BIOMETRICS
by the Harbour

29 Nov – 3 Dec 2015
Hadley's Orient Hotel
HOBART

The International Biometric Society – Australasian Region

PROGRAM and ABSTRACTS

BIOMETRICS
by the Harbour

29 November – 3 December 2015

Hadley's Orient Hotel

Conference of the International Biometric Society (IBS) – Australasian Region

PROGRAM and ABSTRACTS

Cover photo: Tourism Tasmania & Kathryn Leahy

Local Organising Committee (LOC)

Scott Foster (Chair) CSIRO Hobart	scott.foster@csiro.au
Warren Muller (Treasurer) CSIRO Canberra	warren.muller@csiro.au
Yuliya Karpievitch (Website) University Western Australia & Harry Perkins Institute of Medical Research, Perth	yuliya.k@gmail.com
Steve Candy SCandy Statistical Modelling Hobart	burwood70@gmail.com
John McKinlay Australian Antarctic Division Hobart	John.mckinlay@aad.gov.au
Joanne Potts The Analytical Edge Hobart	joanne@theanalyticaledge.com
Miriana Sporcic CSIRO Hobart	Miriana.Sporcic@csiro.au

Scientific Program Committee (SPC)

David Baird (Chair) VSN NZ	david@vsn.co.nz
Ross Darnell CSIRO, Brisbane	ross.darnell@csiro.au
Scott Foster CSIRO, Hobart	scott.foster@csiro.au
Katya Ruggiero University of Auckland	k.ruggiero@auckland.ac.nz
Louise Ryan University of Technology Sydney	Louise.M.Ryan@uts.edu.au
Alan Welsh Australian National University	alan.welsh@anu.edu.au
Rory Wolfe Monash University	rory.wolfe@med.monash.edu.au

Conference Sponsors

The Australasian Region of the International Biometric Society gratefully acknowledges the financial support from the following organisations for this conference.

Prize Sponsors

Contents

About IBS

The International Biometric Society (IBS) is devoted to the development and application of statistical and mathematical theory and methods in the Biosciences, including agriculture, biomedical science and public health, ecology, environmental sciences, forestry, and allied disciplines. It welcomes as members statisticians, mathematicians, biological scientists, and others devoted to interdisciplinary efforts in advancing the collection and interpretation of information in the biosciences.

The Society publishes two journals, Biometrics, reporting communications consistent with the Society's mission, and, jointly with the American Statistical Association, the Journal of Agricultural, Biological, and Environmental Statistics (JABES).

IBS is organised into geographically-defined regions, each with at least 50 members. The Australasian Region is one of 19 regions under the IBS umbrella that provides members of The International Biometric Society who predominantly reside in Australia and New Zealand, a local focus and support for biometrics undertaken in this region. The by-laws for the region date from 1995.

Welcome from the President

As president of the region this year, I would like to welcome you on behalf of the Australasian Region of the International Biometric Society (IBS-AR) to our 2015 biennial conference titled "Biometrics by the Harbour".

I would like to thank Scott Foster and the local organizing committee, and David Baird and the scientific program committee. Please look at the list of committee members on page 1 and 3 and thank them individually for all their hard work in organising this conference.

The IBS-AR conference is always more than a conference, it's a gathering of statisticians working across a wide range of biological science domains throughout the Australasian region and beyond. Conferences are a great opportunity to catch up with colleagues and meet new people and make friends and professional contacts for collaboration. I welcome the invited speakers, and thank those presenting a workshop on Sunday.

This conference we are fortunate to welcome Professor John Hinde, President of the International Biometric Society to our conference. I am sure John would be happy to discuss the role of the broader Society with you during the conference.

I particularly welcome new members so I encourage you to talk to new members and make them feel welcome. One of the great things about these conferences is that they are not so large, giving you the opportunity to get to know and talk to many of the attendees during the week.

I'm pleased to host a student's dinner on the Tuesday night to strengthen links with their peers. The other social events organised by the local committee should allow everyone to connect and catch up.

The program this year has a ecological and environmental bent in keeping with the beautiful local environmental and research institutes, however the program covers the breadth of fields that we work in, from medical, genetics, industry to agricultural.

Finally, my best wishes for you all having a fantastic time at this conference. The committees have started the ball rolling. It's up to you now to make the most of this time.

Ross Darnell
President, Australasian Region
International Biometric Society

INVITED SPEAKERS

Rosemary Bailey (University of St. Andrews)

Sponsored by:

R. A. Bailey is Professor of Mathematics and Statistics at the University of St Andrews. After a doctorate in finite group theory at the University of Oxford, she worked at the Open University for a few years. She embraced statistics while working on restricted randomization and factorial designs as a post-doctoral research fellow at the University of Edinburgh. She spent ten years at Rothamsted Experimental Station designing and analysing agricultural experiments before returning to academia in the University of London. She moved to St Andrews in 2013. She has also held visiting positions and fellowships in France, Australia, New Zealand, the USA and Brazil.

Adrian Bowman (University of Glasgow)

Sponsored by:

Adrian Bowman is a Professor of Statistics in the University of Glasgow. He grew up in the seaside town of Prestwick in Scotland, followed by university education in Glasgow and Cambridge, in Mathematics and then Statistics. Adrian's first academic job was in the University of Manchester but he subsequently moved to Glasgow, where most of his career has developed and where he is now Head of the School of Mathematics & Statistics. He is the joint-author of a book on smoothing techniques and is very active in research, currently on environmental modelling, the analysis of anatomical shapes and of brain images. Adrian also has a longstanding interest in educational technology, particularly in the role of graphics in aiding the understanding of statistical ideas. He is married to Janet, who teaches maths, and they have three grown-up children. A recent fun activity was paddling a sea kayak through the Corryvreckan (you may have to Google that) under strict supervision.

Adrian is supported by the SSAI-WA branch's Frank Hansford-Miller fellowship during his visit.

Richard Emsley (University of Manchester)

Sponsored by:

Richard is a Senior Lecturer in Biostatistics in the Centre for Biostatistics. He is also Visiting Lecturer in the Department of Biostatistics at the Institute of Psychiatry, Psychology and Neuroscience at King's College London. His research aims to answer three key questions: Are treatments effective? How do they work? Which groups are they most effective for? This involves the development of statistical methods for causal inference, efficacy and mechanisms evaluation and stratified medicine. He is the initiator and Chair of the steering group for the UK Causal Inference Meeting (UK-CIM). He is involved in stratified medicine research programmes in schizophrenia, psoriasis, arthritis and cancer.

John Hinde (National University of Ireland)

Sponsored by:

John Hinde is Professor of Statistics at the National University of Ireland Galway. Originally from Oxfordshire, he studied mathematics at the University of Oxford followed by statistics at the University of Kent. John's first post was in the Computing Unit at the University of Newcastle and after a period at the Centre for Applied Statistics at Lancaster University he moved to the University of Exeter. In 2002 he moved to Ireland to take up the newly established Chair of Statistics at Galway. He is the joint author of a book on statistical modelling and has current research interests on extensions of generalized linear models, with particular interest in models for discrete data and overdispersion. He was one of the founding editors of the journal *Statistical Modelling*, has been involved in the running of various statistical societies and is the current President of the *International Biometric Society*. John is a keen hill walker and lover of the outdoors, even when it comes to battling the elements with his wife Kathie to establish a garden at their house in the West of Ireland.

Katherine Lee (Murdoch Childrens Research Institute)

Sponsored by: SSAI Statistical Society of Australia Inc.

Katherine Lee is a senior biostatistician at the Murdoch Childrens Research Institute. She grew up in Bristol in the UK before moving to Nottingham where she obtained a Bachelor of Science in Mathematics from the University of Nottingham. Following this she obtained a Masters of Science in Medical Statistics from the University of Leicester UK and a PhD in Biostatistics from the University of Cambridge. Katherine spent the first 2 years of her working life at the Medical Research Council Clinical Trials Unit in London, before moving to Melbourne. Katherine has over ten years' experience in the design, planning and analysis of randomised trials and observational studies, and is now Associate Director (Biostatistics) at the Melbourne Children's Trials Centre at the Royal Childrens Hospital. She is also an Honorary Fellow in the University of Melbourne. Her methodological interest is in multiple imputation for dealing with missing data.

Thomas Lumley (University of Auckland)

Sponsored by: CSIRO

Thomas Lumley is Professor of Biostatistics at the University of Auckland. He has an undergraduate degree from Monash, an MSc from Oxford, and a PhD from the University of Washington, where he subsequently spent twelve years on the academic staff. His research interests include statistical computing, semiparametric statistics and its connections with survey sampling, cardiovascular epidemiology, and genomics.

Otso Ovaskainen (University of Helsinki)

Sponsored by: Marine Biodiversity Hub, National Environmental Science Programme

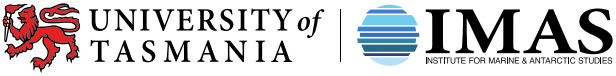
Otso Ovaskainen is an internationally renowned researcher at the interface between mathematics, statistics and biology. His work focusses on connecting general theories with empirical research, which has led to important contributions to metapopulation theory and dispersal theory. He has recently been interested in adapting modern statistical tools to ecology, including diffusion-based stochastic processes for the study of animal movement, and latent variable models for community ecology and the study of species interactions. Otso is a Professor in the Department of Biosciences at the University of Helsinki, Finland as well as in the Norwegian University of Science and Technology in Trondheim, Norway. He has authored over 100 publications, including in Nature and Science. For his contributions to theoretical and population ecology, he won the prestigious 2009 Academy of Finland Award, awarded to only two scientists nationally each year.

Jay Ver Hoef (National Oceanic and Atmospheric Administration (NOAA) Fairbanks Alaska)

Sponsored by: Marine Biodiversity Hub, National Environmental Science Programme

Jay Ver Hoef a statistician for the National Marine Mammal Lab of the National Oceanic and Atmospheric Association (NOAA), U.S. Dept. of Commerce. He obtained a Ph.D. at Iowa State University in Statistics in 1991. Jay develops statistical methods and consults on a wide variety of topics related to marine mammals and stream networks. The Marine Mammal Lab is located in Seattle, Washington, although Jay lives in Fairbanks, Alaska. His main statistical interests are in spatial statistics and Bayesian statistics, especially applied to ecological and environmental data. Jay is a fellow of the American Statistical Association.

David Warton (University of NSW)

Sponsored by: 

David trained jointly between ecology and statistics to postgraduate level (at Sydney and Macquarie), and his research since then has been at the interface of these two disciplines, addressing what he considers to be a significant knowledge gap. He has organised the Eco-Stats conference (this December at UNSW), and its previous iteration in 2013, to help create connections across these disciplines. His research ranges from applied to theoretical translating modern statistical tools to ecology, and developing new statistical methodology motivated by ecological problems, including contributions to high-dimensional data analysis, resampling, point process modelling and model selection.

CONFERENCE PROGRAM

Time Sunday 29 November 2015

8:30am	**Workshop and Conference Registration**	
	Location: CSIRO's Marine Laboratories, Reception Desk, Castray Esplanade	
9:00am - 5:00pm	**Short Courses**	
	Short course 1: **Multiphase experiments: from design to analysis**	**Short course 2:** **An introduction to flexible regression for environmental data**
	Presenters: Rosemary Bailey, Chris Brien and Alison Smith	**Presenter:** Adrian Bowman
	Short course 3: **Causal inference in randomised trials**	**Short course 4:** **Analyzing community data with Hierarchical Modelling of Species Communities (HMSC) with R or Matlab**
	Presenters: Richard Emsley	**Presenter:** Otso Ovaskainen
5:30pm	**Welcome Reception**	
	Location: Hadley's Orient Hotel (Mary Hadley Room)	

Time Monday 30 November 2015

8:15am – 12:30pm	**Conference Registration – Hadley's Orient Hotel**	
9:00	**Welcome Ceremony** Scott Foster, Chair LOC Ross Darnell, President IBS-AR John Hinde, President IBS	
	General Statistics **Room: John Webb** **Chair: Chris Triggs**	**Health** **Room: George Cartwright** **Chair: Brenton Clarke**
9:20	**John Carlin -** Misplaced confidence in confidence intervals	**Sophie Zaloumis -** Application of a Bayesian Markov chain Monte Carlo approach for modelling the dynamics of Plasmodium falciparum parasitaemia in severe malaria patients
9:40	**Alan Welsh -** Evaluating Frequentist Model Averaged Confidence Intervals	**Jake Olivier -** Relative Effect Sizes for Measures of Risk
10:00	**Ruth Butler** – "Myths" in presenting statistical results (or: statistical bees I have in my bonnet…..)	**Rory Wolfe** Health effects of disasters
10:20	**Morning tea (Leadlight Room)**	
	Invited talks in John Webb Room (Chair: David Baird)	
10:50	**John Hinde -** A Medley of Mixtures	
11:40	**Rosemary Bailey -** The design key in single and multi-phase experiments	
12:30	**Lunch (Leadlight Room)**	
	Mixture Models **Roome: John Webb** **Chair: Samuel Mueller**	**Design/Estimation** **Room: George Cartwright** **Chair: Rory Wolfe**
1:30	**Brenton Clarke -** A Comparison of the L_2 Minimum Distance Estimator and the EM-Algorithm when Fitting k-Component Univariate Normal Mixtures	**Graham Hepworth -** Design and analysis of simulation experiments
1:50	**Shirley Pledger -** The fourth corner problem: niche overlap using mixtures	**Tibor Schuster -** Efficient recruitment strategies in randomised controlled trials with continuous outcomes
2:10	**M.A.C.S.S. Fernando -** Limitations of Hierarchical and Mixture Model Comparisons	**Jeff Wood -** How many letters are there in the alphabet?
2:30	**Firouzeh Noghrehchi -** Multiple imputation as a stochastic EM approximation to maximum likelihood	**Jakub Stoklosa -** On quadratic logistic regression models when predictor variables are subject to measurement error

2:50	**Afternoon tea (Leadlight Room)**

Time Monday 30 November 2015 (cont.)

	SSAI Sponsored Session on Multiple Imputation Room: John Webb Chair: Jake Olivier	**Mixed Models I Room: George Cartwright Chair: Mario D'Antuono**
3:20	**Katherine Lee -** Multiple imputation: a miracle cure for missing data?	**Lauren Borg (presented by Alison Smith) -** On spatial model selection for the analysis of field experiments with an application to plant improvement programs
3:40		**Nicole Cocks (presented by Brian Cullis) -** Improving the Computational Efficiency of Model-Based Design with an Application to the Design of Multiphase Experiments
4:00	**Julie Simpson -** Sensitivity analysis within the multiple imputation framework: The pattern-mixture method	**Chong You -** A study of one and two stage analyses for genomic prediction of yield in wheat
4:20	**Cattram Nguyen -** Diagnostic methods for checking multiple imputation models	**Daniel Tolhurst -** Improving the accuracy of genetic predictions for expensive multi-phase traits
4:40	**Margarita Moreno-Betancur -** Multiple imputation and sensitivity analysis for incomplete longitudinal data departing from the missing at random assumption	**Emi Tanaka -** A study of correlated marker effects for dense linkage map in QTL analysis in wheat
5:00	**Poster Presentations (John Webb Room)**	
5:30	**Drinks and poster viewing (Leadlight and Mary Hadley Rooms)**	

Time	Tuesday 1 December 2015	
	Models **Room: John Webb** **Chair: Alan Welsh**	**Genetics** **Room: George Cartwright** **Chair: Thomas Lumley**
9:00	**Russell Millar -** WIC and importance sampling approximations to cross-validation for models with observation-level latent variables	**Louise McMillan -** Visualizing Population Genetics
9:20	**Garth Tarr -** Interactive and data adaptive model selection with mplot	**Daisy Shepherd -** Sliding Through Phylogenetics
9:40	**Alan Huang -** Vector regression without specifying marginal or association structures	**Teresa Neeman -** New Models of Molecular Evolution
10:00	**Chanatda Somchit -** P-Spline Vector Generalized Additive Model and Its Application	**Wei Zhang -** Stationary distribution of the linkage disequilibrium coefficient r^2
10:20	**Morning tea (Leadlight Room)**	
	Invited talks in John Web Room (Chair: Kaye Basford)	
10:50	**Thomas Lumley -** Family-based genetic association modelling in a multistage sample	
11:40	**Richard Emsley -** Stratified medicine: the essential role of mechanisms evaluation	
12:00	**Lunch (Leadlight Room)**	
1:00-onwards	**Excursions (Foyer of Hadley's Orient Hotel)**	

Time Wednesday 2 December 2015

	Multispecies analysis **Room: John Webb** **Chair: Steve Candy**	**Agricultural Modelling** **Room: George Cartwright** **Chair: Alison Smith**
9:00	**Daniel Fernandez -** Categorising Ecological Community Count Data	**Alison Kelly -** Selection of genotypes for resistance and tolerance to pathogens: a multivariate statistical analysis of yield and disease response.
9:20	**Nicole Hill -** Modelling the distribution of sub-Antarctic and Antarctic demersal fish communities: An application of new community –modelling method	**M. Gabriela Borgognone -** Including molecular marker information in the analysis of multi-environment trial data helps differentiate superior genotypes from promising parents: a wheat example
9:40	**Loïc Thibaut -** Inference for multivariate abundance data using generalised mixed effects models and the PIT-trap	**Ky L Mathews -** Selecting environment covariates to explain GxE: A comparison of cyclic forward regression and subset regression approaches.
10:00	**Gordana Popovic -** Approximate likelihood ratio test for multivariate abundance data	**Martin Upsdell -** Fitting quadratic peaks to fluorescence data to identify chemical compounds
10:20	**Morning tea (Leadlight Room)**	
	Invited talks in John Webb Room (Chair: Scott Foster)	
10:50	**Jay Ver Hoef -** Estimating Abundance from Counts in Large Data Sets of Irregularly Spaced Plots using Spatial Basis Functions	
11:40	**Otso Ovaskainen -** Analyzing community data with joint species distribution models: abundance, traits, phylogeny, co-occurrence and spatio-temporal structures	
12:30	**Lunch (Leadlight Room)**	
	Spatial Point Processes **Room: John Webb** **Chair: David Warton**	**Genomics** **Room: Geroge Cartwright** **Chair: Kathy Ruggiero**
1:30	**Ian Renner -** Point process models for presence-only analysis	**Kim-Anh Le Cao -** A multivariate approach for multiple 'omics data integration and biomarker discovery
1:50	**Samantha Peel -** Assessing a species distribution model's performance for sparse and patchy presence data.	**Anne Bernard -** Sparse Multiple Correspondence analysis for selection of Single Nucleotide Polymorphisms
2:10	**Wesley Brooks -** Spatial confounding may cause bias in regression models for presence-only data	**Florian Rohart -** Integrative meta analyses to combine transcriptomics studies
2:30	**Ramethaa Pirathiban -** Eliciting and encoding expert knowledge	**Conrad Burden -** Estimation of the amplicon methylation pattern

	on variable selection into classical or Bayesian species distribution models	distribution from bisulphite sequencing data

Time Wednesday 2 December 2015 (cont.)

2:50	**Afternoon tea (Leadlight Room)**	
	Ecology – Models **Room: John Web** **Chair: Warren Müller**	**Longitudinal Models** **Room: George Cartwright** **Chair: Malcolm Hudson**
3:20	**Natalie Kelly -** Humpback whales in the Great Barrier Reef: model-based inferences on distribution and abundance	**Alastair Scott -** Analysing longitudinal data with outcome-dependent sampling
3:40	**Wen-Hsi Yang -** Mapping Soil Regolith Depth in Large and Censored Spatial Datasets Using Bayesian Hierarchical Models	**Sam Brilleman -** Joint longitudinal and survival models for investigating the association between natural disasters and disability whilst accounting for non-random dropout due to death
4:00	**Melissa Dobbie -** Estimating relative species abundance from partially-observed data	**Pauline Ding -** Bootstrap influence on the variance components in the longitudinal data with multiple source of variation
4:20	**Jan Jansen -** The influence of surface productivity on seafloor food-availability and biodiversity distributions on the George V shelf, East Antarctica	**Liliana Orellana -** A novel methodology for inhomogeneity identification in climate time series
4:40	**Emma Lawrence -** An experimental design for testing Bycatch Reduction and Turtle Excluder Devices in the PNG Prawn Trawl Fishery	**Jessica Kasza -** Assessing the impact of unmeasured confounding: confounding functions for causal inference
5:00	**Australasian Biometrics Society AGM (John Webb Room)**	
6:00	**Bus departs for:** **Conference Dinner at Cascade Brewery Visitor's Centre**	
6:30	**Pre-dinner drinks commences (Cascade Brewery Visitor's Centre)**	

Time	Thursday 3 December 2015	
	Mixed Models 2 **Room: John Webb** **Chair: Brian Cullis**	**Survival Analysis** **Room: George Cartwright** **Chair: John Carlin**
9:00	**Clayton Forknall -** Development of a bivariate linear mixed model to describe yield response due to disease.	**Valérie Garès -** Estimating the correlation between competing risks
9:20	**Francis Hui -** Joint Effects Selection in Mixed Models using CREPE	**Malcolm Hudson -** Factors affecting treatment recurrence and death: a case study with longitudinal hospital retreatment records
9:40	**Vanessa Cave -** Testing significance of random terms in a linear mixed model	**Kasun Rathnayake -** Penalized likelihood parameter estimation for additive hazard model using method of multipliers
10:00	**Hwan-Jin Yoon -** Fitting linear mixed models under misspecification	**Maheswaran Rohan -** Computing Standard Error for Half-life of Rotenone
10:20	**Morning tea (Leadlight Room)**	
	Invited talks in John Webb Room (Chair: Ross Darnell)	
10:50am	**David Warton -** The case of the missing model: the modernisation of multivariate analysis in ecology	
11:40am	**Adrian Bowman-** Visualising the environment: data, models and graphics	
12:30pm	**Lunch (Leadlight Room)**	

Social Program and Transport

Welcome reception
Sunday 29 November 2015
Drinks will be served at Hadley's Hotel from 5.30 – 7.30pm.
This event is free for delegates who are attending the whole conference and delegates registered for the short courses. The cost is $50 for accompanying guests or day delegates.

Poster Presentations
Monday 30 November 2015
Drinks and poster viewing will be at Hadley's Hotel from 5:30pm.

Conference Tours
Tuesday 1 December 2015
Those who have arranged to go on one of the Conference tours will depart from the hotel foyer at 1.00pm.

President's dinner for Young Statisticians
Tuesday 1 December 2015
Traditionally, this is an evening when the President and/or regional council members meet with students and newly graduated statisticians at the conference. We usually meet at the conference venue at 7pm and this conference we will go together to the nearby Custom's House Hotel. It is a good way for young statisticians to get to know each other and older members of the Society. Please let a member of the Local Organising Committee know if you would like to attend.

Conference Dinner
Wednesday 2 December 2015
The conference dinner will be held at the Cascades Brewery from 6:30pm to 11:30pm. Drinks (beer, wine and non---alcoholic) will be provided from 6:30pm (until the tab runs out!). Dinner will commence at 7pm, This event is free for delegates who are attending the whole conference. The cost is $100 for accompanying guests or day delegates.
Transport:

1. The red (open-top, weather permitting) double decker bus. This bus is free and will depart Hadley's at 6 pm. The return trip for the bus is scheduled for 11 pm. The seated capacity is 75 max so if you miss out, don't worry we will have additional transport arranged.
2. For those who have the time and energy the walk from Hadley's to the Cascade Brewery and Visitor Centre is a mostly level 3.5 km including the picturesque Cascade Gardens just before the Brewery.
3. There is a Metro Bus service leaving from Franklin Square (eastern side on Elizabeth St, a few minutes walk from Hadley's). The bus number is 46 or 47 and for those wishing to check out the Gardens without the long walk there are buses departing at 4:10, 4:40, 5:10, 5:30 (for full timetable see http://www.metrotas.com.au/timetables/hobart/fern-tree-via-st-johns/).

Get off the bus at the Brewery bus stop with historic Brewery Building just across the Road. The Cascade Gardens are just east of the Brewery. The Visitor Centre where the Conference Dinner will be held is just up the hill on the opposite side of Cascade Rd to the Brewery. You can purchase a single trip ticket from the driver.

4. Taxis will be booked for groups that do not get the red decker bus and have not made other transport arrangements. So for these dinner guests it would assist us if you can wait outside Hadley's for taxi pickup between 6:10 and 6:20 pm. Similar arrangements will be made for a late return to the city from 11:30 pm.

Airport Transfer
Thursday 3 December 2015

Depending on interest, there wil be a bus from Hadley's Orient Hotel to the airport after lunch on Thursday afternoon. Cost will be no more than the commercial counterpart. Alternatively taxis are readily available and a fare will cost around A$50-A$60.

Abstracts – Invited Speakers

The design key in single and multi-phase experiments

Rosemary Bailey

The University of St Andrews, UK

rab@mcs.st-andrews.ac.uk

Desmond Patterson introduced the design key in 1965 in the context of experiments on crop rotations. It can be used whenever the treatments have factorial structure, the experimental units have a poset block structure, and an orthogonal design is required. The design key gives an algorithm for allocating treatments to experimental units, and another algorithm for identifying which stratum contains which treatment effect. These two properties make it a very useful tool when extended to multi-phase experiments.

Visualising the environment: data, models and graphics

Adrian Bowman

University of Glascow

adrian.bowman@glasgow.ac.uk

This talk aims to reflect a little on the languages we use to explain, discuss and communicate statistical concepts, models and analysis, both within our own community and beyond it. There will be particular emphasis on how we communicate uncertainty. Several types of analysis will be considered, mostly involving flexible regression and focussed largely on different forms of spatiotemporal data. However, there will be a strong focus on the role of graphics, which can provide a powerful means of conceptual communication and give clear expressions of the insights provided by models, while remaining true to the issues associated with uncertainty.

Stratified medicine: the essential role of mechanisms evaluation

Richard Emsley

University of Manchester

richard.emsley@manchester.ac.uk

The idea that a given treatment will have more benefit for some patients than for others is the underlying foundation of what is termed personalised or stratified medicine. If we can identify who these patients are before the decision to give a treatment is made, it will help prevent other patients being exposed to an unnecessary treatment that is likely to be of little or no benefit. As well as the obvious benefits to individual patients, it will also mean that the health care providers can save money on expensive treatment costs by targeting the right treatments to the right patients at the right time. In many clinical scenarios there are hypothesised to be multiple predictive markers that could be combined to target treatments, and often these are of different modalities such as clinical, genetic and imaging markers. In this talk, I will present the underlying concepts of stratified medicine in terms of benefits and harms and describe why the underlying treatment mechanism is fundamental to evaluation of targeted treatments. I will discuss the issue of choosing the appropriate scale of interaction (additive versus multiplicative), methods for evaluating and combining multiple predictive markers, and prospective trial designs which use analysis methods from the causal inference literature.

A Medley of Mixtures

John Hinde

National University of Ireland

john.hinde@nuigalway.ie

Mixture models have now become an important tool in statistical modelling, with the EM algorithm being an important driver in their popularization and applicability. This talk will present the basic ideas and discuss a range of different mixture models and their use in addressing specific applied problems. Examples discussed will include: mixtures of non-linear regression models for fish growth studies; mixtures of smooth curves for clustering time-course microarray data; mixture orthogonal regression models for sugarcane SNP data; regression mixture models for outlier detection/accommodation.

Multiple imputation: a miracle cure for missing data?

Katherine Lee [1,2]

[1] Clinical Epidemiology and Biostatistics Unit, Murdoch Childrens Research Institute, Melbourne, Australia

[2] Department of Paediatrics, Faculty of Medicine, Dentistry and Health Sciences, University of Melbourne, Melbourne, Australia

Researchers conducting clinical and epidemiological studies are inevitably faced with problems of drop-out and missing data. Importantly, the participants who drop out of a study are often those with poor health resulting in selection bias in the available data. Multiple imputation is gaining popularity as a strategy for handling missing data, since it enables all participants to be included in the analysis. However it is far from a panacea for dealing with missing data, and a number of important questions remain regarding its practical application.

In this talk I will provide a brief introduction to multiple imputation and will review some of the research surrounding this powerful and versatile approach for handling missing data conducted by our research group. This will include a discussion of when multiple imputation is likely to be beneficial over a complete case analysis, which imputation procedure to use, and the imputation of skewed, limited range and semi-continuous variables. Importantly I will highlight that multiple imputation is not a miracle cure and that it can introduce bias in the estimation of parameters of interest if the imputation model is not appropriate.

Family-based genetic association modelling in a multistage sample

Thomas Lumley

University of Auckland

t.lumley@auckland.ac.nz

The Hispanic Community Health Study/Study of Latinos (HCHS/SOL) is a US cohort study following a multistage probability sample of approximately 16000 Hispanics/Latinos. One component of HCHS/SOL is a set of large-scale genetic association studies for which the standard analysis would be linear mixed models with terms for ancestry and relatedness as well as shared environment. Currently, methodology is not available for incorporating complex sampling in mixed models except where the sampling structure is nested in the model structure, which is not the case for HCHS/SOL. I will talk about inferential aspects of the modelling, and about computational issues when dealing with hundreds of thousands of genetic variables and the complex sampling and mixed-model correlation structures. This is joint work with Xudong Huang and Alastair Scott.

Analyzing community data with joint species distribution models: abundance, traits, phylogeny, co-occurrence and spatio-temporal structures

Otso Ovaskainen

University of Helsinki

otso.ovaskainen@helsinki.fi

A key aim in community ecology is to understand the factors that determine the identities and abundances of species found at any given locality. Central concepts in this research field include the regional and local species pools, environmental filtering and biotic assembly rules. Typical datasets involve a matrix of presence-absences (or abundances) for a group of species at different sites, some environmental and geographical characteristics of those sites, and possibly information on the ecological traits and phylogenetic relationships of the species. The analysis of such data have been traditionally based on ordination approaches, but there is increasing interest to move to model based approaches, in particular joint species distribution models. I present a joint species distribution model that captures the influences of environmental filtering at the community-level by measuring the amount of variation and covariation in the responses of individual species to various characteristics of their environment. The selection of the local species pool from the regional species pool involves both deterministic (e.g. systematic differences in dispersal abilities) and stochastic (e.g. spatio-temporal randomness in the realized distribution patterns) processes. Biotic assembly rules are reflected in the model with the help of an association matrix, which models positive or negative co-occurrence patterns not explained by the responses of the species to their environment. I use a latent factor approach to enable model parameterization with data on species-rich communities and thus with high-dimensional association matrices. I illustrate the performance of the approach both with simulated and real data.

Estimating Abundance from Counts in Large Data Sets of Irregularly Spaced Plots using Spatial Basis Functions

Jay Ver Hoef

NOAA Fairbanks Alaska

jay.verhoef@noaa.gov

Monitoring plant and animal populations is an important goal for both academic research and management of natural resources. Successful management of populations often depends on obtaining estimates of their mean or total over a region. The basic problem considered in this paper is the estimation of a total from a sample of plots containing count data, but the plot placements are spatially irregular and non-randomized. Our application had counts from thousands of irregularly spaced aerial photo images. We used change-of-support methods to model counts in images as a realization of an inhomogeneous Poisson process that used spatial basis functions to model the spatial intensity surface. The method was very fast and took only a few seconds for thousands of images. The fitted intensity surface was integrated to provide an estimate from all unsampled areas, which is added to the observed counts. The proposed method also provides a finite area correction factor to variance estimation. The intensity surface from an inhomogeneous Poisson process tends to be too smooth for locally clustered points, typical of animal distributions, so we introduce several new overdispersion estimators due to poor performance of the classic one. We used simulated data to examine estimation bias and to investigate several variance estimators with overdispersion. A real example is given of harbor seal counts from aerial surveys in an Alaskan glacial fjord.

The case of the missing model: the modernisation of multivariate analysis in ecology

David Warton

University of NSW

david.warton@unsw.edu.au

For the best part of four decades, multivariate analysis in ecology has diverged substantially from mainstream statistics, perhaps because state-of-the-art in 1980's statistics was not capable of handling the complexity frequently seen in multivariate abundance data simultaneously collected across many species. But the methods developed in the ecological literature, still widely used today, have some serious shortcomings that suggest they are fast approaching their use-by date. The statistical literature appears to be "catching up" with ecology, in part through technologies to fit quite flexible hierarchical models capable of accommodating key date structure. There is a significant movement now to reunify multivariate analysis in ecology with modern statistical practices. Some key developments on this front will be reviewed, and immediate challenges identified.

Abstracts – Contributed Talks

Sparse Multiple Correspondence analysis for selection of Single Nucleotide Polymorphisms

Anne Bernard

QFAB Bioinformatics, Institute for Molecular Bioscience, University of Queensland, QLD 4072, Australia

a.bernard@qfab.org

Coauthors: Christiane Guinot, Arthur Tenenhaus, Derek Beaton, Hervé Abdi, Gilbert Saporta

In a context of high dimensional data (e.g genomic data) statistical methods such as dimension reduction and variable selection are necessary to extract the most relevant information in a set of several hundreds of thousands of variables. Principal Component Analysis and Multiple Correspondence Analysis are well-known multivariate dimension reduction methods for quantitative and categorical data, respectively. However, the components obtained are combinations of all the original variables, making interpretation of results difficult for high dimensional data. To overcome these difficulties, we propose two new unsupervised methods to select groups of variables for data structured by blocks: "Group Sparse Principal Component Analysis" (GSPCA) for quantitative variables and "Sparse Multiple Correspondence Analysis" (SMCA) for categorical variables. GSPCA is a compromise between the sPCA method of Zou, Hastie and Tibshirani and the "group Lasso" developed by Yuan and Lin to select quantitative variables in a unsupervised context. MCA is a special case of PCA for blocks of dummy variables and SMCA is defined as an extension of GSPCA to select categorical variables such as Single Nucleotide Polymorphisms (SNPs). An application of SMCA will be presented on a real data set of 502 women aimed at the identification of genes affecting skin ageing with more than 370 000 SNPs.

On spatial model selection for the analysis of field experiments with an application to plant improvement programmes

Lauren Borg

School of Mathematics and Applied Statistics, University of Wollongong

lb397@uowmail.edu.au

Coauthors: Alison Smith (presenting), Emi Tanaka and Brian Cullis

Field experiments may be affected by spatial variation, so-called because it is linked to the spatial location of the plots in the field. Examples include fluctuations in soil fertility and moisture and variation due to management practices. Hence spatial analysis methods have become popular and have been shown to provide substantial gains in accuracy for treatment comparisons. In this talk we consider the spatial analysis of Gilmour et al. (1997) which builds on from Cullis & Gleeson(1991). The analysis is based on a linear mixed model and involves a sequence of model fitting and checking in order to appropriately accommodate the spatial variation present in the data. Selections of a reasonable spatial model from a sequence of competing models is currently achieved using informal diagnostics or formal tests of significance and information based criteria. The informal diagnostics can be difficult to interpret and the use of formal tests and information criteria can lead to disconnected comparisons. Here we discuss a new criteria for spatial model selection that provides a unified, objective and fast approach. Additionally, we note that information based criteria are more concerned with model selection when the spatial model itself is of primary interest, as is the case in geo-statistics, for example. In the case of the analysis of field experiments, the spatial model plays a secondary role as a means of improving the accuracy of treatment comparisons. The criteria we propose has this aim in mind. We illustrate the technique using data from a cotton breeding program.

Including molecular marker information in the analysis of multi-environment trial data helps differentiate superior genotypes from promising parents: a wheat example

M. Gabriela Borgognone

Department of Agriculture and Fisheries, Toowoomba, QLD, Australia

Gabriela.Borgognone@daf.qld.gov.au

Coauthors: David Butler, Department of Agriculture and Fisheries, Toowoomba, QLD, Australia; Francis Ogbonnaya, International Center for Agricultural Research in the Dry Areas (ICARDA), Beirut, Lebanon, and Grains Research and Development Corporation, ACT, Australia; M. Fernanda Dreccer, CSIRO Agriculture Flagship, Gatton, QLD, Australia

The statistical analysis of multi-environment trial data aims to provide reliable and accurate predictions of genotype performance across the target environments and information on specific performance from the interaction of genotypes with the environments. Genetic gain can be achieved faster when selections are based on predictions from a model that accounts for the relationships among genotypes rather than from a model that assumes unrelated genotypes. Yield and plant height data from 37 international wheat trials were analysed using a linear mixed model that accounted for relationships among the genotypes via a genomic relationship matrix derived from 2487 polymorphic DArT molecular markers for 197 genotypes. The elements of this matrix reflect the actual proportion of the genome that is identical by state between pairs of individuals and including it into the model resulted in generally lower average prediction error variances of individual trials in the analyses. The models were fitted in ASReml 4 (Gilmour et al. 2014). Additionally, the form of the genomic relationship matrix used in these analyses (Van Raden, 2008) allows partitioning the total genetic effects into additive and residual non-additive genetic effects. This partitioning has familiar interpretations for plant breeders and facilitates exploring genotype by environment interactions for additive and total effects. This type of analysis could be readily implemented by plant breeding programs that have access to molecular markers for the genotypes under consideration in order to accelerate genetic gain.

References:

Gilmour AR, Gogel BJ, Cullis BR, Welham SJ, Thompson R (2014) ASReml User Guide Release 4.0 VSN International Ltd, Hemel Hempstead, HP1 1ES, UK www.vsni.co.uk. VanRaden PM (2008) Efficient Methods to Compute Genomic Predictions Journal of Dairy Science 91:4414-4423

Joint longitudinal and survival models for investigating the association between natural disasters and disability whilst accounting for non-random drop-out due to death

Sam Brilleman

Monash University

sam.brilleman@monash.edu

Coauthors: Theodore J. Iwashyna, University of Michigan; Margarita MorenoBetancur, Murdoch Childrens Research Institute; Rory Wolfe, Monash University.

Joint modelling of longitudinal and survival (time-to-event) data has received significant attention in recent years, however much of the literature in this area has been methodological in nature. The use of joint models in applied research has been somewhat less evident. In this study we used a joint modelling approach to investigate the association between individual-level exposures to a natural disaster such as winter storm, flood, etc. and subsequent changes to physical disability, whilst accounting for non-random drop-out due to death.

Data for the study was based on a linked dataset containing 27,790 individuals who were interviewed at least once between 1st January 2000 and 1st December 2010 as part of the longitudinal "Health and Retirement Study" in the United States. Disability was assessed using activities of daily living, measured on a discrete 12-point scale. Individual-level exposure to a natural disaster was identified at the county-level based on disaster funding received from the Federal Emergency Management Agency.

Our joint model consisted of two submodels: (i) a negative-binomial mixed effects model with a log-link function for modelling the repeatedly measured disability scores and (ii) a proportional hazards model for time to death. The association between the two submodels can then be parameterised in various ways. We investigate the association between disaster exposure and disability using a time-varying exposure covariate which can be included in either one or both of the submodels. We fit the joint models using a Bayesian approach, since this provides the greatest flexibility.

This talk will discuss implementation of the joint models, contrasting the possibilities in R with the Bayesian analysis software Stan, as well as compare models based on various exposure characterisations and modelling structures.

Spatial confounding may cause bias in regression models for presence only data

Wesley Brooks

University of New South Wales

wesley.brooks@unsw.edu.au

Coauthors: David Warton

Presence-only data occurs in ecology when, for example, locations are recorded where individuals of some species were observed. A common goal in analyzing presence-only data is to estimate the association between the intensity (i.e. frequency per unit area) of presences and some explanatory covariates. The Cox process is a natural regression model to use, which includes a spatial random effect to model the intuition that nearby locations tend to be alike. However, when estimating a regression model with a spatial random effect, spatial confounding can bias the coefficient estimates. Hodges and Reich (2010) showed that spatial confounding occurs when the covariates are not orthogonal to the spatial structure of the units of observation.

A point process has no clearly defined units of observation. However, evaluating the likelihood of a point process requires estimating the value of an integral by numerical methods. The method of quadrature generalizes Riemann integration to estimate the value of an integral by calculating a sum of the integrand at a large number of points. We show that the arrangement of the quadrature points can induce spatial confounding, and propose a method for estimating the coefficients in a Cox process model where the spatial random effect is not confounded with the regression coefficients.

Reference:

Hodges, J. S. and Reich, B. J., (2010) Adding spatially-correlated errors can mess up the fixed effect you love, The American Statistician, 64(4), 325–334.

Estimation of the amplicon methylation pattern distribution from bisulphite sequencing data

Conrad Burden

Australian National University

conrad.burden@anu.edu.au

Coauthors: Peijie Lin, Sylvain Foret, Susan Wilson

Bisulfite sequencing enables the detection of cytosine methylation. The sequence of the methylation states of cytosines on any given read forms a methylation pattern that carries substantially more information than merely studying the average methylation level at individual positions.

However, the accurate quantification of these DNA methylation patterns is subject to sequencing errors and spurious signals due to incomplete bisulfite conversion of cytosines. We developed a statistical model which accounts for the distribution of DNA methylation patterns at any given locus.

The model incorporates the effects of sequencing errors and spurious reads, and enables a constrained maximum likelihood estimation of the true underlying distribution of methylation patterns. Calculation of the estimated distribution over methylation patterns is implemented in the Bioconductor package MPFE, available at http://www.bioconductor.org/packages/release/bioc/html/MPFE.html. This package has been applied to an analysis of the dynamic nature of methylation patterns in the honeybee brain in an associative learning experiment.

"Myths" in presenting statistical results (or: statistical bees I have in my bonnet.)

Ruth Butler

Plant & Food Research Lincoln, New Zealand
Ruth.Butler@plantandfood.co.nz

There are many commonly held views on the appropriate way to present summaries of data or results of formal analyses that are inconsistent with good statistical practice. Despite a large literature that discusses both the inadequacies of such methods and suggests more statistically sound alternatives, the standard of presentation of results continues to be poor. My impression is that in many biological disciplines, except perhaps in the medical sciences, it would appear that the use of poor presentation methods is expanding. This presentation will discuss these current widespread practices, what is poor about them, and suggest alternatives. It will also explore why it is that this situation has arisen, and ask what the statistical community should do to address the problem.

Misplaced confidence in confidence intervals?

John Carlin

MCRI, University of Melbourne

john.carlin@mcri.edu.au

There is growing concern about a crisis of non-reproducibility in science. For instance, Nosek et al recently reported in Science a large-scale effort to replicate 100 original studies in psychology, in which they found the average effect size reduced by a half in the replication studies. Much of the flakiness of scientific reporting can be put down to excessive dependence on statistical significance as the accepted licence for making scientific claims from noisy data. Confidence intervals have long been promoted as a preferable strategy for scientific inference, by focusing on interval estimation of the parameter of scientific interest rather than dichotomous rejection or acceptance of point hypotheses. However, strictly speaking, a confidence interval is no more than a summary of the parameter values that would not be rejected by a conventional hypothesis test: its use shifts the focus from a single point hypothesis but arguably it does not provide an intuitively understandable inferential summary. Teaching of confidence intervals is often contradictory: careful explanations are provided of the sampling theory surrounding the interval as a random quantity but then when data examples are introduced these are given loose pseudo-Bayesian interpretations ("we can be 95% confident that the true mean difference lies between A and B"). I suggest that classical frequentist interpretations don't pass the common-sense test so we should grasp the nettle and teach Bayesian credible intervals instead. In many standard problems the Bayesian interval will be close to a classical confidence interval but it is important to understand when this holds true (generally under "non-informative" prior distributions), and when it doesn't, which may be more often than we would like.

Testing significance of random terms in a linear mixed model

Vanessa Cave

AgResearch Ltd

vanessa.cave@agresearch.co.nz

Coauthors: Corinne Watts (Landcare Research)

In the simple case of unconstrained and uncorrelated variance parameters, two nested random models (with the same fixed terms) can be compared using the difference between their deviances. Under the null hypothesis, the change in deviance is asymptotically chi-squared with degrees of freedom equal to the number of variance parameters being dropped. However, the situation is more complex if variance parameters are constrained or correlated, and this is an area of ongoing research. In their 2012 Biometrics paper, Lee and Braun proposed a pair of permutation tests for testing significance of random terms in linear mixed models; one based on BLUPs and another on the restricted likelihood ratio test statistic. Both tests involve permuting the weighted marginal residuals. The weights determined by the Cholesky decomposition of the unit-by-unit variance-covariance matrix ensure the marginal residuals are exchangeable under the null hypothesis. In this presentation, we examine the permutation test methodology before applying it to a real-world example assessing whether beetle communities respond differently to two pest control regimes used in New Zealand biodiversity sanctuaries. We model data on beetle species richness and abundance using cubic smoothing splines (formulated as a linear mixed model) and use permutation tests to assess the importance of the random spline terms.

Reference:

Lee, O.E. and Braun, T.M. (2012). Permutation Tests for Random Effects in Linear Mixed Models. Biometrics, 68, 486-493.

A Comparison of the L_2 Minimum Distance Estimator and the EM-Algorithm when Fitting k-Component Univariate Normal Mixtures

Brenton Clarke

Mathematics and Statistics, Murdoch University

B.Clarke@murdoch.edu.au

Coauthors: Thomas Davidson, Australian Bureau of Statistics, Robert Hammarstrand, Mathematics and Statistics, Murdoch University

The method of maximum likelihood using the EM-algorithm for fitting finite mixtures of normal distributions is the accepted method of estimation ever since it has been shown to be superior to the method of moments. Recent books testify to this. There has however been criticism of the method of maximum likelihood for this problem, the main criticism being when the variances of component distributions are unequal the likelihood is in fact unbounded and there can be multiple local maxima. Another major criticism is that the maximum likelihood estimator is not robust. Several alternative minimum distance estimators have since been proposed as a way of dealing with the first problem. This talk deals with one of these estimators which is not only superior due to its robustness, but in fact can have an advantage in numerical studies even at the model distribution. Importantly, robust alternatives of the EM-algorithm, ostensibly fitting t distributions when in fact the data are mixtures of normal distributions, are also not competitive at the normal mixture model when compared to the chosen minimum distance estimator. It is argued for instance that natural processes should lead to mixtures whose component distributions are normal as a result of the Central Limit Theorem. On the other hand data can be contaminated because of extraneous sources as are typically assumed in robustness studies. This calls for a robust estimator.

Improving the Computational Efficiency of Model-Based Design with an Application to the Design of Multiphase Experiments

Nicole Cocks

School of Mathematics and Applied Statistics, University of Wollongong
nac724@uowmail.edu.au

Coauthors: Emi Tanaka, David Butler & Brian Cullis (presenting)

Recent literature in design has focussed on a model-based approach such as Bueno Filho and Gilmour (2003, 2007), Butler (2013), Chan (1999), Chauhan (2000) and Coombes (2002). In this talk we focus on the R package OD (Butler, 2013) for model based design of comparative experiments. OD finds optimal designs within a wide linear mixed models framework, extending this class through the inclusion of known treatment relationship structures. The method is flexible and effective in both classical and emerging novel design problems and can reproduce designs with the same or better efficiency than many designs where theoretical results are available.

There remain computational challenges, however for designs where the order of the mixed model equations is large (say greater than 1000). For example, our motivating example required roughly 8 hours to undertake 4000 random swaps while searching the design space. Martin and Eccleston (1992) present updating formula that only requires calculation of the inverse of the coefficient matrix of the mixed model equations for the initial design. We show that the implementation of this updating formula to the range of designs generated by OD results in significant increases in computational efficiency, making an exhaustive search of the design space more feasible for large problems.

Bootstrap influence on the variance components in the longitudinal data with multiple source of variation

Pauline Ding

The Statistical Consulting Unit, the Australian National University pauline.ding@anu.edu.au

Coauthors: Alan Welsh

Linear mixed models allow to model the dependence among the responses by incorporating random effects. Such dependence inherent in the longitudinal data with multiple source of variation from a complex design can be from the clustering between subjects and the repeated measurements within subjects.

Historically undercoverage is a challenge of the inference on the variance components in the framework of mixed models. We consider three bootstrap estimators defined by the Gaussian quasi-likelihood estimators. A new weighted estimating equation bootstrap, which varies weight schemes for different parameter estimators, gives improved results for the variance component estimators. In addition, we show that in the finite samples, the bootstraps which are not valid under the fixed group size asymptotics perform better in term of the coverage probabilities and the variance estimation for the variance components.

The application of the bootstraps is illustrated through data on the available nitrogen of the New Zealand native tree species, which is from a study monitoring the nutritional quality of the food resources for possums. The study has two 5 kilometer transect lines in the Tararua Mountain region, with 25 equally spaced plot sites within each transect. Leaf samples from various tree species were collected for four times.

Estimating relative species abundance from partially-observed data

Melissa Dobbie

CSIRO Digital Productivity, Brisbane, Australia

melissa.dobbie@csiro.au

Coauthors: Bill Venables, CSIRO Digital Productivity, Brisbane, Australia; Cate Paull, CSIRO Agriculture, Brisbane, Australia; Nancy Schellhorn, CSIRO Agriculture, Brisbane, Australia

For over 10 years, genetically-modified (GM) cotton has been planted in Australia to help control two pest moth species, *Helicoverpa armigera* and *Helicoverpa punctigera*. However, the proportion of a landscape that is planted to GM cotton along with a variety of other crops, some of which are also susceptible to Helicoverpa spp., varies each season. The different factors influencing Helicoverpa spp. dynamics are well known at a field level, but the effects of land use on population dynamics at spatial scales greater than field level are not well understood. Helicoverpa spp. eggs remain for approximately three days throughout summer once laid, thus provide a discrete developmental stage in a moth's life cycle to quantify. Timed longitudinal surveys of plants were undertaken in GM cotton fields over five consecutive summer seasons between 2009 and 2014 during which eggs were located and counted. A subset of eggs was collected at each sampling time, their development was monitored and the fate of each egg was recorded. We develop a two-stage parametric model to estimate relative species abundance using the partially-observed species data. In the first stage we form a generalised linear model for the observed species proportions that takes into account small-scale discreteness and smooths the observation noise. This ancillary model was used to apportion the total eggs counted into imputed egg counts for the individual species. A linear mixed model framework was adopted, given the hierarchical nature of the study, to marry the imputed relative species abundances to important attributes of the landscape.

Categorising Ecological Community Count Data

Daniel Fernandez

Victoria University of Wellington

daniel.fernandez@msor.vuw.ac.nz

Coauthors: Shirley Pledger

Count data sets may involve overdispersion from a set of species and underdispersion from another set which would require fitting different models (e.g. a negative binomial model for the overdispersed set and a binomial model for the underdispersed one). Additionally, many count data sets have very high counts and very low counts. Categorising these counts into ordinal categories makes the actual counts less influential in the model fitting, giving broad categories which enable us to detect major overall patterns. In this talk, a strategy of categorising count data into ordinal data was carried out and measures to compare different cluster structures were implemented. The application of this categorising strategy and a comparison of clustering results between count and categorised ordinal data in ecological community data sets are shown. A major advantage of using our ordinal approach is that it allows for the inclusion of all different levels of dispersion in the data in one methodology, without treating the data differently. This has the important implication of supporting simpler, faster data collection using ordinal scales only.

Limitations of Hierarchical and Mixture Model Comparisons

M. A. C. S. S. Fernando

Department of Statistics, University of Auckland, Private Bag 92019, Auckland 1025, New Zealand sampathf73@gmail.com Coauthors: James M. Curran, Jo-Anne Bright, John S. Buckleton, Renate Meyer

In general, a statistical model is a probabilistic system that involves a probability distribution or a finite/infinite mixture of probability distributions. These models are widely used in explanation, prediction, or making inferences on some real-world phenomena. It is possible to approximate a given phenomenon with more than one model. Therefore, statistical model comparison is essential in the model building process to select the best of two or more competing candidate models. Usually, it is more convenient to build different models based on one particular distribution (e.g., regression models with Normal distribution). Comparison of such models can be easily carried out using an appropriate model comparison criterion. However, in situations where the models are originated from different distributions, the comparisons are quite interesting. This is further complicated when we use hierarchical models, mixture models, and hierarchical mixture models. Bayesian practitioners often use information criteria such as AIC, BIC, DIC, and WAIC for the comparison of models. Although, these criteria are unable to reflect the goodness-of-fit in absolute sense, the differences (in the information theoretic criterion of choice between competing models) can measure the relative performance of the models of interest. However, the use of some of these measures is only valid under certain circumstances. In this talk, I will illustrate the problems and limitations associated with the comparisons of Bayesian hierarchical, non-hierarchical, and mixture models using an example problem from forensic science.

Development of a bivariate linear mixed model to describe yield response due to disease

Clayton Forknall

Queensland Department of Agriculture and Fisheries

clayton.forknall@daf.qld.gov.au

Coauthors: Alison Kelly and Steven Simpfendorfer

Production losses due to foliar and root diseases are a major financial constraint to the production of wheat and barley in Australia with annual losses in wheat alone estimated at $913 m (Murray and Brennan, 2009). A national project to develop response curves that quantify the potential production losses incurred as a result of disease on current Australian commercial wheat and barley varieties is currently funded by the Grains Research and Development Corporation. A case study is presented for crown rot in wheat, where grain yield and crown rot index form the response and explanatory variables respectively. The crown rot index is calculated using a weighted average across disease scores, based on the level of stem browning and is the current industry accepted measure of crown rot infection.

Parameters estimated from a standard linear regression are biased when the explanatory variable is measured with error (Sokal and Rohlf, 2012). Even though alternative methods, such as Errors in Variables regression exist for modelling such data, we require a flexible model that captures the variance and covariance between the variables.

We developed a bivariate linear mixed model based on a random regression approach for the construction of response curves. This model allows for the estimation of heterogeneous variances for the variables, along with the correlation between variables, independently for each of the varieties considered. From this model, an estimate of the yield response to crown rot (slope) and the yield potential in the absence of crown rot (intercept) is determined for each variety. The bivariate linear mixed model is fitted in ASReml-R (Butler et al., 2009). Using this model, the yield response of wheat and barley varieties to various foliar and root diseases can be estimated, assisting growers in quantifying the risks due to disease associated with their selection of varieties.

References:

Butler, D., Cullis, B. R., Gilmour, A. R. & Gogel, B. J. (2009). ASReml-R Reference Manual. Technical Report 3, Queensland Department of Agriculture, Fisheries and Forestry. Murray, G. M. & Brennan, J. P. (2009). Estimating disease losses to the Australian wheat industry. Australasian Plant Pathology 38:558-570.

Sokal, R. R. & Rohlf, F. J. (2012). Biometry. W. H. Freeman and Company, 4th edition.

Estimating the correlation between competing risks

Valérie Garès

NHMRC Clinical Trials Centre — University of Sydney — Australia

Coauthors: Malcolm Hudson and Val Gebski

In problems with a time to event outcome, the issue of competing risks may arise when subjects experience events which prevent the outcome of interest in the study being observed. Methods analogous to the logrank test and proportional hazards models are commonly used to account for the competing risks. Identifying the correlation between the different risks can assist in the understanding and interpretation of the statistical results of subsequent analysis. Our aim is to assess the effects of correlation between two competing risks on the estimator of a hazard ratio (HR) for a treatment versus a control.

The logarithm of the times to event may satisfy assumptions of normality. We propose a bivariate normal censored model to estimate the correlation between two competing events using an EM algorithm. Including a covariate in this model indicating treatment arm provides an estimator of HR for a treatment versus a control. This allows us to impute the survival time for the censored events where censoring was due to the occurrence of a competing event, loss to follow up or termination of the study. Two strategies were used to impute bivariate survivals. Firstly, using a pre-specified correlation, investigate a sensitivity of a HR estimator to the correlation levels. Alternatively, the correlation may itself be included in an EM estimator procedure. Standard methods (Cox and Fine and Gray) are applied to assess the HR for a treatment versus a control and are compared with the proposed method.

These methods were applied to a trial of head and neck cancer. This approach helps to guide both interpretation and usefulness of standard methods for these problems together with the clinical importance of the outcomes. Sensitivity analysis using our approach is also recommended.

Design and analysis of simulation experiments

Graham Hepworth

Statistical Consulting Centre, The University of Melbourne

hepworth@unimelb.edu.au

Some physical processes are very difficult to study using real-world experiments, and there may be little data available from observational studies. It is sometimes possible to model such processes by computer simulation, with both deterministic and stochastic components.

Simulation experiments have much in common with real-world experiments, but some aspects which differ. For example, blocking is unimportant with simulation experiments. Drawing on my experience working with researchers from the Department of Defence, I will describe the similarities and differences between simulation experiments and real-world experiments, and the learning curve that it has involved for all of us.

Modelling the distribution of sub-Antarctic and Antarctic demersal fish communities: An application of new community modelling method

Nicole Hill

Institute for Marine & Antarctic Studies, University of Tasmania

Nicole.Hill@utas.edu.au

Coauthors: Scott D. Foster: Division of Computational Informatics, CSIRO Marine Laboratory, Tasmania, Australia. Guy Duhammel, Philippe Koubbi:Unite Biologie des organismes et ecosystemes aquatiques (BOREA, UMR 7208), Sorbonne Universites, Museum national d'Histoire naturelle, Universite Pierre et Marie Curie, Universite de Caen Basse-Normandie, CNRS, IRD; Paris, France. Dirk Welsford: Southern Ocean Ecosystems Theme, Australian Antarctic Division, Tasmania, Australia.

Demersal fish form an important component of Antarctic and sub-Antarctic ecosystems. We aim to quantify and predict the distribution of fish assemblages along a latitudinal section from the Kerguelen Plateau to Prydz Bay using co-located biological and environmental data. We achieve this by applying a recently developed method, called Regions of Common Profile (RCP), which utilises a mixture-of-experts model to quantify distinct environmental regions within which the vector consisting of all species expected values is relatively constant. Directly modelling species simultaneously (rather than dissimilarities or single species at a time) in relation to the environment offers advantages in the areas of model diagnostics, the interpretability of model outputs, and providing estimates of uncertainty. The RCP method has also been extended to account for sampling artefacts (such as gear type) that may affect the detectability of species, an important aspect for the current study where we have amalgamated several datasets including scientific fisheries data on the Kerguelen Plateau and scientific survey data. By linking demersal fish records with environmental variables in this way, we also aim to understand the environmental factors influencing demersal fish community patterns more broadly in this dynamic and important region.

Vector regression without specifying marginal or association structures

Alan Huang

School of Mathematics and Physics, University of Queensland alan.huang@uq.edu.au

We introduce a flexible yet parsimonious framework for vector regression based on nonparametric multivariate exponential families. The key feature is that underlying exponential family can be left completely unspecified in the model and can be estimated nonparametrically from data along with the usual regression coefficients using a maximum empirical likelihood approach. Its usefulness in practice is demonstrated via various simulations and data analysis examples.

Factors affecting treatment recurrence and death: a case study with longitudinal hospital retreatment records

Malcolm Hudson

Macquarie University and University of Sydney

Malcolm.Hudson@mq.edu.au

Coauthors: M.B. Barton, G. Delaney, Z. Hao, S. Allen

Classical models in competing risks are usually formulated with alternative events considered mutually exclusive and non-recurring. There are a variety of approaches generalizing classical survival analysis models with competing nonrecurrent events to multiply recurring events. For example, for recurring events with competing risks, marginal and multistate models apply survival analysis methods. We present and compare competing risks survival models of recurrent events with observations subject to administrative censoring. Focus is placed on estimation and inference concerning cumulative incidence and mean functions, an area little discussed in the literature. We demonstrate the independence of estimators of cumulative incidence and cumulative mean number of recurrent retreatments from mortality data, even when retreatment is highly dependent on mortality. Models and methods are applied in a large clinical study of radiotherapy cancer treatment in South-West Sydney.

Joint Effects Selection in Mixed Models using CREPE

Francis Hui

The Australian National University

fhui28@gmail.com

Coauthors: Samuel Muller and Alan Welsh

Generalized linear mixed models (GLMMs) are widely used to analyze longitudinal data in ecology and medicine, among other fields. As with all regression models though, a key part of the inference process is variable selection. Furthermore, there is usually an implicit belief that when fitting GLMMs with random intercepts and slope, covariates should be included in the model as either a fixed effect only or a composite effect i.e. both fixed and random effect.

I propose a penalty called CREPE that builds the aforementioned belief into the selection process. CREPE accounts for the hierarchical nature of the covariates, so that the final submodel chosen includes only fixed and composite effects. Asymptotic properties of the CREPE estimator are discussed, including selection consistency and the oracle property. Simulations demonstrate the strong performance of CREPE compared to some other currently available penalized methods for GLMMs.

The influence of surface productivity on seafloor food-availability and biodiversity distributions on the George V shelf, East Antarctica

Jan Jansen

Institute for Marine and Antarctic Studies, University of Tasmania

jan.jansen@utas.edu.au

Coauthors: Michael D. Sumner, Nicole A. Hill, Piers K. Dunstan, Craig Johnson, Alexandra L. Post, Leanne Armand, and Ben K. Galton-Fenzi

Benthic faunal communities below the photic zone are usually dependent on influx of organic matter from distant sources. Food can be transported via water-currents or via sinking from the upper water-layers. Along the Antarctic continental shelf, surface productivity is considered patchily distributed in time and space due to the high seasonality and seasonal distribution of sea ice as well as the varying availability of the limiting nutrient, iron. In this project, we estimated the relative distribution of organic matter influx to the seafloor by proxy satellite-derived measurements of surface-chlorophyll-a and validated the models against diatom abundance and distributions from sediment cores. Further, we quantify the effect of surface productivity on the diversity of benthic communities over the George V shelf (East Antarctica). Applying and comparing the results of a statistical and a mechanistic modelling framework, we show that prevailing shelf currents have a minor influence on the distribution of food-abundance in their transit to the seafloor. In contrast, tidal-currents were found to inhibit the settling of particles on the banks while the basins experience increased sedimentation rates. The significance of these findings in relation to benthic biodiversity will be discussed.

Assessing the impact of unmeasured confounding: confounding functions for causal inference

Jessica Kasza

Department of Epidemiology and Preventive Medicine, Monash University

jessica.kasza@monash.edu

Causal inference is particularly useful in situations in which randomised experiments are unethical or difficult to conduct. A critical assumption of causal inference is that of no unmeasured confounding: for estimated quantities to have valid causal interpretations, all relevant confounders of the exposure and outcome of interest must be measured and correctly included in models. In applications the no unmeasured confounding assumption is often invalid and the sensitivity of conclusions to the violation of this assumption must therefore be assessed, although in practice this is rarely done. Several approaches have been suggested, requiring either the availability of variables with specific properties or assumptions to be made about particular unmeasured confounders. In contrast to these approaches, the confounding function approach allows the impact of unmeasured confounding in its entirety on causal estimates to be assessed, without the need for additional variables or assumptions about the nature of unmeasured confounders. The confounding function describing the degree of unmeasured confounding is defined and used to correct estimates for the impact of unmeasured confounding. In this talk, after briefly reviewing approaches to assessing the impact of unmeasured confounding, I will discuss and demonstrate the confounding function approach.

Selection of genotypes for resistance and tolerance to pathogens: a multivariate statistical analysis of yield and disease response

Alison Kelly

Department of Agriculture and Fisheries, Queensland

alison.kelly@daf.qld.gov.au

Coauthors: Bethany Macdonald and Susan Fletcher

Phenotyping for the effect of disease on genotypes from a plant breeding program requires a measurement of both the growth of the pathogen in the plant and the subsequent effect of the pathogen on grain production in the plant. The ability of the plant to suppress disease expression is known as resistance, and this effect is measured as the severity of the disease response in the plant. The ability of the plant to produce grain in the presence of disease is defined as tolerance, and this is measured through grain yield at harvest. Experiments to test the resistance and tolerance of genotypes typically consist of replicated field trials with a split-plot arrangement of a disease treatment including an untreated control (uninoculated) and an imposed disease level (inoculated). Different genotypes are grown under these two conditions with the aim of selecting genotypes possessing combined traits of resistance and tolerance to disease.

A bivariate linear mixed model (LMM) is presented for the combined analysis of the traits of yield and disease response across inoculated and uninoculated plots. The LMM provides a framework for modelling the variances and covariances associated with the factorial combination of trait by disease at the genetic level, as well as the covariance between traits at the plot level. The model is then extended for field trials conducted at multiple environments. Variance parameters from the model are estimated using Residual Maximum Likelihood (Patterson and Thompson, 1971), and the LMM is fitted in ASReml-R (Butler et al., 2009).

The random regression nature of the model allows us to use the estimated variance parameters and random effects to derive and select for three traits of interest. Yield responsiveness under inoculated and uninoculated conditions can be formulated following an established method for responsiveness (Stevens et al., 2000; McDonald et al.. 2015). Severity of disease symptoms in inoculated plots can be used to select genotypes with superior resistance. Finally, yield advantage in the presence of disease taken from the random regression of yield against

measured disease in the inoculated plots can be used to select for genotype tolerance.

References:

Butler, D., Cullis, B. R., Gilmour, A. R. and Gogel, B. J. (2009). ASReml-R Reference Manual. Technical Report 3, Queensland Department of Agriculture, Fisheries and Forestry.

McDonald G., Bovill, W., Taylor, J. and Wheeler, R. (2015). Responses to phosphorus among wheat genotypes. Crop and Pasture Science 66: 430-444.

Patterson, H.D., and R. Thompson. 1971. Recovery of interblock information when block sizes are unequal. Biometrika 31:100109.

Stevens, M.M., Fox, K.M., Warren, G.N. and Cullis, B.R. (2000) An image analysis technique for assessing resistance in rice cultivars to root-feeding chironomid midge larvae (diptera: Chironomidae). Field Crops Research 66: 25-36.

Humpback whales in the Great Barrier Reef: model-based inferences on distribution and abundance

Natalie Kelly

CSIRO

natalie.kelly@csiro.au

Coauthors: Joshua Smith and David Peel

The population of humpback whales (*Megaptera novaeangliae*) that migrates along the east coast of Australia have their winter breeding and calving areas in and around the Great Barrier Reef World Heritage Area (GBRWHA). However, until recently, it was not really known where within the GBRWHA these animals congregated during their breeding season. Furthermore, with an estimated increase in population of around 10.6% per year, combined with expanding shipping activity within the GBRWHA, there are a range of potential current and future risks (e.g., ship strike, etc) which remain largely unquantified. To help study these potential risks, double-platform aerial surveys for humpback whales were undertaken during the breeding season in 2012 and 2014, covering various locations along the north-south extent of the GBR. These surveys supported model-based inferences (via distance sampling and Generalized Additive Models) of abundance and distribution of humpback whales; inferences which included spatial and environmental covariates, and which allowed some extrapolation to unsurveyed regions. Regions within the GBRWHA where humpback whales congregate were identified, information vital for conservation and management. Furthermore, in combination with data such as location and dynamics of commercial shipping lanes, these model-based inferences on east coast humpback whale distribution now allow for quantification of relative ship strike risk-particularly that which is spatially and temporally delineated-to animals during their breeding season.

An experimental design for testing Bycatch Reduction and Turtle Excluder Devices in the PNG Prawn Trawl Fishery

Emma Lawrence

CSIRO

emma.lawrence@csiro.au

Coauthors: Bill Venables

Experimental design publications typically focus on highly balanced designs where combinatorial theory is often used to arrive at designs with the highest efficiency. It is sometimes difficult to achieve symmetry in design due to logistical constraints, and the standard methods for constructing designs may need to be radically adapted.

In this talk I will discuss an example involving 4 vessels towing quad gear (4 separate, but linked nets) over 18 days. The experiment was designed to assess the effectiveness of 27 combinations of Turtle Excluder Devices (TEDs) and Bycatch Reduction Devices, (BRDs), with a control net without any devices attached. Each vessel carries its own control net, which it must use on all trawls. There is only one copy made of each of the 27 treatment nets, however, requiring the nets to be exchanged at sea. Such exchanges can only take place at specific times during the 18 day trial. These constraints require a highly tailored design.

I will demonstrate how we used simulated annealing to generate a highly efficient design in several stages, meeting all of the logistical constraints. I will then summarise our initial findings and discuss some of the challenges that were faced in the field.

A multivariate approach for multiple 'omics data integration and biomarker discovery

Kim-Anh Le Cao

The University of Queensland Diamantina Institute

k.lecao@uq.edu.au

Coauthors: Amrit Singh, Benoit Gautier, Kevin Chang, Scott Tebbutt, KimAnh Le Cao

The advent of high throughput technologies has led to a wealth of publicly available biological data coming from different sources, the so-called omics data (transcriptomics for the study of transcripts, proteomics for proteins, metabolomics for metabolites, etc). Combining such large-scale biological data sets can lead to the discovery of important biological insights, provided that relevant information can be extracted in a holistic manner.

Starting from the multivariate integrative model from Tenenhaus et al. (2014), we have further improved 'sparse Generalized Canonical Correlation Discriminant Analysis', a projection-based multivariate methodology able to combine multiple data sets originating from different technologicalomics platforms while selecting biological features via the means of Lasso penalisations in the statistical model (Tibshirani, 1997). Our improved approach now includes a classification framework so as to identify a reliable, but also highly correlated molecular signature with a high diagnostic potential.

We applied our multivariate integrative approach to the breast cancer multiomics study from The Cancer Genome Atlas consortium to characterize four breast cancer subtypes from 377 samples with a selection of biological features including mRNA, miRNA, proteomics and methylation data. Our integrative approach gives similar performance to other common approaches to combine different types of data (ensemble approach, concatenation approach), with a classification error rate of 14% on an external cohort of 573 patients. Most importantly, our integrative approach is able to identify a highly correlated (or "connected") signature composed of biological features from different molecular levels which are relevant for the study.

During this presentation we will discuss the benefits of using a multivariate integrative approach for such complex problems arising from high-throughput molecular biology.

References:

A. Tenenhaus, C. Phillipe, V. Guillemot, K-A. Le Cao, J. Grill, V. Frouin (2014), Variable selection for generalized canonical correlation analysis Biostatistics, 15(3): 569-83.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. J. Royal. Statist. Soc B., Vol. 58, No. 1, pages 267-288).

Visualizing Population Genetics

Louise McMillan

Statistics Department, University of Auckland

louise.mcmillan@auckland.ac.nz

Coauthors: Rachel Fewster

Genetic assignment is the process of assigning individuals to source populations based on their genetic profiles. We propose a method for visualizing genetic assignment data by characterizing the genetic distribution of each candidate source population. This method improves upon the assignment method of Rannala and Mountain (1997) by calculating appropriate graph positions for individuals for which some genetic data are missing. Individuals with missing data are plotted at their population quantiles which we obtain using a saddlepoint approximation (Daniels 1954, Lugannani and Rice 1980). The saddlepoint method also provides a way to visualize results from leave-one-out procedures.

We demonstrate our method using simulated data and microsatellite data from ship rats (*Rattus rattus*) captured on Great Barrier Island, New Zealand, and smaller surrounding islands. The visualization method makes it much easier to detect features of population structure and predict the accuracy of assignment results. It improves upon assignment software such as GeneClass, which has no visualization, and STRUCTURE, for which interpretation can be difficult when multiple population memberships are displayed for single individuals.

Selecting environment covariates to explain GxE: A comparison of cyclic forward regression and subset regression approaches

Ky L Mathews

Ky L Mathews Statistical Consulting

Ky.Mathews@gmail.com

Coauthors: Arthur Gilmour, Bangyou Zheng, Scott Chapman

Multi-environment trial analyses generally address the complexity of genotype by environment interaction in order to provide relevant predictions of genotype performance. Environmental covariates (weather and/or soil traits) can assist in interpreting this complexity and consequently provide more informative varietal predictions. Previous work has not assessed environmental covariates based on their contribution to the model in a mixed model framework where the genotype by environment variance matrix was appropriately modelled. We compare two methods for covariate selection in the mixed model framework where a factor analytic, k = 1, model is fitted to the underlying genotype by environment variance matrix. In the first method, a cyclic forward regression approach introduces environmental covariates to the model depending on how much they contribute to the model. In the second method, all possible combinations of environmental covariates are tested and the model that explains the most variation is selected by assessing changes in the AIC and BIC. The Australian National Variety Testing wheat Main Season dataset from 2005-13 is large and unbalanced, containing 1019 trials assessing 163 varieties across 187 locations. Daily weather data from nearby Bureau of Meteorology weather stations and relevant soil data are collected during the crop cycle to provide the environmental covariates. The advantages and disadvantages and similarity/dissimilarity of results from these two methods will be discussed.

WIC and importance sampling approximations to cross-validation for models with observation-level latent variables

Russell Millar

University of Auckland

r.millar@auckland.ac.nz

The Watanabe information criterion (WIC) has formal derivations from the theory of algebraic geometry, and as an approximation to leave-one-out crossvalidation (LOO-CV). Importance sampling (IS) also provides an alternative approximation to LOO-CV. In the context of models with observation-level latent variables, the existing literature provides examples of application of WIC and IS using observation-level likelihood. It is shown here that this approach can be fraught with peril and great care is needed. In particular, the theoretical justifications of WIC are invalid at the observation level, and its behaviour is application specific. In particular, it is shown that observation-level WIC is erroneous when applied to over-dispersed count data. Observation-level IS does continue to be theoretically valid, but in practice can be so numerically unstable as to give misleading results. It is recommended that WIC and IS be applied using likelihood marginalized over the latent variables.

Multiple imputation and sensitivity analysis for incomplete longitudinal data departing from the missing at random assumption

Margarita Moreno-Betancur

Murdoch Childrens Research Institute/Monash University

margarita.moreno@mcri.edu.au

Coauthors: Michel Chavance

Statistical analyses of longitudinal data with drop-outs based on maximum likelihood, and using all the available data, provide unbiased estimates under the assumption that outcomes after drop-out are missing at random. Unfortunately, this assumption can never be tested from the data and sensitivity analyses should be routinely performed to assess the robustness of inferences to departures from it. However, each specific scientific context requires different considerations when setting up such an analysis, no standard method exists and this is still an active area of research. We propose a flexible procedure to perform sensitivity analyses when dealing with continuous outcomes, which are described by a linear mixed model in an initial likelihood analysis. The methodology is based on the pattern-mixture model factorisation of the full data likelihood and its implementation relies on multiple imputation. The approach, which we validated in a simulation study, was prompted by a randomised clinical trial for sleep-maintenance insomnia treatment. This case study illustrated the practical value of our approach and underlined the need for sensitivity analyses when analysing data with drop-outs: some of the conclusions from the initial analysis were shown to be reliable, while others were found to be fragile and strongly dependent on modelling assumptions.

New Models of Molecular Evolution

Teresa Neeman

Statistical Consulting Unit, Australian National University

teresa.neeman@anu.edu.au

Coauthors: Ben Kaehler, Gavin Huttley, John Curtin School of Medical Research, ANU

Molecular evolution, modelled using a continuous time Markov process defined over a phylogenetic tree, takes multiple alignments of sequenced DNA to estimate substitution rate matrices. Nearly all molecular evolutionary models assume time-reversibility and stationarity. Time-reversible means that joint probabilities P(X(t), X(s)) at times t and s, are symmetric. Stationarity is the assumption that the current nucleotide base frequency is in equilibrium over the entire tree. These assumptions have been used to make estimation numerically tractable. Hardware and algorithms now exist to fit a full 12 parameter model to each branch of a phylogenetic tree, opening the door for increasing the complexity and improving the fit of Markov models to evolutionary processes. But increasing model complexity raises the spectre of overfitting. In this talk, I'll discuss our exploration of fitting evolutionary models to molecular data, removing constraints on the one hand in order to discover interesting lower dimensional alternatives.

Diagnostic methods for checking multiple imputation models

Cattram Nguyen

Murdoch Childrens Research Institute

cattram.nguyen@mcri.edu.au

Coauthors: Katherine Lee and John Carlin

Multiple imputation (MI) is a popular method for handling missing data problems. The validity of imputation-based inferences depends on the model that is used to generate the imputed values. Despite the popularity of MI, the checking of imputation models is not widespread. This may be due to the scarcity of guidelines and computational tools for performing imputation diagnostics.

In this presentation, we will provide an overview of diagnostic methods for checking MI models. This includes graphical diagnostics, the Kolmogorov Smirnov test and posterior predictive checking. We will present results from simulation evaluations of these techniques, as well as illustrate the diagnostics using data from the Longitudinal Study of Australian Children. As MI becomes established as a standard missing data method into the future, it will become increasingly important that researchers develop and employ these model-checking techniques to ensure their imputation models are appropriate.

Multiple imputation as a stochastic EM approximation to maximum likelihood

Firouzeh Noghrehchi

UNSW

Coauthors: David Warton and Jakub Stoklosa

Multiple imputation is a popular approach to missing data analysis that is often considered as distinct from maximum likelihood estimation. However, we show that a type of multiple imputation can be understood as a stochastic EM approximation to maximum likelihood, which offers some new insights and opens the doors to the application of a range of likelihood-based tools in a multiple imputation context. For example: we can better understand the consequences of misspecification of the imputation model; how to select an imputation model for a given dataset; we can readily develop alternatives to Wald-type inferences for multiple imputation; and information criteria to use for model selection, given a set of competing models fitted by multiple imputation.

Relative Effect Sizes for Measures of Risk

Jake Olivier

School of Mathematics and Statistics, University of New South Wales

j.olivier@unsw.edu.au

Coauthors: Warren L. May and Melanie L. Bell

Effect sizes are an important component of experimental design, data analysis and interpretation of statistical results. In some situations, an effect size of clinical or practical importance may be unknown to the researcher. In other situations, the researcher may be interested in comparing observed effect sizes to known standards to quantify clinical importance. In these cases, the notion of relative effect sizes (small, medium, large) can be useful as benchmarks. Although there is generally an extensive literature on relative effect sizes for continuous data, little of this research has focused on relative effect sizes for measures of risk that are common in epidemiological or biomedical studies. The aim of this paper, therefore, is to extend existing relative effect sizes to the relative risk, odds ratio, hazard ratio, rate ratio, and Mantel-Haenszel odds ratio for related samples. In most scenarios with equal group allocation, effect sizes of 1.22, 1.86 and 3.00 can be taken as small, medium and large respectively. The odds ratio for a non-rare event is a notable exception and modified relative effect sizes are 1.32, 2.38 and 4.70 in that situation.

A novel methodology for inhomogeneity identification in climate time series

Liliana Orellana

Deakin University

l.orellana@deakin.edu.au

Coauthors: Andres Farall and Jean-phillipe Boulanger

Long time series of climate variables are often affected by sudden changes caused by climatic and/or non-climatic factors (inhomogeneities) which need to be identified as they may hide the true climatic signals and patterns and potentially bias the conclusions of climate studies. The identification of inhomogeneities is technically equivalent to the subdivision of a time series into segments, each one of them characterized by a different data generating process. In this work, we propose a novel robust methodology aimed at identifying breakpoints caused by any type of modifications in the data generating process.

Let *i* identify the monitoring station whose daily data is to be controlled (target station), and IS(i) define a set of monitoring stations, which includes station i and all stations spatially contiguous and related to it, whose daily observations produce a multivariate time series. The methodological proposal involves an exhaustive scan through the multivariate time series in search of an unknown number of regime changes which can occur at any time. The core of the method is based on the notion of depth of multivariate observations, a measure of centrality of the observations with respect to the data set, which transform the data to a center-outward ranking. We consider day d as a breakpoint when the distribution of depths before and after day d change significantly. Transforming the multivariate observations into depths turns a multivariate distribution comparison problem into a univariate one. Therefore, the change in depth distributions before and after d can be evaluated using the standardized Kolmogorov-Smirnov (KS) statistics.

Due to the unknown quantity and location of the breakpoints in the data set, the whole series needs to be screened searching for potential changes in distribution. A recursive partition approach is then applied. The day with the largest KS value (d^1) is chosen as the first breakpoint, and the time series is then split in two segments, before and after d^1. The procedure is recursively repeated until a stopping criterion is reached. The outcome of the procedure renders a hierarchically nested sequence of potential breakpoints. The final step involves a pruning phase which

removes non-significant breakpoints. The distribution of the KS statistic under the null hypothesis (no breakpoints) is obtained using block-bootstrap to account for the lack of independence of the time series data. Evaluation of the methodology under simulation shows a very good performance which increases with the spatial density and number of stations involved in the influence set, and with the magnitude of the inhomogeneites.

Main advantages of this proposal compared with existing methodologies are: 1) does not require metadata or a reference station which has already been controlled; 2) detects multiple breakpoints; 3) can detect any kind of changes in the data (shift, variability, skewness, extremes, etc.); and 4) can be used with high-frequency data (ten minute observations) as for instance those recorded by automatic weather stations. Finally, the methodology was developed in the generic context of time series segmentation and can then be used in several application fields, such as, economics, biology, quality control, engineering or informatics.

Assessing a species distribution model's performance for sparse and patchy presence data

Samantha Peel

University of Tasmania

samantha.peel@utas.edu.au

Coauthors: Nicole Hill, Scott Foster, Simon Wotherspoon, John McKinlay, Piers Dunstan, and Ben Raymond

Natural resource management requires information about the spatial distribution of species. Species distribution models (SDMs) provide this information and do so using presence-absence data or presence-only data. Generally, presence-absence data is preferred as presence-only data can be biased (the number of presences is related to the amount of sampling effort). Unfortunately, in the Southern Ocean, at a circumpolar scale, the existing data is generally presence-only with presence-absence available only at a more localised scale. This raises questions about the reliability of SDMs for many Southern Ocean taxa. Recently, Fithian, Elith, Hastie, and Keith (2015) have proposed an SDM that a) jointly analyses presence-only and presence-absence data, and b) takes into account sampling bias within the data. Their case study uses a large amount of dense presence-absence data with only a small amount of presence-only data. However, the Southern Ocean data has a high proportion of sparsely distributed presence-only data, and the presence-absence data is patchy. We present a simulation study that assesses the robustness of the Fithian et al. (2015) method, with data that more closely reflects our Southern Ocean data. We conclude with some recommendations about the effectiveness of the approach for the Southern Ocean.

Reference:

Fithian, W., Elith, J., Hastie, T. & Keith, D. A. (2015). Bias correction in species distribution models: pooling survey and collection data for multiple species. Methods in Ecology and Evolution, 6(4), 424-438.

Eliciting and encoding expert knowledge on variable selection into classical or Bayesian species distribution models

Ramethaa Pirathiban

Science and Engineering Faculty (SEF), Queensland University of Technology (QUT)

r.jeyapalan@qut.edu.au

Coauthors: Kristen J Williams; Tony Pettitt; Samantha Low-Choy

The quality of species distribution models (SDMs) relies to a large degree on the quality of the input data, from bioclimatic indices to environmental and habitat descriptors (Austin, 2002). Recent reviews of SDM techniques, have sought to optimize predictive performance e.g. Elith et al., 2006. In general SDMs employ one of three approaches to variable selection. The simplest approach relies on the expert to select the variables, as in environmental niche models Nix, 1986 or a generalized linear model without variable selection (Miller and Franklin, 2002). A second approach explicitly incorporates variable selection into model fitting, which allows examination of particular combinations of variables. Examples include generalized linear or additive models with variable selection (Hastie et al. 2002); or classification trees with complexity or model based pruning (Breiman et al., 1984, Zeileis, 2008). A third approach uses model averaging, to summarize the overall contribution of a variable, without considering particular combinations. Examples include neural networks, boosted or bagged regression trees and Maximum Entropy as compared in Elith et al. 2006. Typically, users of SDMs will either consider a small number of variable sets, via the first approach, or else supply all of the candidate variables (often numbering more than a hundred) to the second or third approaches. Bayesian SDMs exist, with several methods for eliciting and encoding priors on model parameters (see review in Low Choy et al. 2010). However few methods have been published for informative variable selection; one example is Bayesian trees (O'Leary 2008). Here we report an elicitation protocol that helps makes explicit a priori expert judgements on the quality of candidate variables. This protocol can be flexibly applied to any of the three approaches to variable selection, described above, Bayesian or otherwise. We demonstrate how this information can be obtained then used to guide variable selection in classical or machine learning SDMs, or to define priors within Bayesian SDMs.

References

M. P. Austin. Spatial prediction of species distribution: An interface between ecological theory and statistical modelling. Ecological Modelling, 157(2):101–118, 2002.

L. Breiman, J. Friedman, C. J Stone, and R.A. Olshen. Classification and regression trees. CRC press, 1984.

J. Elith, C. H. Graham, R. P. Anderson, M. Dudik, S. Ferrier, A. Guisan, R. J. Hijmans, F. Huettmann, J. R. Leathwick, A. Lehmann, J. Li, L. G. Lohmann, B. A. Loiselle, G. Manion, C. Moritz, M. Nakamura, Y. Nakazawa, J. M. Overton, A. T. Peterson, S. J. Phillips, K. Richardson, R. ScachettiPereira, R. E. Schapire, J. Soberon, S. Williams, M. S. Wisz, and N. E. Zimmermann. Novel methods improve prediction of species' distributions from occurrence data. Ecography, 29:129–151, 2006.

T. Hastie, R. Tibshirani, and J. Friedman. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer-Verlag, second edition, fifth reprinting edition, 2002.

J. Miller and J. Franklin Modeling the distribution of four vegetation alliances using generalised linear models and classification trees with spatial dependence. Ecological Modelling, 157:227–247, 2002.

S. Low Choy, J. Murray, A. James, and K. L. Mengersen. Indirect elicitation from ecological experts: from methods and software to habitat modelling and rock-wallabies. In Anthony O'Hagan and Mike West, editors, The Oxford Handbook Of Applied Bayesian Analysis, Handbook of Applied Bayesian Analysis, pages 511–544. Oxford University Press, Oxford, 2010.

H. A. Nix. A biogeographic analysis of Australian elapid snakes. Atlas of Elapid Snakes of Australia.(Ed.) R. Longmore, pages 4–15, 1986.

R. A. O'Leary. Informed statistical modelling of habitat suitability for rare and threatened species. PhD thesis, Queensland University of Technology, 2008. A. Zeileis, T. Hothorn, and K. Hornik. Model-based recursive partitioning.

Journal of Computational and Graphical Statistics, 17(2):492–514, 2008.

The fourth corner problem: niche overlap using mixtures

Shirley Pledger

Victoria University of Wellington, New Zealand

shirley.pledger@vuw.ac.nz

The fourth corner problem has been solved (Brown et al. 2014, Warton et al. 2015). Is there anything left to say?

In the fourth corner problem, there are three matrices available: (i) an n by p matrix of the abundance of p species over n sites, (ii) an n by m matrix of environmental predictors ("site factors", e.g. soil type, altitude), and (iii) an s by p matrix of species traits (e.g. foraging habits and nest-building materials of birds). The aim of the analysis is to find and describe linkages of the s species traits with the m environmental characteristics; this is the matrix in the fourth corner.

If clustering using finite mixtures is included in the analysis, there are some bonuses in (i) detecting and plotting overall patterns, (ii) allowing rare species to borrow strength from more common species by their inclusion in a "species archetype", and (iii) obtaining a unified measure of niche overlap for the species groups.

Approximate likelihood ratio test for multivariate abundance data

Gordana Popovic

University of New South Wales, Sydney

g.popovic@unsw.edu.au

Coauthors: David I. Warton and Francis KC. Hui

When modeling a community of species, there are often a large number of potential species interactions relative to the number of locations where abundances have been observed, making it challenging to build a plausible yet parsimonious model of abundance and co-occurrence. Multivariate models allow us to take species interactions into account when making inferences about treatments, associations between a community and the environment, or potential environmental impacts. Current hypothesis testing in this context utilizes methods, like generalized estimating equations, which do not specify a fully parametric model for the data. This leads to some limitations, for example, a reliance on Wald and Score statistics that have some undesirable properties when data have many zeros and a strong mean-variance relationship, as is often the case for multivariate abundance data. In this talk, we propose building a fully parametric multivariate model based on Gaussian copulas and covariance modelling, and using this to develop an approximate likelihood ratio test. Simulations demonstrate this method has desirable properties.

Penalized likelihood parameter estimation for additive hazard model using method of multipliers

Kasun Rathnayake

Macquarie University

kasun.rathnayake@students.mq.edu.au

Coauthors: Jun Ma, Macquarie University, Sydney, Australia

Here we propose a procedure to simultaneously estimate the baseline hazard and the regression coefficients in additive hazard model. Maximum penalized likelihood (MPL) method is used for parameter estimation and a penalty function is used to smooth the baseline hazard estimate. Here, we have two nonnegativity constraints. First constraint is on the baseline hazard and the second is on the additive hazard function. Alternating Direction Method of Multipliers (ADMM) and Multiplicative Iterative (MI) method are used to estimate baseline hazard and regression coefficients.

For the simulation studies, we used indicator functions as the basis function and both equal bin event count and equal bin observation count as the binning criterion for right censored survival data. The preliminary results of the simulation studies shows that this procedure can be relatively efficient and the standard deviations and bias of the MPL estimates of the regression coefficients decrease with sample size, but increase with the proportion of censoring. In addition, the estimates for the equal bin event count gives more accurate results compared to equal bin observation count scenario.

Point process models for presence-only analysis

Ian Renner

University of Newcastle

Ian.Renner@newcastle.edu.au

Coauthors: Jane Elith, Adrian Baddeley, William Fithian, Trevor Hastie, Steven J. Phillips, Gordana Popovic, and David I. Warton

Species distribution modelling is a branch of environmental statistics in which the distribution of a species is modelled as a function of environmental (and perhaps additional) covariates. A number of different methods have been proposed to model species distributions for presence-only data, in which the available information consists of a list of observed locations. When data arise as point events in this matter, point process models provide a flexible framework for analysis. Recently, many presence-only methods have been linked to point process models, leading to the need to explore the relative advantages and drawbacks of each approach to various aspects of model fitting, prediction, and diagnostic checks. In this talk, I will demonstrate some of the extent of the applications of point process models and compare the capacity of different approaches to fitting point process models in the way of inference, regularisation, diagnostic ability, and accounting for spatial dependence. The approaches considered include: MAXENT Standard GLM approaches (infinitely weighted logistic regression and newly-proposed downweighted Poisson regression) R packages for point process models (spatstat and ppmlasso) R packages for Log-Gaussian Cox processes (R-INLA and lgcp)

Reference:

Point process models for presence-only analysis (2015). Methods in Ecology and Evolution 6(4), pp 366-379. doi: 10.1111/2041-210X.12352

Computing Standard Error for Half-life of Rotenone

Maheswaran Rohan

Department of Biostatistics & Epidemiology, Auckland University of Technology

mrohan@aut.ac.nz

Coauthors: Alastair Fairweather, Natasha Grainger

Rotenone, a natural toxin, is used to eradicate invasive pest fish in New Zealand. Following its use as a piscicide, it is important to monitor the rate of dissipation of rotenone from the water-body and a half-life for the dissipation of rotenone so that normal use activities can re-commence. Rohan et al. (2015) have developed a stochastic gamma model over a traditional dynamic model for determining the half-life of rotenone under various conditions, however, direct computation of the standard error of the half-life estimate is not easy due to the complex form. In this talk, we discuss how to improve the model by computing the standard error of the half-life, using the delta method.

Reference:

Rohan, M., Fairweather, A., Grainger, N., (2015), Using gamma distribution to determine halflife of rotenone, applied in fresh water, Science of the Total Environment, 527-528: 246-251

Integrative meta analyses to combine transcriptomics studies

Florian Rohart

Australian Institute for Bioengineering and Nanotechnology (AIBN), The University of Queensland, Australia

f.rohart@uq.edu.au

Coauthors: Florian Rohart, A. Eslami, S. Bougeard, C. Wells and K-A. Lê Cao

The fast moving nature of experimental developments has led to a wealth of publicly available biological data coming from different sources. Combining such large-scale biological data sets, which are otherwise becoming obsolete, can lead to the discovery of important biological insights provided that relevant information can be extracted. The main statistical challenge when analysing combined data from independent studies is data heterogeneity. The data to be integrated originated from different technological platforms, were assayed in different geographical sites and at different times. Although the data are generated under similar biological conditions, the difference between these series of measurements is large and acts as a confounding factor in the combined analysis. This leads to a spurious relationship between the biological outcome of interest (e.g. disease vs. healthy) and the effect of the different technological platforms. This systematic error is generally referred to as "batch effect". Two main types of analysis can be performed when combining biological data sets. "Meta-analysis" involves analysing each data set separately and combining the results while "Integrative analysis" first combines the samples from all datasets before analysing a single integrated data set. Although Integrative analysis is researchers' favoured option, it is hindered by batch effects that cannot be effectively removed by common normalisation methods. Batch effects come from every step of the biological experiment (organism growing conditions, tissue sampling, RNA processing, hybridisation and data processing) and represent a critical obstacle in the analysis, especially since potential sources of batch effect are rarely fully reported or recorded. We have developed a multi-group approach that classify samples and sits in between meta-analysis and integrative-analysis. Our approach combines transcriptomics data sets by integrating the data after accommodating for batch effects but retains some meta-analysis aspects by keeping the group structure of the data. We applied our sparse multi-group approach on hundreds of samples coming from fifteen independent experiments looking at human cells (Fibroblasts, hESC, hiPSC) with more than 13, 000 genes, our

approach identified 17 genes leading to a Balanced Error Rate (BER) of 7. Compared to a two-step procedure for batch removal and classification of the samples, our approach gave a superior performance demonstrating its promising potential as BER were higher than 9 for all methods and around 10 when the popular ComBat normalisation method was involved.

Efficient recruitment strategies in randomised controlled trials with continuous outcomes

Tibor Schuster

Clinical Epidemiology and Biostatistics Unit, Murdoch Childrens Research Institute, Royal Childrens Hospital, Melbourne, Victoria

Tibor.Schuster@mcri.edu.au

Coauthors: John B. Carlin, and Katherine J. Lee

At the planning stage of a clinical trial, the assumed variance of the primary outcome variable is a key parameter in determining the required sample size. Established factors known to explain outcome data dispersion are often used to define stratification variables to be considered in the randomisation procedure and, after completion of the trial, as covariates in the statistical analysis. However, the representation of stratification subgroups in the study sample commonly remains unspecified so that the ultimate strata distribution depends on the natural course of patient accrual. If effect homogeneity across strata can be assumed, for a fixed total sample size, precision of the effect estimate can be expressed as a function of within and between-subgroup variance components and the prevalence of each subgroup in the study sample. It follows that different representations of subgroups may lead to different required sample sizes in order to satisfy pre-specified requirements for precision of estimation and statistical power. We demonstrate how variance decomposition can be used to assess the extent to which the precision of an effect estimate depends on sample proportions of subpopulations. This approach is shown to be useful for the identification of efficient recruitment strategies which lead to minimum sample sizes under simultaneous consideration of feasibility of recruitment and desired subgroup representation to support generalisability. Applications to published study data on atopic dermatitis in infants and a large set of simulated trial scenarios confirm substantial possible gains in efficiency using the proposed approach.

Analysing longitudinal data with outcome-dependent sampling

Alastair Scott

Department of Statistics, University of Auckland

a.scott@auckland.ac.nz

Coauthors: Chris Wild

We look at methods for analysing data from longitudinal studies in which the chance of being included in the study depends on quantities that are correlated with the outcome of interest. For example, in a study of Attention Deficit Hyperactivity Disorder (ADHD), children were selected for the study with probabilities based on a care-giver's assessment of their ADHD status. Such biased sampling can invalidate standard methods of analysis. We develop efficient likelihood-based methods for fitting generalised linear mixed models to longitudinal data arising from outcome-dependent sampling, and compare the results with alternatives that have been suggested. We apply the results to a couple of examples, including the ADHD study mentioned above.

Sliding Through Phylogenetics

Daisy Shepherd

The University of Auckland

dshe078@aucklanduni.ac.nz

Coauthors: Steffen Klaere

Phylogenetics focuses on a critical problem in biology the ability to derive the evolutionary history between groups of organisms. Statistical models are used to describe the changes in DNA that occur during evolution, to help determine how closely related these groups are. As a result, our ability to accurately explain the evolutionary relationships depends heavily on the use of an appropriate statistical model.

Unfortunately, the complexity of the evolutionary mechanisms presents a number of challenges to the modelling process. One such problem concerns the presence of varying rates of evolution across our DNA alignment. Failure to account for this rate heterogeneity leads to poorly fitting models, and poor estimation about the evolutionary relationships.

When selecting which model to use, traditional approaches look at the DNA alignment as a whole. However, generalising behaviour across the complete alignment could affect how we detect these heterogeneous rates.

Therefore, we proposed an alternate technique to analysing our DNA sequences the sliding window (SW) approach. This method involved partitioning the alignment into windows of sites, before iteratively fitting a model to each subset. We wished to apply this existing statistical technique to our particular problem, and test whether the SW approach improved our ability to detect variation in evolutionary rates.

Whilst the sliding window approach was more computationally intensive, the ability to profile and apply multiple inferences across an alignment allowed more insightful and detailed heterogeneity detection within the phylogenetic analysis. Results indicated the approach is undoubtedly a useful tool in detecting rate heterogeneity. Applying this approach has the potential to improve the statistical models we use, and more accurately model the evolutionary relationships between groups of organisms.

Sensitivity analysis within the multiple imputation framework: The pattern-mixture method

Julie A Simpson

Centre for Epidemiology and Biostatistics, Melbourne School of Population and Global Health, The University of Melbourne

julieas@unimelb.edu.au

Coauthors: Panteha Hayati-Rezvan and Katherine J Lee

Multiple imputation (MI) is a well-recognised statistical technique for handling missing data. As usually implemented in standard statistical software, MI assumes that data are "Missing at random" (MAR); an assumption that cannot be tested using the data available and in many settings may be implausible. Two model-based methods, the weighting approach and the pattern mixture method, have been proposed within the MI framework for performing sensitivity analyses to assess departures from the MAR assumption. In previous work, we demonstrated that the weighting approach provides biased parameter estimates, even when a large number of imputations are performed. For this presentation, we show using a simulation study that the pattern mixture method produces unbiased and consistent estimates across a varying number of imputations. We extend the pattern mixture method to handle missing data in more than one variable and present an application of the method to data from the Longitudinal Study of Australian Children.

P-Spline Vector Generalized Additive Model and Its Application

Chanatda Somchit

Department of Statistics, The University of Auckland

csom017@aucklanduni.ac.nz

Coauthors: Thomas Yee and Chris Wild

Vector generalized additive models (VGAMs) are an extension of the class of generalized additive models (GAMs) to include a class of multivariate regression models by using vector smoothing. The class of VGAMs is now very large and includes many statistical distributions and models. For example, these models are univariate and multivariate distributions, categorical data analysis, quantile and expectile regression, time series, survival analysis, extreme value analysis, and nonlinear least-squares models. Parameter estimation can be achieved by using a modified vector backfitting algorithm. The main issue is that it is not easy to efficiently integrate smoothness estimation methods with the backfitting approach.

In our research study, we aim to develop new efficient methods based on penalized regression splines for estimating parameter coefficients for the VGAM models, and to integrate the automatic numerical procedure used to determine the shape of non-linear terms from the data to the VGAM framework. To achieve these, we develop VGAMs based on penalized regression splines using P-spline smoothers, which we term 'P-spline VGAMs'. By using P-spline smoothers for every smooth component, VGAMs can be transformed into the VGLM framework. P-spline VGAMs can be then fitted by penalized likelihood maximization. In practice, this maximization can be achieved by penalized iteratively re-weighted least squares (P-IRLS). Importantly, we implement the penalized iteratively reweighted least squares approach for the full range of VGAM models involving complications like constraints on model terms. With our approaches, the multiple smoothing parameters can be estimated by minimization of the approximate unbiased risk estimator (UBRE). The computational procedure for the automatic and stable multiple smoothing parameter selection based on the pivoted QR decomposition and singular value decomposition is employed to minimize this criterion.

In this talk, some theoretical and practical aspects of P-spline VGAMs are described, the practical performance of the proposed method is shown through an extensive simulation study, and the new approach is illustrated on two data sets using several statistical models, which

include the multinomial logit, proportional and non-proportional odds models, bivariate logistic model, and the LMS method for quantile regression. It is attempted to show that P-Spline VGAMs offer greater scope for additive modeling and an advantage of automatic smoothing parameter selection is conveyed to a very large class of models.

On quadratic logistic regression models when predictor variables are subject to measurement error

Jakub Stoklosa

The University of New South Wales

j.stoklosa@unsw.edu.au

Coauthors: Yih-Huei Huang, Elise Furlan and Wen-Han Hwang

Owing to its good properties and a simple model fitting procedure, logistic regression is one of the most commonly used methods applied to data consisting of binary outcomes and one or more predictor variables. However, if the predictor variables are measured with error and the functional relationship between the response and predictor variables is non-linear (e.g., quadratic) then consistent estimation of model parameters is more challenging to develop. To address the effects of measurement error in predictor variables when using quadratic logistic regression models, two novel approaches: (1) an approximated refined regression calibration; and (2) a weighted corrected score method are developed to adjust for the bias inherent in naive estimators. Both proposed approaches offer several advantages over existing methods in that they are computationally efficient and are straightforward to implement. A simulation study was conducted to evaluate the estimators' finite sample performance. The proposed methods are also applied on real data from a medical study and an ecological application.

A study of correlated marker effects for dense linkage map in QTL analysis in wheat

Emi Tanaka

University of Wollongong

etanaka@uow.edu.au

Coauthors: Alison Smith and Brian Cullis

The detection of quantitative trait locus (QTL) is often an important early step for identification of genes that cause trait variation. The analysis of QTL is often done making use of a bi-parental population and genetic markers such as single nucleotide polymorphisms (SNPs). The linear mixed model approach is often used as it accommodates well to account for non-genetic sources of variation as well as complex designs. Verbyla et. al (2007) considered the use of all markers for QTL analysis however assume that marker effects are identically and independently distributed (iid). This approach may have been more suitable at a time when the marker covariates were sparse but with the availability of dense marker map, the assumption of independent marker effects is rather weak. A number of different spatial covariance structure to the marker effects within chromosomes were proposed (Gianola et al., 2003, Smith and Cullis, 2011, Yang and Templeman, 2012, Morota and Gianola, 2014) including modelling correlation of marker effects by first order autoregressive process; Gaussian decay; and bayesian antedependence model. We apply various spatial covariance structure to the marker effects within chromosomes in comparison with the iid marker effects using real data.

References:

Gianola, D., Perez-Enciso, M. & Toro, M. A. (2003) On marker-assisted prediction of genetic value: Beyond the ridge. Genetics. 163: 347-365.

Verbyla, A. P., Cullis, B. R. & Thompson, Robin (2007) The analysis of QTL by simultaneous use of the full linkage map. 116: 95-111

Smith, A. & Cullis, B. R. (2011) Detecting QTL for photoperiod sensitivity in a doubled haploid Brassica napus population. Statistics for the Australian Grains Industry Technical Report Series

Morota, G. & Gianola, D. (2014) Kernel-based whole-genome prediction of complex traits: a review. 5: 1-13

Yang, W. & Tempelman, R. J. (2012) A bayesian antedependence model for whole genome prediction. 190: 1491-1501

Interactive and data adaptive model selection with mplot

Garth Tarr

University of Newcastle

garth.tarr@newcastle.edu.au

Coauthors: Samuel Meuller and Alan Welsh

This talk introduces the mplot R package which provides a collection of functions to aid exploratory model selection (Tarr et al., 2015). We have developed routines for modified versions of the simplified adaptive fence procedure (Jiang et al., 2009) as well as other graphical tools such as variable inclusion plots and model selection plots (Mueller and Welsh, 2010; Murray et al., 2013). A browser based graphical user interface is provided to facilitate interaction with the results. These variable selection methods rely heavily on bootstrap resampling techniques. Fast performance for standard linear models is achieved using the branch and bound algorithm provided by the leaps package. Reasonable performance for generalised linear models and robust models can be achieved using sensible default tuning parameters and parallel processing. The methods implemented in mplot allow us to better explore the stability of model selection criteria.

References

Jiang J., Nguyen T. and Rao J.S. (2009). A simplified adaptive fence procedure. Statistics and Probability Letters, 79(5), 625-629. DOI: 10.1016/j.spl.2008.10.014

Mueller S. and Welsh A.H. (2010). On Model Selection Curves. International Statistical Review, 78, 240-256. DOI: 10.1111/j.1751-5823.2010.00108.x

Murray K., Heritier S. and Meuller S. (2013). Graphical tools for model selection in generalized linear models. Statistics in Medicine, 32, 44384451. DOI: 10.1002/sim.5855

Tarr G., Mueller S. and Welsh A.H. (2015). mplot: Graphical model stability and model selection procedures. R package. https://github.com/garthtarr/mplot

Inference for multivariate abundance data using generalised mixed effects models and the PIT-trap

Loïc M. Thibaut

School of Mathematics and Statistics and Evolution & Ecology Research Centre. The University of New South Wales. NSW 2052

Loic.Thibaut@unsw.edu.au

Coauthors: David I. Warton

In ecology, multivariate abundance data are widely used to study how community structure changes along environmental gradients and to test hypotheses about the impact of some environmental variable or experimental treatment. Typically, distance-based approaches have been used to perform such analyses: to reduce the dimensionality of the data, a distance is calculated among pairs of multivariate observations and the analysis proceeds on the resulting distance matrix. Model-based methods provide a modern alternative to distance-based approaches by explicitly modelling the multivariate responses as a function of the predictors. Because the number of observations is usually not largely greater than the number of response variables, specific resampling procedures are required for inference. The PIT-trap is such a resampling procedure, based on the probability integral transform residuals. Simulations studies show that model-based approaches using the PIT-trap have better power properties than distance-based approaches. However it is unclear whether the good properties of the PIT-trap for generalised linear models extend to models with random effects. In this paper, we extend the PIT-trap procedure to mixed effects models and compare it to distance-based approaches. Our bootstrapping scheme proceeds in two steps: we first resample the spherical random effects, then we resample the PIT-residuals to generate the response variable. Overall, we found that model-based approaches perform better than traditional approaches such as permanova for generalised linear mixed effects models.

Improving the accuracy of genetic predictions for expensive multiphase traits

Daniel Tolhurst

University of Wollongong

tolhurst@uow.edu.au

Coauthors: Alison Smith, Steve Jefferies and Brian Cullis

The provision of accurate information on end-use capabilities of wheat varieties is crucial for both breeding and commercial purposes. The measurement of these traits involves multi-phase experiments, where a series of field trials (Phase I) is followed by one or two laboratory phases (Phase II and III). The laboratory phases are typically expensive (comparative to phase I) and time consuming. Consequently, laboratory processing has historically been carried out using composite samples for each variety (by combining all field replicates) in the absence of sound experimental design techniques (including randomisation and replication). This has precluded an efficient statistical analysis and, hence, compromised the accuracy of predictions of genetic effects of interest.

In this talk we demonstrate the approach of Smith et al. (2015) using an Australian wheat quality project, where traits of interest include a range of grain, flour, dough and end-product characteristics. The approach incorporates sound design techniques in every phase of testing so that an efficient analysis can be conducted on the resultant data, including (for the first time) quantification of non-genetic sources of variation and a valid estimate of error. Replication from the field phase is achieved by using a mixture of individual replicate and composite samples that are processed separately. Replication in each laboratory phase is done by splitting a sample from the previous phase to form two subsamples, again for separate processing. In addition to considering these ideas, we show how the approach of Smith et al (2015) has satisfied a number of budgetary, time and practical constraints; in terms of complying with commercial laboratory practice.

Finally we present the results of a simulation study that quantifies the improvements in the accuracy of genetic predictions using the Smith et al (2015) approach compared with the historical method.

References:

Smith, A. B., Butler, D. G., Cavanagh, C. R. & Cullis, B. R. (2015). Multiphase variety trials using both composite and individual replicate samples: a model-based design approach. The Journal of Agricultural Science 153, 10171029.

Fitting quadratic peaks to fluorescence data to identify chemical compounds

Martin Upsdell

AgResearch

martin.upsdell@agresearch.co.nz

Certain chemical compounds absorb light at a particular wavelength and emit light at a longer wavelength. The wavelengths that are absorbed and emitted can be used to identify the compound. The amount of light emitted, the fluorescence, is proportional to the amount of chemical present.

A data set of 108 wool samples will be used to identify some of the compounds influencing the coloration of wool. Light was shone at each sample at wavelengths from 220nm to 490nm in 10nm steps. The amount of light emitted was measured at each wavelength between 330nmm to 600nm in 10nm steps. Only amounts coming from points where the emitted wavelength lies between the excited wavelength and 2 times the excited wavelength are used. This results in 468 data points for each sample, giving a total of 50,544 data points over all 108 wool samples. The size of the data set means that there is a strong tendency to overestimate the number of chemical compounds present.

Evaluating Frequentist Model Averaged Confidence Intervals

A.H. Welsh

Australian National University

Alan.Welsh@anu.edu.au

Coauthors: Paul Kabaila, La Trobe University, Melbourne, Australia P.Kabaila@latrobe.edu.au Waruni Abeysekera, La Trobe University, Melbourne, Australia W.Abeysekera@latrobe.edu.au

We evaluate frequentist model averaging procedures by considering them in a simple situation in which there are two nested linear regression models over which we average. We obtain exact expressions for the coverage and the scaled expected length of the model-averaged intervals and apply them to compare model averaged profile likelihood confidence intervals and model averaged tail area confidence intervals. We show that the profile-likelihood confidence intervals are not better than the standard confidence interval used after model selection but ignoring the model selection process. The tail-area confidence intervals perform better than the profile likelihood and post-model-selection confidence intervals but, for the examples that we consider, offer little over simply using the standard confidence interval under the full model, with the same nominal coverage.

Health effects of disasters

Rory Wolfe

Monash University

Rory.Wolfe@monash.edu

Tsunami, earthquake, bushfire. World Trade Centre, Hiroshima, Chernobyl. Disasters come in many forms and their impact on human health is typically well-characterised as a death toll in the immediate aftermath. But what about longer-term effects on health? Study design for this question is challenging and usually retrospective as will be illustrated with a study of the long-term health effects of the Hazelwood open-cut coalmine fire in Victoria. A range of statistical methods need to be called on for analysis with multilevel regression and poststratification (Wang 2015) offering particular promise. Study design can also be based on surveillance, for example using a component of the health system such as routine reporting of disease. This approach will be illustrated with a surveillance-based study of the impact of natural disasters in the US on the health of older people. Joint modelling of disability and time to death enables the impact of disasters to be quantified. We adopt a Bayesian approach to estimating the joint model parameters noting that the implementation of this approach is becoming increasingly feasible (e.g. Rizopoulos 2015). Aspects of the model's implementation and interpretation of results will be discussed.

References:

Rizopoulos D (2015). JMbayes: Joint Modeling of Longitudinal and Time-toEvent Data under a Bayesian Approach. R package version 0.7-2. http://CRAN.Rpro ject.org/package=JMbayes Wang W, Rothschild D, Goel S, Gelman A (2015). Forecasting elections with non-representative polls. International Journal of Forecasting. 31:980-991.

How many letters are there in the alphabet?

Jeff Wood

Fenner School of Environment and Society, Australian National University

jeff.wood@anu.edu.au

The problem referred to in the title of this talk occurred to me when on holiday in Greece. In Greek, as you probably already know, there are some upper case letters which you do not see in English, for example Λ and Ξ, as well as old friends like A and B. However if you look at car number plates you only see Latin letters. The first question that occurred to me was how many plates would you have to see before you felt confident that purely Greek letters were not used. The second question is how many letters do they actually use. I saw a poster advertising a body-building competition in Crete called MR KPHTH, so they are not above using non-Greek letters when it suits them.

The number plate problem is an example of what is often called species richness estimation. If an ecologist goes to a site and counts all the butterflies he sees, how many more species were present that he did not see. The earliest reference seems to be a paper by Fisher, Corbet and Williams (1943).

The same issue occurs in many other disciplines including linguistics, numismatics, genomics, computer science, etc. The data is often presented as a frequency of frequencies, the number of species which were seen once, the number of species seen twice, etc. The interest, of course, is in the number of species which were seen zero times. The problem cannot be solved without making some assumptions about the occurrence of very rare species. In the number plate example I feel that there will not be very rare letters, so we would hope that any sensible method would give a reasonable result, so it seems to me it is a good test bed for the various methods which have been proposed. And we know the answer, or we can find it out.

I collected some data from vehicles in Canberra and in Macedonia and applied a number of methods which have been proposed. The results were disappointing.

Mapping Soil Regolith Depth in Large and Censored Spatial Datasets Using Bayesian Hierarchical Models

Wen-Hsi Yang

CSIRO

Wen-Hsi.Yang@csiro.au

Coauthors: Searle, Ross (CSIRO Land and Water, Dutton Park, QLD, Australia), Clifford, David (The Climate Corporation, San Francisco, California, USA), Wilford, John (Geoscience Australia, Canberra, ACT, Australia)

Regolith is a layer from the Earth's surface down to unweathered bedrock at depth. This layer plays an important role in controlling hydrological and landscape processes which are essential for plant growth, aquifer recharge and landscape salinity. As such, understanding regolith depth is important for activities such as agriculture, forestry and mining. In Australia, regolith depth has been measured at thousands of locations in the state of Queensland. These measurements are typically derived from estimates of bedrock depth recorded when bores are dug in search of groundwater resources. Hence, some measurements may be right censored as groundwater may be encountered before bedrock is reached. In this study, we propose Bayesian hierarchical spatial models for large and censored spatial datasets. Importantly, our model extracts features of geographic environmental covariates to improve prediction rather than to explain variation of regolith depth. Finally, we illustrate the effectiveness of our approach through simulation and by applying our model to the Queensland dataset.

Fitting linear mixed models under misspecification

Hwan-Jin Yoon

Statistical Consulting Unit, The Australian National University

hwan-jin.yoon@anu.edu.au

Coauthors: Alan Welsh

A popular model for the analysis of clustered data is the linear mixed model (LMM). The effect of the cluster structure in the response variable is incorporated by including random effects in the model; cluster structure in the explanatory variables is often ignored, possibly because interpreting the model conditionally makes it seem unnecessary to consider this structure. At least when the explanatory variable has a normal distribution, cluster structure in the explanatory variable can be incorporated into LMM by centering the explanatory variable about the cluster means and then also including the cluster means in the model as a cluster-level covariate. This is called the contextual effects linear mixed model (CLMM) given as

$$y_{ij} = \beta_0 + \beta_b \bar{x}_i + \beta_w \left(x_{ij} - \bar{x}_i\right) + \delta_i^* + \varepsilon_{ij}^*,$$

where the subscripts *w* and *b* are within and between-cluster slopes.

The CLMM is an extension of LMM to allow cluster structure in the explanatory variables. When within and between-cluster effects are equal, the CLMM reduces to the LMM. Therefore, if the CLMM can be treated as the true model, the LMM can be viewed as a misspecified submodel of the CLMM (Berlin et al., 1999; Chao et al., 1997). Not considering the cluster structure in the explanatory variable can be lead to very misleading assessments of the association between the response and explanatory variables and misleading estimates of the variance components.

In this talk, we explore the effect of fitting LMM under misspecification (i.e. when the CLMM holds with $\beta_b \neq \beta_w$) by both maximum likelihood (ML) and restricted likelihood (REML) estimation.

References:

Berlin, J.A., Kimmel, S.E., Ten Have, T.R., Sammel, M.D.(1999) An empirical comparison of several clustered data approaches under confounding due to cluster effects in the analysis of complications of coronary angioplastry. Biometrics 55: 470 476

Chao, W.H., Palta, M., Yound, T. (1997) Effect of omitted confounders on the analysis of correlated binary data. Biometrics 53: 678 689

A study of one and two stage analyses for genomic prediction of yield in wheat

Chong You

University of Wollongong

chongy@uow.edu.au

Coauthors: Emi Tanaka, Alison Smith, Brian Cullis

Historically, prediction of additive genetic effects was based on the pedigree based additive infinitesimal model (Fisher, 1918), however molecular markers paved ways to marker-assisted selection and now, with the availability of dense genome-wide molecular markers, to the use of genomic selection (Meuwissen et al., 2001).

In crop breeding, the use of line replication allows for genetic effects to be partitioned into additive and non-additive (or residual) effects using a linear mixed model approach. The additive effects are further partitioned into marker effects and marker lack of fit (or polygenic) effects with the joint use of marker and pedigree information (Garrick et al., 2009).

In contrast to this joint analysis, i.e. the so-called one-stage analysis, an alternative analysis can involve two steps: step 1 to incorporate pedigree information as well as experimental design terms and step 2 that uses the (deregressed) estimated breeding value from step 1 as the response variable and incorporate marker information in a linear mixed model framework. This so-called two-stage analysis is employed widely in animal breeding (e.g. Ostersen et al. 2011) with some usage in plant breeding (e.g. Zhoe et al. 2013, Rutkoksi et al. 2014).

In this talk, we present a comparative study of single-environment one-stage analysis to two-stage analysis wheat breeding trials using a linear mixed model approach. Our results clearly show the effectiveness of the one-stage analysis over two-stage analysis in the genomic selection context.

Application of a Bayesian Markov chain Monte Carlo approach for modelling the dynamics of Plasmodium falciparum parasitaemia in severe malaria patients

Sophie Zaloumis

Centre for Epidemiology and Biostatistics, Melbourne School of Population and Global Health, University of Melbourne, Victoria 3010

sophiez@unimelb.edu.au

Coauthors: Julie Simpson and PKPD-IVARS Study Group

The largest pooled dataset of drug concentration and parasite count data collected from severe malaria patients receiving intravenous artesunate (IV-ARS) has been assembled. The results of pharmacokinetic modelling of this dataset were used as evidence to revise the WHO guidelines for the treatment of severe malaria (3 mg/kg for children weighing less than 20kg and 2.4 mg/kg for larger children and adults). We aim to complement this work by performing a pharmacodynamic analysis of the corresponding pooled parasite counts. This will provide insights into drug action, the killing rate constant and in vivo IV-ARS concentration at which half the maximal killing constant of the drug is achieved (referred to as the EC50 concentration), and the distribution of the age of the parasites, in particular, the proportion of sequestered parasites, within a severe malaria patient.

The pooled dataset consists of 70 adults (age range 16 to 75 years) and 195 children (age range 6 months to 9 years) with severe malaria who were administered IV-ARS. There was a mixture of sparse and rich sampling designs over 12 hours with, on average, 6 parasitaemia measurements available per patient (range 1 to 14). The pharmacodynamic model is a mechanistic model relating antimalarial drug concentration to the clearance of parasites in the body over time and nonlinear mixed-effects (NLME) modelling was used to allow for betweenand withinpatient variability in the clearance of parasites over time. Drug concentrations for each patient at a particular time point, corresponding to when a parasite count was recorded, were predicted using parameters that govern the within patient processes of drug absorption, distribution and elimination obtained from NLME modelling of the observed drug concentration data. Bayesian inference and Markov chain Monte Carlo methods (such as the multiple-block Metropolis-Hastings algorithm and parallel tempering) were used to derive point

and interval estimates for the pharmacodynamic parameters of interest. All analyses were performed in the open source software package R.

Stationary distribution of the linkage disequilibrium coefficient r^2

Wei Zhang

Department of Statistics, The University of Auckland, New Zealand
wzha217@aucklanduni.ac.nz

Coauthors: J. Liu, B. S. Weir and R. M. Fewster

The linkage disequilibrium coefficient r^2 is a measure of statistical dependence of the alleles possessed by an individual at two different genetic loci. It is used to find the positions of disease-causing genes on the chromosomes. For this reason, seeking the statistical properties of r^2 is an important and meaningful issue. The maximum entropy principle is a useful tool to approximate the density function of an unknown distribution, given a sequence of the distribution's moments. Here I use this method to approximate the density function of r^2 using some stationary moments computed under models for genetic drift. In order to obtain the sequence of moments, I generalize an analytic method that was originally used to compute the expectation of r^2.

Abstracts – Poster Session

Robust analysis of Poisson and binary data using a mixture model approach

Ken Beath

Macquarie University, Australia

ken.beath@mq.edu.au

The method typically used for robust analysis of Poisson or binary data is an M-estimator which down weights unusual observations. This has a number of limitations, in particular that it is not likelihood based which prevents the use of some techniques in model building and may result in a loss of efficiency. One approach to robustness in linear models is to assume that observations are from a mixture of two groups, either standard or outlier, with different error variance. This approach is modified for Poisson and binary data by assuming that observations come from a mixture of a standard and overdispersed distribution, where the overdispersion is modelled by including a random effect. This method is compared to an M-estimator using simulation for Poisson data. The method also allows for a statistical test for the presence of outliers and identification of outliers.

Varying coefficients regression with local variable selection

Wesley Brooks

University of New South Wales

wesley.brooks@unsw.edu.au

Coauthors: Jun Zhu, University of Wisconsin-Madison Zudi Lu, University of Southampton

Varying coefficients regression is a flexible technique for modeling data where the coefficients are functions of some effect-modifying parameter, e.g. time or location. While there are a number of methods for variable selection in a varying coefficient regression model, the existing methods are for global selection, which includes or excludes each covariate over the model's entire domain. Presented here is local adaptive grouped regularization (LAGR), a new method for local variable selection in varying coefficients regression. Through an adaptive group Lasso procedure tailored to locally linear regression, LAGR identifies the covariates that are associated with the response at any point in space and simultaneously estimates their coefficients.

The method of LAGR possesses the theoretical "oracle" properties, tending (with increasing sample size) to identify the correct local covariates and to estimate their coefficients as accurately as if their identities were known in advance. The R package lagr can be used to estimate a model by LAGR and is available from github.com/wrbrooks/lagr. The method is illustrated by application to an ecological data set and the finite sample properties of LAGR are assessed in a simulation study.

Piecewise mixed regression models: A tool to detect trajectory divergence between groups of longitudinal outcomes in long-term observational studies

Marie-Jeanne Buscot

Menzies Institute of Medical Research, University of Tasmania

M.Buscot@utas.edu.au

In epidemiological prospective cohort studies, especially those designed to study adult-onset disorders such as heart disease, investigators tend to follow study participants over relatively long periods, sometimes several decades, to determine how a number of risk factors, their interactions and normal aging may impact the onset and the progression of disease in the population over time. Subjects are often stratified in groups, to assess the association between the longitudinally collected outcome and the categorical factor of interest. In these instances, the main scientific interest often lies in being able to determine whether the outcome is stationary over time between and within groups of subjects, and whether the different groups are changing in a similar or different fashion overtime. Change-point models have become a useful alternative to linear models when the interest lies in determining if and when changes have taken place in an event window. However, the use of a multilevel random change point model to test for the existence of a trajectory divergence between two (or more) groups of longitudinal observations has not been explored so far.

Here, through a simulation study, we investigate the properties of a hierarchical random broken stick change-point model formulated to provide an estimate of the time at which the trajectories of a continuous outcome start diverging between two groups of subjects. By comparing such a model to set of candidate hierarchical models formulated with no change-point between groups, we propose a method to formally test the hypothesis that the response trajectory of one group of observations departs from the other(s), and determine the point in time at which the divergence occurs. We adopt a sampling-based Bayesian procedure using Markov chain Monte Carlo (MCMC) methods.

The colourful life of a biometrician

Mario D'Antuono

Department of Agriculture and Food, Western Australia

mario.dantuono@agric.wa.gov.au

Coauthors: Robin Jacob

Life is full of colour but it can be a very subjective matter in the eyes of the beholder. We describe some recent work where knowing the colour on a mathematical scale may help colour blind people in the world as well as the majority in providing some consistency to quality standards in many industries in agriculture such as the meat industry.

A statistical analysis of the images (photographs) from the samples of meat from an experiment was performed to compare various treatments. The magic number 42 appears to be relevant in defining the colour of meat on a particular mathematical scale.

Applying the factor analytic mixed model to meta-analyses to assess legume yield performance.

Ross Darnell

CSIRO

ross.darnell@csiro.au

Coauthors: Lindsay Bell and Justin Fainges

A meta-analysis was undertaken to explore genotype by environmental interactions to compare the relative performances of forage legumes in northern Australia. The data was sourced from historical evaluation of legume accessions at multiple sites and over several years. Data were represented as the accession- averages for a particular site at a particular assessment time, so it was not possible to estimate inter-plot variability. Not all accessions were grown at all sites so the analysis is moderately unbalanced.

The meta-analysis aimed to use this data set to identify genotypes (species and/or accessions) that were ”standouts”, those that persisted and produced well across a range of conditions and those that were persistent and productive across specific environments.

We follow the analysis as shown in Smith (2015) which is typically used for multi-environment trials where plot information is available. For cases in our legume database for which there was sufficient species data, we used factor analytic mixed models to perform these analyses to explore whether these models are useful where plot data is not available.

Identification of pre-clinical Alzheimers Disease in a longitudinal study of ageing with overfitted finite mixture models.

James Doecke

ACEMS, Queensland University of Technology

James.Doecke@csiro.au

Coauthors: Zoe van Havre, Paul Maruff, Victor Villemagne, Kerrie Mengersen, Judith Rousseau, Nicole White

Alzheimers Disease (AD) is a long protracted pathophysiological process, extending for more than two decades. During the early stages of the disease process, little evidence of the building pathology is identifiable without either CSF and/or PET imaging analyses. Clinical manifestation of AD does not present until late stages of pre-clinical or early stages of prodromal disease. From a large body of neuropsychological tests, six composite scores have been proposed to encapsulate different aspects of the disease. The current study investigates longitudinal composite score data from the Australian Imaging, Biomarkers and Lifestyle (AIBL) study of ageing to identify those elderly cognitively nor- mal participants with strong probability of having pre-clinical AD. Using a recently published Bayesian method for overfitted finite mixture models, we define an aggregate measure of posterior probabilities (AMPP score) establishing the likelihood of pre-clinical AD. From Baseline through 54 months, visuo-spatial function had the greatest contribution to the AMPP score, followed by visual memory, verbal memory and language. The AMPP score identified increasing neo-cortical β-amyloid (Aβ) and decreasing hippocampus volume over 54 months for those participants in the highest category of the aggregate, com- pared to those in the lowest category with stable Aβ-burden. In conclusion, we demonstrate here the use of a Bayesian clustering method to define probabilities of pre-clinical AD in a cognitively normal elderly control population over a 54-month time period.

Food flavour compound analysis using multivariate methods

Lindy F.G. Guo

Plant and Food research, 120 Mt Albert Road, Sandringham, Auckland

lindy.guo@plantandfood.co.nz

Flavor is one of the most important aspects concerning the acceptance of food by the consumer. Our scientists suspect that certain groups of chemical compounds may largely contribute to some special fruit flavors. By studying 157 different kinds of chemical compound extracted from a particular fruit, it is attempted to find out if any compounds could be a major source for determining the fruit species and hence determining flavors. This presentation outlines a couple of multivariate techniques used in an initial look at the data and some findings.

Maximising the benefits of a plant breeding data-base

Peter Jaksons

Plant & Food Research Lincoln, New Zealand

Peter.Jaksons@plantandfood.co.nz

Coauthors: Peter Alspach, Plant & Food Research, Motueka, New Zealand
Carmel Woods, Plant & Food Research, Lincoln, New Zealand

Plant breeding is an important part of Plant & Food Research's work. Recently we have started using an SQL plant breeders database application (E- Brida) in most of our breeding programmes. To maximise the benefit of this application, we have been developing an R package. This R-package has two main purposes. The first is pre-data base automated daily data checking. I will discuss this process and the statistical issues involved, as well as the practical advantages for the breeders. Secondly, R-functions are created to facilitate fit- ting of appropriate models to the data collected in the database. Furthermore, the package displays analysis results in a manner which enhances selection de- cisions. I will demonstrate some graphical applications for this.

Estimating Sea Level Change

Terry Koen

NSW Office of Environment and Heritage

Terry.Koen@environment.nsw.gov.au

Trend analysis is ideally a simple affair, especially when one desires to report a single summary statistic such as a linear rate of change. However, long term data series are unlikely to be purely linear in nature, so flexible regression techniques such as Generalised Additive Models may be worth considering. Also, covariates may need to be added to the model, and the assumption of independent errors tested. An example is presented where monthly Sea Level time series data from 13 NSW tidal stations are modelled for time trends. The approach needs to be capable of accommodating inter-annual sinusoidal seasonality, cli- mate covariates (eg Southern Oscillation Index), a flexible trend component, modelled as a spline, and autocorrelated error terms.

An efficient stratified sampling strategy for assessing crown rot in wheat

Bethany Macdonald

Queensland Department of Agriculture and Fisheries

Bethany.Macdonald@daf.qld.gov.au

Coauthors: Alison Kelly

Crown rot is a fungal pathogen occurring naturally in the soil that infects cereal crops grown in Australia. Assessing the severity of crown rot in wheat is a central component of the National Crown Rot Initiative Project, funded by the Grains Research and Development Corporation. This project involves running replicated field trials in which genotypes are grown in inoculated plots, with the aim of comparing the performance of genotypes in the presence of crown rot. In assessing performance, it is important to gain an accurate estimate of the severity of the crown rot infection. This has typically been done by sampling ten plants from every plot and then rating a number of tillers on each plant for crown rot severity. Multiple tillers are sampled to produce an accurate measure of genotype infection. Sampling is a very time consuming, hence expensive, process with the number of tillers rated for each plant ranging from between 5 and 63.

The first step to improve sampling efficiency was to gain a comprehensive understanding of the sources of variability across the strata of plots, plants and tillers. A number of data sets were analysed using a simple variance component model. The majority of the variability was located at the tiller level and the plant level. In addition, tillers can be classified into three types; primary, secondary, and tertiary. Utilising this information, a simulation study was carried out to determine whether the sampling strategy in crown rot trials could be performed in a more resource efficient way.

Three data sets were simulated in which ratings were recorded for tillers on plants within plots using estimates of variance components from inoculated crown rot trials. The variance components used to generate the three data sets were chosen to represent different scenarios for the proportion of variance associated with the plant and tiller strata. Each of the three data sets contained one experiment with three replicates of 25 genotypes. Ratings were simulated for five tillers (primary, 2 x

secondary, 2 x tertiary) for twenty plants within each plot. Each data set was then sub-sampled using a number of sampling scenarios, and the sampled data set was analysed. Comparisons between the estimated genetic effects for each of these sampled data sets and the "true" genetic effects used to simulate the data were undertaken for 500 simulations.

From the findings of this simulation study, recommendations can be made regarding stratified sampling to reduce the sampling effort for the assessment of crown rot severity, while maintaining accuracy of genotype comparisons.

Survival analysis with multiple causes of death: Reconsidering the competing risks model

Margarita Moreno-Betancur

Murdoch Childrens Research Institute/Monash University/Inserm Epidemiology Centre on Medical Causes of Death

margarita.moreno@mcri.edu.au

Coauthors: Hamza Sadaoui, Clara Piffaretti, and Gregoire Rey

Cause-specific mortality statistics are usually based on the so-called 'underlying cause of death', which is selected from the diseases declared on the standardized death certificate according to international rules. However, the assumption that each death is caused by exactly one disease is debatable, particularly with an aging population in an era where infectious diseases are replaced by chronic and degenerative diseases. The need to consider multiple causes of death has been acknowledged in epidemiologic research, with a growing body of literature producing statistics based on any mention of a disease on the death certificate. Yet, to date there has not been a formal framework proposed for the statistical modelling of death arising from multiple causes. We propose a model for multiple-cause mortality grounded on an empirical approach that assigns weights to each cause on the death certificate. This new modelling framework extends the standard survival analysis model for all-cause death and is an alternative to the classic competing risks model, which assumes that each death has exactly one cause. We identify two goals of high relevance for public health and epidemiology research, describe how the new framework can be used to address these goals, and discuss the limitations of the classic and 'any mention' approaches in this regard. We describe estimation in the new framework, particularly Cox regression. A simulation study and an application to a study of socioeconomic inequalities in mortality are used for illustration.

Analysis of community data - how well do different statistical frameworks predict species occurrences and community patterns?

Anna Norberg

Department of Biosciences, University of Helsinki, Finland

anna.norberg@helsinki.fi

A large array of statistical frameworks has been developed for analyzing com- munity ecological data sets describing the occurrences or abundances of species in a number of sampling units, and the environmental attributes of those sampling units. We compare various statistical frameworks (species distribution models and joint-species distribution models) in terms of their predictive performance. We parameterize the statistical frameworks with training data, and then use them to predict independent validation data representing both similar and different environmental conditions under which the training data were acquired. We consider both real and simulated data. The simulated data are generated by dynamical models mimicking communities with either competitive, mutualistic or trophic interactions. The real data sets involve diatoms, fungi, trees and butterflies, and they are contrasting in terms of the numbers of species, the fraction of rare species, the number of sampling units, and the spatial context in which the sampling units are acquired. We compare the model predictions to the validation data both at the species and community levels, the latter including measures of species richness, community dissimilarity, and co-occurrence.

Delegate List

Nerea Abrego	nerea.abrego@ntnu.no
Rosemary Bailey	rab@mcs.st-andrews.ac.uk
David Baird	david@vsn.nz
David Balding	david.balding@unimelb.edu.au
Kaye Basford	k.e.basford@uq.edu.au
Ken Beath	ken.beath@mq.edu.au
Anne Bernard	a.bernard@qfab.org
Martin Bland	martin.bland@york.ac.uk
Gabriela Borgognone	gabriela.borgognone@daf.qld.gov.au
Adrian Bowman	adrian.bowman@glasgow.ac.uk
Chris Brien	chris.brien@unisa.edu.au
Sam Brilleman	sam.brilleman@monash.edu
Wesley Brooks	wesley.brooks@unsw.edu.au
Conrad Burden	conrad.burden@anu.edu.au
Marie-Jeanne Buscot	m.buscot@utas.edu.au
Stephen Bush	stephen.bush@uts.edu.au
Kym Bulter	kym.butler@unimelb.edu.au
Ruth Butler	ruth.butler@plantandfood.co.nz
Catherine Cameron	c.cameron@agresearch.co.nz
Steve Candy	burwood70@gmail.com
John Carlin	john.carlin@mcri.edu.au
Vanessa Cave	vanessacave@gmail.com
Brenton Clarke	b.clarke@murdoch.edu.au
Brian Cullis	bcullis@uow.edu.au
Jisheng Cui	jisheng.cui@dhhs.vic.gov.au
Mario D'Antuono	mario.dantuono@agric.wa.gov.au
Ross Darnell	ross.darnell@csiro.au
Pauline Ding	pauline.ding@anu.edu.au
Melissa Dobbie	melissa.dobbie@csiro.au
James Doecke	james.doecke@csiro.au
Richard Emsley	richard.emsley@manchester.ac.uk
Paige Eveson	Paige.eveson@csiro.au
Daniel Fernandez	daniel.fernandez@msor.vuw.ac.nz
Sampath Fernando	sampathf73@gmail.com
Clayton Forknall	clayton.forknall@daf.qld.gov.au
Scott Foster	scott.foster@csiro.au
Siva Ganesh	siva.ganesh@agresearch.co.nz
Beverley Gogel	beverley.gogel@adelaide.edu.au

Lindy Guo	lindy.guo@plantandfood.co.nz
Brent Henderson	brent.henderson@csiro.au
Graham Hepworth	hepworth@unimelb.edu.au
Rob Herbert	r.herbert@neura.edu.au
Nicole Hill	nicole.hill@utas.edu.au
John Hinde	john.hinde@nuigalway.ie
Hans Hockey	hans@biometricsmatters.com
Alan Huang	alan.huang@uq.edu.au
Malcolm Hudson	malcolm.hudson@mq.edu.au
Francis Hui	fhui28@gmail.com
Peter Jaksons	peter.jaksons@plantandfood.co.nz
Jan Jansen	jan.jansen@utas.edu.au
Peter Johnstone	peter.johnstone@agresearch.co.nz
Hugh Jones	hugh.jones@environment.nsw.gov.au
Jessica Kasza	jessica.kasza@monash.edu
Barbara Kachigunda	bkachigunda@gmail.com
Alison Kelly	alison.kelly@daf.qld.gov.au
Natalie Kelly	natalie.kelly@csiro.au
Terry Koen	terry.koen@environment.nsw.gov.au
John Koolaard	john.koolaard@agresearch.co.nz
Emma Lawrence	emma.lawrence@csiro.au
Kim-Anh Lê Cao	k.lecao@uq.edu.au
Alan Lee	lee@stat.auckland.ac.nz
Katherine Lee	katherine.lee@mcri.edu.au
Catherine Lloyd-West	catherine.lloyd-west@agresearch.co.nz
Sama Low-Choy	s.low-choy@griffith.edu.au
Thomas Lumley	t.lumley@auckland.ac.nz
Dongwen Luo	dongwen.luo@agresearch.co.nz
Bethany Macdonald	bethany.macdonald@daf.qld.gov.au
Ky Mathews	ky.mathews@gmail.com
Patrick McElduff	patrick.mcelduff@newcastle.edu.au
John McKinlay	John.mckinlay@aad.gov.au
Louise McMillan	louise.mcmillan@auckland.ac.nz
Russell Millar	r.millar@auckland.ac.nz
Margarita Moreno-Betancur	margarita.moreno@mcri.edu.au
Samuel Mueller	samuel.mueller@sydney.edu.au
Warren Muller	warren.muller@csiro.au
Terry Neeman	teresa.neeman@anu.edu.au
Cattram Nguyen	cattram.nguyen@mcri.edu.au
Sharon Nielsen	snielsen@csu.edu.au

Anna Norberg	anna.norberg@helsinki.fi
Jake Olivier	j.olivier@unsw.edu.au
Liliana Orellana	l.orellana@deakin.edu.au
Otso Ovaskainen	otso.ovaskainen@helsinki.fi
Toby Patterson	toby.patterson@csiro.au
Debra Partington	debra.partington@ecodev.vic.gov.au
Samantha Peel	samantha.peel@utas.edu.au
Ramethaa Pirathiban	r.jeyapalan@qut.edu.au
Shirley Pledger	shirley.pledger@vuw.ac.nz
Gordana Popovic	g.popovic@unsw.edu.au
Joanne Potts	joanne@theanalyticaledge.com
Kasun Rathnayake	kasun.rathnayake@mq.edu.au
Ian Renner	ian.renner@newcastle.edu.au
Maheswaran Rohan	mrohan@gmail.com
Florian Rohart	f.rohart@uq.edu.au
Kathy Ruggiero	k.ruggiero@auckland.ac.nz
Ken Russell	krus9437@bigpond.net.au
Tibor Schuster	tibor.schuster@mcri.edu.au
Alastair Scott	a.scott@auckland.ac.nz
Daisy Shepherd	dshe078@aucklanduni.ac.nz
Julie Simpson	julieas@unimelb.edu.au
Alison Smith	alismith@uow.edu.au
Chanatda Somchit	csom017@auckland.ac.nz
Miriana Sporcic	Miriana.Sporcic@csiro.au
Jakub Stoklosa	j.stoklosa@unsw.edu.au
Emi Tanaka	etanaka@uow.edu.au
Garth Tarr	garth.tarr@newcastle.edu.au
Loic Thibaut	loic.thibaut@my.jcu.edu.au
Daniel Tolhurst	tolhurst@uow.edu.au
Chris Triggs	triggs@stat.auckland.ac.nz
Berwin Turlach	berwin.turlach@gmail.com
Martin Upsdell	martin.upsdell@agresearch.co.nz
Andrew Van Burgel	andrew.vanburgel@agric.wa.gov.au
Chikako Van Koten	chikako.vankoten@agresearch.co.nz
Jay Ver Hoef	jay.verhoef@noaa.gov
You-Gan Wang	you-gan.wang@qut.edu.au
David Warton	david.warton@unsw.edu.au
Sean Watson	sean.watson@daf.qld.gov.au
Alan Welsh	alan.welsh@anu.edu.au
Rory Wolfe	rory.wolfe@monash.edu

Jeff Wood	jeff.wood@anu.edu.au
Wen-Hsi Yang	wen-hsi.yang@csiro.au
Jin Yoon	hwan-jin.yoon@anu.edu.au
Chong You	chongy@uow.edu.au
Sophie Zaloumis	sophiez@unimelb.edu.au
Nanxi Zhang	n.zhang2@uq.edu.au
Wei Zhang	wzha217@aucklanduni.ac.nz
Alec Zwart	alec.zwart@csiro.au

Index to Abstracts