[ABSTRACTS draft only 12 Nov 2013]

Biometrics by the Canals

The Sebel Mandurah Hotel
Mandurah, Western Australia
December 1–5, 2013

Conference of the Australasian Region of the International Biometric Society

I B S

http://www.BiometricSociety.org.au

Contents

Introduction and acknowledgements

This document provides the abstracts from the invited speakers and the speakers presenting a contributed talk or a poster. The content is sorted in order of Family-name but the table of contents shows ”First-name and then Family-name”.

The photograph of the Mandurah canals is kindly provided by Lisa Gardiner of the Mandurah Visitor Centre, and Dee Ann Walker kindly from the International Business Office of the International Biometric Society (IBS) kindly provided the IBS logo.

I would like to thank the scientific programme committee comprised of David Baird, Renate Meyer, Brenton Clarke and Nick de Klerk for their valuable time and input over the structure of the scientific programme and reviewing the submitted abstracts. I would like to thank Elliott Pearl, the moderator at the Topology Atlas abstract processing service http://at.yorku.ca/topology/ for allowing us to use the system and for helping me with all of my queries.

Finally, a special thanks to the sponsors and kind supporters : VSNi, SAS, NIASRA (University of Wollongong), Technical University of Dortmund, Murdoch University, University of Western Australia, and the Department of Agriculture and Food, Western Australia for their valuable contributions to making the conference a success. Please note that is a draft only and a final version is due for release soon after November 18, 2013.

Mario D'Antuono
Chair, Scientific Programme and Local Organising Committees
Perth, Western Australia
12 November, 2013

Invited Speakers

Option x Context interaction and the design of multi-environment trials

Richard Coe
World Agroforestry Centre, Nairobi, Kenya and University of Reading, UK
`r.coe@cgiar.org`

Smallholder farming systems of developing countries have been characterised by their high heterogeneity, being variable both ecologically and socially.

Research aims to identify options for these farmers which range from components (e.g. crop varieties) to resource management strategies and institutional changes. When looking at performance of these options, there tends to be a strong interaction with the variable contexts in which they are used. What are the implications for research approaches and methods?

Breeders have long recognised GxE interaction and developed research designs to investigate it. Concepts and principles pulled from that work can be exploited to improve the design of studies to investigate more general option x context interaction.

These designs are often implemented in situations in which the researcher has only partial control such as farmer research networks or research embedded in development activities which can limit the scope of research but can also be exploited if the principles are understood.

Spatio-temporal smoothing of CO_2 retrievals

Noel Cressie
National Institute for Applied Statistics Research Australia (NIASRA), University of Wollongong
ncressie@uow.edu.au

Processing massive amounts of spatio-temporal data to provide estimates of the current (hidden) spatial state is challenging, even for the Kalman filter. A large number of spatial locations observed through time quickly leads to an overwhelmingly high-dimensional statistical model.

We demonstrate how a spatio-temporal random effects (STRE) component can reduce the model to one of fixed dimension, resulting in a method we call Fixed Rank Filtering (FRF). To show FRF's capability for fast computation, the unknown parameters of the model are handled through EM-estimation, although a (slower) Bayesian version of FRF has also been developed.

A global remote sensing dataset of mid-tropospheric CO_2 from the AIRS (http://airs.jpl.nasa.gov/) instrument on the Aqua satellite is analyzed, which results in optimally interpolated, optimally filtered predictions of CO_2 values along with their prediction standard errors. This talk presents joint research with Matthias Katzfuss, Texas A&M University.

Host-viral interactions: Some statistical immunogenetic issues

Ian James
Institute for Immunology & Infectious Diseases, Murdoch University
i.james@murdoch.edu.au

The ability of rapidly replicating viruses such as HIV and Hepatitis C to mutate and adapt to evade human immune defences in part underpins their large pandemics. Identification of associations of viral adaptation (or escape mutations) with host genetic characteristics, particularly the host human leukocyte antigen (HLA) alleles, is important for intrinsic biological understanding of the diseases and may have implications for drug resistance studies and vaccine design.

We briefly review a number of statistical issues which arise in analyses of these associations at the population level, considering in particular some exploratory methods to assess where HLA-driven immune pressure may impact viral escape mutations and to quantify via a Kullback-Leibler measure the degree of viral diversity which might be attributable to carriage of different host HLA alleles.

The issues and methods are illustrated using host-viral data from the Western Australian HIV Cohort Study.

Prediction of growth processes

Christine Müller
Technische Universität Dortmund, Germany
cmueller@statistik.tu-dortmund.de

The prediction of a growth process is an important topic in biology, economics and the engineering sciences. We consider two motivating examples on the growth of a kidney dialysis factor, and of a crack in construction (concrete/steel). In this context, prediction intervals are of interest to express the precision of the prediction.

We discuss several approaches for prediction intervals for growth models. Some of these approaches use only the observations of the growth process for which the prediction should be done. Especially, approaches of nonlinear models and stochastic differential equation (SDE) models are considered and a new approach based on data depth is presented.

Additionally, approaches are studied which use also the information of other growth processes. Here a special mixed linear model is considered.

Model selection in linear mixed models

Alan Welsh
Centre for Mathematics and its Applications, Australian National University
alan.welsh@anu.edu.au
Coauthors: Janice Scealy; Samuel Müller (School of Mathematics and Statistics, University of Sydney)

Linear mixed effects models are able to handle a broad range of data types and are therefore widely used in applications. A key part in the analysis of data is model selection, which often aims to choose a parsimonious model with other desirable properties from a possibly very large set of candidate statistical models.

Over the last 5-10 years the literature on model selection in linear mixed models has grown extremely rapidly. The problem is more complicated than in linear regression because selection on the covariance structure is not straightforward due to computational issues and boundary problems arising from positive semidefinite constraints on covariance matrices.

We review four major approaches to selecting mixed models, namely information criteria such as AIC or BIC, shrinkage methods based on penalized loss functions such as LASSO, the Fence procedure and Bayesian techniques.

Generalized linear models: Some thoughts and work 40 years on

Thomas Yee
Department of Statistics, University of Auckland
t.yee@auckland.ac.nz

The year 1972 saw the publication of two regression models that would change the subject of statistics enormously. One was the class of generalized linear models (GLMs), which was due to Nelder and Wedderburn and appeared in Series A of the Journal of the Royal Statistical Society. Since then GLMs have been developed and extended in almost every conceivable direction.

This talk will look at some of the ideas behind GLMs that now play a central role in applied regression analysis. These include the exponential family, iteratively reweighted least squares, link functions and linear predictors, diagnostics, and flexible software (GLIM). Some GLM-based work by the speaker will be described, as well as some thoughts towards the future.

Biometrics/JABES Showcase

Model-averaged profile likelihood intervals

David Fletcher
University of Otago
dfletcher@maths.otago.ac.nz
Coauthors: Daniel Turek

Model-averaging is commonly used as a means of allowing for model uncertainty in parameter estimation. In the frequentist framework, a model-averaged estimate of a parameter is the weighted mean of the estimates from each of the candidate models, the weights typically being chosen using an information criterion. Current methods for calculating a model-averaged confidence interval assume approximate normality of the model-averaged estimate, i.e., they are Wald intervals. As in the single-model setting, we might improve the coverage performance of this interval by a one-to-one transformation of the parameter, obtaining a Wald interval, and then back-transforming the endpoints.

However, a transformation that works in the single-model setting may not when model-averaging, due to the weighting and the need to estimate the weights. In the single-model setting, a natural alternative is to use a profile likelihood interval, which generally provides better coverage than a Wald interval.

We propose a method for model-averaging a set of single-model profile likelihood intervals, making use of the link between profile likelihood intervals and Bayesian credible intervals. We illustrate its use in an example involving negative binomial regression, and perform two simulation studies to compare its coverage properties with the existing Wald intervals.

Predicting additive and non-additive genetic effects from trials where traits are affected by inter-plot competition

Colleen Hunt
Department of Agriculture, Fisheries and Forestry, Queensland
Colleen.Hunt@daff.qld.gov.au
Coauthors: David Jordan; Alison Smith and Brian Cullis (University of Wollongong)

There are two key types of selection in a plant breeding program, namely selection of hybrids for potential commercial use and the selection of parents for use in future breeding.

Oakey *et al.* (*TAG*, 2006) showed how both of these aims could be achieved using pedigree information in a mixed model analysis in order to partition genetic effects into additive and non-additive effects. Their approach was developed for field trial data subject to spatial variation.

I will extend the approach for data from trials subject to inter-plot competition. I will show how the approach may be used to obtain predictions of pure stand additive and non-additive effects.

The methodology is developed in the context of a single field trial using an example from an Australian sorghum breeding program.

A framework for the joint modeling of longitudinal diagnostic outcome data and latent infection status: Application to investigating the temporal relationship between infection and disease

Geoff Jones
Massey University
g.jones@massey.ac.nz
Coauthors: Wesley Johnson (UC Irvine), Daan Vink and Nigel French

For many diseases the infection status of individuals cannot be observed directly, but can only be inferred from biomarkers that are subject to measurement error. Diagnosis of *infection* based on observed symptoms can itself be regarded as an imperfect test of infection status. The temporal relationship between infection and marker outcomes may be complex, especially for recurrent diseases where individuals can experience multiple bouts of infection.

Given repeated measures of a biomarker for infection and apparent disease status of a number of individuals at multiple time points, together with relevant covariates, we propose and estimate a model in which the unobserved infection status is a correlated latent process.

This model can be used to investigate the temporal dynamics of infection, and to evaluate the usefulness of the biomarker for monitoring purposes. Our work is motivated and illustrated by a longitudinal study of Bovine Digital Dermatitis on commercial dairy farms in Cheshire, United Kingdom.

Equivalence of MAXENT and Poisson point process models for species distribution modeling in ecology

Ian Renner
School of Mathematics and Statistics and Evolution & Ecology Research Centre, The University of New South Wales
Ian.Renner@unsw.edu.au
Coauthors: David Warton

Modeling the spatial distribution of a species is a fundamental problem in ecology. A number of modeling methods have been developed, an extremely popular one being MAXENT, a maximum entropy modeling approach.

In this talk, I will show that MAXENT is equivalent to a Poisson regression model and hence is related to a Poisson point process model, differing only in the intercept term, which is scale-dependent in MAXENT.

I will illustrate a number of improvements to MAXENT that follow from these relations. In particular, a point process model approach facilitates methods for choosing the appropriate spatial resolution, assessing model adequacy, and choosing the LASSO penalty parameter, all currently unavailable to MAXENT. Functions for fitting point process models with LASSO penalties are available in the R package ppmlasso.

Contributed Talks

The coverage probability of confidence intervals in one-way analysis of covariance after two F tests

Waruni Abeysekera
Department of Mathematics and Statistics, La Trobe University
W.Abeysekera@latrobe.edu.au
Coauthors: Paul Kabaila and Oguzhan Yilmaz

Volume 3 of *Analysis of Messy Data* by Milliken & Johnson (2002) provides detailed recommendations about sequential model development for the analysis of covariance. In his review of this volume, Koehler (*JASA*, 2002) asks whether users should be concerned about the effect of this sequential model development on the coverage probabilities of confidence intervals for comparing treatments. We answer Koehler's question in the context of the two-stage model selection procedure that uses two F tests and is proposed in Chapter 2 of this volume. There, we consider a one-way analysis of covariance model and the parameter of interest θ is a specified linear contrast of the expected responses, for a given value of the covariate. We present a general methodology for the examination of the effect of the two-stage model selection procedure based on two F tests, on the coverage probability of a subsequently-constructed confidence interval for θ, with nominal coverage $1 - \alpha$. We apply this methodology to an illustrative example from this volume and show that these coverage probabilities are typically very far below nominal. This lead to our conclusion that users should be very concerned about the coverage probabilities of confidence intervals for comparing treatments constructed after this two-stage model selection procedure.

Key words: Confidence interval; Coverage probability; F test; One-way analysis of covariance; Preliminary hypothesis test

Statistical confidence measures for genetic mapping

Daniel Ahfock
Department of Mathematics, University of Queensland
daniel.ahfock@uqconnect.edu.au
Coauthors: Ian Wood; Emma Huang (CSIRO Computational Informatics, Brisbane) and Colin Cavanagh (CSIRO Plant Industry, Canberra)

Genetic studies are a key element of modern crop breeding, facilitating more rapid and economical plant improvement than traditional approaches. Through these studies, associations detected between genetic markers and traits such as quality and yield can be incorporated into the breeding process. Constructing a genetic map is the starting point in finding marker-trait associations, and as such, errors in this stage can complicate further analysis and lead to spurious results. Many map construction approaches have been developed which focus on estimating the order and position of a set of genetic markers. However, little work has been done on quantifying map uncertainty, and map verification

is often an ill-defined labour-intensive task. Statistical confidence measures for genetic maps are typically obtained using non-parametric bootstrap or Bayesian methods. Both of these approaches are computationally expensive, and do not provide explicit mathematical results for uncertainty estimates.

We develop a computationally efficient resampling approach based on the derivation of the asymptotic joint distribution of maximum likelihood estimators for inter-marker distances. This approach allows us to assess the quality of an estimated map and gives closed form expressions for estimates of order instability. We investigate the effects of marker density, marker informativeness and population size on map confidence in inbred pedigrees. We compare our theoretical predictions to results from simulation studies, and discuss the application of the techniques to a large mapping population in wheat.

A fully Bayesian nonparametric alternative to multiple imputation for randomly missing covariates in generalized regression models

Murray Aitkin
Department of Mathematics and Statistics, University of Melbourne
murray.aitkin@ms.unimelb.edu.au
Coauthors: Irit Aitkin

This talk describes an extension of multiple imputation which provides a fully Bayesian analysis without requiring a parametric model for the covariates with missing values. The extension is illustrated with simple normal regression models; the computational approach can be extended to GLMs and GLMMs.

Do baseline P-values follow a Uniform distribution in randomised trials?

Martin Bland
University of York, United Kingdom
martin.bland@york.ac.uk

The theory has been put forward that if a null hypothesis is true, P-values should follow a Uniform distribution. This can be used to check the validity of randomisation.

The theory was tested by simulation for two sample t tests for data from a Normal distribution and a Lognormal distribution, for two sample t tests which are not independent, and for chi-squared and Fisher's exact test using small and using large samples.

For the two sample t test with Normal data the distribution of P-values was very close to the Uniform. When using Lognormal data this was no longer true, and the distribution had a pronounced mode. For correlated tests, even using data from a Normal distribution, the distribution of P-values varied from simulation run to simulation run, but did not look close to Uniform in any realisation. For binary data in a small sample, only a few probabilities were possible and distribution was very uneven. With a sample of two groups of 1,000 observations, there was great unevenness in the histogram and a poor fit to the Uniform.

The notion that P-values for comparisons of groups using baseline data in randomised clinical trials should follow a Uniform distribution if the randomisation is valid has been found to be true only in the context of independent variables which follow a Normal distribution, not for Lognormal data, correlated variables, or binary data using either chi-squared or Fisher's exact tests. We should not use this as a check for valid randomisation.

Time series analysis of daily maximum and minimum temperatures

Ross Bowden
Mathematics and Statistics, School of Engineering and Information Technology, Murdoch University
ross.bowden@iinet.net.au
Coauthors: Brenton Clarke

This talk will analyse over 60 years of daily maximum and minimum temperatures for Perth, Western Australia. A technique for simultaneous time series estimation called interleaving will be presented. It will be used to explore the effect on mean temperatures of climate change and of movements in recording sites. The lagged relationships between daily maximum and minimum temperatures will also be examined.

A tale of randomization

Chris Brien
School of Information Technology and Mathematical Sciences, University of South Australia
chris.brien@unisa.edu.au

At one time I was a disciple of randomization analysis. Where am I now? Randomization analysis for experiments whose designs employ a single randomization is reviewed. For the case when treatment effects are partially confounded, a combined test statistic is proposed. Its performance, in comparison to a mixed-model analysis, is investigated using data from (i) two experiments laid out using balanced incomplete block designs, (ii) a split-plot experiment, and (iii) another that employed a latinized row-column design.

Spreading your effort: a balanced sample design for n-dimensions

Jennifer Brown
Department of Mathematics and Statistics, University of Canterbury
jennifer.brown@canterbury.ac.nz
Coauthors: Blair Robertson, Chris Price, Marco Reale and Peter Jaksons; Trent McDonald (Western Ecosytems Technology, Inc., Wyoming, USA)

Spatially balanced sampling is gaining popularity for biological and environmental studies. The most common design is called GRTS (Generalized Random Tessellation Stratified sampling) where sample effort is spread evenly over the target region. The term spread evenly in this context means having fairly consistent survey effort over the region, without having to revert to fixed interval, regularly spaced, systematic sampling.

We have extended the idea of GRTS to higher dimensions other than spreading effort over a (2D) map. The motivation for the n-dimension design came from wanting to achieve not only balance in geographic space but balance and randomness in the time interval between repeat surveys for long term monitoring. We have achieved a design that can spread effort in any dimension, for example in geographic space with longitude and latitude on a map, over time with repeated surveys, in environmental space with linear combinations of environmental features, and in space related to population dynamics and species conservation. Selection of sites in the sample can be with equal probability sampling, or if it is desirable to target effort to special parts of the sample space, unequal probability sampling.

qPCR data: some statistical problems and possible solutions

Ruth Butler
Plant & Food Research, Lincoln
ruth.butler@plantandfood.co.nz

Quantitative Polymerase Chain Reaction (qPCR) analyses have become an essential tool for a wide range of science areas. PCR is a cyclical process whereby a single copy of a target molecular sequence (gene) is replicated at each cycle. qPCR is an extension of this, which allows an estimate of the original number (or quantity) of molecules of a target sequence to be made. Relative quantitation is where estimates of a target sequence and a reference sequence are made, and the ratio of the target to the reference quantity is of interest.

In most qPCR instruments, up to 96 samples are run on an individual plate, usually with a single target gene assessed for all samples on the plate. Results are analysed on the logarithmic scale. Two statistical problems that arise with standard qPCR and relative quantitation are (i) how to deal with undetermined ($\tilde{0}$) data, and (ii) how to adjust for random effects generated from the original sample generation experiments and (where there are more samples than can be assessed in a single run), from the qPCR plates. This is especially a problem for relative quantitation. The (statistical) design of the qPCR phase is also important to consider.

This talk explores the statistical problems associated with qPCR data, particularly for relative quantitation, and in the context of designed or semi-designed experiments. A Poisson-Gamma hierarchical generalized linear model approach for the analysis of such data will be described.

Modelling interval-censored survival times in toxicological studies using generalized additive models

Steve Candy
Wildlife Conservation and Fisheries, Australian Antarctic Division, Kingston (Hobart)
steve.candy@aad.gov.au
Coauthors: Bianca Sfiligoj, Catherine King and Julie Mondon

A method for combining a proportional-hazards survival time model with a bioassay model where the log-hazard function is modelled as a linear function of log-concentration combined with a cubic smoothing spline function of time is described.

The combined model is fitted to mortality numbers resulting from interval-censored survival times using a generalized additive model (GAM) with the assumption that mortalities are conditional binomials combined with an approximation to the log of the integral of the hazard function using freely-available, general software for fitting GAM.

Extensions of the GAM to allow random effects to be fitted and to allow for time-varying concentrations by replacing time with a calibrated cumulative exposure (CCE) variable with the calibration parameter estimated using profile likelihood are described. Random effects defined by multi-level sampling units such as replicate bioassay or defined by lack-of-fit within bioassay via the interaction of concentration levels with time intervals can be incorporated. Numerical methods are used to obtain lethal concentration values under fixed concentrations over time.

The models are demonstrated using data from a study of a marine and, a previously published paper on freshwater taxa. The marine study involved two replicate bioassays of the effect of zinc exposure on survival of an Antarctic amphipod, *Orchomenella pinguides.* The other example modelled survival of the daphnid, *Daphnia magna*, exposed to potassium dichromate and was fitted by both the GAM and the process-based DEBtox model. The GAM gave a 30% improvement in fit to the daphnid data compared to DEBtox as measured by the deviance.

Keywords: dose-response model, interval-censored survival times, generalized additive model

Anaesthesia clinicians estimate blood pressure by feeling the radial pulse: a randomised control trial

Brenton Clarke
Mathematics and Statistics, School of Engineering and Information Technology, Murdoch University
B.Clarke@murdoch.edu.au
Coauthors: Betty Mouchel (Statistics and Business Intelligence, University Paul Sabatier Toulouse III, France) and David Simes (Department of Critical Care Medicine, Fremantle Hospital)

We report here the statistical analysis of a trial that was carried out by a single investigator, David Simes, as a part of a Registrar project designed in 1994, with express written approval of the Director of Clinical Services. Referral to the hospitals Ethics Committee was waived as the trial did not involve any action outside of normal medical practice. The end point was to determine the accuracy of anaesthesia clinicians in estimating an anaesthetised patients systolic blood pressure (SBP) by feeling the radial pulse.

The volunteer medical participants were sequentially randomised to one of four groups: one group given no help (control), the second allowed to feel the pulse, the third given pre- and peri-operative clinical information, the fourth given both. Allowing for possible attrition approximately 60 observations were recorded in each group.

This expose will discover the trials pluses and minuses in carrying out classical and robust analysis of variance, and illustrates the usefulness of carrying out a variant of the Bland-Altman plot. In the end the assessment of clinical relevance proceeded with repeated use of chi-squared tests on $2x2$ contingency tables.

Key Words: One way analysis of variance, Bland-Altman plot, robustness, Levenes Test, randomised control trial.

Semi-parametric risk prediction models for recurrent cardiovascular events in the LIPID study

Jisheng Cui
Victorian Department of Health, Melbourne
jisheng.cui@health.vic.gov.au

Traditional methods for analyzing clinical and epidemiological cohort study data have been focused on the first occurrence of a health outcome. However, in many situations, recurrent event data are frequently observed. It is inefficient to use methods for the analysis of first events to analyse recurrent event data.

We applied several semi-parametric proportional hazards models to analyze the risk of recurrent myocardial infarction (MI) events based on data from a very large randomized placebo-controlled trial of cholesterol-lowering drug. The backward selection procedure was used to select the significant risk factors in a

model. The best fitting model was selected using the log-likelihood ratio test, Akaike Information and Bayesian Information Criteria.

A total of 8557 persons were included in the LIPID study. Risk factors such as age, smoking status, total cholesterol and high density lipoprotein cholesterol levels, qualifying event for the acute coronary syndrome, revascularization, history of stroke or diabetes, angina grade and treatment with pravastatin were significant for development of both first and subsequent MI events. No significant difference was found for the effects of these risk factors between the first and subsequent MI events. The significant risk factors selected in this study were the same as those selected by the parametric conditional frailty model. Estimates of the relative risks and 95% confidence intervals were also similar between these two methods.

Our study shows the usefulness and convenience of the semi-parametric proportional hazards models for the analysis of recurrent event data, especially in estimation of regression coefficients and cumulative risks.

A hidden Markov model of cattle herd age-class structure.

Ross Darnell
CSIRO Computational Informatics, Brisbane
ross.darnell@csiro.au

The Carbon Farming Initiative is a legislative scheme that provides for credits to be issued in return for the abatement of greenhouse through activities in the land sector. Currently the data needed to estimate necessary key performance indicators such as number and class of cattle on the property, is difficult to obtain because of the extensive nature of cattle production across much of northern Australia. Without these estimates it is not possible to determine GHG emissions for the herd. This talk presents one approach to build a tool for estimating number and class of livestock in a production system where accurate head counts are not feasible. We have developed a hidden Markov model based on proposed monthly transitions of the age-class groups in herds and observations of cattle numbers observed during biannual musters. The aim of the study is to estimate the number of cattle in the herd that are absent from one or more musters.

Analysis of near infared spectra for genetic components in barley

Dean Diepeveen
Department of Agriculture and Food, Western Australia
dean.diepeveen@agric.wa.gov.au
Coauthors: Peter Clarke (retired) and Chengdao Li; Matthew Bellgard and Rudi Appels (Murdoch University)

An analytical approach using NIR spectra (Diepeveen *et al.* 2012, *J. Cereal Science*) that removes environmental and experimental variability to produce genetic fingerprint spectra has been applied to several genetically structured double haploid barley populations over 2 years. The technique requires selecting regions of the spectra that are diagnostic for a particular combination of traits and in turn using these spectral-regions as traits in subsequent spatial mixed models. This analysis associates spectral regions with plant development and plant harvest height traits. This analysis has identified several issues as to the best approach to address inconsistent variability between trial-year environments. Some preliminary work will be presented on better characterizing grain through the use of flour related traits.

Improved estimation of intrinsic growth: integrating matrix models and allometry

Peter Dillingham
School of Science and Technology, University of New England
pdillingham@une.edu.au
Coauthors: Jeffrey Moore (Southwest Fisheries Science Center, National Marine Fisheries Service, National Oceanographic and Atmospheric Administration, USA)

In order to better understand the ability of marine megafauna, such as sharks, seabirds, sea turtles, and marine mammals, to withstand directed or incidental catch in fisheries, better understanding of intrinsic growth (r_{max}) is needed.

Traditionally, estimating r_{max} can be done by modelling population trends over time or through the use of matrix population models, where individual demographic parameters are estimated and growth is projected. Data challenges and/or model assumptions for either approach are substantial, with empirical estimates of r_{max} being limited to a small subset of populations.

An alternative approach based on allometric relationships between body mass, generation time, and r_{max} that has been applied to birds allows estimation of r_{max} with fewer parameters, allowing productivity estimates for more populations. However, this approach relies heavily on estimates of adult survival. For populations where survival estimates are unavailable (notably sharks), none of the current methods are satisfactory.

We develop a method that integrates matrix models and allometry, allowing us to estimate r_{max} (and survival) given fecundity data only. We take advantage

of knowledge about population structure and fecundity from the matrix models, as well as across-taxa relationships between intrinsic growth and generation time from the allometric model. Our method provides estimates for idealised populations, but can also be extended to account for population variation from the ideal.

Finally, we apply the method to estimate r_{max} for number of shark populations, particularly focussing on the white shark (*Carcharodon carcharias*).

Evaluating two-factor experimental results for RNA-Seq data using simulation

Margaret Donald
School of Mathematics & Statistics, University of New South Wales
merricks.merricks@gmail.com
Coauthors: Susan Wilson

In a recent experiment consisting of 15 patients, eight suffering from Myelodysplastic Syndrome (MDS) and the other seven from Chronic Myelomonocytic Leukaemia (CMML), all were treated with the DNA hypomethylating drug, Vidaza (AZA). RNA-Seq data were obtained both before and after treatment. In all, 35868 genes or gene variants were obtained. Details of initial analyses can be found in Donald *et al.* (IWSM, 2013).

We use data cloning as described in Lele (Ecology, 2010) of the experimental data to create simulations of RNA-seq data using the posterior predictive distributions for 30 sets of non-replicated RNA-seq counts. In order to create realistic counts, the posterior distributions of counts are found from a paired data two-factor (with interaction) model fitted in WinBUGS to the experimental data, using the negative binomial distribution and allowing for differing library sizes for each experimental unit.

The results for differentially expressed genes found from the marginal distributions are compared with the DE genes found in the R-packages, DEseq, PoissonSeq, edgeR, and QuasiSeq. We used posterior predictive distributions for the simulations on the basis that they satisfy the assumptions of negative binomial distributions with tagwise dispersions and varying library rates.

We found that QuasiSeq performed well for these data, confirming the findings of Burden *et al.* (2013 submitted). QuasiSeq gave consistent results between cloned and observed data. Surprisingly, given the simulation method, edgeR did not.

An improved exact confidence interval for proportions using group testing with groups of variable Size

Konstantine Dres
Department of Mathematics and Statistics, University of Melbourne
ddres@student.unimelb.edu.au
Coauthors: Graham Hepworth and Ray Watson

Group (or pooled) testing is the pooling of individual samples, and examining them as a group for the presence of an attribute. It is often used in either the identification of a rare disease, or to estimate the prevalence of such a disease.

Situations involving the testing of insect vectors (e.g. mosquitoes) often give rise to many distinct group sizes, rather than a common group size, and this results in a multi-dimensional event space for the number of positive groups.

We propose an improved exact confidence interval procedure for estimation of proportions by group testing, when multiple group sizes are used. Exact two-sided procedures such as Clopper-Pearson and mid-P type intervals based on MLE ordering, and the Sterne interval based on probability ordering have been previously considered, but Blaker type intervals have not.

We introduce the Blaker interval (or Combined Tail method) based on MLE ordering. We illustrate this procedure with some examples, and show that it possesses some very attractive properties when compared to the alternatives. Applications of an improved algorithm for computing the intervals are included.

Likelihood-based finite mixture models for ecological ordinal data

Daniel Fernandez
Victoria University of Wellington
daniel.fernandez@msor.vuw.ac.nz
Coauthors: Shirley Pledger and Richard Arnold

Many of the methods to deal with dimension reduction in matrices of data are based on mathematical techniques such as distance-based algorithms or matrix decomposition and eigenvalues. In general, it is not possible to use statistical inference with these techniques because there is no underlying probability model. Recent research has developed a set of likelihood-based finite mixture models for a data matrix of binary or count data (Pledger and Arnold 2013, to appear).

My current research extends this earlier work by formulating likelihood-based multivariate methods for ecological ordinal data. My talk will introduce the first results from this research, which applies fuzzy clustering via finite mixtures to the ordered stereotype model. Data from areal ecological example will be shown to illustrate the application of this approach.

Finally, I will present the results of a simulation study conducted to determine which information criteria are most appropriate for these particular mixture models when applied to ordinal data.

Latent identity models in capture-recapture

Rachel Fewster
Department of Statistics, University of Auckland
r.fewster@auckland.ac.nz
Coauthors: Peter Jupp (School of Mathematics and Statistics University of St Andrews, Scotland) and Richard Vale (Department of Mathematics, University of Auckland)

Capture-recapture studies to estimate the size of wild animal populations increasingly rely on indirect methods of 'capture' that do not involve handling the animals. Individual identity can be established using photographs of natural markings, DNA samples, or a combination of both.

These approaches greatly widen the scope of capture-recapture, enabling population monitoring to take place for animals that are too elusive, fragile, or frightening to allow manual captures. At the same time, they create a goldmine of interesting statistical problems.

When identity is indirectly observed, it is prone to error, so two samples from the same animal might wrongly be thought to come from different animals. The researcher is left with the problem of estimating the number of animals not seen, when they can't even be sure how many animals they have seen.

We discuss two models that examine different aspects of this problem: one that enables photographic and DNA samples to be combined into the same analysis, and the other a more general model for dealing with identity error. In the first case, the combination of information from different sources necessarily proves stronger than taking either source in isolation. However, in the second case the elegant modelling tactics arguably leave the researcher worse-off as far as information goes, and it appears to be more cost-effective to use labour-intensive methods to correct errors before applying the model, rather than to take care of errors within the model.

Exploring Genotype x Environment in linear mixed models

Arthur Gilmour
NIASRA, University of Wollongong
arthur.gilmour@cargovale.com.au
Coauthors: Ky Mathews and Scott Chapman (CSIRO Plant Industry, Brisbane)

In cereal crops, genetic effects are moderated by the environment, and we would like to select varieties based on sensitivity to the particular environmental characteristics. In this context, Cullis *et al.* (TAG, 1996) reported that 70% of genotype variance was specific to year and location. They looked at how certain genotype characteristics like maturity contributed to the interaction.

The Grains Research and Development Corporation has instigated a project to further explore how particular environmental variables such as climate and soil measures contribute to Genotype x Environment (GxE) interactions in the

Australian Wheat Variety Testing database. We present and discuss the statistical models being used for investigations looking at the effects of 100 variables on 80 varieties in 274 trials conducted in West Australian between 2005 and 2010.

The analysis is of the corrected variety mean yield obtained from trial specific analysis of each trial, weighted according to its associated precision measure. There are mean yields for 42% of the 21920 cells of the full GxE table. The analysis uses a factor analytic 1 variance structure across environments to model the common genotype effect, scaled for each environment. In this base model, the specific variances represent the GxE interaction.

A modified forward selection process is then used to identify environmental variables, which contribute most to this interaction.

Climate variables reflecting water availability dominated. The final stage is to identify the varieties which were interacting most with these variables.

A general Bayesian framework for working with imperfect data

Patrick Graham
Statistics New Zealand
patrick.graham.br@gmail.com
Coauthors: John Bryant

Research in many applied fields, including public health, economics and, increasingly, official statistics, often depends on re-use of data originally collected for purposes other than statistical research.

These data can suffer from defects such as unrecorded confounding covariates, unrecorded or unavailable identifiers which would enable linkage between datasets, under and/or over coverage of the population of interest as well as more conventional types of missingness and misclassification.

These data errors can be accommodated within a generic Bayesian inference framework which produces simulations from the posterior predictive distribution of a corrected dataset. For all but the simplest problems, the Gibbs sampler plays a key role and often requires specification of a model for the observed data given the corrected data. Specification of this model can be straightforward in some problems. Information from validation studies can be incorporated within the Gibbs sampler. The overall framework is an elaboration of Rubin's approach to missing data but also has links to the emerging field of bias modelling in epidemiology and biostatistics.

In this paper we discuss the general Bayesian approach to inference from imperfect data, show how it can accommodate several types of data errors and discuss illustrative examples including problems of under and over coverage of the population of interest.

Getting the right data - a lesson in communication and perseverance

Kathy Haskard
Data Analysis Australia, Nedlands
kathy@daa.com.au
Coauthors: John Henstridge

This talk illustrates the value of persevering to get the ”right” data. Data from a transect-based ecological survey, unequally replicated over time, was initially presented as estimated population numbers per transect, then as counts adjusted using a ”visibility” measure, and later as raw counts. Zero-inflated and/or over-dispersed generalised linear mixed models appeared to be indicated, but this was complicated, difficult, and no model could be found that fitted well. After further probing and discussion with the client, it was discovered that a different measure of ”visibility”, initially believed unavailable, could be provided. With the fourth version of the data, and a slight modification of the form of the model, the analysis simplified immediately, yielding a good outcome for both statistician and client. The moral of the story is that good communication and perseverance in finding the right solution can pay great dividends.

A pollination and fertilisation model for kiwifruit

Harold Henderson
AgResearch Ruakura, Hamilton
harold.henderson@agresearch.co.nz
Coauthors: Mark Goodwin (Plant & Food Research, Ruakura)

Lescourret *et al.* (*Agricultural Systems*, 1998) developed a model describing flower pollination and ovule fertilisation. The model operates at the flower level and computes the number of ovules in a flower developing into mature seeds, and the success of fruit set given this number. Aspects of the model may be considered as occupancy problems.

We explore this model with different cultivars of kiwifruit.

Confidence intervals for proportions estimated by inverse binomial group testing

Graham Hepworth
Statistical Consulting Centre, The University of Melbourne
hepworth@unimelb.edu.au

Group testing (or pooled testing) arises when units are pooled together and tested as a group for the presence of an attribute, such as a disease. It has been applied to fields such as blood testing, plant disease assessment, genetics, fisheries, and transmission of viruses by insect vectors. The purpose of group testing is either to identify the positive units, or, as in our study, to estimate the proportion (p) of positive units in the population. Cost savings can be considerable if p is small.

Most work on group testing has assumed a binomial model, in which the number of groups is fixed. However, there are some situations where it is useful to employ inverse sampling, for which the testing process is stopped when a fixed number of positive groups has been observed. An example of where this could be adopted is the monitoring of a disease outbreak following a natural disaster. When equal group sizes are used, a negative binomial model results.

We propose two interval estimators which improve on existing methods and have excellent coverage properties. One is a score-based method with a correction for skewness, and the other is an exact method with a mid-P correction.

Never fit Sequence - The design and analysis of multi-period clinical trials

Hans Hockey
Biometrics Matters Ltd, Hamilton
hans@biometricsmatters.com

Examples are given of the design and analysis of multi-period studies, typically cross-overs, in early phase studies in the pharmaceutical industry.

The aim is to explore the usefulness of the inclusion of the between-subject sequence term typical in linear models used for analysis and the effect this has on design.

Starting with the simple two-period, two-treatment, two-sequence design, various multi-period studies are introduced and alternative designs and analyses considered. The usefulness for inference of different designs and analyses using models with and without sequence are contrasted.

It is hoped to be able to show that omitting the sequence term from analysis models allows a broader range of designs, more useful models, and can give more information with the same experimental resources.

Doubly-nonparametric generalized linear models

Alan Huang
School of Mathematical Sciences, University of Technology Sydney
alan.huang@uts.edu.au

Generalized linear models (GLM, McCullagh & Nelder, 1989) have become indispensable tools for analyzing agricultural, engineering and biomedical data.

In this talk, we look at some (recent) extensions of GLM, including classical nonparametric GLM in which the mean-curve is nonparametric (e.g. Green & Silverman, 1994), semiparametric GLM in which the error distribution is nonparametric (Huang, JASA 2013), and current work on doubly-nonparametric GLM in which both the mean-curve and error distribution are nonparametric.

Some interesting data analysis examples and graphical tools will also be presented.

Variable selection in multi-species mixture modeling

Francis Hui
School of Mathematics and Statistics, University of New South Wales
fhui28@gmail.com
Coauthors: David Warton; Scott Foster (CSIRO Computational Informatics, Hobart)

Species Archetype Models (SAMs) are a recently developed tool for modeling species assemblages. SAMs use a finite mixture model to cluster species with similar environmental responses to form archetypal response groups, and models these archetypal responses using a GLM (for instance).

In this talk, I consider penalized likelihood methods for variable selection in finite mixture models and SAMs. Since the coefficients in finite mixture models and SAMs exhibit an inherent grouped structure, then I propose using penalties which take this feature into account.

First, I consider a novel adaptation of the group lasso called MIXGL2. Due to its group sparsity property, MIXGL2 can remove a covariate simultaneously from all mixture components (archetypes). I then propose a new penalty called MIXGL1, which has both group and individual coefficient sparsity. This allows MIXGL1 to remove covariates from some but not necessarily all of the mixture components. I show that both MIXGL2 and MIXGL1 possess the oracle property, and are variable selection consistent at a covariate and coefficient level respectively.

Simulations show that MIXGL2 and MIXGL1 outperform other penalties which do not take into account the grouped nature of the coefficients. I consider the application of our methods to a multi-species dataset collected from the Great Barrier Reef off the north-east coast of Australia.

Model selection for the Multimix class of mixture models

Lynette Hunt
University of Waikato, Hamilton
lah@waikato.ac.nz
Coauthors: Kaye Basford (University of Queensland)

The mixture approach to clustering is a model based approach to clustering that requires the specification of the form of the component distributions and the number of groups that are to be fitted to the model. The Multimix class of mixture models (Hunt & Jorgensen, ANZJS 1999) enables the clustering of data that have both categorical and continuous attributes. However with this approach to clustering data, the user also has to decide on the correlation structure that is to be incorporated into the model.

We investigate the performance of some commonly used model selection criteria in the selection of an appropriate model when using the finite mixture model to cluster data containing mixed categorical and continuous attributes. The performance of these criteria in selecting both the form of the correlation structure and the number of components to be used in the model is assessed using simulated data and a medical data set.

We found that the criteria based on the Bayesian information criterion and the integrated classification likelihood perform in a similar manner in selecting both the form of the component distributions and the number of groups to be fitted to the model while the AIC and CLC perform in a less satisfactory way.

Robustifying the geographic relative risk function

Khair Jones
Massey University, New Zealand
khairjones@gmail.com
Coauthors: Martin Hazelton

In the analysis of the geographic distribution of disease, a popular method to estimate risk for data in the form of spatial coordinates of cases and controls, is to use the spatial relative risk function, which is the log ratio of the case and control densities estimated by bivariate kernel smoothing.

One issue with this method is that the risk estimate may be very unstable in areas of sparse population density. This can be seen as analogous to the problem of low cell counts in contingency tables.

Our proposed solution involves adding pseudo data to the control and case densities, with consideration given to how much pseudo data should be added, and where. We analyse the use of pseudo data through simulations, before looking at some real world applications, including the famous Chorley-Ribble cancer dataset.

Assessment and representation of co-occurrence of species

Peter Lane
Fenner School of the Environment & Society, Australian National University
peter.lane@anu.edu.au
Coauthors: David Lindenmayer and Philip Barton

The pattern of occurrence and co-occurrence of species characterizes many aspects of the ecology of a habitat. Ideas such as focal species and indicator species are examples of this, where surrogate measurements of fauna are hoped to indicate the level of biodiversity and ecological status.

We have investigated records of presence and absence of birds in a project which is monitoring types of restoration of woodlands in the South-West Slopes of New South Wales. We developed methods to assess and display the patterns, allowing in particular the comparison of different types of habitat.

One method is the use of logistic regression to measure the potential of individual species to act as indicators of species richness, together with a visual display of the results.

A second method is the use of odds ratios to characterize patterns of co-occurrence, and the construction of network diagrams to visualize the inter-association of species. The same type of diagram can also be used to assess differences between patterns, for example to compare habitats.

[Peter Lane is a senior researcher and statistical consultant at the ANU, working with David Lindenmayer on the use of surrogacy in ecology. He was previously in GlaxoSmithKlines Research Statistics Group in Stevenage, UK, working mostly on statistical issues associated with clinical trials, and before that he worked at Rothamsted in Harpenden, UK, as a statistical consultant in agricultural research, and developer of the GenStat software.]

Permutation tests for fixed effects of linear mixed models

Dongwen Luo
AgResearch Ltd, Grasslands Research Centre, Palmerston North
dongwen.luo@agresearch.co.nz
Coauthors: Siva Ganesh and John Koolaard

Inference regarding the fixed effects in linear mixed models may be questionable in various situations such as when we have insufficient information about the distribution of the data (or are not comfortable making assumptions about the distribution) or if the distribution of the test statistic is not easily identified.

However, we may avoid the problem of non-normal data by finding a suitable transformation or by using a generalized linear (mixed) model with a non-normal error structure specified explicitly. An alternative approach is to use permutation tests, and the approach has become practical due to the availability of fast computational capability.

In this presentation, a restricted permutation test is suggested, which considers the grouped structure of random effects in the model. The t and F statistics produced from linear mixed model are the test statistics for permutation, and multiple comparison procedures can be performed based on permutation results. We will demonstrate the method with examples using associated R functions.

Artificial bee diets

Taryn Major
Data Analysis Australia, Nedlands
taryn@daa.com.au
Coauthors: Linda Eaton, Kathy Haskard and Philip Vlaskovsky; Rob Manning (Department of Agriculture and Food, Western Australia)

Investigating the viability of supplementing honey bee diets with artificial diets, especially during times when pollen is scarce in the natural environment, is of particular interest to bee keepers.

During these periods, bees use up stored resources within the hive and can survive for a period on the protein and lipids stored in their bodies. However the queen stops laying eggs and the colony weakens.

In 2011, an experiment was conducted at the Department of Agriculture and Food to explore the effect of a number of artificial diets on honey bees. We analysed these data, exploring the effect of bee diet on longevity, head weights and the fatty acid, lipid and mineral composition of bee bodies, using various statistical techniques including survival analysis, linear mixed models and analysis of variance.

In this presentation we will discuss the statistical techniques applied and results of the analysis.

Simulation evaluations of spatio-temporal imputations for fishing catch rate standardisation.

Ross Marriott
School of Mathematics and Statistics, University of Western Australia
Ross.Marriott@fish.wa.gov.au
Coauthors: Kevin Murray and Berwin Turlach

Analysis of trends in catch rate data reported by commercial fishers is a key feature of routine assessments of exploited fish populations. Catch rate data are often used to develop an index of historical fish abundance. Generalized linear models are typically used to analyse fishery dependent or environmental effects influencing catch rates and predicted means from these models are used to produce standardised catch rates for assessment.

Missing catch rate data from any area at any time often occur due to the highly mobile nature of commercial fishing and can bias the resulting index of fish abundance. Spatio-temporal imputation is often advocated as an appropriate method to fill the missing cells. However there are many alternative

imputation methods available and it is not clear which may produce the least biased and most precise results.

This work explores and evaluates the suitability of alternative imputation methods for catch rate standardisation. The study focuses on catch rate standardisation for demersal scalefish species targeted by commercial hook and line fisheries operating in the West Coast and Gascoyne Coast Bio-regions of Western Australia.

Results will be presented from simulations, based on species targeted by the West Coast Demersal Scalefish Interim Managed Fishery (WCDSIMF), evaluating the performance (i.e. bias and precision) under a range of alternative hypothetical harvest scenarios.

Orthogonalised multivariate linear mixed models: a significance test for more complex multivariate data

Brian McArdle
Department of Statistics, University of Auckland
b.mcardle@auckland.ac.nz

I introduce a simple but effective approach to the testing of multivariate hypotheses not usually tractable due to complexities in the data (e.g. nested sampling, autocorrelation, or heterogeneity).

I introduce the generic orthogonalised multivariate test (using one way ANOVA) and show that while it is conservative at small numbers of observation per variable, it becomes useable at between 5 to 10 observations per variable (a very moderate requirement).

A bootstrap bias correction to compensate for this conservativeness at smaller sample sizes is described. In power it is sometimes better and sometimes weaker than Pillai's trace (MANOVA), depending largely on the number of variables and the degree of correlation between them.

The test is then applied with simple mixed models (nesting, heterogeneity and autocorrelation) and is shown to perform well, in situations where there are no alternatives to compare with.

Exploring informative diversity in SNP-based association studies

Elizabeth McKinnon
Centre for Clinical Immunology and Biomedical Statistics, Murdoch University
e.mckinnon@murdoch.edu.au
Coauthors: Ian James

Univariate analyses of single nucleotide polymorphisms (SNPs) within the major histocompatibility complex yield strong associations with Type 1 diabetes (T1D) susceptibility that are dominated by markers in the Class II HLA-DR/DQ region but extend across the whole complex. Whilst many of the associations can be attributed to linkage disequilibrium between the SNPs and the dominant HLA-DRB/DQA/DQB loci there is also evidence of additional modifying effects. Establishing independence of effects and quantifying their relative importance remains a difficult problem in studies of multigenic diseases such as T1D.

Here we illustrate a novel approach for providing visual assessment of patterns of SNP diversity, based on clustering of fitted values obtained from conditional logistic regression models. We apply the method to case-control type modelling of sibling data collected by the T1D Genetics Consortium and highlight regions of both independent and co-effects. Model output informs familial clustering.

The approach provides a useful tool for mapping informative diversity across regions of interest. It is easy to apply using standard software and is readily extendable to other applications where many non-independent tests are carried out. Furthermore its application can highlight either independent effects or events that occur together, depending on the underlying structure of the data.

Bayesian analysis of recurrent failure time data using copulas

Renate Meyer
Department of Statistics, University of Auckland
meyer@stat.auckland.ac.nz
Coauthors: Jose Romeo

Recurrent event data, i.e. sequentially observed survival times, occur in many different areas; for instance, times between insurance claims in finance, successive failures of a technical system in engineering, or tumour recurrences in medicine.

Although the individual subjects are assumed to be independent, the times between events of one subject are neither independent nor identically distributed, and dependent censoring is induced.

Apart from the marginal approach by Wei et al. (JASA, 1989), the shared frailty model, see e.g. Duchateau and Janssen (The frailty model, 2008), has been used extensively to analyse recurrent event data, where the correlation between sequential times is implicitly taken into via a random effect. Oakes

(JASA, 1989) showed the equivalence of frailty models for bivariate survival data to Archimedean copulas.

Whereas copula-based models have been used to model clustered survival data, their application to recurrent failure time data has only recently been suggested by Lawless and Yilmaz (Biometrical Journal, 2011) for the bivariate case.

Here we extend this to more than two recurrent events and model the joint distribution of recurrent events explicitly using parametric copulas within a Bayesian framework. We illustrate the flexibility of this approach using data from an asthma prevention trial in young children.

A one-step-ahead pseudo DIC for comparison of Bayesian state-space models

Russell Millar
Department of Statistics, University of Auckland
r.millar@auckland.ac.nz
Coauthors: Sam McKechnie

The deviance information criterion (DIC) has become very popular for comparison of competing Bayesian models. However, it is well known that it must be used with discretion. In the state-space model context, conventional DIC evaluates the ability of the state-space model to predict an observation at time t given the latent state at time t.

Motivated by the failure of conventional DIC to clearly choose between competing Bayesian state-space models for coho salmon population dynamics, this work proposes a one-step-ahead DIC where prediction is conditional on the state at the previous time point.

Simulations revealed that one-step-ahead DIC worked well for choosing between state-space models with different process equations. In contrast, conventional DIC could be grossly misleading, with a strong preference for the wrong model. This can be explained by the failure of conventional DIC to account for increased process error arising from process model mis-specification.

The one-step-ahead DIC is not based on a true likelihood, but it is shown to have interpretation as a pseudo DIC in which the compensatory behaviour of the process errors is eliminated. It can be easily calculated using the DIC monitors within popular BUGS software when the process and observation equations are conjugate.

The improved performance of the one-step-ahead DIC is demonstrated by application to the multi-stage modeling of coho salmon abundance in Lobster Creek, Oregon.

Cluster identification in internal nitrogen use efficiency field data

Isabel Munoz-Santa
Biometry Hub, University of Adelaide
sabela.munozsanta@adelaide.edu.au
Coauthors: Olena Kravchuk; Petra Marschner (Soils, University of Adelaide) and Stephan Haefele (Australian Centre for Plant Funtional Genomics, University of Adelaide)

Internal nitrogen use efficiency (NUE) is the ratio of grain yield (GY) to nitrogen uptake (NU). The relationship between GY and NU follows a saturation curve, with a sharp increase in GY at low levels of NU and a plateau at higher levels of NU. The effect of flattening is due to a change in the correlation between both variables. The described relationship is caused by complex plant physiological processes driven by plant genetics and environmental conditions.

The current most common statistical techniques to approach the GY and NU field data are: 1) univariate analyses of GY and NU, which disregard the correlation between both variables; 2) analyses of the ratio, whose distribution exhibits non-normal heavy tails; or 3) simple least square regressions to model the GY trend on NU, which ignore the environmental factors affecting the GY and NU relationship.

We argue that a more suitable approach is a bivariate analysis which considers the mutual correlation and avoids the ratio distributional problems. Different environments and genotypes affect plant physiology and therefore may produce clusters on the data. The bivariate mixture model is an appropriate and useful statistical tool for identifying subpopulations in the presence of factors unknown a priori. A case study will be presented to illustrate the application of the mixture model technique.

Estimating sample size in high-throughput experiments

Teresa Neeman
Statistical Consulting Unit, Australian National University
teresa.neeman@anu.edu.au

A lot of attention has been given to controlling family-wise error rates and false positive rates (FDR) in high throughput experiments with thousands of hypotheses. In contrast, there has been little guidance on the number of samples needed to be able to reliably observe interesting results.

It is not unusual to see sample sizes as low as 1-3 biological samples per condition, and although statisticians often cringe at these small sample sizes, they have no standard tools for guiding researchers on appropriate sample sizes (such as power calculations) for these experiments.

As a consequence, effects of interest may go undetected, and chance outcomes may not be easily differentiated from interesting ones. Ideally, one would like the observed interesting hypotheses (rejected false nulls) to be both a high proportion of the all rejected hypotheses (1-FDR), as well as a high proportion of all interesting hypotheses (false null hypotheses).

In this talk, I will show how one can generalise power calculations for a single hypothesis to the high throughput setting. This tool can be used effectively to compare the expected performance across proposed sample sizes using metrics defined by (1-FDR) and expected proportion of false nulls that are rejected.

Dissecting genetic and non-genetic sources of long-term yield trend in German official cultivar trials

Hans-Peter Piepho
University of Hohenheim, Stuttgart, Germany
hans-peter.piepho@uni-hohenheim.de
Coauthors: Friedrich Laidig, Thomas Drobek, Uwe Meyer (Bundessortenamt, a federal authority responsible for granting of Plant Breeders' Rights, the registration of varieties and for variety and seed affairs; Hannover, Germany)

Long-term yield trends of crop cultivars may by studied to identify genetic components due to breeding efforts and non-genetic components due to advances in agronomic practices.

Many such studies have been undertaken, and most of these inspect trends either by plotting means against years and/or by some form of regression analysis based on such plots.

Dissection of genetic and non-genetic components is a key challenge in such analyses. In the present paper we consider mixed models with regression components for identifying different sources of trend.

We pay particular attention to the effect of disease breakdown, which is shown to be confounded with long-term genetic and non-genetic trends. The models are illustrated using German multi-environment trial (MET) data on yield and mildew susceptibility for winter wheat and spring barley.

Can we trust the interpretation of past climates in New Zealand gained from kauri trees?

Maryann Pirie
AgResearch Ruakura, Hamilton
maryann.pirie@agresearch.co.nz

Knowledge of past climates needs to increase to allow a coherent picture of current and past climate systems to be developed.

Kauri tree rings are a significant resource that allows us to better understand past climates within New Zealand. The variation in the ring widths for kauri has previously been shown to be related to the El Niño-Southern Oscillation (ENSO) phenomenon, where the evolving chronology variance can be interpreted as a reconstruction of the past activeness of ENSO events.

Using this interpretation, the 2010 kauri master chronology (AGAUc10c), containing both archaeological and modern (living tree) source material of kauri ring widths sequences suggests that ENSO was less active in the 1500s and activity appears to be increasing from the 1500s to the present.

Another key feature of the reconstruction is the inter-annual to decadal-scale periodicities. These interpretations rely on the assumption that the observed trends are a true representation of past climate variation. However, there are concerns that the change in the size and/or age composition of the cores making up the chronology may be affecting the observed trends.

Visualising ecological matrix data

Shirley Pledger
Victoria University of Wellington, New Zealand
shirley.pledger@vuw.ac.nz
Coauthors: Richard Arnold

Ecologists frequently collect data in the form of nxp matrices of binary or count data. Examples include (i) ecological community data (presence/absence or counts of n species over p samples), and (ii) bipartite networks (e.g. plant-pollinator networks, the occurrence or counts of visits of n insect species to p flowering plant species).

Dimension reduction is used to detect and visualise overall patterns. This is traditionally done by mathematical methods involving either a distance metric (e.g. multidimensional scaling) or eigenvalue analysis (e.g. correspondence analysis, principal component analysis).

Our new method (in press, Computational Statistics and Data Analysis, 2013) exploits redundancy in the matrix (e.g. groups of species with similar occurrence patterns) by (bi)clustering with finite mixtures. This statistical method, based on likelihoods, provides model comparison techniques and a surprisingly rich array of visualisations.

Comparison of two nonparametric regression curves: test of superiority and non-inferiority

Suman Rakshit
Monash University
sumanprobability@gmail.com
Coauthors: Mervyn Silvapulle (Department of Econometrics and Business Statistics, Monash University)

Methods are developed for testing the hypothesis that a new treatment is better than the standard when there is a covariate and the response is represented by a nonparametric regression curve. Thus, these are related to analysis of covariance, but without assuming parametric regression models.

First, to introduce the definition, we choose two curves called the non-inferiority and superiority bounds. The former lies below and the latter lies above the mean regression function for the standard treatment. Now, the new treatment is defined to be overall better than the standard if the mean regression function for the new treatment lies above the non-inferiority bound at every value of the covariate and above the superiority bound at least at some values of the covariate.

As an example, it may be desired to test the hypothesis that a new proposed government policy is non-inferior for every income level and that it is superior for some in the low-income group. The foregoing non-inferiority and superiority bounding curves need not be rigidly specified for statistical inference. More specifically, for any given shapes of these curves, their best locations corresponding to acceptance of simultaneous non-inferiority and superiority, can be estimated.

The asymptotic version of the proposed test can be implemented by using the asymptotic critical values that are easily accessible. Because the asymptotic test is conservative, a less conservative bootstrap test is also proposed and is shown to be asymptotically valid. The null hypothesis turns out to be the union of two intersecting convex cones in a space of functions.

We identify the least favorable configuration in the null hypothesis, and compute the critical values at these configurations. Thus, the results of this paper provide nonparametric extensions to some results in parametric order restricted inference.

Countermatched design as a two-phase sampling design

Claudia Rivera
Department of Statistics, The University of Auckland
`clriverarodriguez@gmail.com`
Coauthors: Thomas Lumley

When studying risk factors for disease incidence in a cohort, some of them could become costly relative to their collection. For example, laboratory analysis of blood and urine samples or records of diet and exercise exposure. Additionally, if the exposure is rare, it would be needless and vain to sample many subjects.

Langholz and Borgan (Biometrika, 1995) proposed to use *countermatching* as a technique to sample a group of controls for each of the cases in a cohort study. At each failure time, strata are formed by variables of interest and subsequently a stratified sample is drawn in each risk set. Expressing the countermatched design as a two phase sampling design is of interest when fitting Cox model, either parametrically or semiparametrically.

In order to use weighted likelihood or weighted partial likelihood, the inclusion probabilities are found for the final two-phase sample. It leads to an extension of the methods proposed by Samuelsen (Biometrika, 1997) for nested-case control designs.

The controls from subjects in the risk set are reused for each case so that efficiency of estimators might be improved.

Keywords: countermatching, nested-case control designs and Cox model.

A stable statistical tool to find a signature of mesenchymal stem cells

Florian Rohart
Australian Institute for Bioengineering & Nanotechnology (AIBN), The University of Queensland
`f.rohart@uq.edu.au`
Coauthors: Celena Heazlewood, Kim-Anh Le Cao and Christine Wells

Gene expression signatures can be found in most experimental series, but signatures which are reproducible across many experimental series are difficult to find. This is partly because of the technical barriers preventing large-scale meta-analysis of different expression experiments generated across different laboratories, or on different experimental platforms.

We will present a new gene-signature tool which works across multiple expression series. The performance of this tool was assessed on mesenchymal stem cells (MSCs). MSCs have been found in almost all tissues in the body and, although some biological differences exist between MSCs populations derived from different tissue sources, we hypothesized that all MSCs share a common gene expression signature, irrespective of their origin.

The gene-signature tool has been implemented using 680 samples of different cell types from the large-scale Stemformatics stem cell expression database (www.stemformatics.org). Our approach integrates the YuGene cross-platform normalisation method, which is effective at reducing experimental batch effects and integrating data from different commercial platforms, while retaining genuine biological variation between samples.

A sparse partial least-squares differential analysis (sPLS-DA) method was used to identify an accurate classifier of mesenchymal cell types. A stability analysis provided an average of more than 96% classification accuracy.

Our approach highlights the robustness of biological signatures when experimental variables such as platform or batch can be reduced. The tool as well as a visualisation of the signature will be available on CRAN and at Stemformatics, respectively.

Hierarchical mixture failure time regression for classification of the immune response of Atlantic salmon

Jose Romeo
University of Santiago of Chile and Department of Statistics, University of Auckland
jose.romeo@usach.cl
Coauthors: Renate Meyer (University of Auckland) and Felipe Reyes-Lopez (University of Santiago of Chile)

This work presents a Bayesian hierarchical model with the dual objective to analyze stratified survival data and to automatically classify each stratum into a finite number of groups.

This is achieved by specifying stratum-specific baseline hazards and a finite mixture distribution for the stratum-specific shape parameters. A proportional hazards or accelerated failure time regression component allows to identify the influence of covariates on the survival distribution.

We illustrate the model using a dataset of Atlantic salmon, stratified by families, that have been challenged with infectious pancreatic necrosis virus (IPNV). The main objectives are to model the survival time in terms of certain covariates as well as to classify the salmon families into either an IPNV susceptible or resistant group with the ultimate goal of improving resistance to IPNV through a selective breeding programme.

We compare the fit of different models that include stratum-specific baselines and covariate effects and show a certain degree of robustness of the classification.

Deterministic modelling of whole-body sheep metabolism

Emma Smith
Data Analysis Australia, Nedlands
emma@daa.com.au
Coauthors: Volker Rehbock (Department of Mathematics and Statistics, Curtin University)

The livestock industry is one of the most enduring and prolific in Australia; however there is a tangible lack of whole-body approaches to research in this area.

A variety of techniques across statistics, mathematics, biochemistry and livestock analysis are employed to develop a deterministic model for sheep growth from 12 weeks of age through to maturity. The model tracks body protein, fat, water, circulating metabolites and DNA pools to simulate whole-body metabolism.

The work extends an existing growth model, and is implemented into the MISER3 software to assess parameter selection, optimal growth trajectories, and stability into adulthood. There is a particular focus on producing accurate feed intake, nutrient absorption, energy expenditure and wool growth representation, as well as defining appropriate initial conditions.

P-Spline vector generalized additive models

Chanatda Somchit
Department of Statistics, The University of Auckland
csom017@aucklanduni.ac.nz
Coauthors: Chris Wild and Thomas Yee

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 vector generalized linear models (VGLMs) and VGAMs is very large and encompasses many statistical distributions and models.

The underlying algorithm is iteratively re-weighted least squares (IRLS) and modified vector backfitting using vector splines. Backfitting provides the advantage in terms of allowing the component functions of an additive model to be represented almost any smoothing or modelling technique. But implementing estimation of the degree of smoothness of a model cannot be done easily in this approach.

We aim to develop automatic smoothing parameter selection, which is difficult in a classic backfitting approach. We develop VGAMs using a more amenable penalized likelihood-based approach. This can be achieved by using penalized regression smoothers, based on P-splines, and transforming VGAMs into the vector generalized linear model framework. Then the P-spline VGAM penalized likelihood can be maximized by penalized iteratively re-weighted least squares (P-IRLS).

In terms of smoothing parameter selection, we aim to fit the smooth components simultaneously. This can be performed by using well-founded criteria such as generalized cross-validation (GCV) or the Akaike information criterion (AIC). Additionally, a very important facility which allows the smooth component functions to be constrained is provided.

Statistical modeling for clinical trials

Gwynn Sturdevant
Department of Statistics, The University of Auckland
sstu011@aucklanduni.ac.nz
Coauthors: Thomas Lumley

Recently, trials have been used to evaluate pharmaceutical treatments and their ability to delay the onset of diabetes and hypertension. Criticism of one trial focuses on a simulated 80% Type I error rate.

Diagnosis of hypertension occurs when a noisy measurement exceeds a threshold, and diagnosis results in treatment, which censors subsequent blood pressure measurements.

Approaches to design were studied by simulation using simple comparison of cumulative incidence as the analysis. As an approach to analysis, we studied mixed models treating post-diagnosis measurements as data missing at random. Neither design strategy reliably controls Type I error.

The mixed-model analysis approach does control Type I error and give unbiased estimates of treatment effect. Carryover effects on incident hypertension or diabetes require care in design and analysis. We recommend an analysis based on mixed models.

A model-based clustering method for viral quasispecies identification with the Illumina Platform

Olivier Thas
Ghent University, Belgium and University of Wollongong
Olivier.Thas@UGent.be
Coauthors: Bie Verbist (Ghent University, Belgium)

A patients viral population tyically exists of several hundreds of variants of a particular virus, caused by mutations. Each variant (quasispecies) shows a different sensitivity to antiviral treatment.

The introduction of massively parallel sequencing (MPS) technologies has fundamentally altered genomics research. In a virology research environment, this technology creates opportunities for studying viral quasispecies in HIV-1 and HCV-infected patients, which is essential for the understanding of pathways to resistance and can substantially improve individualized treatment.

Whereas standard genotyping only provides information on the most abundant variants, the MPS technologies allow in-depth characterization of sequence variation in more complex populations, including low-frequency viral strains.

One of the challenges in the detection of low-frequency viral strains concerns the errors introduced during the MPS process. Technology-associated errors may occur at a higher frequency than the truly present mutations, impeding a powerful assessment of low-frequency virus mutations.

Phred-like quality scores, which are associated with the base-calls, quantify the base-call reliability. However, we learnt that the quality scores reported by the Illumina MPS platform do not always reflect the true error probabilities. Often the error rates are underestimated.

We propose to model the error probabilities as a function of the reported quality scores and an extra covariate (read direction). Additionally, we model the error probabilities of the second-best base-calls. These probabilities form the basis of a multinomial structure in a model-based clustering approach which allows for the identification of viral quasispecies while reducing the number of false positives drastically, and thereby lowering the limit of detection.

Evaluating methods of estimating missing values for three-way three-mode multi-environment trial data

Ting Tian
School of Mathematics & Physics, The University of Queensland
t.tian2@uq.edu.au
Coauthors: Geoff McLachlan; Mark Dieters and Kaye Basford (School of Agriculture & Food Sciences)

The analysis of three-way three-mode multi-environment trial data has played a critical role in understanding multi-attribute genotype response. However, the presence of missing observations in such trial data makes the analysis difficult.

We propose modifications of imputation methods based on principal component analysis and clustering analysis, where the nature of three-way three-mode datasets is emphasized. The modifications are used with singular value decomposition imputation, probabilistic principal component analysis and maximum likelihood principal component analysis, as well as finite normal mixture clustering and hierarchical clustering.

These methods were compared by randomly deleting various proportions of the values from full data sets (two real and one simulated), and then comparing the estimated with the actual values using the normal root mean square error. The results indicated that the maximum likelihood principal component analysis and hierarchical clustering imputation processes had lower root mean square error than the other methods.

Decomposing smooths to identify structure

Martin Upsdell
AgResearch Ruakura, Hamilton
martin.upsdell@agresearch.co.nz

Generalised additive models form an overall smooth by adding individual smooths. These enable the different effects of the individual factors to be examined by inspecting the smooth curve associated with the factor.

Smoothing spline ANOVA models fit a smooth surface to the whole data and then attempt to assess the effects of each factor by computing mean levels for the different values of the factor.

The first approach suffers from the need for all effects to be additive and so limits the shape of the curves to be fitted. The second approach suffers from combining both the effect of the factor with the choice of weighting given to the various levels of the other factors.

This talk will show how the variance approach to smoothing allows complex models to be built by the two standard model building tools of adding effects together and crossing effects together. The curves produced can be individually plotted by the factors involved allowing the factors contributions to the overall model to be understood without having to worry about the weighting to give to other factors. Nuisance factors typically are allowed to have rather more complex shapes.

General model comparison tools can be used to identify the needed complexity of the curves to describe the factors of interest. Examples will be given.

Fitting a mixed-effects model with a multivariate nonparametric distribution of random effects

Xuxu Wang
Department of Statistics, The University of Auckland
demiwang363@gmail.com
Coauthors: Yong Wang

A parametric distribution is commonly used to model random effects in a mixed-effects model. This, e.g., based on a normal distribution, may work well in many cases, but it gives biased estimation when the assumption is incorrect.

Alternatively, one can use a nonparametric distribution to avoid model misspecification, but, owing to computational difficulty, there is very little progress in this regard in the literature, especially when there exist multiple random effects and thus a multivariate nonparametric distribution has to be used.

In this talk, I will describe a new, efficient likelihood-maximising algorithm that fits a mixed-effects model with a multivariate nonparametric distribution assumed for random effects. The availability of this algorithm makes it feasible to apply these models in practice.

Using logistic regression as an example, our empirical studies show that these nonparametric/semiparametric models perform consistently better than their parametric counterparts on both simulated and real-world data.

Optimality and contrasts in block designs

Emlyn Williams
Statistical Consulting Unit, Australian National University
emlyn.williams@anu.edu.au
Coauthors: Hans-Peter Piepho (University of Hohenheim, Stuttgart, Germany)

A-, E-, and D-optimality criteria are used to compare incomplete block designs with the same set of basic parameters and thereby determine design optimality. Authors have used these criteria to compare the optimality of equally and unequally replicated designs and have observed inconsistencies. In this talk we will revisit some of this work and discuss the importance of treatment contrasts in assessing design optimality.

Design and analysis of experiments to detect differential gene expression

Susan Wilson
Australian National University and University of New South Wales
sue.wilson@anu.edu.au

To measure differential gene expression, microarrays have been widely used. Design issues, as well as the statistical challenges associated with analysing such data, are relatively well-understood.

More recently, RNA-sequencing technologies have been replacing microarrays for the detection of differential expression. These new technologies present both similar and different experimental design issues.

Further, although they share with the older microarray technology many of the statistical challenges, they pose others that are fundamental, and unique, and for which further research is needed.

Comparison between linear mixed model (LMM) with and without centering for separation of within- and between-subject effects

Hwan-Jin Yoon
Statistical Consulting Unit, The Australian National University
hwan-jin.yoon@anu.edu.au
Coauthors: Alan Welsh (Centre for Mathematics and Its Applications, The Australian National University)

Generally, hierarchically structured data with lower level observations nested within higher level(s) arises frequently in practice, particularly in biology and in the study of animal behaviour. This hierarchically structured data may be due to gathering repeated measurements on experimental units as in longitudinal studies or may be due to subsampling the primary sampling units.

An important feature of such hierarchically structured data is that the observations within a level of aggregation are often not independent. Observations in the same individual are generally more similar than are observations from other individuals.

One of the statistical models to account for interdependency and structuring of data is linear mixed model. Linear mixed models (LMM) not only take into account the heterogeneity between such sources of aggregation but also allow the partitioning of total variation among these sources. However, researchers do not distinguish within- and between-group effects and so implicitly assume that the these effects are the same.

In this talk, we show that within- and between-group effects can be very different; as a consequence, models that incorrectly assume common effects can lead to very misleading assessments of the association of explanatory variables with response.

A hierarchical Bayesian model for analyzing clinical proteomic data with non-random missingness

Irene Zeng
Department of Statistics, The University of Auckland
i.zeng@auckland.ac.nz
Coauthors: Thomas Lumley, Kathy Ruggiero, Ralph Stewart, Martin Middleditch, See-Tarn Woon, Patrick Gladding, Wikke Koopman and Rohan Ameratunga

Introduction Proteomics is emerging as a new stream in medical studies for investigating hundreds and thousands of molecular biomarkers simultaneously. The high-throughput data from proteomic study brings challenge to its data analysis. The challenges originate from the hierarchical levels of the relative quantity of the protein expressions, the complexity of the experiment, the large amount of information, and the non-random missingness of the peptide quantification data. Method We use multivariate multilevel models to analyze the hierarchical protein expression data. This proposed method takes into account the different types of variations from the experimental factors such as the physical features of the quantitative Mass Spectrometer, labels of ITRAQ, and potential run effects. It is demonstrated to be reliable for deriving the study parameters at the protein level comparing to using unadjusted protein ratio. Under this multivariate philosophy, a Bayesian hierarchical approach was used to handle the abundance-dependent missingness of the protein expression data and to provide shrinkage of overly-variable estimates. Gibbs sampling and Hamiltonian MC/No U-Turn Sampling were compared for evaluating the posterior joint distributions of the study parameters. Results The proposed methods were assessed in a simulated proteomic study and two clinical proteomics studies. The proposed multivariate multilevel model and the missing data approach enable us to cope with the large heterogeneity in the relative peptides intensity, from which the protein intensities are derived. It is shown to be an improvement compared to the protein ratios approach. The multivariate protein model utilizes experimental information across all proteins and this enabled those proteins with small number of peptide information to be derived while adjusting for experimental effects. The HMC/NUTS sampler was substantially more efficient, as expected for a smooth, high-dimensional posterior distribution.

Constrained ordination analysis with an increased number of bell-shaped response functions with applications in metagenomics

Yingjie Zhang
Ghent University, Belgium
yingjie.zhang@ugent.be
Coauthors: Olivier Thas (Ghent University, Belgium, and University of Wollongong)

Constrained ordination analysis, e.g. canonical correspondence analysis (CCA), is a class of popular techniques for analyzing species abundance studies in ecology.

These methods aim to find an environmental score or gradient (linear combination of environmental variables) along which the species abundances are maximally separated. The species response functions, which describe the expected abundance as a function of the environmental score, are only ecologically meaningful if they are bell-shaped.

Many classical model-based ordination methods, however, use quadratic regression models without imposing the bell-shape and thus allowing for meaningless U-shaped response functions. The analysis output (e.g. a biplot) is therefore misleading and the conclusion are prone to errors. The problem becomes even more severe in metagenomics studies in which hundreds of species abundances are measures simultaneously.

We present a modification that enforces an increase in the number of bell-shaped response functions by penalising U-shaped solutions. This is accomplished by introducing a convenient prior distribution on the regression parameters. The performance of the new method is studied in a simulation study and the method is applied to a metagenomics dataset from microbial ecology.

Posters

Hidden Markov analysis of high bandwidth mechanosensitive ion channel gating data.

Ibrahim Almanjahie
School of Mathematics and Statistics, University of Western Australia,
20104542@student.uwa.edu.au
Coauthors: Robin Milne, Nazim Khan, and Boris Martinac (School of Medicine and Pharmacology, University of Western Australia)

Hidden Markov models (HMMs) are used to describe the gating behaviour of single ion channel and form the basis for statistical analysis of patch clamp data. Extensive high bandwidth (25 kHz, 50 kHz) data from the mechanosensitive channel of large conductance (MscL) in *E. Coli* were analysed using HMMs and HMMs with a moving average adjustment for filtering. The analysis aimed to determine the number of conductance levels, together with mean current, mean dwell time and proportion of time for each level. Comprehensive data analyses and comparisons across all our high bandwidth data sets have consistently shown seven conductance levels for this channel.

Doubly Resolvable Row-Column designs

David Baird
VSN NZ Ltd
david@vsn.co.nz

These designs generalise Latin squares, with replicates made up of groups of rows and columns. Replicates in both directions are useful when spatial effects and trial operations can happen in both row and columns directions. For example, the trial may be sown row by row, and irrigation performed down the columns. The replicates may be made up of full rows and part rows, but are contiguous in both directions. The allocation of treatments is also made in a manner to optimize the normal row-column analysis of the trial. The advantage over the standard row-column design is that, if only the row and column replicates are needed to control the spatial effects, then these are orthogonal to the treatments means, so no treatment information is lost and the means require no adjustment. In some special cases, where both the rows and columns are divisible by the number of replicates, then a third level of resolvability can be added to the design with quadrants forming replicates.

The design of spatial and temporal experiments to estimate pest populations

Jennifer Bramwell
Data Analysis Australia, Nedlands
jennifer@daa.com.au
Coauthors: Maryann Pirie (AgResearch Ruakura, Hamilton)

Pests can cause damage to pastures resulting in the loss of profitability within the primary industry sector. Therefore, it is of interest to use the prior season adult population to estimate future populations. The pests are often clustered in their preferred habitat and their abundance determined by environmental conditions.

Therefore, we want to design an experiment that will allow us to estimate the scale and location of damage based on the number of pests trapped. The idea is that environmental data can be incorporated to produce a map of the area including soil and moisture gradients, and land use. From this map we can choose trap locations based on habitat preference, and fewer traps placed in areas likely to get zero or small catch. The trap data can then be measured over a period of time, and this trap data added to the spatial map to provide predictions of damage to pastures.

Investigation of a robust adaptive method for identification of outliers

Ken Bredemeyer
Mathematics and Statistics, School of Engineering and Information Technology, Murdoch University
kbredemeyer@iinet.net.au
Coauthors: Brenton Clarke

We investigate a method mooted by Clarke (2012) at the Australian Statistical Conference (ASC2012) to identify outliers in regression. The method correctly identifies the outliers in famous data sets. We now compare its use by simulation in a simple linear regression, to that intimated by Cook (1977). The method we examine uses residuals from MM estimation. Multiple outliers can be identified without the need to specify the number of outliers to test for. Simulations suggest that this method is suitable for data with sample sizes up to 1000.

Confidence interval for a negative binomial mean

Andrew van Burgel
Biometrics Unit, Department of Agriculture and Food, Western Australia
andrew.vanburgel@agric.wa.gov.au

Faecal worm egg counts are typically overdispersed with many zero or low counts and are often modelled using the negative binomial distribution. Worms are a major cost to farmers and the mean faecal worm egg count from a subsample of about ten sheep from a mob is commonly reported. A very simple yet accurate formula is proposed for the confidence interval of the mean faecal worm egg count. One can't simply log transform to normalise the data, calculate a standard confidence and then back-transform, because this gives a confidence interval centered on the median which is considerably lower than the mean. The proposed confidence interval formula has potential for broader application where data displays a similar overdispersed distribution.

Modelling changes in beetle communities following pest control

Vanessa Cave
AgResearch, Hamilton
vanessa.cave@agresearch.co.nz
Coauthors: Corinne Watts, Danny Thornburrow, and John Innes (Landcare Research, New Zealand)

Small introduced mammals, particularly rodents and hedgehogs, are important predators of indigenous New Zealand invertebrates. Despite the high conservation value and important functional roles of invertebrates, they are frequently overlooked in conservation management and research, where the priority is for charismatic vertebrate species. This case study investigated how ground-dwelling beetle communities responded to two pest control regimes used in New Zealand biodiversity sanctuaries: (a) eradication or near eradication of mammals within a fenced exclosure (Zealandia), and (b) sustained trapping and poisoning of introduced mammals (Otari-Wiltons Bush).

Using pitfall traps, data on beetle abundance and species richness was collected at 1-3 fixed seasonal visits at both sanctuaries for several years. Within a sanctuary, sampling effort was standardised between years, however it differed between sanctuaries. This difference precluded the use of the data for comparing beetle abundance and richness at the two sanctuaries directly: with reduced sampling effort it is likely fewer rare species are detected and fewer beetles are caught. However, the data could be used to compare annual trends between the two sanctuaries. As the measures of beetle abundance and richness were expected to be sanctuary-dependent, to exhibit consistent seasonal effects, and to vary smoothly with time, cubic smoothing splines formulated as linear mixed-effects models were used to model the data.

Linear and non-linear calibration of levels of aflatoxin in maize using NIR spectroscopy.

Ross Darnell
CSIRO Computational Informatics, Brisbane
ross.darnell@csiro.au
Coauthors: Bill Venables

Aflatoxin in maize represents a serious health risk for eastern Africans who rely on a maize-based diet. Near infrared (NIR) spectroscopy is a possible non-destructive technology to measure levels of aflatoxin in maize kernels and flour. This work compares the more commonly used partial least squares approach with random forest methods to the general problem of inverse regression.

Fitting a mixture model with univariate normal components: A comparison between MLE via the EM algorithm and a robust approach

Thomas Davidson
Mathematics and Statistics, School of Engineering and Information Technology, Murdoch University
thomas.davidson88@gmail.com
Coauthors: Brenton Clarke

Mixture models with univariate normal components have a long history in the application of representing subpopulations within the overall population. Often, the fitting of the model can be seen as an incomplete data problem as it is not possible to determine which component distribution generated each observation.

With this in mind, the usual approach of parameter estimation involves maximum likelihood estimation (MLE) via the expectation maximisation (EM) algorithm. This however, for the case of univariate normal components, is flout with difficulty as is evident by the unboundedness of the influence function for the MLE. This implies the MLE is easily perturbed by potential gross outliers.

We investigate the robust L_2 estimator of Clarke and Heathcote (1994) as a useful alternative to the MLE. Illustration of the stability of the L_2 estimator compared to the MLE as obtained from the EM algorithm is made for various examples of contamination. Finally we will apply the robust methods of estimation in the context of normal mixtures in the field of Biometrics. An example is provided.

An exploration of historical changes in pasture growth in relation to climate change

Siva Ganesh
AgResearch Grasslands, Palmerston North
siva.ganesh@agresearch.co.nz
Coauthors: Paul Newton, Mark Lieffering, Frank Li, and Mike Dodd

The detection and attribution studies on historical trends in biological systems in relation to changes in climate are particularly challenging in agricultural systems where other factors (such as management) are changing over time.

Our study considered changes in pasture yield (net herbage accumulation (NHA)) over the period 1960-2004 in data collected from a trial where management (grazing protocol and fertiliser application) was kept constant over time.

We looked for trends in, and correlations between, NHA and climate variables. There was a notable positive trend for NHA in spring over the period, and positive trends in rainfall, atmospheric CO_2 concentration and soil fertility. In this presentation, we highlight the approach we took, including a time series regression model, and briefly discuss our findings.

Analysis of spatial point patterns on the celestial sphere

Tom Lawrence
School of Mathematics and Statistics, The University of Western Australia
tjlawrence@bigpond.com
Coauthors: Adrian Baddeley, Gopal Nair and Robin Milne

The New General Catalogue of Nebulae and Clusters of Stars (NGC) is a historically important dataset giving the locations (sky positions) of nonstellar astronomical objects as points on the celestial sphere.

Most statistical research on spatial point patterns has focussed on data in two- and three-dimensional Euclidean space. Methods for point patterns on the sphere are only now being developed.

In this presentation, I outline several statistical methods and statistical models for point patterns on the sphere. Methods include the spherical counterpart of Ripley's K function. Models include the analogue of the stationary Thomas process; we sketch the construction of this process and the derivation of Ripley's K function for the process.

We apply the methods to the spatial pattern of galaxies in the revised NGC.

Stochastic generation of harmony

Steven Miller
Department of Statistics, University of Waikato
smiller@waikato.ac.nz

Leibniz described music as the pleasure the human soul experiences from counting without being aware that it is counting. Music is, in essence, an application of mathematics. There are clear rules for generating tones that sound harmonious, and tones that are discordant.

Even untrained ears are able to detect differences between music from different periods, music in different styles, and music from different cultures. There must be a mathematical logic that underlies these distinctions.

Skilled musicians are able to improvise simultaneously, harmonising with each other with no reference point, other than perhaps a common key centre, or cadence structure. This suggests an online system of step-ahead prediction of states, possibly governed by some prior distribution, informed by data as it becomes available.

We describe a project that will examine these three cases: producing harmonies stochastically that sound good; identifying features of distinct styles of music, and attempting to replicate them; and producing an automated accompanist for improvised performances.

Stochastic methods have been able to generate text that at a cursory glance or to the uninformed reader seems to make sense. We would like to be able to replicate this achievement in a musical medium that is, generate music stochastically that to the casual listener sounds natural.

MM algorithms for biclustering models

Duy Vu
Department of Mathematics and Statistics, University of Melbourne
duy.vu@unimelb.edu.au
Coauthors: Murray Aitkin

Biclustering is an important tool in statistical exploratory analysis which can be used to detect latent row and column groups of different response patterns. However, few studies include covariate data directly into their biclustering models to explain these variations.

In this study, we describe a model-based biclustering framework that considers both stochastic block structures and covariate effects. This introduction of covariate data together with a large number of latent variables makes the estimation task challenging. We address this problem by proposing approximation estimation algorithms derived from the variational generalized expectation-maximization (EM) framework where the goal is to increase, rather than maximize, the likelihood lower bound in both E and M steps. More specifically, we not only extend the minorization-maximization (MM) update in the E step to model both discrete and continuous responses, but also propose new MM updates for the M step to handle high-dimensional covariate structures.

The advantage of the MM principle is then demonstrated through a comparison experiment between the new MM update in the E step with the fixed-point update on two large datasets. Finally, the utility of the proposed biclustering framework and estimation algorithms is demonstrated through three block modelling applications in model-based collaborative filtering, network modelling, and microarray analysis.

Investigation of the performance of the trimmed likelihood of lifetime distributions with censoring

Karuru Wamahiu
Mathematics and Statistics, School of Engineering and Information Technology, Murdoch University
31245726@student.murdoch.edu.au
Coauthors: Brenton Clarke

Trimmed likelihood estimation is a relatively new approach to estimation, having been developed initially for location and scale estimation in normal parametric families.

In 2000, Clarke, Gamble and Berdnarski gave a form of the estimator for the negative exponential distribution. Since then Brenton Clarke and Christine Müller have investigated the equations to describe the estimator in the case of censoring. We investigate the performance of such an estimator in terms of efficiency and robustness through simulation.

This is a simulation study of the performance of such an estimator. The investigation will involve the study of the bias and efficiency in comparison to the usual maximum likelihood approach.

Publication patterns: a co-citation network for Biometrics 1993-2013

Jennifer Wilcock
Department of Statistics, The University of Auckland
j.wilcock@auckland.ac.nz

The poster presents a network graph of co-citations in papers published in Biometrics over the past 20 years. The graph illustrates the areas of study that have dominated in the journal, the extent to which these areas of study are connected and by which groups of cited papers, how tightly clustered these papers are, and which papers have been most highly co-cited. Whilst the analysis is essentially descriptive I hope it will be of intrinsic interest to anyone working in any area of biometrics and familiar with published literature in our field.

The nodes of the graph are papers cited in Biometrics since 1993, and the edges of the graph connect two papers cited together, or co-cited, in the same

Biometrics paper. Edges are weighted by the frequency of co-citation occurring over the 20 year period and the graph is laid out using a force-directed algorithm.

In order to restrict the size of the graph and make it more readable, only papers (the nodes) that have been cited at least ten times in Biometrics are included, and included co-citations (the edges) must have appeared in at least five different Biometrics papers. The nodes in the network can be papers from any year and any journal however, since the nodes are based on the articles being cited in Biometrics rather than the papers published in Biometrics themselves.

The End

*** The End of the Abstracts - thank you to all for contributing to the conference! In statistical methodology, the 3 asterisks indicate a highly significant achievement on behalf of everyone.
MD.***