: Xihong Lin*

MDS-OPTIMAL SUPERSATURATED DESIGNS

Arden Miller* and Boxin Tang
*University of Auckland**
E-mail: a.miller@auckland.ac.nz

A minimal dependent set (MDS) is a set of vectors that are linearly dependent but if any one of them is removed the resulting subset is independent. This talk will discuss the relationship between the minimal dependent sets of the column vectors of the design matrix for a 2-level supersaturated design and the resolvability of the design. It will introduce the concepts of MDS-resolution and MDS-aberration as criteria for comparing supersaturated designs. Results concerning supersaturated designs that have minimum MDS-aberration will be presented.

*****

SENSITIVITY OF EWMA CONTROL CHARTS

Saddam Akber Abbasi* and Arden Miller
University of Auckland
E-mail: sabb025@aucklanduni.ac.nz

Control chart is the most important Statistical Process Control (SPC) tool used to monitor reliability and performance of manufacturing processes. Variability EWMA charts are widely used for the detection of small shifts in process dispersion. For ease in computation all the variability EWMA charts proposed so far are based on asymptotic nature of control limits. It has been shown in this study that quick detection of initial out-of-control conditions can be achieved by using exact or time varying control limits. Moreover the effect of fast initial response (FIR) feature, to further increase the sensitivity of variability EWMA charts for detecting process shifts, has not been studied so far in SPC literature. It has been observed that FIR based variability EWMA chart is more sensitive to detect process shifts than the variability charts based on time varying or asymptotic control limits.

*****

DESIGNING A TWO-PHASE EXPERIMENT FOR MANY TREATMENTS AND FEW REPLICATES IN BLOCKS OF SIZE TWO

Kathy Ruggiero*, Richard G. Jarrett
*University of Auckland**
E-mail: k.ruggiero@auckland.ac.nz

Measuring the abundance of gene products – gene transcripts, proteins and metabolites – requires two phases of experimentation: an intervention applied to the experimental material at Phase 1 and subsequent laboratory processing of the material harvested from the Phase 1 experiment at Phase 2. The use of a dual-labelling technology at Phase 2 enables the simultaneous analysis of the gene products in two biological samples (i.e. blocks of size two) so that abundances are measured under homogeneous conditions. While generating connected design for a two-phase experiment for a small number of treatments in blocks of size two might be straightforward, achieving the same objective for a large number of treatments poses more of a challenge. We will show how we consider the experimental error introduced at Phases 1 and 2 in designing a large microarray experiment involving in excess of 140 recombinant inbred lines (RILs) of Arabidopsis thaliana. Further, we will show how our approach achieves connectedness (i.e. all treatment, or line, contrasts estimable) without the use of a common pooled reference RNA on all arrays.

*****

COMPLETE ALLOCATION SAMPLING: AN EFFICIENT AND EASILY IMPLEMENTED ADAPTIVE SAMPLING DESIGN

Jennifer A. Brown*, Mohammad Salehi M., Bardia Panahbehagh and Mohammad Moradi
*University of Canterbury**
E-mail: jennifer.brown@canterbury.ac.nz

Adaptive sampling designs are becoming increasingly popular in environmental science particularly for surveying rare and aggregated populations. There are many different adaptive survey designs that can be used to estimate animal and plant abundances. The appealing feature of adaptive designs is that the field biologist gets to do what innately seems sensible when working with rare and aggregated populations' field effort is targeted around where the species is observed in the first wave of the survey. However there are logistical challenges of applying this principle of targeted field-effort whilst remaining in the framework of probability-based sampling. We propose a simplified adaptive survey design where entire strata are sampled. This design incorporates both ideas of targeting field effort and being logistically feasible. We show with a case study population of rockfish that complete allocation stratified sampling is very efficient design.

*****

1100- 1220	**WEDNESDAY, 31 AUGUST** **Contributed Session: Categorical data analysis 2 [OGGB5]** *Chair: Beatrix Jones*

BICLUSTERING AND PATTERN DETECTION FOR BINARY AND COUNT DATA

Shirley Pledger*
Victoria University of Wellington
E-mail: Shirley.pledger@vuw.ac.nz

Patterns of association in matrices of binary or count data, for example the occurrence or abundance of different species over multiple samples, are traditionally found by multivariate methods such as multidimensional scaling, cluster analysis and correspondence analysis. These give graphical and descriptive results, but usually not statistical conclusions. By introducing statistical mixture models, we may switch to fuzzy clustering, in which species and/or samples are allocated to groups probabilistically. This is a likelihood-based approach, which provides statistical inference in addition to graphical pattern analyses. We give examples to show how the mixture approach provides (i) model selection by information criteria, (ii) dimension reduction, and (iii) biplots which usually, but not always, appear similar to plots from multidimensional scaling and correspondence analysis.

*****

BICLUSTERING MODELS FOR ORDINAL DATA

Eleni Matechou*, Ivy Liu, Shirley Pledger and Richard Arnold
*Victoria University of Wellington**
E-mail: Eleni.Matechou@vuw.ac.nz

Questionnaires with ordinal responses are widely used and can provide information on a range of topics, including, but not limited to, politics, consumer views on products or level of satisfaction with one's working environment. Despite the wide-spread collection of ordinal data sets, the methods traditionally used for cluster analysis for ordinal data wrongly treat the ordinal score as continuous, are not based on statistical likelihoods and do not fully incorporate the structure of the data within a probability model. Our approach models associations in ordinal data sets using fuzzy clustering based on finite mixtures. We introduce likelihood-based models for pattern detection which simultaneously cluster the rows (individuals) and the columns (questions/samples) by recasting the proportional odds model to include fuzzy clustering. We demonstrate the methods using the course evaluations for a second year applied statistics course at Victoria University of Wellington.

*****

GOODNESS-OF-FIT TESTS FOR LOGISTIC REGRESSION MODELS USING STOCHASTIC PROCESSES

Ivy Liu* and Estate Khmaladze
*Victoria University of Wellington**
E-mail: iliu@msor.vuw.ac.nz

Traditional methods to detect lack of fit for a simple logistic regression model use either likelihood-ratio or Pearson chi-squared tests. The test statistics follow asymptotically a chi-squared distribution when data are not sparse. It applies when the explanatory variable is categorical. When the model contains a continuous explanatory variable, these goodness-of-fit tests are not valid. Furthermore, we can partition observed and fitted values according to the predicted probabilities of success using the original data, and then apply a Pearson statistic, which is known as the Hosmer-Lemeshow method. Unfortunately, methods based on the grouping strategy do not have good power. This talk provides an alternative goodness-of-fit method using a process that converges in distribution to a Brownian motion. The Kolmogorov-Smirnov statistics are constructed to assess the adequacy of the model. For various cases, we will show that empirical distributions of statistics are very close to the limiting distribution under the null.

*****

MODELLING STRATEGIES FOR REPEATED MULTIPLE RESPONSE DATA

Thomas Suesse* and Ivy Liu
*University of Wollongong**
E-mail: tsuesse@uow.edu.au

Agresti and Liu (2001) discussed modelling strategies for a multiple response variable, a categorical variable for which respondents can select any number of outcome categories. This talk discusses modelling strategies of a repeated multiple response variable, a categorical variable for which respondents can select any number of categories on repeated occasions. We consider each of the responses as a binary response and model the mean binary responses with two approaches: a marginal model approach and a mixed model approach. For the marginal model approach, we consider a generalised estimating equations (GEE) method to account for different correlations over time and between items as an alternative to standard GEE, which only allow relatively simple correlation structures. We illustrate the different approaches using The Household, Income and Labour Dynamics in Australia (HILDA) Survey, a household-based panel study.

*****

1100- 1220	**WEDNESDAY, 31 AUGUST** **Contributed Session: Bayesian statistics 2 [OGGB3]** *Chair: Kate Lee*

BENCHMARKING WINBUGS AND OPENBUGS TO INDEPENDENT METROPOLIS-HASTINGS WITH EHAVY-TAILED CANDIDATE FOR GLMS: PART 1

Toufiq Al Gheilani*
Waikato University
E-mail: toufiqalgehilan@hotmail.com

This paper will discuss a brief introduction to Bayesian inference and some computational methods especially Gibbs sampling. Moreover, I will show the BUGS project, its place in Bayesian inference and some ways to set up problem in BUGS project.

*****

BENCHMARKING WINBUGS AND OPENBUGS TO INDEPENDENT METROPOLIS-HASTINGS WITH EHAVY-TAILED CANDIDATE FOR GLMS: PART 2

Bill Bolstad*
Waikato University
E-mail: bolstad@waikato.ac.nz

In this paper we compare the performance of WinBUGS and OpenBUGS to independent Metropolis-Hastings for a heavy tailed Generalized Linear Model.

*****

EFFECTS OF INCORPORATING GPS ROUTING INFORMATION INTO CURRENT TRAFFIC MODELS

Katharina Parry*, M. L. Hazelton
Massey University
E-mail: k.parry@massey.ac.nz

We present an initial investigation into the effects of incorporating extra information in the form of GPS data into Bayesian traffic models. Our model parameters of interest are the mean route flows, λ and the probability of vehicles being equipped with GPS on any given route in the network, **p.** We examine the estimation of the parameters separately. A first visual study of the profile likelihood functions reveal unusual features. A critical aspect is that these likelihoods can be ridged or very flat. This makes calculation of maximum likelihood estimates challenging, and renders unstable derivative-based algorithms. A solution was to use a more general optimiser, the subplex minimisation method, which was found to be more robust. We found that the method of moments delivers a good statistic for estimating the probability of being GPS equipped, given this is constant across all routes in the network. We show an example of identifiability issues we encounter when the parameter **p** is not a scalar. This work has the important finding that parameter estimation and other statistical inference is significantly improved by the inclusion of routing information even for sparse coverage of GPS equipped vehicles.

*****

A PRINCIPLE FOR QUALITY CONTROL IN BAYESIAN ANALYSES

Robin Willink*

E-mail: robin.willink@gmail.com

Bayesian analyses seem inappropriate to many statisticians. One reason for this may be a perceived lack of criteria or techniques by which the methodology can be held to account. What actually are Bayesian statisticians claiming about the performance of their methods, and how can those claims be assessed? This talk will introduce the idea that for a routine Bayesian methodology to be acceptable the mean value of the probability integral of the posterior distribution evaluated at the corresponding prior median must be close to 0.5. Adherence to this principle can be tested using records of prior and posterior distributions.

*****

1330- 1410	**WEDNESDAY, 31 AUGUST** **Plenary Session [OGGB4]** *Chair: Alastair Scott* **Alan Welsh** **Australian National University**

HANDLING NONRESPONSE WHEN FITTING MODELS TO SURVEY DATA

Alan Welsh*
Australian National University
E-mail: Alan.Welsh@anu.edu.au

Nonresponse is a pervasive and difficult problem with sample survey data. It is often treated by implicitly making strong assumptions which produce a simple analysis, but it is better to formulate explicit models for the nonresponse mechanism and to try to collect data so we can check the models. Within this framework, there are various ways (raising different issues) we can approach the problem. The relationships between different approaches are not always clear. In this talk, we will consider fitting simple models (for analytic inference) to sample survey data which is subject to informative nonresponse. We will use the maximum likelihood approach implemented by the Missing Information Principle (Orchard and Woodbury, 1972, Proceedings of the 6th Berkeley Symposium on Mathematical Statistics) and show how it works. We will discuss the use of follow up data, selection models and imputation methods, focussing on the relationships between them and the issues they raise.

*****

1410- 1510	**WEDNESDAY, 31 AUGUST** **Contributed Session: Applied statistics 2 [OGGB4]** *Chair: Alastair Scott*

ATTRITION IN THE LONGITUDINAL IMMIGRATION SURVEY: NEW ZEALAND (WAVE1 TO WAVE3)

Maoxin Luo*
Statistics New Zealand
E-mail: joe.luo@stats.govt.nz

The Longitudinal Immigration Survey: New Zealand (LisNZ) provides information on the initial settlement experiences of migrants in New Zealand. It has been widely used in policy research and academic studies. However LisNZ, like almost all longitudinal surveys, is subject to attrition. Approximately 15 percent of respondents who were interviewed in wave 1 could not be re-interviewed in wave 2; and about another 15 percent of respondents who were interviewed in wave 2 could not be re-interviewed in wave 3. There are two aspects we investigate in this report (1) the characteristics of attritors (People who left the survey) and non-attritors (People who stayed in the survey); and (2) whether the attrition in LisNZ leads to selection bias in both cross-sectional and dynamic models. Four tests are applied: (i) Are the distributions of characteristics between non-attritors and attritors significantly different? (ii) Is attrition in wave 3 related, after controlling for standard explanatory variables, to outcomes in wave 1 and wave 2 respectively? (iii) Does the relationship between outcome and explanatory variables differ between non-attritors and the complete sample? (iv) Is attrition in wave 3 related to the change of social and economic statuses from wave 1 to wave 2 by using dynamic model? The first test describes the distinct characteristics of non-attritors and attritors. All the four tests determine the existence of selection bias. However, the selection bias is not significant according to our test results. Hence, attrition is unlikely to compromise research and analysis using LisNZ data.

*****

AN ASSUMPTION-FREE SMALL-SAMPLE PROCEDURE FOR THE DIFFERENCE IN MEDIANS

Robin Willink*

E-mail: robin.willink@gmail.com

The Wilcoxon-Mann-Whitney and two-sample-median procedures are valid only under the location-shift model, which may be criticised as being unrealistic, especially in the non-parametric context. This talk presents a small-sample distribution-free two-sample test of medians that requires no such model. A confidence-interval procedure for a difference in medians follows from this test. The principle that null and alternative hypotheses in a test are to be logically complementary is stressed.

*****

MEASURING THE PRICE MOVEMENTS OF USED CARS AND RESIDENTIAL RENTS IN THE NEW ZEALAND CONSUMERS PRICE INDEX

Frances Krsinich*
Statistics New Zealand
E-mail: frances.krsinich@stats.govt.nz

Price movements of second hand cars and residential housing rentals in the New Zealand Consumers Price Index are both based on large samples of data that lend themselves well to hedonic regression methods, where the price index is derived from the parameters for time controlling for other price-determining characteristics. In 2000 the existing stratification method for second hand cars was replaced by a hedonic method and this is being updated with an improved hedonic model in 2011. We will present the updated model and explain how it is used in production. The rental index, which currently uses a matched-sample approach, was recently assessed using a hedonic index as a benchmark. The longitudinal nature of the rental survey lets us control for unobserved characteristics in the hedonic model, but questions were raised about how well this hedonic formulation deals with the newly rented dwellings. We will discuss where this investigation has led.

*****

1410- 1510	**WEDNESDAY, 31 AUGUST** **Contributed Session: Statistical computing [OGGB5]** *Chair: Robert Gentleman*

WHAT’S IN A NAME?

Paul Murrell*
University of Auckland
E-mail: p.murrell@auckland.ac.nz

The plots produced by statistical graphics systems are complex images, consisting of a carefully designed arrangement of many individual shapes. Providing fine control over all aspects of the individual shapes and their arrangement leads to complex interfaces. In the case of R plotting functions, this means very long argument lists. This talk will discuss a possible solution to this problem: implementing a naming scheme for statistical graphics so that each shape in an image has its own label. There will be a discussion of several benefits that flow from this approach.

*****

INFODECOMPUTE: AN R PACKAGE FOR INFORMATION DECOMPOSITION IN TWO-PHASE EXPERIMENTS

Kevin C. Chang*, Richard G. Jarrett, Chris M. Triggs, and Katya Ruggiero
*University of Auckland**
E-mail: kcha193@aucklanduni.ac.nz

Studies in which an experimental unit's response to treatment cannot be measured directly are said to be two-phase. In such cases, material harvested from the experimental units requires further processing in a subsequent experiment before measurements can be made. Consequently, each experimental phase introduces different sources of variation and how these interact with one another depends on the experimental designs for each phase, e.g. they may not yield a valid F-test in the analysis of variance. To assess the properties of competing designs for two-phase experiments, it is necessary to examine their theoretical ANOVA tables, which can be a very time-consuming exercise to perform manually. We will introduce our very flexible R package, infoDecompuTE, which for a given single- or two-phase experiment will quickly construct the ANOVA table, showing any existing strata, expected mean squares for all sources of variation and average efficiency factors, as appropriate.

*****

SOFTWARE FOR DISTRIBUTIONS

David Scott*
University of Auckland
E-mail: d.scott@auckland.ac.nz

I will describe the R packages I have created for dealing with distributions: DistributionUtils, GeneralizedHyperbolic, SkewHyperbolic, VarianceGamma, and NormalLaplace. I will first outline the common design of the packages. I will then concentrate on recent developments already incorporated into the packages, and ongoing development work.

*****

1410-1510

WEDNESDAY, 31 AUGUST
Contributed Session: Mixture models and methods 2 [OGGB3]
Chair: Yong Wang

FITTING MIXTURE MODELS MADE EASY

Murray Jorgensen*
University of Waikato
E-mail: maj@waikato.ac.nz

I will outline a general strategy for fitting mixture models, and show why it turns out not to be too hard to do in R and some other statistical software.

*****

GARCH MODEL WITH SCALE NORMAL MIXTURE ERRORS

Michael Kao* and Yong Wang
University of Auckland
E-mail: kobegoya@hotmail.com

This paper presents an extension to the widely used volatility model GARCH to account for the empirical heavy tails. The ARCH model was first developed by Engle in 1982 and then further generalised to GARCH by Bollerslev in 1986. It is often critisied for the inadequacy to capture the natural heavy tail of the typical financial time series. In the past, improvements have been achieved through the adoption of heavy tailed distributions such as t and Generalised Error distributions, or via the use of simulation based methods. We extend the previous work of Yong (2007, 2009) and Chang (2010) by relaxing the assumption of the error distribution to a finite mixture of scale normal distributions. There are several advantages associated with this method. Firstly, the mixture distribution is much more flexible than the other heavy tail distributions mentioned above. Secondly, this approach has an implicit interpretation; the risk can be decomposed into time dependent market risk results from market turbulence and also time independent risk such as financial crisis and the risk of default. Moreover, if the time independent risk consists of multiple component densities then it is possible to further correspond each risk driver to a specific source, thus gaining an increased understanding of the risk structure confronted.

*****

MODE-BASED CLUSTERING USING NONPARAMETRIC MIXTURE MODELS

Xuxu Wang*
University of Auckland
E-mail: xwan302@aucklanduni.ac.nz

Most existing model-based clustering methods have difficulties with revealing irregular multi-dimensional clusters which are overlapping and not linearly separable by using only single normal distribution models. A new clustering approach based on mode identification and nonparametric mixture models is proposed to solve such problems. The new approach estimates the density function through multivariate nonparametric mixture models and recognizes clusters by finding modes of a resulting mixture models. Because of the likelihood function involved, theoretic-information model selection criteria can be readily used to reduce variable dimensions. Experiments on simulated and real data sets demonstrate that the new approach tends to solve the irregular high- dimensional clustering problem well. By comparing the estimated clustering results with the true confirmed groups of observations, studies show that the new method has a satisfactory performance with very low classification error.

*****

Poster Presentation Abstracts

1220- 1330	TUESDAY, 30 AUGUST Poster session (LISTED ALPHABETICALLY BY SUBMITTING AUTHOR)

BAYESIAN INFERENCE ON EMRI SIGNALS IN LISA DATA

Asad Ali*, Renate Meyer, Nelson Christensen and Christian Roever
*University of Auckland**
E-mail: asad.ali@aucklanda.c.nz

Extreme mass ratio inspirals (EMRIs) are one of the most exciting sources of gravitational waves (GWs) that laser interferometer space antenna (LISA) is expected to observe. These inspirals are formed when a stellar mass compact object (CO) with mass 1-10 x Solar Mass is captured in a strong orbit of a spinning super massive black hole (SMBH) with mass 10^5-10^7 x Solar Mass and subsequently, under the influence of gravitational radiation, inspirals gradually into SMBH through the emission of GWs. The detection and characterization of signals from such sources will help to understand the structure and formation of SMBHs and the characteristics of the space-time around them, such as lense-thirring or frame-draging effects. Bayesian approach equipped with sophisticated Markov chain Monte Carlo (MCMC) algorithms has been used for GWs signal processing with great successes.

*****

TWO-COMPONENT BAYESIAN TIME-SERIES MODEL AND THE AUTOREGRESSIVE INTERGRATIVE MOVING AVERAGE (ARIMA) MODEL IN THE SURVEILLANCE OF DENGUE

Arul Earnest*, Tan Say Beng and Annelies Wilder-Smith
*Duke-NUS Graduate Medical School, Singapore**
E-mail: arul.earnest@duke-nus.edu.sg

We compared two common statistical models that can be used in the surveillance and forecast of notifiable infectious diseases, namely the Autoregressive Integrated Moving Average (ARIMA) model and the Bayesian two-component model. The Mean Absolute Percentage Error (MAPE) was used to compare between the models. We performed external validation of the models using data on notifiable dengue fever in Singapore from January 2001 till May 2008. The two-component model resulted in a slightly lower MAPE value (17.21 versus 17.54 for the ARIMA model). We conclude that the models' performances are similar. However, the Bayesian model has added advantages, including its ability to incorporate prior information, and the extension of complex terms (e.g. spatial random effects) in the model.

*****

MULTIVARIATE GAUSSIAN PROCESSES

Robin Hankin*
AUT University
E-mail: robin.hankin@aut.ac.nz

The "emulator" technique allows one to predict the output of complex computer codes without actually running them. Based on the statistical theory of Gaussian Processes, emulators have been used in climate science, econometrics, and oceanography.

In the standard univariate case, a finite number of observations of a single type is made; here, observations of different types are considered. A multivariate generalization of the emulator technique is presented in which an arbitrary multivariate function may be assessed. The representative example chosen is temperature and rainfall which are predicted over the surface of the Earth using a climate model.

The technique has the property that marginal analysis (that is, considering only a single observation type) reduces exactly to the univariate theory. The associated software is used to analyze datasets from the fields of economics and climate change. This work is currently under review at the Journal of Statistical Software.

*****

NONPARAMETRIC ESTIMATION AND TEST OF CONDITIONAL KENDALL'S TAU UNDER SEMI-COMPETING RISKS DATA AND TRUNCATED DATA

Jin-Jian Hsieh* and Wei-Cheng Huang
National Chung Cheng University
E-mail: jjhsieh@math.ccu.edu.tw

In this article, we focus on estimation and test of conditional Kendall's tau under semi-competing risks data and truncated data. We apply the Inverse Probability Censoring Weighted (IPCW) technique to construct an estimator of conditional Kendall's tau, τ_c We provide a test statistic for$H_0\colon \tau_c = \tau_0$, where $\tau_0 \in (-1,1)$. When two random variables are quasi-independent, it implies $\tau_c = 0$. Thus, $H_0\colon \tau_c = 0$ is a proxy for quasi-independence. Tsai (1990), and Martin and Betensky (2005) also considered the testing problem for quasi-independence. We compare the three test statistics for quasi-independence in simulation studies. Furthermore, we provide the large sample properties for our proposed estimator. We also examine the finite-sample performance of the proposed estimator and the suggested test statistic via simulations. Finally, we provide two real data analyses for illustration.

*****

SELECTING SUBSETS OF EXPERIMENTAL POPULATIONS TO MAXIMIZE GENETIC DIVERSITY

Emma Huang*, David Clifford, and Colin Cavanagh
*CSIRO**
E-mail: emma.huang@csiro.au

Selective phenotyping (i.e. phenotyping a subset of individuals based on their genetic background) is a cost-effective means of capturing information about gene-trait relationships within a population. In particular, we discuss its application to multi-parent advanced generation intercross (MAGIC) populations. In these studies, although genotypes are collected on thousands of lines to provide a large population resource, much smaller samples are phenotyped to investigate individual traits. The diversity within the sample gives an indication of the efficiency of this information capture; less diversity implies greater redundancy of the genetic information and greater risk that the sample does not fully represent the source population. We propose a method to maximize genetic diversity within the selected sample. Our method is applicable to general experimental designs, and robust to common problems such as missing data and dominant markers. Through simulation, we compare our method to simple random sampling. Our method has improved power and results in a much more diverse sample of genotypes. When applied to real data from a four-parent MAGIC population, our method detects known QTL more accurately than in random samples.

*****

BAYESIAN UNIT-ROOT TESTS IN STOCHASTIC VOLATILITY MODELS FOR FINANCIAL TIME SERIES

Christian Hubschneider* and Renate Meyer
Karlsruhe Institute of Technology
E-mail: christian.hubschneider@student.kit.edu

Stochastic volatility models are important models used in the field of mathematical finance to evaluate derivative securities, such as options. They are one approach to resolve a short- coming of the Black-Scholes model, namely the assumption that the underlying volatility is constant over time, by treating the volatility as a random process. They are an alternative to commonly used ARCH/GARCH models. Jacquier, Polson and Rossi (1994) developed frequentist and Bayesian parameter estimation techniques and showed that the Bayesian estimates were more efficient. A very controversial topic in econometrics is the existence of a unit root in the volatility of financial assets. Several Bayesian approaches to unit-root testing have appeared in the recent literature, e.g. So and Li (1999), Koop (2006). Kalyli- oglu and Ghosh (2009), Li and Yu (2010). The aim of this project was to compare, contrast and potentially extend these approaches as well as implement efficient Markov chain Monte Carlo methods such as for instance specified in Yu and Li (2010) and Li and Zhang (2011) for the posterior computation of Bayes factors. Frequentist properties of the procedures like ‘size’ and ‘power’ were investigated in a simulation study. The approaches were also applied to financial time series data in a case study.

*****

UNDERSTANDING SOIL FERTILITY CHANGES AFTER SUSTANTIAL LAND MODIFICATION ON THE WEST COAST OF NEW ZEALAND

Esther Meenken*, Steve Thomas, Abie Horrocks, Mike Beare, Craig Tregurtha and Richard Gillespie
*Plant and Food Research**
E-mail: esther.meenken@plantandfood.co.nz

Humping & hollowing is a land development practice used in the West Coast region of New Zealand to improve drainage and dry matter production (DMP) of soils characterised by cemented gravels and hydraulically impermeable pans. Recent research has shown that the soil organic matter (SOM) and nutrient content of these greatly modified soils is initially very low, but increases markedly during the first 7?15 years following modification. We report the results of Nitrogen (N) fertiliser trials aimed at determining the effects of N rate on DMP and N use efficiency. Factors considered were years since modification, position (Hollow, Slope, Hump) and applied nitrogen fertiliser. Since there is no natural implementation of randomising some factors, there were limits on possible designs. These and the consequences for analysis are discussed. In general our results show that DMP increases linearly with N fertiliser rate and that N use efficiency increases markedly with development time and position, with Humps having the highest DMP and Hollows the lowest. Our findings are being used to develop fertiliser management practices that improve N use efficiency and mitigate N losses during development of these highly modified soils.

*****

EARTHQUAKES AND STATISTICS: THE HLFS EXPERIENCE

Emma Bentley, Chris Hansen, Michelle Smith and Nathan Young
Presented by: Richard Penny*
Statistics New Zealand
E-mail: richard.penny@stats.govt.nz

The HLFS is designed to provide estimates of New Zealand's labour force that are representative of the total working-age population. The representativeness of the sample was affected by the earthquakes as we were unable to interview households in affected areas for some time after each earthquake. The September earthquake did not introduce significant bias into the sample, however the February earthquake introduced significant bias at a regional level. With significant bias in the sample, an inaccurate picture of New Zealand's labour force is presented. To preserve the quality of HLFS estimates the methodology was changed for the March 2011 quarter. This change resulted in March 2011 quarter estimates reflecting a picture of the labour market as if the earthquake never occurred. Through imputation, revisions to the March quarter estimates may be made to incorporate some of the effects of the earthquake. Looking forward, alterations to the usual methodology may be required for the June 2011 quarter. The method will not be finalised until after collection for the quarter is complete and an assessment has been made on the representativeness of the dataset.

*****

SPATIO-TEMPORAL DISEASE MAPPING OF FOOT AND MOUTH DISEASE IN VIETNAM

Kate Richards*, Martin Hazelton, Nguyen van Long, Mark Stevenson
Massey University
E-mail: kkrichards@hotmail.com

Foot and mouth disease is a virus that can be transmitted by direct and indirect animal contact as well as by airborne means. It can affect all types of cloven-hoofed animals, but is found principally in cattle, sheep and pigs. Foot and mouth disease worldwide is listed as one of the top 10 agricultural diseases, with an economic impact estimated to be well into the billions of dollars (US), through loss of stock, trade losses, vaccine costs etc. Foot and mouth disease is of critical importance in Vietnam where the main agricultural animals are buffalo, cows and pigs which are all susceptible to this disease. There is considerable animal migration present, creating the potential for devastating spread of the disease over the whole country.

In this poster we describe the application of modern disease mapping techniques to better understand the spatio-temporal distribution of foot and mouth disease in Vietnam, and in particular the patterns in the distribution of its three major serotypes. The available data comprise monthly disease counts by province from March 2006 to January 2009. The data for serotype are not complete, with variable rates of serotype testing present across the country. This provides problems in modelling the data, but also provides the opportunity for our model to be used in the prediction of local disease strains in areas where this information is not directly available. We employ Poisson log-linear models for the disease counts, incorporating conditional autoregressive processes in the linear predictor to account for spatial and temporal dependency in the data. We fit them using MCMC methods implemented in the WinBUGS package. Our models identify two areas of Vietnam with significantly elevated risk, and suggest an overall decrease in the risk of disease from 2006 to 2009. This type of information could inform schemes for vaccination distribution. At present these vaccines are rather expensive, requiring booster vaccine to maintain immunity.

*****

GENETIC CLUSTERING WITH UNKNOWN K: ANALYSIS OF A SET OF TRAFFIC ROAD ACCIDENTS BINARY DATA

Sabariah Saharan*, Jennifer Brown and Marco Reale
Department of Mathematics and Statistics, University of Canterbury, Private Bag 4800, Christchurch 8140, New Zealand
E-mail: sbs40@uclive.ac.nz

Analysis of traffic accidents severity has been an ongoing research topic for many researchers world-wide. Study of traffic accidents becomes increasingly important when the implications of accident's costs and improving public road safety awareness are considered. The aim of most traffic accidents research is to understand the underlying relationships between the factors that are associated with accidents and the effects of these factors on the severity of injuries. Nonparametric and statistical regression models are the common methods used in traffic accidents injury analysis. However, as data collection and storage methods improve, very large datasets are now available. Large data sets combined with the fact that often the data is overlapping and highly imbalanced, means that these standard techniques may not be suitable for analysis. We deal with the data set of the traffic road accidents recorded in Christchurch, New Zealand, from 2000 to 2009. The data is a binary set and for each and every accident the severity level and factors (a single one or, more often, more than one) are registered out of a list of 50. We used genetic algorithms for the analysis because we are in the presence of a large unbalanced data set with overlapping data and standard techniques may not be suitable for the task. The genetic algorithms based clustering have been used to identify the factors associated with accidents of different levels of severity.
The results provided us with an interesting insight into the relationships between factors and accidents severity level and suggest that the two main factors that contributed to fatal accidents are speed more than 60 km h and inattention while driving (driver did not see other people until it was too late). A comparison with the k-means algorithm is performed to validate the results.

Keywords: Binary data; Cluster analysis; Genetic Algorithms; Traffic accidents

*****

A HPD INTERVAL FOR THE PROPORTION DIFFERENCE

Kawasaki Youhei* and Etsuo Miyaoka
*Tokyo University of Science**
E-mail: yk.sep10@rp.mtwave.com

The statistical inference for the difference between two independent binominal proportions has been frequently discussed in papers. This interest arises from the fact that the actual confidence level does not correspond to the nominal con?dence level. The approximate confidence interval is constructed by using the asymptotical argument. Therefore, the finite sample size, particularly small sample sizes, creates difficulties. Many authors have shown that the actual confidence level is lower than the nominal confidence level. On the other hand, the Bayesian approach also has been applied to statistical inference of the binominal proportion. Agresti and Cafio (2000) and Pan (2002) attempted the improvement of the Wald interval by using the Bayes estimator. Agresti and Min (2005) studied the frequentist performance of Bayesian intervals for comparing the proportion of two independent binominal samples and found that the Jeffereys prior performance is as good as the score interval. However, these works have not used an accurate posterior pdf for the difference between two independent binominal proportions. We show the expression of the accurate posterior pdf in this study. We calculate an HPD credible interval by using this expression. In addition, we calculate the approximate credible interval, and compare both the HPD credible interval and the approximate credible interval.

*****

PROBABILISTIC MODELING OF GENE TREES GIVEN SPECIES NETWORKS

Sha Zhu*, James Degnan and Mike Steel
University of Canterbury
E-mail: joe.zhu@pg.canterbury.ac.nz

In phylogenetic studies, trees are used for describing evolutionary histories. In particular, a species tree presents population divergences, and a gene tree indicates the times that genes start to differentiate within populations. Often, the inconsistency between gene trees and species trees makes describing species relationships very difficult. Common causes of the conflict include gene duplication, horizontal gene transfer, incomplete lineage sorting, and hybridization. The coalescent process is used in genetics studies; it starts from the bottom of a species tree, and traces the gene history backwards in time. The coalescent process allows us to calculate the probabilities that gene trees differ from the species tree using times between speciation events. This research will introduce a new way of probabilistic modeling of the coalescent with lineage sorting in hybridized species. In these models, relationships between species are represented by a network rather than a tree, while relationships at the gene level are still represented by trees.

*****

Notes:

Notes:

Notes:

Notes: