Australasian Applied Statistics Conference, 2018
and
Pre-Conference Workshops


2-7 December 2018, Rotorua

www.aasc.nz

VSNi

THE UNIVERSITY OF AUCKLAND
NEW ZEALAND

agresearch

Welcome

We are pleased to welcome you to Rotorua for the 2018 Australasian Applied Statistics Conference (AASC18). Our conference is devoted to providing you with the opportunity to liaise with fellow statisticians within the agricultural, biological and environmental sciences and to keep abreast of the most recent developments in statistics within this context.

This conference is one of a longstanding series, originating with the initial Australasian Genstat conference in Canberra in 1979. A previous Genstat conference was held in Rotorua in 1992, so this is our second conference in Rotorua. The Genstat conference changed to the Australasian Applied Statistics Conference in 2011 to encompass the wider community of applied statisticians, and AASC conferences have been held in Palm Cove (Aus), Queenstown (NZ), Port Lincoln (Aus) and Bermagui (Aus). We are glad to have long time attendees of these conferences (including some who were at the first conference) and newcomers join us. One of the positive things of this conference is that it's not too large, so that you can meet most people, and we hope you will make new friends and opportunities for collaboration.

The themes of AASC18 are big data analytics, reproducible research, history of ANOVA and REML, experimental design, statistical consultancy in the biosciences, and applied statistics. Our exciting group of invited speakers and workshop presenters (Chris Auld, Peter Baker, Salvador Gezan, Roger Payne, Alison Smith, Robin Thompson, Linley Jesson, Ruth Butler, Gabriela Borgognone and Helene Thygesen) will help explore these themes in various contexts.

Our social program will provide you with a great opportunity to catch-up with old friends and to make new ones. We hope also that you enjoy exploring Rotorua, one of the most unique tourism destinations in the world: a geothermal paradise and a Māori cultural heartland. The conference dinner will provide great views over the town and lake and a sumptuous banquet.

We will be awarding a prize for the best poster (voted for by the audience on Tuesday evening) and a prize of the best student presentation and these will be presented at the end of the conference.

We are very grateful to our sponsors for their support of AASC18. Thanks to VSNi, our principal sponsor, for generously funding two of our invited international speakers. Thanks also to the University of Auckland and AgResearch for their sponsorship.

We thank you, and all conference delegates, invited speakers, workshop presenters, speakers, poster presenters, sponsors and the local staff for helping create this exciting conference program.

David Baird (Chair, VSN NZ Ltd)
Chris Triggs (Treasurer, University of Auckland)
Vanessa Cave (Secretary, AgResearch Ltd)

Australasian Applied Statistics Conference, Rotorua NZ, 3-7 December 2018

Time	Tuesday morning	Wednesday morning	Thursday morning
8:45	Welcome	Notices	Notices
8:50	Chris Auld Large scale & real time data analytics	Alison Smith Design Tableau: An aid to specifying the linear mixed model for a comparative experiment	Peter Baker Computing tools for a Don't Repeat Yourself data analysis workflow and reproducible research
9:50	Benoit Liquet Statistical Methods for the Analysis of High-Dimensional and Massive Data using R	Nicole Cocks, Alison Smith, David Butler and Brian Cullis Efficient designs for early generation variety trials using genetic relationships	Janet Chaseling, Kyle James and Kirsty Wright Beyond statistical consulting: What role should statisticians play in ensuring good practice in the application and interpretation of statistics by disciplinary practitioners?
10:10	Peter Jaksons Ceci n'est pas une pipe… yet: Building data & analysis pipelines	Chris Lisle, Alison Smith, Carole Birrell Connectivity, does it impact genetic variance parameter estimates in a Multi Environment Trial analysis?	Kerry Bell and Rao Rachaputi Answering the research question by identifying balanced embedded factorials in messy combined trials
10:30	Morning tea	Morning tea	Morning tea
11:00	David Baird Updates to Genstat 19	Graham Hepworth Statistical issues arising from the Australian Royal Commission into Institutional Responses to Child Sexual Abuse	Ruth Butler "From cradle to grave": making an impact from conception to publishing
11:20	Donghui Ma and Darren Murray Www.MixedModel.Academy as a SaaS learning tool	Ken Dodds, John McEwan, Timothy Bilton, Rudiger Brauning and Shannon Clarke A depth-adjusted Hardy-Weinberg test for low-depth sequencing data	Helene Thygesen Statistical inference and management decisions Linley Jesson The view from the other side: a biologist's view of communicating statistics
11:40	Robin Turner, Claire Cameron, Ella Iosua and Ari Samaranayaka Establishing a Biostatistics Unit at the University of Otago	Sarah Sonal, Philip Schluter, Martin Lee and Jennifer Brown Deprivation, Hospital admissions and previous dental appointment records in Early Childhood: A comparative use of traditional statistical modelling with machine learning	Gabriela Borgognone From heaven to hell… and how to find a way back!
12:00	Daisy Shepherd and Steffen Klaere Assessing model adequacy in phylogenetics - are the tools powerful?	Hans Hockey and Kristian Brock Hockey sticks and broken sticks – a design for a single-treatment, placebo-controlled, double-blind, randomized clinical trial suitable for chronic diseases	Discussion

Australasian Applied Statistics Conference, Rotorua NZ, 3-7 December 2018

Time	Tuesday afternoon		Thursday afternoon
12:20	Lunch	Lunch	Lunch
13:20	Robin Thompson Desert island papers - a life in variance parameter and quantitative genetic parameters estimation reviewed using ten papers	Conference Trips	Salvador Gezan Getting the most of my mixed model (and specially ASReml): applications in quantitative genetics and breeding
14:20	Arthur Gilmour ASReml moving forward		Jess Meza, Gururaj Kadkol, Steven Simpfendorfer, Steve Harden and Ky Mathews Crown Rot Tolerance in Durum Wheat
14:40	Chris Brien Mimicking anova in reml mixed modelling of comparative experiments using the R-package asremlPlus		Emi Tanaka, Pauline O'Shaughnessy, Chong You and Chris Brien Don't be so negative about negative estimates of variance components
15:00	Afternoon tea		Afternoon tea
15:30	Charlotte Jones-Todd Shared latent fields for mark-location dependence in a log Gaussian Cox process		Jelena Cosic, Steffen Klaere, Matthew Goddard and Bruno Fedrizzi Bayesian Network as a Modelling Tool for Increasing Knowledge on the Factors Influencing Vineyard Longevity and Sustainability
15:50	James O'Malley, Pablo Martinez-Camblor, Todd MacKenzie, Doug Staiger and Philip Goodney Instrumental variable estimation in the Cox Proportional Hazards Model		Kyle James, Janet Chaseling, Kirsty Wright and Albert Gabric Statistical theory of rare event detection applied to forensic database establishment
16:10	Aswi Aswi, Susanna Cramb, Wenbiao Hu, Gentry White and Kerrie Mengersen The impact of covariates on the grouping structure of a Bayesian spatio-temporal localised model		Rodelyn Jaksons, Elena Moltchanova, Beverley Horn and Elaine Moriarty Estimating the Extent of Underreporting in Disease Counts
16:30	Pierre Lafaye de Micheaux, Pavlo Mozharovskyi and Myriam Vimond A notion of depth for curve data		Oliver Stevenson and Brendon Brewer Finding your feet: modelling the batting abilities of cricketers using Gaussian processes
16:50	Kouji Yamamoto A one-sided test to simultaneously compare the predictive values		Rory Ellis, Daniel Gerhard, Elena Moltchanova and Mike Trought Using Bayesian Growth Models to Predict Grape Yield
17:10	Poster Session		Conference Dinner
17:30	Drinks and viewing of posters		

Time	Friday morning
9:00	Notices
9:10	Poppy Miller, Chris Jewell, Peter Diggle and Kate Hacker Identifying hotspots of rat activity and how they affect the risk of leptospirosis in urban slums
9:30	Julie Mugford, Alex James and Elena Moltchanova Developing methods to improve the accuracy of classification based crowdsourcing
9:50	Ari Samaranayaka, Rohana Kumara De Silva, B.S.M.S. Siriwardena, W.A.M.U.L. Abeyasinghe and W.M. Tilakaratne A predictive model for nodal metastases among oral cancer patients
10:10	Jillian Haszard, Kim Meredith-Jones, Sheila Williams and Rachael Taylor Analysing compositional time-use data in paediatric populations
10:30	Morning tea
11:00	Roger Payne 50 Years of Genstat ANOVA
12:00	Awards and Farewell
12:10	Lunch

Contents

Pre-Conference Workshops

Monday Morning

Scaling R in the Cloud with doAzureParallel

Monday
Workshop
8:30-12:30

Chris Auld
chauld@microsoft.com
Microsoft

This workshop will equip attendees with the skills to scale their R workloads into the cloud. Cloud computing offers the promise of near infinite computing power on tap. The *doAzureParallel* package is a parallel backend that integrates with the native *parallel* support provided by the R runtime from v2.14. With *doAzureParallel*, each iteration of the loop runs in parallel on a pool of Virtual Machines (VM) in the cloud, allowing users to scale up their R jobs to tens to thousands of CPU cores.

In this workshop we will cover:

- Simulation and optimization at scale. Using monte carlo methods, attendees will learn general purpose approaches to executing and reproducing simulation workloads at scale.
- Parallelism of data intensive tasks such as ETL and feature engineering. This will include approaches to scaling *plyr* and *data.table* based manipulations.
- The use of *doAzureParallel* with the two key machine learning meta-frameworks; *caret* and *mlr*. We will cover the use of parallel computing for k-fold cross-validation and hyper-parameter optimization.

This is a **hands on workshop**. Attendees should have had some experience working with the R language before and should bring along a machine (Windows, Mac, Linux all great) running a recent build of R and the editor of their choice. Cloud computing access will be provided on the day.

Monday
Workshop
8:30-12:30

Genstat 19ed Masterclass: how to get the analyses you need, the output you want, and the graphs you prefer

Roger Payne, David Baird and Vanessa Cave
roger.payne@vsni.co.uk
VSN International, Hemel Hempstead, United Kingdom

In this workshop we will help you to master recent new features that make it easier for you to identify and validate the most appropriate model, produce graphs in the styles that you prefer, and archive results for convenient future use.

Topics will include:

- manipulating and validating complicated data sets
- finding the best REML model for single and series of trials
- meta-analysis of series of trials
- checking the validity of your analysis
- saving and archiving results
- exploiting new flexibility in the graphics menus
- customizing the plots
- producing publication-quality graphs
- producing multi-paged PDF files directly from batch jobs
- use of the new foreign language/Unicode text facilities in Genstat 19.2

The sessions will involve a mixture of examples and practicals. So please bring your laptops (ideally with Genstat 19ed already installed).

Lunch: 12:30-13:30

Monday Afternoon

Tools for Efficient Data Analysis Workflow and Reproducible Research

Monday
Workshop
13:30-17:30

Peter Baker
p.baker1@uq.edu.au
School of Public Health, University of Queensland, Herston, Australia

Researchers and statistical consultants are drowning in data. In my early career, computer processing power and storage were limited and so we spent a considerable amount of time formulating strategies to efficiently manipulate data and focus on the analysis of relatively small data sets or subsets of larger ones. The day I started as a biometrician with NSW Agriculture, my boss told me that 80% of a biometrician's time was spent organising and cleaning data for analysis. Given today's powerful computing environments we might think that things would be easier but it appears that the 80% figure is still the case. Indeed, due to larger data sets and competing demands for our time, many data analysts face problems with organising their workflow. This tutorial provides a hands-on introduction to computing oriented strategies for the workflow of research data management and data analysis. The ideas presented in this tutorial follow Long's 2009 book on the workflow of data analysis using STATA which provides a useful guide to managing the workflow for data analysis in large projects. However, Long's approach concentrates on manual methods for implementing steps in the workflow. Efficient computing solutions are available for most steps in the process. Programming tools like GNU Make for managing workflow and regenerating output, GNU git for version control, GNU R functions and packages for repetitive tasks and finally R Markdown for producing reports directly from analyses. Generic and statistical software specific tools are presented. These tools are incorporated into hands on exercises.

Keywords: Data analysis, DRY (Don't Repeat Yourself) workflow, version control, Make, R, project management

---

Monday
Workshop
13:30-17:30

Modelling correlations between observations in agricultural and environmental sciences with ASReml-R

Salvador Gezan
sgezan@ufl.edu
University of Florida, United States of America

This half day workshop will concentrate on the statistical analysis of observations that present some form of temporal, genetical or spatial correlations as found on field or laboratory studies. The aim is to illustrate how to analyze this data using the framework of linear mixed models with the software ASReml-R.

This workshop will consist of several small sessions starting with a brief introduction to the command syntax for ASReml as implemented in R, followed by topics related to repeated measures (random regression), genetic relatedness, and spatial analyses using many examples from agricultural and environmental sciences. Some theoretical aspects will be presented for the construction of the models, but the focus will be on the practical issues and interpretation of the results.

Welcome Reception: 17:30-18:30

Abstracts - Tuesday Morning

Welcome: 8:45

Large Scale & Real Time Data Analytics

Tuesday
Invited
8:50-9:50

Chris Auld
chauld@microsoft.com
Microsoft

Computing capacity and storage capacity exploded in recent years. This has allowed us to deal with larger datasets and in more recent years focus has moved to approaches for processing data in near real time. In this session we'll discuss approaches to large analytics jobs that we run inside Microsoft and LinkedIn as well as learnings taken from engagements with customers globally.

Biography
Hi, I'm Chris and I'm a Principal Software Engineering manager for Microsoft living in Singapore. My team of engineers is located across Asia and Europe and we work with Microsoft's largest global customers to build advanced analytics and machine learning solutions. I studied Information Science and Law at the University of Otago back when R was S and Margaret Wilson was the Attorney General. Rotorua is my home town and I'm on the board of our tourism organization so ask me to point you in the direction of all the awesome activities. Finally, I'm not much into the 3rd person to be honest...

Tuesday
9:50-10:10

Statistical Methods for the Analysis of High-Dimensional and Massive Data using R

Benoit Liquet
benoit.liquet@qut.edu.au
University of Pau and Pays de L'Adour, France

In this talk, I will first concentrate on a class of multivariate statistical methods called Partial Least Square (PLS). They are used for analyzing the association between two blocks of 'omics' data, which bring challenging issues in computational biology due to their size and complexity. These powerful approaches can be applied to data sets where the number of variables is greater than the number of observations and in the presence of high collinearity between variables. Different sparse versions of PLS have been developed to integrate multiple data sets while simultaneously selecting the contributing variables. Sparse modeling is a key factor in obtaining better estimators and identifying associations between multiple data sets. A unified algorithm is proposed to perform four types of PLS including their regularized versions. We present various approaches to decrease the computation time and show how the whole procedure can be scalable to big data sets.

The second part of the talk will focus on a massive data setting. In this context, we focus on a semiparametric regression model involving a real dependent variable Y and a p-dimensional covariate X (with p>1). This model includes a dimension reduction of X via an index X'b. The Effective Dimension Reduction (EDR) direction cannot be directly estimated by the Sliced Inverse Regression (SIR) method due to the large volume of the data. To deal with the main challenges of analyzing massive data sets which are the storage and computational efficiency, we propose a new SIR estimator of the EDR direction by following the "divide and conquer" strategy. A simulation study and an illustration on a massive airline data set using our edrGraphicalTools R package shows that our approach enables us to reduce the computation time and conquer the memory constraint problem posed by massive datasets.

Ceci n'est pas une pipe... yet: Building data & analysis pipelines

Tuesday
10:10-10:30

Peter Jaksons
peter.jaksons@plantandfood.co.nz
Plant & Food Research, New Zealand

Modern research projects often rely on several data sets and several types of data analysis are required to be undertaken at various steps throughout the project. To meet these challenges, we are building numerous data and analyses pipelines within Plant and Food Research so that we may increase the efficiency, reliability and reproducibility of our science projects.

In this talk I will discuss the process of setting up these pipelines and the tools that we use, illustrating them in two case study projects. The first is about soil nitrogen modelling to promote accuracy in fertilising in agriculture, and the second case study is about genomic marker selection in red kiwifruit.

Morning Tea: 10:30-11:00

Tuesday
11:00-11:20

Updates to Genstat 19

David Baird* and Roger Payne
david@vsn.co.nz
VSN NZ Ltd, New Zealand

New features in Genstat 19 will be covered, including graphics updates, writing multi-paged PDFs from the server, using Unicode symbols and ease of use features.

---

www.MixedModel.Academy as a SaaS learning tool

Tuesday
11:20-11:40

Donghui Ma* and Darren Murray
donghui@vsni.co.uk
Biosci Thailand

For statisticians and experienced mixed model users the advantage of applying REML (Restricted Maximum Likelihood) analysis over classical ANOVA is obvious both as a practical tool to overcome real-life problems such as missing values or as a theoretical framework to loosen the assumption of the classical i.i.d. response. The mixed model framework also provides more flexibility of modelling the factors and their correlations.

At the same time the complexity both for learning and applying REML grows. There are many challenges for introducing mixed models to the research world and for new adopters to learn mixed model applications syntax; reproduce analytical methodology from academic publication; understand the mixed model theorem with the language of matrix algebra; find examples for different types of analysis and analytical methods in different areas of research; interpret results and draw conclusions.

To anticipate the challenges MMA is a platform that can generate syntax for major commercial mixed model applications and relevant free R packages using a simple user interface. It will help support migration such as ASReml-R3 to ASReml-R4, SAS to ASReml-R or Vice versa. The platform will provide an opportunity for users to run their analysis using ASReml-R. The MMA will provide access to datasets and associated code from influential mixed model literature to help support teaching. Its knowledge base will allow user generated examples to be stored and managed within individual projects. Users will be able share their projects with others and can be summarised and recorded into a video to make the analysis reproducible to other users. It will include videos to help interpret each step of an analysis and will cover a wide range of topics. The mixed model anatomy section of the platform will provide material on the mixed model algebra and intermediate results such as design matrices for deep learners.

Tuesday
11:40-12:00

Establishing a Biostatistics Unit at the University of Otago

Robin Turner*, Claire Cameron, Ella Iosua and Ari Samaranayaka
robin.turner@otago.ac.nz
University of Otago, New Zealand

Cogent design and analyses are essential elements of good quality health related research. Accordingly, biostatisticians are critical team members for any health related research contributing to the design and analysis of a study but, also, having a relatively holistic approach to the research process. While there are a number of different models for employing and supporting biostatisticians worldwide there are also many challenges. These challenges include perception of their discipline and workforce shortages[1].

The University of Otago has employed biostatisticians in teaching, research and consulting roles over many years. The biostatisticians were primarily departmentally managed and several were employed (originally by the Health Research Council) to support health related research across the region. These roles were seen as consulting roles. More recently, the focus has become Divisional within the university and to align with this, a Biostatistics Unit was established at the end of 2017 drawing together the divisionally funded biostatisticians in Dunedin.

This presentation will discuss the challenges of establishing a biostatistics unit in part of a large division and the benefits that come from taking this approach. The unit offers a model of setting in place structures to support the growth and development of biostatistics, particularly those with consulting roles. Future research will be undertaken to assess this model and the impact it has on biostatistics at the university.

1. Cameron, C., Iosua, E., Parry, M., Richards, R., & Jaye, C. (2017). More than Just Numbers: Challenges for Professional Statisticians. Statistics Education Research Journal, 16(2).

Assessing model adequacy in phylogenetics - are the tools powerful?

Tuesday
12:00-12:20

Daisy Shepherd* and Steffen Klaere
dshe078@aucklanduni.ac.nz
The University of Auckland, New Zealand

Phylogenetic inference focuses on a critical problem in biology - deriving the evolutionary history between groups of organisms. Statistical models are used to describe the changes in DNA that occur over evolutionary time, to help infer the best estimate of their phylogeny. Our ability to accurately explain the evolutionary relationships depends heavily on the use of an appropriate statistical model.

Tests to establish the 'best-fitting' model have been studied extensively, with a wide range of tools available. However, tests for model-to-data fitness have received much less attention in the last decade. The complex nature of phylogenetic data and models make adapting popular goodness of fit (GOF) tools to the phylogenetic framework difficult. As a result of this, classical distributions for standard GOF test statistics are no longer suitable.

Previous research has suggested the use of simulation approaches to assess model adequacy to overcome this issue. These methods remove the dependence on a specified distribution, and rather build the null hypothesis using a simulation approach. Both the maximum likelihood (ML) and Bayesian frameworks have adopted these methods. However, the power of simulation-based tests has often been overlooked, with little knowledge on how the tests actually perform when faced with messy data.

We conducted an exploratory study into the power of these simulation-based approaches within both the ML and Bayesian frameworks. We explored whether the methods rejected the null hypothesis of adequate fit under the presence of noisy data. We also investigated to what degree our data needs to be 'messy' in order for the tests to fail and produce a false negative result.

Lunch: 12:20-13:20

Abstracts - Tuesday Afternoon

Desert island papers - a life in variance parameter and quantitative genetic parameters estimation reviewed using ten papers.

Tuesday
Invited
13:20-14:20

Robin Thompson
robin.thompson@rothamsted.ac.uk
Rothamsted Research, United Kingdom

When I was recently asked to review the history of REML I thought of adapting the device used in Desert Island discs when 'castaways' are invited to discuss their life and suggest eight record tracks they would like if stranded on a desert island. I have replaced eight discs by ten papers to make a more coherent story. This includes discussion of the initial motivation for REML, links with other methods, the development of the Average Information algorithm, the development of computer software and I will end with some open questions.

Biography

Robin Thompson is a pioneering leader in the fields of statistics, quantitative genetics and animal and plant breeding. He started his career in Edinburgh in the late 1960's in the then Agricultural Research Council Unit of Statistics, later moving to the Animal Breeding Research Organisation, which ultimately became part of the Roslin Institute. He remained there until the mid 1990's when he moved to the Institute of Arable Crops Research at Rothamsted as the head of the prestigious department of statistics, established by R.A. Fisher, that laid the foundation for much of modern statistics.

In the 1970's, while based at the University of Edinburgh, Robin and Desmond Patterson proposed and developed a new statistical method which came to be called REML. It now dominates in several fields including statistics, genetics, breeding, and field trial analysis. Data collected in many real-life settings are inherently unbalanced and REML provides optimized statistical methodology for such data. The foundation paper from 1971, "Recovery of inter-block information when block sizes are unequal" is a citation classic with more than 3,700 citations to date. These days, REML is implemented in most widely used statistical analysis packages.

In addition to inventing REML, Robin has made significant contributions to the development of computationally efficient algorithms to facilitate the application of REML to large datasets. Of these, the most important is the Average Information

algorithm, developed in the 1990s. Robin, together with colleagues Arthur Gilmour, Brian Cullis and Sue Welham developed the versatile and efficient software package called ASReml that is the most widely used in animal and plant breeding across the globe today.

Robin has made a broad range of contributions to the development of rigorous science underpinning UK and global animal and plant breeding programmes. His collaborations with the various Edinburgh groups had, and continue to have, particular impact in UK dairy, beef and sheep breeding.

Finally, Robin has made a major input to post-graduate education in Edinburgh. For many years, he taught components of the MSc in Animal Breeding and Quantitative Genetics. He was a formal supervisor of more than twenty research students and an informal mentor of many more. Robin has been incredibly generous with his ideas to both students and established researchers (perhaps because he realised he did not have the skills to see things to fruition). Several of his former students now have high international reputations. He has been awarded honorary Doctor of Science hons causa degrees by the university of Edinburgh and the Technical University of Valencia, and had a street named after him in Valencia.

Tuesday
14:20-14:40

ASReml moving forward

Arthur Gilmour
arthur.gilmour@cargovale.com.au
Private consultant, Australia

Arthur Gilmour wrote ASReml while working for NSW Agriculture (DPI) as a biometrician to facilitate mixed model analysis of research data. He worked closely with Brian Cullis who proposed some of the models we needed to analyse, especially in relation to variety testing, and with Robin Thompson who proposed the main statistical ideas incorporated in ASReml. ASReml implemented the REML algorithm using the Average Information (AI) algorithm and sparse matrix methods.

Software development is an ongoing process. ASReml was built on previous programs developed by Arthur over the preceding 20 years, especially REG, then BVEST and then AIREML. The development includes changes in syntax as well as algorithmic development, changes to exploit improving hardware, extensions to model flexibility and changes to the development environment.

DPI and Rothamsted Research (RR) contracted VSN to retail ASReml. After Robin and Arthur had 'retired', and Brian had left DPI in 2010, DPI and RR sold ASReml to VSN. However, all three continued to support and develop ASReml in association with VSN.

The core code of ASReml is now over 20 years old. Among many other changes, the model demands on ASReml have grown to now include more dense models, especially those based on genomic relationship matrices. This is a fundamental change from the types of models ASReml was geared for and we have begun writing the next generation of mixed model software. Of course, it must still handle the sparse models effectively, but needs to be much better at the dense models. ASReml 4.2 has made some progress in this regard.

Arthur has begun a new program under the name Echidna. In the first instance, it will just replicate 80% of what ASReml can do, but with much more coherent source code. It will then be extended to incorporate new methods geared for dense models. As ASReml was initially a platform for scientific research in fitting new models, so this is the goal for Echidna. Once proven, the new methods will be made available to the commercial world in collaboration with VSN.

Echidna is available from EchidnaMMS.org for non-commercial use (education and science).

Mimicking anova in reml mixed modelling of comparative experiments using the R-package asremlPlus

Tuesday
14:40-15:00

Chris Brien
chris.brien@unisa.edu.au
University of South Australia, Australia

Fisher introduced anova in 1918 and it was the dominant analysis method for comparative experiments for much of the 20th Century. Eisenhart put forward the idea of a mixed model in 1947 and the next 50 years saw the theoretical development of mixed modelling, including Patterson and Thompson's 1971 formulation of reml. However, it took the arrival of software for reml mixed modelling in the 1990s for its widespread use and the replacement of anova to be feasible.

Principles established for the use of anova in analysing comparative experiments during the 20th Century include:

(i) The random terms in the analysis should be the unit terms appropriate to the randomization and they should never be omitted from the analysis.

(ii) Fixed terms correspond to treatment factors and, while they can be subjected to hypothesis testing, they should not be dropped from the analysis so as to ensure that a pure estimate of error is obtained.

(iii) Only use the hypothesis tests that are needed to select a marginality-compliant (or hierarchical) model.

While anova is based on fitting a mixed model, the maximal mixed model, it is usually a single-fit method of analysis. On the other hand, reml mixed modelling is inherently a model selection procedure in which multiple models are fitted and the anova principles can be overlooked in using it.

A number of functions implemented in asremlPlus[‡] facilitate the observance of these anova principles when asreml is used to do the reml mixed modelling. Their use will be illustrated with an example that uses a three-factor factorial experiment laid out using generalized randomized complete block design.

[‡]http://chris.brien.name/rpackages

Afternoon Tea: 15:00-15:30

Tuesday
15:30-15:50

Shared latent fields for mark-location dependence in a log Gaussian Cox process

Charlotte Jones-Todd
Charlotte.JonesTodd@niwa.co.nz
NIWA, New Zealand

The locations of objects or events in space form point patterns. Understanding the processes that drive a point pattern's spatial structure is typically of interest. Additional information associated with the points, called marks, may provide an improved understanding of these spatial processes. Furthermore, the marks themselves may exhibit spatial structure that depends on the process generating the point locations. This structure may itself be of interest, as might the relationship between the marks and the points. Here, the spatial structures of both the point locations and their marks, along with the relationship between them, are relevant and interpretable in the context of the application, rather than being characterised by 'nuisance parameters'. This talk considers a marked spatiotemporal log-Gaussian Cox process which accounts for the dependence between the spatial structures driving the points and their marks. In particular, this talk will focus on the flexibility of using shared Gaussian random fields in modelling a range of data sets from very different scientific fields.

Instrumental variable estimation in the Cox Proportional Hazards Model

Tuesday
15:50-16:10

James O'Malley*, Pablo Martinez-Camblor, Todd MacKenzie, Doug Staiger and Philip Goodney

James.OMalley@Dartmouth.edu

Geisel School of Medicine at Dartmouth, United States of America

Instrumental variable (IV) methods are widely used for estimating average treatment effects in the presence of unmeasured confounders. However, the capability of existing IV procedures, and most notably the two-stage residual inclusion (2SRI) algorithm recommended for use in nonlinear contexts, to account for unmeasured confounders in the Cox proportional hazard model is unclear. We show that instrumenting an endogenous treatment induces an unmeasured covariate, referred to as an individual frailty in survival analysis parlance, which if not accounted for leads to bias. We propose a new procedure that augments 2SRI with an individual frailty and prove that it is consistent under certain conditions. The case of a non-proportional hazards (homogeneous treatment effect) is also considered. The finite sample-size behaviour is studied across a broad set of conditions via Monte Carlo simulations. Finally, the proposed methodology is used to estimate the average effect of carotid endarterectomy versus carotid artery stenting on the mortality of patients suffering from carotid artery disease. Results suggest that the 2SRI-frailty estimator generally reduces the bias of both point and interval estimators compared to traditional 2SRI.

Tuesday
16:10-16:30

The impact of covariates on the grouping structure of a Bayesian spatio-temporal localised model

Aswi Aswi*, Susanna Cramb, Wenbiao Hu, Gentry White and Kerrie Mengersen
aswi@hdr.qut.edu.au
Queensland University of Technology, Australia

A variety of Bayesian models have been used to describe spatial and temporal patterns of disease, where the data are aggregated at a small area level. A relatively recent approach is the spatio-temporal conditional autoregressive localised model introduced by Lee and Lawson (2016) [1] which allows for spatial autocorrelation between adjacent areas within discontinuous groups. In this paper we use a case study approach to evaluate the impact of covariates on the groups identified in such a model. The study focuses on the influence of climate on annual dengue fever cases in 14 geographic areas of Makassar, Indonesia, during the period 2002-2015, where the climatic factors are measures of temperature, rainfall and humidity. All subsets of the covariates are considered and different spatio-temporal formulations of the model are compared with respect to three metrics: the overall goodness of fit (Watanabe-Akaike Information Criterion (WAIC)), the group-specific coefficients and the proportion of areas included in the groups. The evaluations are complemented by a range of innovative visualisations of these performance metrics. The results show that inclusion of climatic predictors causes group size and structure to alter in the localised model and that examination of these changes gives greater understanding of their influence. The study also provides more general insight into the behaviour of the Bayesian spatio-temporal conditional autoregressive localised model in the presence of covariates.

Reference
1. Lee D, Lawson A. Quantifying the Spatial Inequality and Temporal Trends in Maternal Smoking Rates in Glasgow. The Annals of Applied Statistics. 2016;10(3):1427-46.

A notion of depth for curve data

Tuesday
16:30:16:50

Pierre Lafaye de Micheaux*, Pavlo Mozharovskyi and Myriam Vimond
lafaye@unsw.edu.au
The University of New South Wales, Australia

Following the seminal idea of John W. Tukey, statistical data depth is a function that determines centrality of an arbitrary point w.r.t. a data cloud or a probability measure. During the last decades, data depth rapidly developed to a powerful machinery proving to be useful in various fields of science. Recently, implementing the idea of depth in the functional setting attracted a lot of attention among theoreticians and applicants. We suggest a halfspace-based notion of data depth suitable for data represented as curves, or trajectories, which inherits both Euclidean-geometry and functional properties but overcomes certain limitations of the previous approaches. It can be shown that the Tukey curve depth satisfies the requirements posed on the general depth function, which are meaningful for trajectories. Application of the Tukey curve depth is illustrated on brain imaging and written patterns recognition.

Tuesday
16:50-17:10

A one-sided test to simultaneously compare the predictive values

Kouji Yamamoto
yamamoto.phd@gmail.com
Osaka City University, Japan

Positive and negative predictive values are important measures of a medical diagnostic test performance. The positive predictive value is the probability of disease when the diagnostic test result is positive, and the negative predictive value is the probability of no disease when the test result is negative.

There are several methods to compare the predictive values of two diagnostic tests separately. However, there are many cases where is not only one endpoint but multiple endpoints are required to evaluate the accuracy among two diagnostic tests. In this presentation, we consider the following case: the effectiveness of a new diagnostic test is confirmed only when the superiority of the new test to the other test is showed in at least one measure and non-inferiority is showed in the other measure. For this type of trials, we propose a new one-sided test statistic. Also, we evaluate the performance of the proposed method via simulation studies.

Poster Session: 17:10-18:30

Abstracts - Poster Session

Poster presenters are invited to give a 1 minute lightning presentation. All posters will remain on prominent display throughout the conference.

How do you like them predictions? Experiences with obtaining predictions from GLMMs

Tuesday
Posters
17:10

Michael Mumford* and Kerry Bell
michael.h.m@hotmail.com
Department of Agriculture and Fisheries, Queensland, Australia

Generalised linear mixed models (GLMMs) are known to suffer from estimation biases, particularly when the data is binomially distributed with structural terms fitted as random effects. A glasshouse experiment that explored the rate of germination of wheat genotypes under two soil surface crust treatments provides the motivating example for an investigation of such biases from a GLMM.

From the analysis of this experiment, it was noted that the predicted values for both the strong and weak crust were higher than the raw data means when random blocking effects were included in the analysis. This upwards bias resulted from the fixed effects being on the logit scale, while the random effects were on the original scale (i.e. assumed to be normally distributed).

The proposed solution involved the use of a conjugate hierarchical generalised linear model (HGLM) which relaxes the assumption of normality for random effects. In the motivating example, we assumed that the fixed effects followed a binomial distribution with a logit link function and the random effects followed a beta distribution with a logit link function. By performing the analysis using a HGLM as opposed to a GLMM, the predictions for the strong and weak crust matched closely with the raw data means, as opposed to the GLMM predictions.

Tuesday
Posters
17:10

Associations between symptoms and colorectal cancer outcome in GP/hospital e-referrals

Malgorzata Hirsz*, Lynne Chepulis, Lyn Hunt, Ross Lawrenson and Michael Mayo
mh331@students.waikato.ac.nz
Department of Statistics, University of Waikato, New Zealand

The incidence of colorectal cancer (CRC) in New Zealand (NZ) and Australia is amongst the highest in the World, however, 5-year survival is generally poor in NZ. One of the reasons for this is because CRC is often diagnosed at later stages (Duke's stages C and D) when treatment is less effective. In order to diagnose more patients in earlier stages of the disease, identification of symptoms associated with early presentation are being investigated. Knowledge of such symptoms could be helpful for physicians when making decisions about further investigation of a patient (e.g. colonoscopy).

Diagnostic performance of CRC symptoms has not been investigated in New Zealand before. One of the obstacles for conducting such research is the lack of coding of symptoms included in the e-referral system. Initially, we extracted symptoms from all referrals made to the Gastroenterology and General Surgery departments at Waikato Hospital in Hamilton in the years 2015-17 (n=31,059). Cancer cases were then identified from the cancer registry (n=503). The extraction of symptoms involved identifications of different spelling variants and medical terminology used in the e-referrals. Diagnostic performance of commonly studied CRC symptoms was described using the following measures: sensitivity, specificity, PPV, NPV and odds ratio (OR). The values were compared to the results from a meta-analysis by Jellema et al. (2010). In general, the ORs were similar but the prevalence of the symptoms in the NZ secondary care patients was lower compared to the values from the meta-analysis, leading to lower sensitivity and higher specificity. The data will be used to assess the associations between stage of CRC and the extracted symptoms.

Multiple endpoint test for both left-censored and arbitrarily distributed endpoints

Tuesday
Posters
17:10

Ludwig Hothorn
hothorn@biostat.uni-hannover.de
Leibniz University Hannover, Germany

The demonstration of an association between a selected clinical covariate and multiple analytes from diverse metabolomic platforms is a recent problem in metabolomics.

The clinical covariate can be modeled either quantitatively, i.e. in a regression context, or qualitatively i.e. multiple contrast tests (by design, or post-hoc classification). In the regression context, a maximum test on three regression models for the arithmetic, ordinal, and logarithmic-linear dose metameters is used (Tukey et al., 1985), whereas the distribution is available via multiple marginal model approach (mmm) (Pipper, 2012). Multiple contrast test for ordered alternatives are available, depending in the categorization. Here, a joint approach for modeling the quantitative covariate both quantitatively and qualitatively using the CRAN package tukeytrend.

When considering multiple endpoints, the problem of different distribution of the many analytes arises. Some publications ignore this problem (and assume implicitly normal distribution throughout) or use the same transformation for all analytes, such as log-transformation. But it is not a realistic assumption that all analytes follow the same distribution. Metabolomic data contain measurement of complete analytes and those with detection limits. Analytes with detection limits are considered as left-censored variables assuming some data points are below a certain limit but it is unknown how many. Therefore, the joint consideration of analytes without or with detection limits is needed. The concept of most likely distribution (mlt) (Hothorn, 2015) is used, where models for the unconditional or conditional distribution function of any univariate response variable can be estimated by choosing an appropriate transformation function and related parametrisation. Fortunately, most left-censored variables belong to the class of distributions within the mlt-framework.

A multiple endpoint test will be demonstrated using a max-max-test (maximum on endpoints, maximum on models) where the correlation between the models (not the data) is achieved by the multiple marginal model approach. Using the function mmm within library(multcomp) up to 333 endpoints can be analysed simultaneously today.

Using part of the KarMeN cross-sectional data (Rist, 2017) and the CRAN packages multcomp (for mmm), tukeytrend and mlt, the association between age and selected analytes without and with detection limits are shown in detail.

References

J. W. Tukey, et al. Testing the statistical certainty of a response to increasing doses of a drug. Biometrics, 41(1):295-301, 1985.

Hothorn, T., Moest, L., Buehlmann, P.: Most likely transformations. eprint arXiv:1508.06749, 2015

C. B. Pipper, et al. A versatile method for confirmatory evaluation of the effects of a covariate in multiple models. Journal of the Royal Statistical Society Series

C-Applied Statistics, 61:315-326, 2012

Rist, M. et al. Metabolite patterns predicting sex and age in participants of the Karlsruhe Metabolomics and Nutrition (KarMeN) study PLOS ONE 12;8; e0183228. 2017

DNA-MAP: Statistics as an aid to decision making in ancestry prediction of unidentified historical military remains

Tuesday
Posters
17:10

Kyle James*, Janet Chaseling, Kirsty Wright, Andrew Bernie and Albert Gabric
kyle.james@griffithuni.edu.au
Griffith University, Australia

DNA-Military Ancestry Predictor (MAP) is a simple Knowledge Based Decision Support System (KBDSS) developed to assist the Unrecovered War Casualties – Army (UWC-A) predict ancestry of historical military remains. Ancestry prediction involves the application of the Bayesian formula using conditional probabilities. A theoretical approach to studying the effects of the parameters involved becomes complex and unwieldy with a large number of inputs, each with a different degree of reliability. This approach is also unsuitable for end users; UWC-A Investigators and members of the Australian Defence Force Identification Board.

DNA-MAP uses combined genetic and historical data to evaluate the probability of a set of remains belonging to either an Australian or a Japanese World War II (WWII) soldier. DNA-MAP utilises ***What-If*** scenarios developed from realistic possibilities. The prior expectation of relative numbers of Australian soldiers missing in geographical areas may be available from historical army records, but actual values from different sources may vary considerably. Identification of appropriate reference populations is also difficult. The composition of the population of Australian WWII soldiers is quite different from today's Australian population. It is also unlikely the DNA probability estimates from individual countries are appropriate for people of these nationalities during WWII. Migrants to Australia do not represent random samples from their respective countries. Variation and unreliability also stem from huge variation in sample sizes available to estimate DNA probabilities. Emphasis is on using DNA areas which are rare in one nationality and common in another and it is important that samples are of sufficient size, ensuring that if a rare event does exist it will be detected.

The ***What-If*** scenarios enable the user to see the effects varying these different parameters have on the outcome, namely, the probability the recovered remains are of Australian ancestry. This paper presents the first stage of DNA-MAP's development.

Tuesday
Posters
17:10

Unbiased growth traits from noninvasive phenotypic data produced in high-throughput controlled environments

Chris Brien, Bettina Berger, Nathaniel Jewell* and Trevor Garnett

nathaniel.jewell@adelaide.edu.au

Australian Plant Phenomics Facility, School of Agriculture, Food and Wine, University of Adelaide, Australia

A revolutionary aspect of imaged-based high-throughput phenotyping is its noninvasive nature, so that the growth of individual plants can be tracked over time. The challenge is to process the large amount of resulting data to answer biologically relevant questions.

A typical phenotyping experiment at The Plant Accelerator®, Australian Plant Phenomics Facility, involves daily imaging and watering of each plant. The resulting longitudinal data is then transformed into biological growth traits along the lines described by Al-Tamimi et al. (2016) as follows:

1. Calculate and plot the observed projected shoot area (PSA, kilopixels) for each plant, together with implied absolute growth rate (AGR) and relative growth rate (RGR).

2. Smooth the trend over time for each plant to remove transient influences, then calculate and plot smoothed AGR and RGR.

3. Check for anomalous plants.

4. Using the smoothed data, identify and calculate biologically relevant growth parameters (one value per plant) – for example, average AGR and RGR for each plant in a set of subintervals characterised by near-constant acceleration/deceleration in the growth rates.

5. For each parameter in Step 4, perform a mixed-model analysis and produce spatially adjusted estimates for each experimental line and/or condition.

Steps 1-4 are assisted by the R package imageData (Brien, 2018). The process is flexible and computationally undemanding, with a minimum of a priori assumptions (e.g. no assumption of logistic growth). It is illustrated using data from a large-scale rice salinity experiment from Al-Tamimi et al. (2016) and refinements to their method are outlined.

References

Al-Tamimi, N., Brien, C., Oakey, H., Berger, B., Saade, S, Ho, Y. S., Schmöckel, S. M., Tester, M. & Negrão, S. (2016) Salinity tolerance loci revealed in rice using high-throughput non-invasive phenotyping, Nature Communications, 7: 13342

Brien, C. J. (2018). imageData: aids in processing and plotting data from a Lemna-Tec Scananalyzer. http://cran.at.r-project.org/package=imageData.

Finite Mixture Clustering via Adjacent-Categories Logit model

Tuesday
Posters
17:10

Lingyu Li
lilingyunz@hotmail.com
Victoria University of Wellington, New Zealand

Traditional analysis of ordinal data treats the outcome either as nominal or continuous variables. The nominal approach ignores the ordinal property whereas the continuous approach introduces assumptions about the ordinal level spacing - thus these traditional approaches can lead to a loss of statistical power or can introduce bias.

This talk presents cluster analysis of ordinal data utilising the natural order information of ordinal data. Three models usually used in ordinal modelling are discussed: the proportional odds model, the adjacent categories model and the ordered stereotype model.

In our research, the data take the form of a matrix where the rows are subjects, and the columns are a set of ordinal responses by those subjects to, say, the questions in a questionnaire. We implement model-based fuzzy clustering via a finite mixture model, in which the subjects (the rows of the matrix) and/or the questions (the columns of the matrix) are grouped into a finite number of clusters. We will explain how to use EM (Expectation–Maximisation) algorithm to estimate the model parameters. Specifically, we illustrate the details of using Adjacent-Categories logit model to perform row/column and bi-clustering. This clustering method differs from other typical clustering methods such as K-means or hierarchical clustering, because it is a likelihood-based model, and thus statistical inference is possible.

Tuesday
Posters
17:10

Recent Development of R package 'predictmeans'

Dongwen Luo
dongwen.luo@agresearch.co.nz
AgResearch Ltd, New Zealand

R package 'predictmeans' is used to calculate predicted means for linear models (including 'aov', 'lm', 'glm', 'gls', 'lme', and 'lmer'). It provides functions to diagnose and make inferences such as predicted means and standard errors, contrasts, multiple comparisons, permutation tests and graphs. This talk will highlight some major improvement of the package by examples, especially, output presentation and permutation test.

Covariate selection using penalized regression analysis in mill mud data analysis

Tuesday
Posters
17:10

Muyiwa Olayemi* and Joanne Stringer
molayemi@sugarresearch.com.au
Sugar Research Australia

Data from an experiment designed to investigate the effect of mill mud on grower revenue per tonne, cane yield per hectare, commercial cane sugar and sugar yield per hectare when combined with other beneficial by-product from the mills before its distribution over farms as an organic soil conditioner, an important source of plant nutrients were analysed. The data consist of five trials from the Herbert region, Queensland, Australia. Electrical conductivity (EC) mapping obtained at different depths of 0.5, 1.0, 1.6 and 3.2m (a trait known to be highly dependent on one another) were fitted as covariates for each trial. The variance inflation factor (VIF) indicates that there is multi-collinearity amongst these covariates. Therefore, fitting all the four covariates in the statistical model to analyse this data would be inefficient since inference for multivariable analysis assumes that all predictive variables are uncorrelated, which is not true for our data. A penalised regression approach using adaptive lasso with AIC as the tuning method was applied to the data for covariate selection, using proc glmselect of SAS. This approach placed a constraint on the size of the regression coefficients by shrinking them toward zero. Covariates with a zero coefficient were removed from the model. A mixed model was then fitted to the data with treatment and selected covariates as fixed and replicate as the random effects. The effect of treatment was observed in four trials for grower revenue, in three trials for cane yield and in two trials for commercial cane sugar. Covariates were not significant for all the measured traits in all trials except for grower revenue in one of the trials.

Tuesday
Posters
17:10

Bootstrapping F test for testing random effect in Linear Mixed model

Pauline O'Shaughnessy
poshaugh@uow.edu.au
University of Wollongong, Australia

Testing the significance of the random effects in the mixed models remains a crucial step in data analysis practise and a topic of research interest. Hui et al. (2018) revisited the F-test, which was originally proposed by Wald (1947), and generalized the application to test subsets of random effects in a mixed model framework. They allow correlation between random effects and showed that F test is an exact test when the first two moments of the random effects are specified.

We extended the investigation of the F test in the following two aspects: firstly, we examined the power of the F test under non-normality of the errors. Secondly, we consider a bootstrap counterpart to the F test, which could potentially offer improvement for the cases with small cluster size or for the cases with non-normal errors. This is inspired by the results shown in Hui et al. (2018) that bootstrap helps with correcting power and Type I error rate for the standard likelihood ratio test when the assumption of the test is violated.

References

Hui, FKC, Mueller, S & Welsh, AH. (2018). Testing random effects in linear mixed models: another look at the F-test. Australian and New Zealand Journal of Statistics.

Wald, A. (1947). A note on regression analysis. Annals of Mathematical Statistics. 18 (4), 586-589.

On comparison between two square tables using index of marginal inhomogeneity

Tuesday
Posters
17:10

Kouji Tahata
kouji_tahata@is.noda.tus.ac.jp
Tokyo University of Science, Japan

The present paper treats an issue of comparison between two square contingency tables with ordinal categories. For example, a contingency table for treatment group versus a contingency table for placebo group. An index, which measures the degree of departure from marginal homogeneity where the marginal homogeneity indicates the equality of two marginal distributions, is proposed. The index is expressed by using Matusita distance. Mathematical properties of the index are (1) the value of index lies between 0 and 1, (2) it equals 0 if and only if the marginal homogeneity holds, and (3) it equals 1 if and only if the degree of departure from the marginal homogeneity is the maximum under certain conditions. The difference between each of indexes for two contingency tables can be used to test the hypothesis that the degree of marginal inhomogeneity for one table is equal to that of marginal inhomogeneity for the other table. The proposed index is useful to compare between two contingency tables about the degree of departure from marginal homogeneity.

Tuesday
Posters
17:10

Examining the association between telomere length and periodontal attachment loss in the Dunedin Study

Jiaxu Zeng*, Murray Thomson, Jonathan Broadbent, Lyndie Foster Page, Idan Shalev, Terrie Moffitt, Avshalom Caspi, Sheila Williams, Antony Braithwaite, Stephen Robertson and Richie Poulton

jimmy.zeng@otago.ac.nz

University of Otago, New Zealand

Aim: Telomere length is a marker which is associated with both ageing and adverse exposures through life. There has been recent interest in studying the relationship between telomere length and periodontitis. This study was aimed to examine the association between telomere length and periodontitis in a long-standing prospective cohort study of New Zealand adults.

Materials and Methods: Periodontal attachment loss and telomere length data were collected at ages 26 and 38 in the Dunedin study. A generalised estimating equation (GEE) approach was used to examine the association between telomere length and attachment loss. A set of smoking exposure and sociodemographic variables was considered in this study as potential confounding variables.

Results: After controlling for the confounders, the mean telomere length among the participants reduced by 0.15 T/S ratio, and the mean periodontal attachment loss increased by 10% from age 26 to 38. However, we found no association between telomere length and periodontal attachment loss.

Conclusions: Although both periodontitis and telomere length are dependent, they do not appear to be linked, suggesting that determination of leucocyte telomere length may not be a promising clinical approach for identifying people who are at risk for periodontitis.

Abstracts - Wednesday Morning

Notices: 8:45

Design Tableau: An aid to specifying the linear mixed model for a comparative experiment

Wednesday
Invited
8:50-9:50

Alison Smith* and Brian Cullis

alismith@uow.edu.au

Centre for Bioinformatics and Biometrics, University of Wollongong, Australia

The design and analysis of comparative experiments has changed dramatically in recent times. There has been a move away from text-book designs towards complex, non-orthogonal designs. At the same time, analysis of variance (ANOVA) techniques have been superseded by the use of linear mixed models (LMM). The latter have the advantage of accommodating non-orthogonality and of allowing more complex variance models, which may be beneficial for improving the efficiency of treatment comparisons or for providing a more plausible structure. However, this flexibility has come at a cost, since in the transition from ANOVA to LMM, many practitioners overlook some of the fundamental principles in the analysis of comparative experiments. In order to address this we have developed "Design Tableau" (DT), which is a simple but general approach for specifying the LMM for a comparative experiment. The approach is based on the seminal work of John Nelder and Rosemary Bailey. It can accommodate a wide range of experiment types, including multi-environment trials, multi-phase experiments and experiments with longitudinal data. DT comprises a series of straight-forward steps aimed at removing the subjectivity in model specification and ensuring that the key principles in the analysis of comparative experiments are taken into consideration.

Biography

Alison has worked as a biometrician for more than 25 years and is currently an Associate Professor within the Centre for Bioinformatics and Biometrics at the University of Wollongong. Her main interest is the use of linear mixed models for the analysis of data from plant breeding and crop improvement programs. Her early work focussed on the analysis of genotype by environment interaction and the methods she developed are now used in all major plant breeding programs in Australia. Alison has also extensively researched improved methods of experimental design and analysis

for plant quality traits that require multi-phase testing. Most recently she has been involved in the development of the Design Tableau approach for specifying linear mixed models for comparative experiments. Alison has published over 50 refereed journal articles and has presented her research at a number of national and international statistical and scientific conferences. She has active links with industry, including most private and public plant breeding programs in Australia.

Efficient designs for early generation variety trials using genetic relationships

Wednesday
9:50-10:10

Nicole Cocks*, Alison Smith, David Butler and Brian Cullis
ncocks@uow.edu.au
University of Wollongong, Australia

Cullis et al. (2006) introduced a new class of designs for early generation variety trials known as p-rep designs. They showed that efficient p-rep designs could be obtained using a model-based approach as implemented in DiGGeR (Coombes, 2002). Using a simulation study, they showed that the realised genetic gain of p-rep designs was higher than that of comparable grid-plots designs.

P-rep designs have become widely used in many plant improvement programs in Australia for the design of early stage variety selection trials, however, most analyses of these trials follow the approach of Oakey et al. (2006), which incorporates genetic relatedness of test lines through the so-called numerator relationship matrix (NRM). The p-rep designs used for these analyses are still largely based on the original version of p-rep designs, which did not accommodate correlated test line effects.

Building on the work of Eccleston and Chan (1998), Butler et al. (2014) developed a model-based approach, which produces efficient (p-rep) designs for correlated treatment (i.e. test line) effects. Their approach can incorporate such through the NRM or the genetic relationship matrix if genotypic data is available.

In this paper we present the results of a simulation study designed to investigate the realised genetic gain of p-rep designs for a range of replication obtained under two scenarios, either using the uncorrelated or correlated test line effect approaches. The degree of partial replication for each design type is obtained by either considering a fixed number of plots and varying the number of test lines which are phenotyped, or by considering a fixed number of genotypes, selected in an optimal manner (see Huang et al., 2013) and varying the number of plots.

References

Butler, D.G., Smith, A.B. & Cullis, B.R. JABES (2014) 19: 539. https://doi.org/10.1007/s13253-014-0191-0

Coombes, N. (2002). The Reactive Tabu Search for Efficient Correlated Experimental Designs. PhD thesis, Liverpool John Moores University.

Cullis, B.R., Smith, A.B. & Coombes, N.E. JABES (2006) 11: 381. https://doi.org/10.1198/108571106X154443

Eccleston J., Chan B. (1998) Design Algorithms for Correlated Data. In: Payne R., Green P. (eds) COMPSTAT. Physica, Heidelberg.

Huang B. Emma, Clifford D. and Cavanagh, C. (2013). Selecting subsets of genotyped experimental populations for phenotyping to maximize genetic diversity. Theoretical and Applied Genetics 116, 379-388.

Oakey, H., Verbyla, A. P., Pitchford, W. S., Cullis, B. R., and Kuchel, H. (2006). Joint modeling of additive and non-additive genetic line effects in single field trials. Theoretical and Applied Genetics 113, 809–819.

Wednesday
10:10-10:30

Connectivity, does it impact genetic variance parameter estimates in a Multi Environment Trial analysis?

Chris Lisle*, Alison Smith, Carole Birrell and Brian Cullis

clisle@uow.edu.au

The University of Wollongong, Australia

Crop variety trial testing is conducted in Australia through the Grains Research and Development Corporation funded National Variety Trials (NVT). Annually, approximately 700 trials across eleven crops are established in locations across Australia. The key objective of these trials is to provide grain growers the best information to allow informed decisions on which varieties to grow in their conditions. To provide this information, one stage Multi Environment Trial (MET) analyses are performed using the methods of Gogel et al. (Euphytica, 2018). Here a Factor Analytic (FA) mixed model is used to model the genetic covariance between trials. However, there is concern that the accurate estimation of these FA parameters is impacted by the degree of varietal connectivity between trials.

In this talk, I consider simulation studies to inspect these effects over a range of varietal connectivity and scenarios. Here, the variance components to derive data were obtained using summaries of actual estimated parameters from four NVT MET analyses. In total, 36 datasets (scenarios) are simulated 5000 times, each with differing levels of varietal connectivity (poor to complete). Scenarios constitute the combination of four trial sizes (small to large) and a two-way factorial structure of low, medium, and high values of between trial genetic correlation and trial reliability. Trials were designed following the same principles to those in NVT. Presented in this talk are results including accuracy, bias, variance, and MSE.

Morning Tea: 10:30-11:00

Statistical issues arising from the Australian Royal Commission into Institutional Responses to Child Sexual Abuse

Wednesday
11:00-11:20

Graham Hepworth

hepworth@unimelb.edu.au

Statistical Consulting Centre, The University of Melbourne, Australia

One of the institutions investigated by the Royal Commission was concerned about the methodology used by the Commission to calculate the proportion of alleged perpetrators. One concern was the Commission's reporting of proportions which had been weighted by the length of time of an individual's membership. Longer-serving members of the institution were more likely to be perpetrators, resulting in a higher estimated proportion than the corresponding unweighted proportion. As a result, there was a concern that the headline figure would be misunderstood by many members of the public and the media.

Another concern related to members who had moved from part of the institution to another. It was suspected that such individuals had been counted more than once, resulting in an overestimate of the proportion.

These concerns were examined, and following correspondence with the Commission, some of their methodology was altered, resulting in a revision of their report.

Wednesday
11:20-11:40

A depth-adjusted Hardy-Weinberg test for low-depth sequencing data

Ken Dodds*, John McEwan, Timothy Bilton, Rudiger Brauning and Shannon Clarke

ken.dodds@agresearch.co.nz

AgResearch Ltd, New Zealand

The Hardy-Weinberg test is a test for association between the two alleles at a genotype. It is often used to detect population structure, selection and non-Mendelian inheritance. For genotypes derived from (low-depth) DNA sequencing there are two issues. Firstly, only one of an individual's alleles might have been observed. Secondly, these projects generate thousands or even millions of markers, and it is important to be able to screen these markers quickly. We have developed an adjustment to the Hardy-Weinberg goodness-of-fit test using inferred counts rather than observed counts. The method does not require optimisation and is therefore efficient. The theory and performance of the test are discussed.

Deprivation, hospital admissions and previous dental appointment records in Early Childhood: A comparative use of traditional statistical modelling with machine learning

Wednesday
11:40-12:00

Sarah Sonal*, Philip Schluter, Martin Lee and Jennifer Brown

sarah.sonal@gmail.com

University of Canterbury, New Zealand

Early Childhood Caries (ECC) are a preventable chronic disease that has been increasing in prevalence in New Zealand in recent years. This study utilizes routinely collected dental record data, hospital admission data and deprivation data from approximately 22,000 5 year olds in the Canterbury region from 2014 to 2017. The initial aim of the study is to identify potential contributing factors to ECC's. Secondly, the study aims to compare a statistical technique routinely used in health studies with a supervised machine learning technique, to establish which method of modeling is better for these data.

The best traditional statistical technique identified to model this dataset is a zero inflated negative binomial model. This is compared to the best supervised learning model identified, a random forest regression model. The data is split into 3 datasets, a training dataset to build the model, a tuning dataset to tune the best model of each type, and a testing dataset to compare the models on their predictive abilities.

Routinely collected dental data, hospital admission data and public deprivation data together are good predictors of Early Childhood Caries. Machine learning, specifically random forests, are a faster approach to modeling routinely collected dental data, with greater precision and accuracy.

Wednesday
12:00-12:20

Hockey sticks and broken sticks – a design for a single-treatment, placebo-controlled, double-blind, randomized clinical trial suitable for chronic diseasese

Hans Hockey* and Kristian Brock
hans@biometricsmatters.com
Biometrics Matters Ltd, Hamilton, New Zealand

This work is motivated and exemplified by a rare genetic disorder causing early onset diabetes and also blindness and deafness, which is extremely rare, inevitably fatal and has no current direct treatment. While the standard two-arm, placebo-controlled RCT is the gold standard required by the funding agency for a new proposed drug study, it is conjectured that potential study participants will prefer a design which guarantees that they are always assigned to the drug under study. A design is proposed which meets this patient need and hence probably increases recruitment and compliance. At the same time, it meets the requirement for full randomization. Analyses which follow naturally from this design are also described.

Lunch: 12:20-13:20

Optional Excursions: Afternoon

Abstracts - Thursday Morning

Notices: 8:45

Computing tools for a Don't Repeat Yourself data analysis workflow and reproducible research

Thursday
Invited
8:50-9:50

Peter Baker
p.baker1@uq.edu.au
School of Public Health, University of Queensland, Herston, Australia

Is there a crisis in reproducible research? Some studies such as Ioannidis et al. (2009) have estimated that over fifty percent of published papers in some fields of research are not reproducible.

The data analysis cycle starts a lot earlier than many researchers appreciate. Planning study design and organising workflow before the first data are collected is important. Complex data analysis projects often consist of many steps that may be repeated any number of times. Like many statistical consultants, I have often found myself repeating the same steps when analysing data for different projects. Standardising your approach, reusing statistical software syntax, writing your own functions or procedures (and even incorporating them into R packages when using GNU R) improves efficiency and saves time. Employing computing tools like GNU Make to regenerate output when syntax or data files change, GNU git for version control, writing GNU R functions or packages for repetitive tasks and R Markdown for reproducible reporting. In addition to providing tools and strategies for data analysis projects, these Don't Repeat Yourself (DRY) approaches also aid reproducible research and reporting.

Since the early 90s, I've employed version control systems and Make to project manage data analysis using GENSTAT, BUGS, SAS, R and other statistical packages. Also, as an early adopter of Sweave and R Markdown for reporting, I have found these approaches invaluable because, unlike the usual cut and paste approach, reports are reproducible. My overall strategy will be briefly described and illustrated. For GNU Make pattern rules, preliminary R packages and examples see https://github.com/petebaker.

Biography

Peter has worked as a statistical consultant and researcher in areas such as agricultural research, Bayesian methods for genetics, health, medical and epidemiological

studies for thirty years. He is a Senior Lecturer in Biostatistics at the School of Public Health, UQ where he also acts as a senior statistical collaborator and adviser to several research projects in the Faculty of Medicine.

Beyond statistical consulting: What role should statisticians play in ensuring good practice in the application and interpretation of statistics by disciplinary practitioners?

Thursday
9:50-10:10

Janet Chaseling*, Kyle James and Kirsty Wright
j.chaseling@griffith.edu.au
Griffith University, Australia

Determining the question of interest and ensuring correct and understandable interpretations of results, are fundamental tasks for statistical consultants. Students 'fear' studying statistics wanting only to pass; researchers regard statistics with respect but trepidation; the public see statistics as something that can be manipulated to prove anything. Why is 'statistics' feared, difficult and potentially misunderstood? How can, and should, statisticians redress the problem? Two real life situations are presented that demonstrate potential dangers when experts with little real statistical background embrace their own statistical needs and use them for decision making.

Following a ruling from the Queensland Anti-Discrimination Tribunal, the insurance company, Cerebus, were to provide: 'reasonable actuarial or statistical data', to justify the age discrimination practised by them in travel insurance premiums. Following statistical challenge of inadequate and manipulated data provided to the hearing before QCAT, Cerebus elected to settle out of court. The court-endorsed settlement included admission of the company's age discrimination and failure to provide adequate evidence for justification, together with a requirement that methodologies be reviewed, and revisions fully disclosed to the complainant.

Presenting statistics required for forensic DNA evidence in Australian courts has always been difficult. New statistical methodology is being actively pursued in which Bayesian models are used with priors based on 'guesstimates' rather than sound experimentation. Forensic scientists will rely on purchased software which produces a single likelihood ratio (LR) as evidence. There is little understanding of the methodology, or of effects of assumption violations and propagation of errors on the outcome. Preliminary surveys suggest that the LR is poorly understood by potential jury members; the question they want answered is: 'What is the probability of . . . '. If not challenged, this use of statistics may lead to the miscarriage of justice due to poor statistical knowledge and poor understanding of the evidential outcome.

Thursday
10:10-10:30

Answering the research question by identifying balanced embedded factorials in messy combined trials

Kerry Bell* and Rao Rachaputi

Kerry.Bell@daf.qld.gov.au

Department of Agriculture and Fisheries (DAF), Queensland, Australia

The aim of this presentation is to explore how the research questions influence the selection of embedded factorials chosen in combined trials with incomplete factorial structures.

The case study in this presentation is based on designed field trials of a pulse crop run over several seasons and sites as part of a grower invested GRDC project. The treatments chosen for each trial were influenced by client needs and emerging research questions identified by the project researchers. This resulted in a meta-data set with an incomplete factorial structure, with core factors of interest and supplementary factors in some trials. The levels of the core factors, such as variety, were also inconsistent across trials partly due to their suitability at particular sites.

The combined trial analyses can have a multitude of possible embedded complete factorials. The revisited research questions to identify specific embedded factorials for this case study were:

1. Were there yield advantages in using different row spacings (narrow, medium or wide) across environments, exploring each variety separately and sown at an industry standard plant density?

2. What was the yield response to actual plant density across environments (in a regression approach), exploring each variety separately at specific row spacings?

3. What was the predicted yield adjusted at the industry standard plant density for each environment, exploring each variety separately at specific row spacings? These predictions were then related to environmental descriptors such as starting moisture, in-crop water and strategic weather measurements.

These analyses from points 1 and 2 resulted in significant interactions with environment. For researchers to make recommendations to growers, the environments were then clustered together into groups that showed similar trends, so that factors or regression effects did not significantly interact with environment within a group. The relationship between clusters and environmental descriptors could then be explored.

Morning Tea: 10:30-11:00

Consulting in the real world: Communicating statistics to scientists

For this special session, there will be four 15 min talks followed by a discussion. To assist the discussion, we intend to use an internet tool (evpoll), which will allow anonymous comments, with voting on those comments. So, please bring along your smart phones, tablets or laptops!

"From cradle to grave": making an impact from conception to publishing

Thursday
Invited
11:00-11:15

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

Biological research has statistical aspects from initial conception through all stages to final publication and sometimes beyond. Good foundations – trial planning and design – are the bed-rock for quality science results, thus requiring good statistical input from the very beginning. Over my time as a biometrician, I have increasingly been involved at all stages of a project, developing approaches that facilitate that, leading to increases in efficiency and efficacy of the research through good statistical practice. I will present some approaches that I use, from trial design, to data management, through to presentation of results.

Biography
Ruth Butler has worked as a consulting biometrician since 1987, first in the UK at Long Ashton Research Station for seven years, and subsequently in New Zealand for Plant & Food Research and its preceding organizations. She is based near Christchurch. She primarily works with bio-protection scientists, but has worked with a range of mainly plant-based science disciplines across her career.

Thursday
Invited
11:15-11:30

Statistical inference and management decisions

Helene Thygesen
hthygesen@doc.govt.nz
Department of Conservation, New Zealand

When working for a bookmaker, the role of the statistical inference in the management decision process is fairly straight forward. We are asked to identify the strategy that maximizes expected profits – maybe taking risk averseness into account. Identifying similar objectives when working for government or health care organizations can feel like opening a can of worm – different stakeholders will have different objectives, and there will be legal or cultural barriers to non-standard statistical approaches. Nevertheless, I will argue that one should generally make an effort to operationalize management objectives. In this talk I will give examples from biosecurity, health care, criminal justice and wild life protection to illustrate how the statistician can try to adapt the type of statistical inference to the needs of different sectors.

Biography
Helene Thygesen studied mathematics in Copenhagen and received her Ph.D. in biostatistics from Amsterdam. She has worked as a consulting statistician for various research organizations in the UK, Netherlands and Denmark. She is currently principal statistical science adviser at the Department of Conservation, New Zealand. Helene's interests are primarily in modelling of biological processes and in making statistical inference relevant to decision processes.

The view from the other side: a biologist's view of communicating statistics

Thursday
Invited
11:30-11:45

Linley Jesson
linley.jesson@plantandfood.co.nz
Plant & Food Research, New Zealand

Biological research has changed immensely in the last twenty years and for most researchers handling and interpreting data has become their mainstay activities. Asking good questions in biology requires a fundamental understanding of the philosophy of science, of which statistics is the key underpinning. I will outline some fundamental statistical concepts that I argue all biologists should have a clear understanding of, and discuss ways in which we can increase the integration of statistics and biology.

Biography
Linley Jesson has come from a biological research background in evolution and ecology, and taught statistics to undergraduate Biology students for over 15 years. She joined Plant and Food Research in 2016 as a Biometrician and is currently group leader of the Data Science group.

Thursday
Invited
11.45-12:00

From heaven to hell... and how to find a way back!

Gabriela Borgognone
Gabriela.Borgognone@daf.qld.gov.au
Queensland Department of Agriculture and Fisheries, Australia

In many cases, the statistician is considered an integral part of the research team, who helps define clear objectives and ensures that correct statistical procedures are used and sound inferences are made. In many other cases, the statistician is seen as an outsider, a service provider, or someone that simply crunches the numbers at the end. Then, the challenge for the statistician is to slowly educate the researchers on the importance of statistics in science, earn their trust, and skillfully become an integral part of their projects.

Biography
Gabriela Borgognone has worked as statistical consultant at the Queensland Department of Agriculture and Fisheries, Australia, for the last 13 years. Throughout this time, she worked as part of state and national plant breeding programs. For the last two and a half years, she has mainly been working with the food science group of the Department. Before coming to Australia, she worked in Argentina for 10 years, both lecturing at university and as statistical consultant at the National Institute of Agricultural Technology.

Discussion: 12:00-12:20

Lunch: 12:20-13:20

Abstracts - Thursday Afternoon

Getting the most of my mixed model (and specially ASReml): applications in quantitative genetics and breeding

Thursday
Invited
13:20-14:20

Salvador Gezan
sgezan@ufl.edu
University of Florida (USA) and Trigen Consulting (Canada)

Quantitative genetic analyses have benefited from the increase in computational power and access to better and more efficient algorithms. Over the last few years, the increasing availability of large amounts of phenotypic and molecular data has allowed us to perform completer and more complex genetical analyses that exploit better the potential of the linear mixed model (LMM) framework with flexible variance-covariance structures. ASReml has been a particularly important tool to face many of these problems for many public and private breeding programs.

In this presentation, we will present several illustrations of the implementation of complex problems that require fitting LMM to support important operational decisions for many genetic improvement programs in agriculture, forestry and aquaculture. These cases will focus on the use and incorporation of large pedigree- and molecular-based relationship matrices for multi-trait and multi-environment analyses. Further extensions of the use of genomic selection (GS) on the estimation of additive and dominant effects also will be presented. Finally, the importance of considering these genetic relationships, in the context of design and analyses of experiments, is also illustrated in the context of augmented designs.

Biography

Salvador Gezan is statistician/quantitative geneticists with more than 20 years of experience in breeding, statistical analysis and genetic improvement consulting. He currently is an Associate Professor at the University of Florida, USA. He started his career at Rothamsted Research UK as a biometrician where he worked with GenStat and ASReml. Over the last 15 years he has taught for companies and university researchers many ASReml workshops around world

Dr. Gezan has worked on agronomy, aquaculture, forestry, entomology, medical, biological modelling, and with many commercial breeding programs applying traditional and molecular statistical tools. His research has led to more than 90 peer review publications, and he is one of the coauthors of the textbook "Statistical Methods in Biology: Design and Analysis of Experiments and Regression".

Thursday
14:20-14:40

Crown Rot Tolerance in Durum Wheat

Jess Meza*, Gururaj Kadkol, Steven Simpfendorfer, Steve Harden and Ky Mathews

jmeza@uow.edu.au

Centre for Bioinformatics & Biometrics, Australia

Crown rot (CR) is an important disease in durum wheat production in Australia and breeding for CR resistance/tolerance is desirable. While the best advances in bread wheat breeding have been achieved by screening breeding lines for their ability to yield under CR disease pressure, there is limited information regarding the extent of genetic variation for resistance/tolerance in durum breeding material. The aim of this research was to investigate genetic variation for CR resistance/tolerance within the elite material of the Northern Australian Durum Breeding programme.

Six trials were conducted annually at Tamworth between 2012 and 2017. Early trials had a two-level factorial treatment structure of genotypes by CR treatment regimes, while later trials had a split-plot treatment structure with CR treatment regimes allocated to main plots, and genotypes allocated to subplots. A one-stage Multi-Environment Trial (MET) analysis was conducted accounting for the three-way interaction of Genotype by Treatment by Environment (G×T×E), predicting genotype effects for each treatment level at each environment. Several models were tested, and it was determined that a Factor Analytic (FA) model applied to a factor indexing Treatments nested within Environments provided the most appropriate genetic covariance structure based on the Akaike Information Criterion (AIC). Tolerance is thus calculated for each genotype in each environment as the difference in Best Linear Unbiased Predictions (BLUPs) between the two CR treatment regimes. Due to the G×T×E structure, it was necessary to utilise the conditional distribution for a multivariate normal to calculate the matrix describing the regression between treatment regimes, and thus the lack-of-fit effects. Subsequently, it is possible to visualise how treatment level impacts varietal performance within each environment, and demonstrate genetic diversity due to the presence or absence of CR.

Don't be so negative about negative estimates of variance components

Thursday
14:40-15:00

Emi Tanaka*, Pauline O'Shaughnessy, Chong You and Chris Brien
emi.tanaka@sydney.edu.au
The University of Sydney, Australia

Linear mixed models are widely used across many disciplines due to its flexibility to model many complex data. The flexibility includes taking into account a hierarchical or correlated structure and the consideration of the different sources of variability. These complex structures can give rise to many different variance parameters. The estimation of variance components can be negative under certain methods and these have long been baffling to practitioners. These cases are sometimes unnoticeable due to the software's inherent restriction for variance components to remain positive. In such a case, it is often that the variance components are close to the zero boundary. The occurrence of negative variances is attributed to a range of reasons. We present examples, based on agricultural experiments, of the occurrences of negative estimates of variance components and the consequences of ignoring random effects with negative variance components with respect to the aim of the analysis.

Afternoon Tea: 15:00-15:30

Thursday
15:30-15:50

Bayesian Network as a Modelling Tool for Increasing Knowledge on the Factors Influencing Vineyard Longevity and Sustainability

Jelena Cosic*, Steffen Klaere, Matthew Goddard and Bruno Fedrizzi
j.cosic@auckland.ac.nz
The University of Auckland, New Zealand

The long-term project "Resilient and Profitable NZ wine industry" has the objective to study the impact of different vineyard management techniques on the vineyard longevity and profitability, and to increase the knowledge of the factors influencing longevity and profitability. To find meaningful answers appropriate quantifiable outcomes need to be obtained. Profitability of a vineyard can be quantified by its yield and quality of the end product, while health will be studied in a more holistic way by developing a vineyard ecosystems model incorporating the data obtained from different areas of interest. The empirical nature of data collection makes a computational ecosystem modelling approach the most suitable. Such approaches are quite common and popular in ecology and are promising for this project. Of particular interest are Bayesian Networks (BNs) which have received increased attention throughout several research fields for their ability to incorporate prior knowledge and to handle incomplete data. BN have also been shown to efficiently avoid overfitting the data and avoiding the observation of "chimeric" effects. We will use BN to model vineyard ecosystems incorporating microbial, fungal and eukaryotic molecular data, chemical profiles, meteorological information, and other markers at different points in the life cycle of vineyards and discover the differences vineyard managements make with respect to resilience and profit. Some of the challenges that we see are: variables that have been measured on different time scales, a large amount of microbial data and uncertainty of the interactions of components included in our ecosystem.

Statistical theory of rare event detection applied to forensic database establishment

Thursday
15:50-16:10

Kyle James*, Janet Chaseling, Kirsty Wright and Albert Gabric
kyle.james@griffithuni.edu.au
Griffith University, Australia

Ancestry analysis depends on assigning an individual to a population based on the presence or absence of specific genetic traits believed to be carried by individuals in that population. Previous data of varying sample sizes are used to determine the 'specific traits' relevant for a particular population. An important aspect which appears to receive little, if any, attention is the possibility of rare events. An individual may be assigned to Population A because they have a trait commonly seen in that population but not seen at all in Population B. However, the failure to observe this trait in Population B may simply reflect the sampling protocol used. The correct ancestry for this individual could be Population B.

In other research areas such as ecology, epidemiology and veterinary science considerable research on detecting rare events has been performed. A review found most methods to be unsuitable for forensic application, but methodology provided by Green and Young[1] is appropriate. If a rare event occurs at the rate of 1 in 200 (as seen in Poulsen[2]), then to be sure of a 95% power of detection, a minimum sample size of 600 is required. In forensic science there are studies using sample data with as few as fifty samples; these should be treated with caution. Other research in a range of areas including the physical, natural and social sciences stresses that increasing the number of variables (in forensic science the number of markers) will also increase the sample size needed for valid statistical analysis.

Many published studies in forensic science are made with samples which have low power to detect the rare events which could be present. The detection of rare events can lead to dire implications if ignored. A greater awareness of methods to account for rare events is required.

References

1: Green, R. H. and Young, R. C., 1993, "Sampling to detect rare species", Ecological Applications, 3, 351-356.

2: Poulsen, F. 2015, 'Construction of an Australian Y-haplogroup database to assist with ancestry identification of historical military remains', Bachelor of Forensic Science with Honours thesis, Griffith University.

Thursday
16:10-16:30

Estimating the Extent of Underreporting in Disease Counts

Rodelyn Jaksons*, Elena Moltchanova, Beverley Horn and Elaine Moriarty
rodelyn.avila@pg.canterbury.ac.nz
University of Canterbury, New Zealand

In disciplines such as epidemiology and ecology, researchers are often interested in estimating the true but unknown population size of interest. Often, we have count data associated with a region for which known and representative predictor(s) are available. However, the count data in many cases underreports the true size of the population. This problem is not new, and many researchers have come up with novel methods to estimate the size of the true population. In this presentation I will discuss how Bayesian Hierarchical models can be used to account for the underreporting and to estimate the true but unknown population size.

Finding your feet: modelling the batting abilities of cricketers using Gaussian processes

Thursday
16:30-16:50

Oliver Stevenson* and Brendon Brewer
o.stevenson@auckland.ac.nz
The University of Auckland, New Zealand

In the sport of cricket, variations in a player's batting ability can usually be measured on one of two scales. Short-term changes in ability that are observed during a single innings, and long-term changes that are witnessed between matches, over entire playing careers. To measure short-term, within-innings variation, a Bayesian survival analysis method is derived and used to fit a model which predicts how the batting abilities of international cricketers evolve during an innings. The results from the within-innings model provide evidence to support the cricketing belief of 'getting your eye in', whereby batsmen are more vulnerable early in their innings, but improve as they adapt to the specific match conditions. A second model is then fitted to explain how player batting ability changes between-innings, from match to match. To account for both recent performances and the element of randomness associated with cricket, the model uses a Gaussian process to measure and predict current and future batting abilities. Generally speaking, the results from the between-innings model support an anecdotal description of a typical sporting career. Young players tend to begin their careers with some raw but undeveloped ability, which improves over time as they gain experience and participate in specialised training and coaching regimes. Eventually players reach the peak of their careers, after which ability tends to decline. However, continual fluctuations in ability are commonly observed for many players, likely due to external factors such as injury and player form, which can lead to multiple peaks during a long career. The results provide more accurate quantifications of a player's batting ability at any given point of their career, compared with traditional cricketing metrics, and have practical implications in terms of player comparison, talent identification and team selection policy.

Thursday
16:50-17:10

Using Bayesian Growth Models to Predict Grape Yield

Rory Ellis*, Elena Moltchanova, Daniel Gerhard, Mike Trought
rorryaellis@gmail.com
University of Canterbury, New Zealand

For grape growers, it is necessary to be able to accurately predict the final yield. These predictions need to be made early in the growing season in order to make informed decisions about production and sales. In this study, we apply a double sigmoidal growth curve model within a Bayesian modelling framework to produce such estimates and assess the value of information. The latter refers to the tradeoff between making decisions early in the growing season using less accurate predictions vs. having more accurate predictions but at a later date.

The data was collected from Rowley Crescent vineyard in the Marlborough region for the 2016/2017 and 2017/2018 growing seasons. For each season, the data consists of 13 measurements of bunch weights for 15 apical and 15 basal bunches taken at weekly intervals.

The model was fitted using both vague priors and the so-called historical priors informed by the previous growing season using WinBUGS software. The accuracy was assessed via the posterior predictive probability of the final yield being within 20% of the actually observed one.

Our initial results show, that for the vague priors, the proportion of estimates within the pre-defined threshold was consistently around 20-25% in the first half of the growing season. This increased to 40% around the eighth week of the growing season, after which it levelled off. For historical priors, the results were less consistent. This highlights the importance of the choice of priors allowed and illustrates the importance of continuing research into feasibility of accurate early prediction. In the future, we will also consider the impact of climate factors on the yield estimates, as well as analyse data from other sites.

Conference Dinner: 17:30-22:00

Abstracts - Friday Morning

Notices: 9:00

Identifying hotspots of rat activity and how they affect the risk of leptospirosis in urban slums

Friday
9:10-9:30

Poppy Miller*, Chris Jewell, Peter Diggle and Kate Hacker
poppy.p.miller@gmail.com
AgResearch Ltd, New Zealand

Leptospirosis is much more prevalent in urban slums than in most other parts of the world. It is suspected that one of the key predictors of leptospirosis risk is exposure to rats. However, rodent control has so far been largely ineffective at reducing the burden of leptospirosis in urban slum environments where Norway rats (Rattus norvegicus) are the primary reservoir hosts.

Our study aims to quantify the risk attributable to rat exposure compared to other known risk factors in an urban slum. As part of this, we estimate a spatio-temporal rat activity surface over the study area which can be used to target rodent control more effectively.

The study design was a spatially continuous constrained random sample (340 points) at locations throughout a Brazilian urban slum community. An additional 100 random points were added at close range, to distinguish between short range spatial variation and underlying noise, allowing identification of hotspots at very small ranges. Tracking boards at each study location detect if a rat crosses them. All residents of the study area were asked to participate in the study, and their leptospirosis infection status was measured every 6 months.

At all points, we performed environmental surveys and used satellite imagery to derive spatially relevant covariates. We analyzed the rat activity data using a generalised linear spatial model to evaluate the association between rat activity and nearby environmental characteristics, and to create high-resolution predictive maps of relative rat activity/abundance. This allowed identification of defined environmental features of slum communities, which predict rat abundance.

We then predicted rat abundance near each study participants homes and used these predictions as a covariate in a generalised linear model where human leptospirosis infection status was the response, allowing quantification of the risk attributable to nearby rat abundance compared to other risk factors.

Friday
9:30-9:50

Developing methods to improve the accuracy of classification-based crowdsourcing

Julie Mugford*, Alex James and Elena Moltchanova
julie.mugford@gmail.com
University of Canterbury, New Zealand

Crowdsourcing is a widely used method to classify large amounts of images or objects. However, due to the openness of crowdsourcing, participants may contribute low quality responses. To improve the accuracy of classifications, multiple participants identify each object and consensus methods are used to decide the classification of the object. Commonly, simple consensus methods, e.g. majority vote, are used. However, majority vote weights the contributions from each participant equally but the participants may vary in accuracy with which they can label objects. We show that using Bayes' Rule to classify images based on participants responses, participants accuracies and relative frequencies of classes improves the accuracy of classifications compared to using majority vote. We show methods for estimating participants accuracies for varying levels of prior information about true image identities and participant characteristics.

A predictive model for nodal metastases among oral cancer patients

Friday
9:50-10:10

Ari Samaranayaka*, Rohana Kumara De Silva, B.S.M.S. Siriwardena, W.A.M.U.L. Abeyasinghe and W.M. Tilakaratne

ari.samaranayaka@otago.ac.nz

University of Otago, New Zealand

Oral squamous cell carcinomas (OSCCs) is the sixth leading cancer worldwide, accounting <5% of cancers in Europe, but about 45% of cancers in India. The most important prognostic factor is the presence of nodal metastases (spread of cancer to the neck lymph nodes) which associates with 50% reduction in survival rate. Nodal metastases are difficult to be diagnosed until late stages. Electronic scannings (CT/MRI/PET) are commonly used to diagnose, although their reliability is controversial given none of them can identify micro metastases (<1 mm). Current gold standard treatment for OSCC is the surgically removal of neck lymph node metastases with or without radiotherapy but identifying patients with metastases is challenging. Most surgeons now prefer operating patients only if the risk of nodal metastasis is above a threshold. This is because, surgery is unnecessary for three-quarters of patients as they do not have nodal metastases, and surgery creates a series of morbidity issues. A wait-and-rescan approach avoids neck dissection, but it raises risks of occult metastases. Also, routine scanning and rescanning is not practical in most developing countries due to lack of facilities.

Therefore, researchers tend to develop statistical models to predict metastasis using clinical, histopathological and molecular parameters related to cancers [17-20]. Relevant clinical parameters are not common universally. In European countries they include smoking, alcohol, and HPV, but chewing of tobacco, betel, and areca plays a major role in South Asian countries. Therefore, empirical models are not generalisable between different aetiological backgrounds.

We addressed this issue by developing an empirical model to predict the likelihood of nodal metastasis for patients in the South Asian region [1]. The model uses easily available clinical and histopathological parameters to rank OSCC patients on the risk of having nodal metastasis, allowing prioritising higher risk patients for surgeries using user-defined risk thresholds.

Reference: https://doi.org/10.1371/journal.pone.0201755

Friday
10:10-10:30

Analysing compositional time-use data in paediatric populations

Jillian Haszard*, Kim Meredith-Jones, Sheila Williams and Rachael Taylor
jill.haszard@otago.ac.nz
University of Otago, New Zealand

Understanding the influencing factors on the growth and health of children allows us to develop effective programmes and policies to protect and enhance their lives. Evidence to date suggests that physical activity and sleep are beneficial for the healthy growth of children whereas sedentary behaviour may be detrimental. However, much of this literature examines these behaviours in isolation or with inappropriate statistical analyses.

The different types of activity in a 24-hour period can be classified into: sleep, sedentary, light activity, moderate activity, and vigorous activity. As time spent in these components necessarily adds to 24 hours, they are interdependent. For example, if time spent asleep is increased, this must decrease time spent in one or more of the other activity levels. This is referred to as compositional time-use data. Statistical approaches for compositional data were originally developed for use in geology by John Aitchison and methods more recently refined for activity time-use data by several researchers including Sebastien Chastin and Dorothea Dumuid.

We have used these compositional data analysis methods to analyse actigraphy data from two large studies of New Zealand children. We examined associations with measures of body fat and also mental health indicators. Key difficulties in analysing this compositional data appropriately have included: normalising the data to 24 hours when there is missing time, presenting the results in a meaningful way for health researchers, and longitudinal analysis. In particular, we argue that when presenting estimates of association these should use the activity in question relative to all other activites, not relative to one single other activity. Without this, misleading conclusions can occur especially when the composition of the data is not evenly distributed.

Morning Tea: 10:30-11.00

50 Years of Genstat ANOVA

Friday
Invited
11:00-12:00

Roger Payne

roger.payne@vsni.co.uk

VSN International, Hemel Hempstead, United Kingdom

Graham Wilkinson's ANOVA algorithm has been a key part of Genstat ever since its early days. It was one of the motivations for the creation of Genstat. It was also the reason why I myself originally became involved with Genstat - initially to take up the responsibility for ANOVA ready for Graham's departure from Rothamsted in 1974. The algorithm provides a very efficient method of analysis that, even after more than 50 years, is unmatched elsewhere.

Analysis of variance and the associated design of experiments had been a Rothamsted speciality since the establishment of the Statistics Department under Sir Ronald Fisher in 1919. This was not a merely theoretical interest, but was motivated by the many experiments that needed to be designed, and then analysed, for the Rothamsted biologists. Fisher retired to Adelaide, and had a strong influence on Graham's statistical views. The Rothamsted connection was strengthened in 1965, when John Nelder visited the WAITE Institute in Adelaide, and began his collaboration with Graham. This laid the foundations for Genstat. Work on Genstat began in earnest, when John was appointed as head of the Rothamsted Statistics Department in 1968, and Graham joined the Department in 1971.

The original ANOVA algorithm was described by Wilkinson (1970), and its theoretical underpinnings by James & Wilkinson (1971). Payne & Wilkinson (1977) described the more efficient method for determining the structure of the design, that was my first task to get working when I took over. The relationship between the first-order balanced designs, that ANOVA analyses, and Nelder's (1965) general balance was explained by Payne & Tobias (1992), together with their algorithm that extended ANOVA, to estimate variance components and calculate estimates of treatment effects that combine information from all the strata in the design. Payne (2004) described how to obtain degrees of freedom for these combined effects.

The algorithm involves a sequence of sweeps in which effects are estimated, and then removed, from a working variate. There is also special sweep, known as a pivot, that projects the working variate into a specific stratum of the design. Matrix inversion is thus required only for the estimation of covariate regression coefficients and, as a result, the algorithm is very efficient in its use of workspace and computing time. Even 50 years on, this remains an important consideration.

References

James, A.T. & Wilkinson, G.N. (1971). Factorisation of the residual operator and canonical decomposition of non-orthogonal factors in analysis of variance. Biometrika, 58, 279-294.

Nelder, J.A. (1965). The analysis of randomized experiments with orthogonal block structure. I Block structure and the null analysis of variance. II Treatment structure and the general analysis of variance. Proceedings of the Royal Society, Series A, 283, 147-178.

Payne, R.W. & Wilkinson, G.N. (1977). A general algorithm for analysis of variance. Applied Statistics, 26, 251-260.

Payne, R.W. & Tobias, R.D. (1992). General balance, combination of information and the analysis of covariance. Scandinavian Journal of Statistics, 19, 3-23.

Payne, R.W. (2004). Confidence intervals and tests for contrasts between combined effects in generally balanced designs. COMPSTAT 2004 Proceedings in Computational Statistics, 1629-1636. Physica-Verlag, Heidelberg.

Wilkinson, G.N. (1970). A general recursive algorithm for analysis of variance. Biometrika, 57, 19-46.

Biography

Roger Payne is the Company Secretary at VSN, now working part-time after 15 years as its Chief Science and Technology Officer. He has a degree in Mathematics and a PhD in Mathematical Statistics from University of Cambridge, and is a Chartered Statistician of the Royal Statistical Society. He has led the development of Genstat since 1985, at Rothamsted prior to joining VSN. Roger was a statistical consultant and researcher at Rothamsted, becoming their expert on design and analysis of experiments, as well as leader of their statistical computing activities. His other statistical interests include generalized and hierarchical generalized linear models, linear mixed models, the study of efficient identification methods (with applications in particular to the identification of yeasts). Roger's statistical research has resulted in 9 books with commercial publishers, as well as over 100 scientific papers. He has a visiting professorship at Liverpool John Moores University, and also retains an honorary position at Rothamsted, to help him keep in touch with practical statistics.

Awards and Farewell: 12:00-12:10

Lunch: 12:10-13:10

Instructions for Presenters

Posters

Please hand in your poster at the AASC18 registration desk on Monday 3 Dec. Posters will be on prominent display throughout the conference, so please don't remove your poster before Friday lunchtime.

On the evening of Tuesday 4 Dec there will be a special poster event during which you are invited to give a 1 minute lightning presentation. Should you wish to present slides, give them to the registration desk well in advance of the poster session.

If you want to make an electronic copy of your poster available on the conference website, email it to Vanessa Cave (vanessa.cave@agresearch.co.nz).

Contributed Talks

Please provide a copy of your talk to the AASC18 registration desk well before your time to present. Each contributed talk has been allocated 20 minutes. You're asked to speak for a maximum of 15 minutes and allow 5 minutes for discussion.

Presentations will be made available on the AASC18 website following the conference. Should you not want your slides posted on the website, please inform Vanessa Cave (vanessa.cave@agresearch.co.nz).

Social Events

Welcome Reception

Monday 3rd
17:30 - 18:30, including a pōwhiri at 17:45
Poolside, Millennium Hotel

Poster Session and Drinks

Tuesday 4th
17:10 - 18:30
Millennium Hotel

Conference Dinner

Thursday 6th
17:30 - 22:00
Skyline Rotorua

The conference dinner will be held at Skyline Rotorua on the evening of Thursday 6th December. The dinner is included as part of registration, however additional tickets may be purchased ($135, see website).

The evening will begin with a gondola ride to the Skyline Rotorua complex, followed by welcome drinks (1 hr) and a sumptuous buffet. Transport to and from the venue will be provided.

For those wishing to enjoy the activities Skyline has on offer (luge, zipline, skyswing), there will be the option of an early arrival. Note: all activities are additional and will need to be booked and paid for upon arrival at Skyline.

Tentative Schedule
17:30: First bus
17:45-18:30: Optional luging ($29.90 for 3 rides) and sightseeing
18:00: Second bus
18:30-19:30: Happy hour drinks
19:30: Buffet
21:30: First bus
22:00: Second bus

Optional Excursions

Wednesday 5th
For more information and booking enquiries, contact Vanessa Cave.

Rotorua Walking Tour
Departing: 13:30

During this leisurely walking tour, you will visit some of Rotorua's most scenic points of interests, including Motutara (Sulphur Bay), the lakefront promenade, and the Sanatorium reserve. The walk will allow you to enjoy the city's geothermal areas, native wildlife reserves and local historic sites. For those feeling more energetic, there is also the option of completing the 26km Rotorua Walkway.

Cost: Free
Duration: 2-4 hrs
Bring: walking shoes, sunhat, sunscreen, rain jacket, water bottle

Wai-O-Tapu Thermal Wonderland
Departing: 13:00

Sculptured out of volcanic activity and thousands of years in the making, Wai-O-Tapu Thermal Wonderland is considered to be New Zealand's most colourful and diverse geothermal sightseeing attraction. During this excursion, you will see unique volcanic features as you explore the park's well-defined walking tracks.

Cost: $55
Duration: 4-5 hrs
Bring: walking shoes, sunhat, sunscreen, rain jacket, water bottle

Kaitiaki White Water Rafting
Departing: 14:15

This white water rafting excursion will take you on an epic journey down the Kaituna river, through 14 awesome rapids and over 3 waterfalls; including the 7m Tutea Falls – the largest commercially rafted waterfall in the world! Infused with touches of Maori culture, this trip offers it all, for both white water rafting enthusiasts and first timers. No experience is necessary – comprehensive training is provided.

Cost: $85
Duration: 3 hrs
Bring: swimwear, towel

Canopy Tours Ziplining
Departing: 13:30

Considered by TripAdvisor users as an 'Adventure activity with a brain,' this is Rotorua's number 1 ranked outdoor activity! During this excursion, you will explore ancient native forest on a network of ziplines, swingbridges and treetop platforms. There are 6 ziplines, the highlight for most being the jaw-dropping 220m tui song zipline that departs from 22 metres up a 1000 year old tree. The tour also includes 2 treetop swingbridges and numerous treetop platforms at varying heights above the forest floor. During the excursion, knowledgeable guides will share fascinating conservation and eco-tour information. Note: You must weigh no more than 120kg.

Cost: $139
Duration: 3 hrs
Bring: appropriate/comfortable clothing, close-toes shoes, warm layers

Delegates

Aswi, Aswi	Queensland University of Technology
Auld, Chris	Microsoft
Baird, David	VSN NZ Ltd
Baker, Peter	The University of Queensland
Bell, Kerry	DAF, Queensland
Borgognone, Gabriela	Queensland Department of Agriculture and Fisheries
Brien, Chris	University of South Australia/University of Adelaide
Butler, Kym	Agriculture Victoria Research
Butler, Ruth	Plant & Food Research
Cameron, Claire	University of Otago
Cameron, Catherine	AgResearch
Cave, Vanessa	AgResearch
Chaseling, Janet	Griffith University
Cocks, Nicole	University of Wollongong
Cosic, Jelena	University of Auckland
Dodds, Ken	AgResearch
Ellis, Rory	University of Canterbury
Gezan, Salvador	University of Florida
Gilmour, Arthur	Private Consultant
Green, Peter	AgResearch
Guo, Lindy	Plant & Food Research
Haszard, Jill	University of Otago
Hea, Shen	AgResearch
Hepworth, Graham	The University of Melbourne
Hirsz, Malgorzata	University of Waikato
Hockey, Hans	Biometrics Matters Ltd
Hothorn, Ludwig	Leibniz University Hannover (retired)
Iosua, Ella	University of Otago
Jaksons, Peter	Plant & Food Research
Jaksons, Rodelyn	University of Canterbury & ESR
James, Kyle	Griffith University
Jesson, Linley	Plant & Food Research
Jewell, Nathaniel	The University of Adelaide
Jones-Todd, Charlotte	NIWA
Lafaye de Micheaux, Pierre	University of Sydney
Li, Lingyu	Victoria University of Wellington
Liquet, Benoit	University of Pau Et Pays De L'adour
Lisle, Chris	University of Wollongong
Luckman, Maree	Fonterra
Luo, Dongwen	AgResearch
Ma, Donghui	Biosci Thailand
Maclean, Paul	AgResearch
McKenzie, Catherine	Plant & Food Research
McLachlan, Andrew	Plant & Food Research

Meza, Jess	Centre for Bioinformatics and Biometrics
Miller, Poppy	AgResearch
Mugford, Julie	University of Canterbury
Mumford, Michael	DAF, Queensland
Murray, Darren	VSNi
Olayemi, Muyiwa	Sugar Research Australia
O'Malley, James	Geisel School of Medicine at Dartmouth
O'Shaughnessy, Pauline	University of Wollongong
Partington, Debra	Agriculture Victoria Research
Payne, Roger	VSNi
Samaranayaka, Ari	University of Otago
Shepherd, Daisy	University of Auckland
Smith, Alison	University of Wollongong
Sonal, Sarah	University of Canterbury
Staincliffe, Maryann	AgResearch
Stevenson, Oliver	University of Auckland
Tahata, Kouji	Tokyo University of Science
Tanaka, Emi	University of Sydney
Taylor, Julian	University of Adelaide
Thomasen, Lisa	Fonterra
Thompson, Robin	Rothamstead Research
Thygesen, Helene	Department of Conservation
Triggs, Chris	University of Auckland
Turner, Robin	University of Otago
Wohlers, Mark	Plant & Food Research
Yamamoto, Kouji	Yokohama City University
Zeng, Jimmy (Jiaxu)	University of Otago

Index of Authors