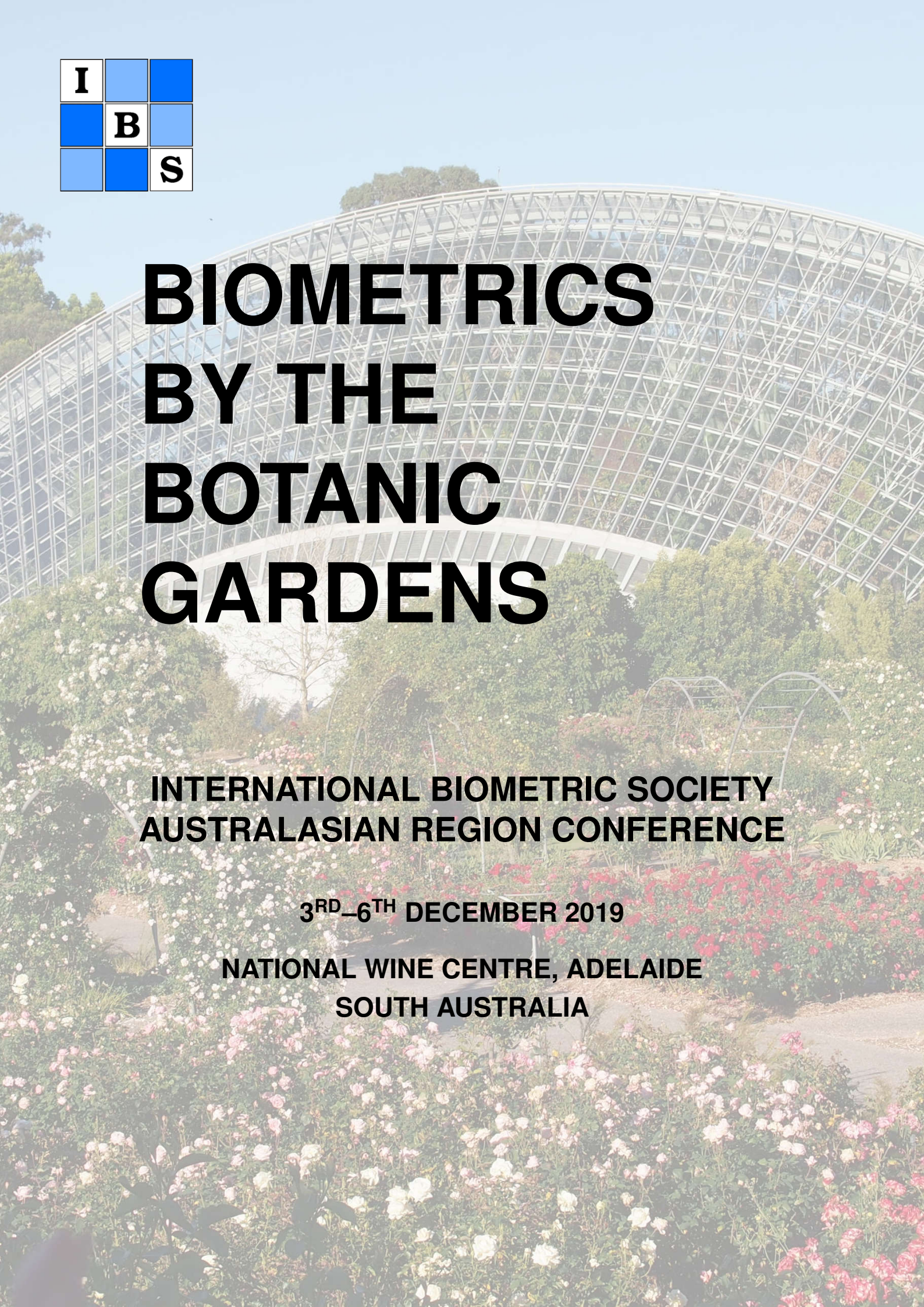


BIOMETRICS BY THE BOTANIC GARDENS

INTERNATIONAL BIOMETRIC SOCIETY AUSTRALASIAN REGION CONFERENCE

3RD–6TH DECEMBER 2019

NATIONAL WINE CENTRE, ADELAIDE
SOUTH AUSTRALIA

Contents

Biometrics by the Botanic Gardens IBS Conference 2019 **7**

Venue Map **15**

Conference Schedule **18**
Tuesday 23
Wednesday 25
Thursday 26
Friday 28

Talk Abstracts **29**
Multilevel multiple imputation for health and survey data: your flexible (and robust) friend
James Carpenter 29
Hockey sticks and broken sticks continued - enhancing the gold standard RCT for chronic diseases
Hans Hockey 30
Towards causal inference for spatio-temporal data: adjusting for time-invariant latent confounders
Rune Christiansen 31
From the stepped wedge to the staircase: the information content of stepped wedge trials
Jessica Kasza 32
Spatial Confounding in GEEs - Why the Working Correlation Matters (well, sort of)
Francis Hui 33
Estimation of proportions by group testing with retesting of positive groups
Graham Hepworth 34
Modelling species communities using presence-only data
Ian Renner 35
The effect of number of clusters and cluster sizes on multiple imputation in multilevel models
Nidhi Menon 36
Fitting species distribution models with uncertain species labels using point process models
Emy Guilbault 37

Multiple imputation for missing outcome data in trials involving independent and paired observations
Thomas Sullivan 38
`clustglm` and `clustord`: R packages for clustering with covariates for binary, count, and ordinal data
Louise McMillan 39
Pursuing the cancer-schizophrenia disassociation paradox: genomes, phenomes and intimate conversations with inconclusive evidence
Max Moldovan 40
Dependent selection and observation schemes in life history studies
Richard Cook 41
Error variance bias in neighbour balance and evenness of distribution designs
Emlyn Williams 42
Prof Ken Russell and the Design of Experiments for Generalized Linear Models
David Steel 43
Designs with many singly replicated treatments
Richard Jarrett 44
Confidence intervals centred on bootstrap smoothed estimators
Paul Kabaila 45
Performance of factor analytic models in multi-environment trial data with small variety numbers
Bethany Macdonald 46
Towards distributional analyses of biomarker data
Fernando Marmolejo-Ramos 47
An evaluation of separable variance structures for highly genetically correlated environments
Bethany Rognoni 48
Simulation-Selection-Extrapolation: Estimation in High Dimensional Errors-in-Variables Models
Linh Nghiem 49
Multi environment trial analysis to determine the tolerance of cereal varieties to cyst nematode
Isabel Munoz-Santa 50
Trimmed Estimators - and a Hybrid-Censored Data Approach to Estimation
Brenton R Clarke 51

An across trials random regression approach to describe tolerance of wheat cultivars to disease
Clayton Forknall 52
Robust scale estimation under small measurement errors
Michael Stewart 53
Clustering environments in a combined trial analysis of yield response to plant population
Michael Mumford 54
Using Deep Neural Models to Facilitate Statistical Modeling of Complex Spatio-Temporal Dynamics
Christopher K. Wikle 55
Quasi-random spatially balanced sampling
Blair Robertson 56
A new Liu estimator under a linear constraint
Takeshi Kurosawa 57
Big Biometric Data: A Biostatistical Perspective
Susan Wilson 58
Model selection and principle of parsimony in statistical modelling in agriculture
Zhanglong Cao 59
Bridging the gap between science and statistics
Teresa Neeman 60
On the effect of dependencies between regressors and random effects when analysing hierarchical structured data
Hwan-Jin Yoon 61
Statistics for agronomists: a constructive synthesis of workplace learning and community of practice
Sharon Nielsen 62
Modelling and Parameterization of Soil-Water Retention Curves
Warren Muller 63
Uncertainty in digital agriculture: An interdisciplinary perspective
Esther Meenken 64
Partial automation of sampling and data collection to yield better estimates for faba and canola seed emergence project
Peter Kazprzak 65
Vine copulas and health applications
Claudia Czado 66
Copula models for ecological community data
Marti J. Anderson 67

Determination of Indirect Reference Intervals for Immunoglobulin in an Australian Population
Alice Richardson 68
Statistical Efficiency of Distance Sampling
Robert Clark 69
The Propensity Score with Semi-continuous Exposures
Tugba Akkaya-Hocagil 70
Being Random and Efficient for Transect-Based Ecological Surveys
Scott Foster 71
Compositional Rank Methods for Testing for Differential Abundance in Microbiome Studies
Olivier Thas 72
Two-Stage Cluster Samples with Judgment Post-Stratification
Omer Ozturk 73
Comparison of Methods for the Detection of Outlier and Associated Biomarker in Mislabeled Omics Data
Tong Wang 74
Optimal design for monitoring groundwater quality using geostatistical modelling
Dan Gladish 75
Multiclass Hexbin Plots
Thomas Lumley 76
Estimation of mutation rate matrices from genomic site frequency data
Conrad Burden 77
Visualisation of model stability information for better prognosis based feature extraction
Connor James Smith 78
A multi-locus variable selection strategy for association mapping analysis
Beata Sznajder 79
mcvis: A new framework for collinearity discovery, diagnostic and visualization
Kevin Wang 80
Variable selection for genome wide association analysis in plant breeding
Anabel Forte 81
`Vizumap`: An R package for visualizing uncertainty in spatial data
Petra Kuhnert 82
Genetic dissection of phytophthora root rot resistance in chickpea using modern statistical methods
Julian Taylor 83

Improving Estimates of Fried's Index from Mating Competitiveness Experiments
Dan Pagendam 84
Visualization and measures of population differentiation based on the saddlepoint approximation
Louise McMillan 85
Generalised latent variable models for multivariate abundances in ecology
David Warton 86
Opportunities from combining statistical and crop growth models to predict Genotype × Environment interations over time
Daniela Bustos-Korts 87
Statistical Methods for data from High Throughput Phenotyping in Agriculture
Joanne De Faveri 88

Poster Abstracts **89**
Cross-validation of pasture biomass predictions from handheld NDVI sensor measures
Angela Anderson 89
Bayesian Approaches to Carbon Cycle Modelling
Mohammad Javad Davoudabadi 90
Analysis of continuous water use data for wheat plants grown on a Droughtspotter platform
Nathaniel Jewell 91
Comparative study of probability models for agriculture field index data
Olena Kravchuk 92
Fitting a nugget variance to the analysis of National Variety Trials
Chris Lisle 93
Yield response curves using fixed effects vs random coefficient regression across environments
Yao Lu 94
Multivariate analysis for Greenhouse Gas Emissions from New Zealand Sheep and Beef Farm Systems
Esther Meenken 95
Is the SpATS model as good as we would like it to be for the spatial analysis of field trials?
Lucas Peitton 96
Computation of the expected value of a function of a chi-distributed random variable
Nishika Ranathunga 97

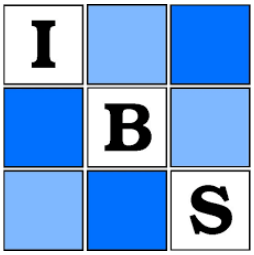
A comparison of linear mixed model packages in R for analysis of plant breeding experiments
Sam Rogers 98
Exploration of trait relationships in mungbean using a multivariate linear mixed model
Eugenia Settecase 99
Random thoughts on p-values
Alan Welsh 100
NIRS Classification
Carole Wright 101

Abstracts Index **102**

Biometrics by the Botanic Gardens
IBS Conference 2019

The International Biometric Society proudly welcomes you to the Biometrics by the Botanic Gardens Conference for 2019! More details about this conference may be found on the conference website, available at `ausbiometric2019.org`.

The International Biometric Society



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

Conference Organising Committee

Chair of Conference Organising Committee: Prof. Alan Welsh
Chair of Scientific Committee: Dr Petra Kuhnert
Chair of Local Organising Committee: Dr Olena Kravchuk
Treasurer: Mr Warren Muller
Website Developer: Mr Sam Rogers

Scientific Committee

Chair of Scientific Committee: Dr Petra Kuhnert
Dr Scott Foster, Dr Emi Tanaka, Mr Clayton Forknall, Dr Kevin Murray

Local Organising Committee

Chair of Local Organising Committee: Dr Olena Kravchuk
Conference Dinner Organisation: Dr Helena Oakey
Abstract Booklet Design: Mr Russell Edson
Mrs Wendy Li, Mrs Lisa Dansie, Mr Peter Kasprzak, Ms Annie Conway, and other members of the Biometry Hub, School of Agriculture, Food and Wine, University of Adelaide

Conference IBS-AR Code of Conduct Safety Officers

Dr Emi Tanaka, Dr Ian Renner, Dr Vanessa Cave

Invited Speakers

Marti Anderson
Massey University, New Zealand
m.j.anderson@massey.ac.nz

Distinguished Professor Marti J. Anderson is an ecological statistician whose work spans several disciplines, from ecology to mathematical statistics. A Fellow of the Royal Society of New Zealand, and a recent recipient of a prestigious James Cook Fellowship, she holds the Professorial Chair in Statistics in the New Zealand Institute for Advanced Study (NZIAS) at Massey University in Auckland. Her core research is in community ecology, biodiversity, multivariate analysis, models of ecological count data, experimental design and resampling methods, with a special focus on creating new applied statistics for ecology that can yield new insights into global patterns of biodiversity. Marti is also the Director of PRIMER-e (Quest Researcher Limited), a boutique research and software development company that creates user-friendly software (PRIMER and PERMANOVA+) to implement robust multivariate statistical methods for ecological analysis and synthesis.

Daniela Bustos-Korts
Wageningen University, The Netherlands
daniela.bustoskorts@wur.nl

Daniela Bustos-Korts works as a researcher at Biometris, Wageningen University (NL). She graduated with a BSc in Agriculture and an MSc in Crop Physiology from the Universidad Austral de Chile. In 2017, she obtained her PhD from Wageningen University. Her PhD thesis was about the use of statistical and crop growth models for phenotype prediction across multiple environments. Her research interests are in the development of strategies that combine statistical and crop growth models to help breeders designing effective phenotyping and prediction schemes that will help them improving response to selection. These strategies involve the use of crop growth models like APSIM and linear mixed models for multi-trait and multi-environment genomic prediction.

James Carpenter
London School of Hygiene & Tropical Medicine, UK
james.carpenter@lshtm.acuk

James Carpenter is professor of medical statistics at the London School of Hygiene & Tropical Medicine, and has a 50% secondment to the MRC Clinical Trials Unit at UCL. Research interests include: missing observations (both outcomes and covariates), in particular the method of multiple imputation and sensitivity analysis; meta-analysis; multilevel modelling and bootstrap methods, with social and medical applications.

Richard Cook
University of Waterloo, Canada
rjcook@uwaterloo.ca

Richard Cook is Professor of Statistics in the Department of Statistics and Actuarial Science at the University of Waterloo. His research interests include the analysis of life history data, the design and analysis of clinical and epidemiological studies, and statistical methods for the analysis of incomplete data. He collaborates extensively with researchers in rheumatology, transfusion medicine and cancer. He co-authored The Statistical Analysis of Recurrent Events (Springer, 2007) and Multistate Models for the Analysis of Lifetime History Data (Chapman and Hall, 2018) with Jerry Lawless and together they have given many short courses on these topics. Richard is a recipient of the CRM-SSC Prize from the Centre de recherches mathématiques (CRM) and the Statistical Society of Canada (SSC) in recognition for contributions within 15 years of a doctorate degree, and in 2018 he was awarded the Gold Medal of the Statistical Society of Canada.

Claudia Czado
Technical University of Munich, Germany
cczado@ma.tum.de

The research activities of Professor Claudia Czado are in statistics. Her special interests lie in the modeling of complex dependencies including regression effects and time/space structures. For this she uses a copula approach and especially the flexible class of vine copulas. This allows for different non-symmetric dependencies for pairs of variables. For model selection and estimation in high dimensions computer-aided methods are developed. Applications are in finance, insurance and engineering.

After studying mathematics in Göttingen, Germany, Professor Claudia Czado obtained her doctorate in 1989 at Cornell University, USA, in Operations Research and Industrial Engineering. She then became an Assistant Professor and 1995 Associate Professor of Statistics at York University, Toronto, Canada. In 1998 she was appointed to the Technical University of Munich, Germany, in the field of Applied Mathematical Statistics. Professor Claudia Czado is the author or co-author of more than 120 publications. She is also co-founder/coordinator of the junior research program "Global Challenges for Women in Math Science" at the Technical University of Munich.

Joanne De Faveri
CSIRO Data61 & SAGI-North/DAF, Australia
Joanne.DeFaveri@data61.csiro.au

Dr Joanne De Faveri is an applied statistician with interest in the application of statistical methods to agricultural research, especially plant breeding. She has worked most of her career in the Department of Agriculture and Fisheries as a Biometrician in Far North Queensland working with Horticulture breeding programs such as strawberry, pineapple, macadamia, mango and citrus. More recently she joined CSIRO as part of SAGI-North, the Statistics for the Australian Grains Industry project, where she develops and applies statistical methods to GRDC funded grains projects. She completed her PhD at the University of Adelaide in

"Spatial and Temporal Modelling for Perennial Crop Variety Selection Trials" and has research interests in Linear Mixed Models, spatio-temporal modelling in field trials, statistical genetics and more recently statistics for High Throughput Phenomics (HTP). Currently she is most interested in developing statistical methodology to best integrate HTP (including aerial image, sensor and hyperspectral) data into crop breeding programs for better variety predictions. She enjoys the collaborative side of statistics in agriculture and applying new ideas to get the most out of research data.

Max Moldovan
Registry of Senior Australians, Australia
max.moldovan@sahmri.com

Max obtained his PhD in computational statistics from the University of Melbourne in 2008, with his doctoral thesis dedicated to the design and implementation of novel exact inference procedures. Being passionate about identifying hidden functional patterns contained in empirical datasets, Max is experienced in applying statistical learning and network analyses to -omics and medical/healthcare records datasets.

Max is currently employed as a Senior Data Scientist by the Registry of Senior Australians (ROSA), focusing on the analysis of large complex empirical datasets possessed by ROSA and the partners.

Blair Robertson
University of Canterbury, New Zealand
blair.robertson@canterbury.ac.nz

Blair Robertson is a Senior Lecturer at the University of Canterbury, New Zealand. He obtained his PhD in mathematics at the University of Canterbury in 2011. Before taking his current position, he was an Assistant Professor at the University of Wyoming, United States. Blair was awarded the Worsley Early Career Research Award by the New Zealand Statistical Association in 2015 for outstanding research

from a statistician in the early stages of their career. His recent research has concentrated on spatially balanced sampling designs, classification trees and random search optimization.

Christopher K. Wikle
University of Missouri, USA
wiklec@missouri.edu

Christopher K. Wikle is Curators' Distinguished Professor and Chair of Statistics at the University of Missouri (MU), with additional appointments in Soil, Environmental and Atmospheric Sciences and the Truman School of Public Affairs. He received a PhD co-major in Statistics and Atmospheric Science in 1996 from Iowa State University. He was research fellow at the National Center for Atmospheric Research from 1996–1998, after which he joined the MU Department of Statistics. His research interests are in spatio-temporal statistics applied to environmental, ecological, geophysical, agricultural and federal survey applications, with particular interest in dynamics. His work has been concerned with formulating computationally efficient deep hierarchical Bayesian models motivated by scientific principles, with more recent work at the interface of deep neural models in machine learning. Awards include elected Fellow of the American Statistical Association (ASA), elected Fellow of the International Statistical Institute (ISI), Distinguished Alumni Award from the College of Liberal Arts and Sciences at Iowa State University, ASA Environmental (ENVR) Section Distinguished Achievement Award, co-awardee 2017 ASA Statistical Partnership Among Academe, Industry, and Government (SPAIG) Award, the MU Chancellor's Award for Outstanding Research and Creative Activity in the Physical and Mathematical Sciences, the Outstanding Graduate Faculty Award, and Outstanding Undergraduate Research Mentor Award. His book Statistics for Spatio-Temporal Data (co-authored with Noel Cressie) was the 2011 PROSE Award winner for excellence in the Mathematics Category by the Association of American Publishers and the 2013 DeGroot Prize winner from the International Society for Bayesian Analysis. His latest book, Spatio-Temporal Statistics with R, with Andrew Zammit-Mangion and Noel Cressie, was published in 2019 and is free to download at `spacetimewithR.org`. He is Associate Editor for several journals and is one of six inaugural members of the Statistics Board of Reviewing Editors for Science.

Acknowledgements

The Australasian Region of the International Biometric Society gratefully acknowledges the support offered by the following organisations to the Biometrics by the Botanic Gardens conference:

Grains Research & Development Corporation (GRDC)
grdc.com.au

CSIRO Australia
www.csiro.au

University of Adelaide Biometry Hub
ua.edu.au/biometryhub

Australian National University Research School of Finance, Actuarial Studies and Statistics
www.rsfas.anu.edu.au

ARC Centre of Excellence for Mathematical and Statistical Frontiers (ACEMS)
acems.org.au

University of Adelaide Waite Research Institute
www.adelaide.edu.au/
waite-research-institute

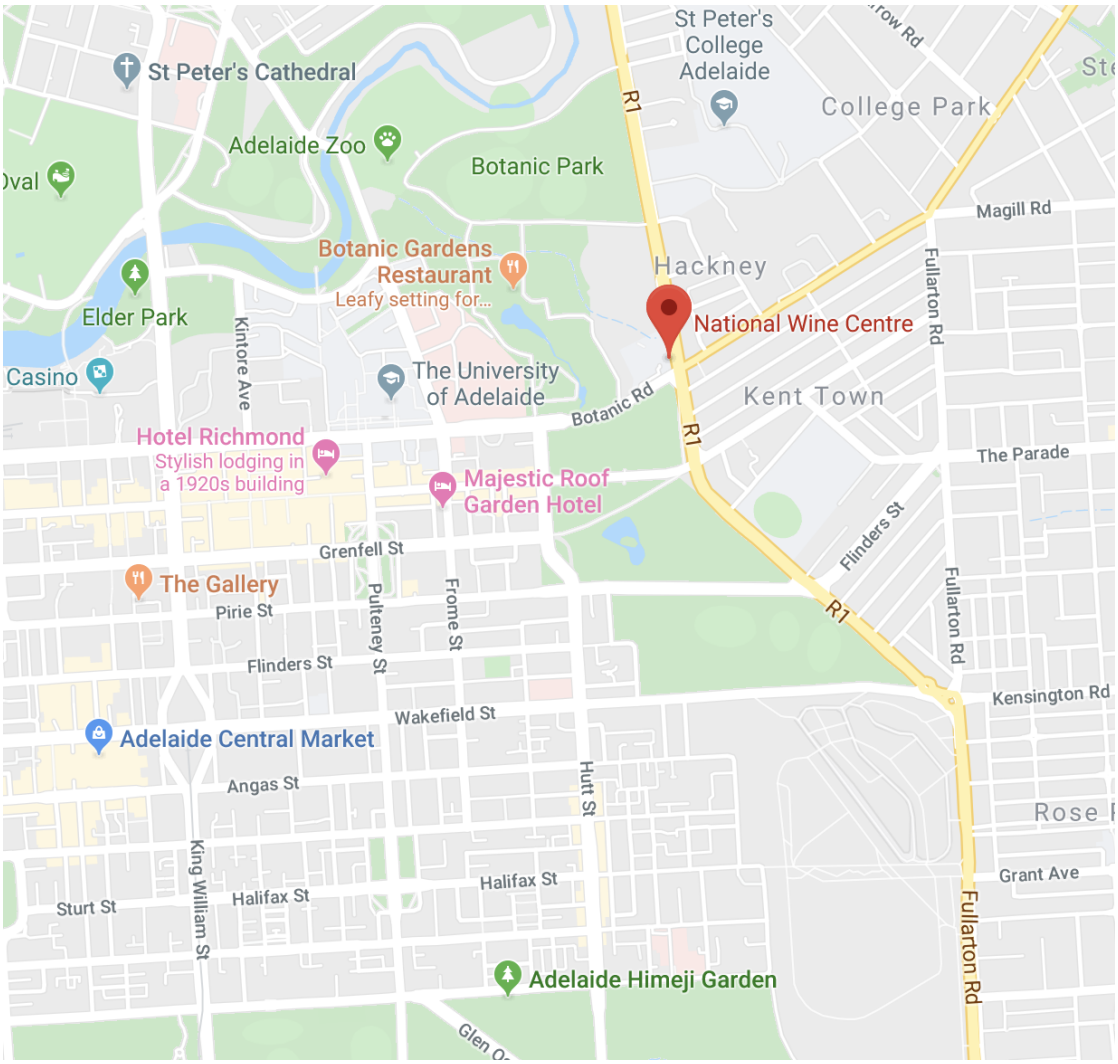
South Australian Health and Medical Research Institute (SAHMRI)
`www.sahmri.org`

Australian Grain Technologies
`www.agtbreeding.com.au`

Statistical Society of Australia Inc.
`www.statsoc.org.au`

Venue map

Our venue for this conference is the National Wine Centre in Adelaide, South Australia. The National Wine Centre is located on the corner of Botanic Road and Hackney Road, adjacent to the Adelaide Botanic Gardens.



More details about travelling around Adelaide and navigating to the conference venue are available on the website at `ausbiometric2019.org/travel/`.

National Wine Centre

The University of Adelaide's National Wine Centre provides important links with the wine industry, is an important platform for wine education and research, and provides a world class venue for food and wine. It contains an interactive permanent exhibition of winemaking, introducing visitors to the technology, varieties and styles of wine. It also has a wine tasting area, giving visitors the opportunity to taste and compare wines from different areas of Australia.

The National Wine Centre was recently awarded the below three categories—each for the second consecutive year.

- Best Wedding Caterer,
- Best Function/Convention Centre Caterer,
- Caterer of the Year.

Ground Floor

On the Ground floor are the **Vines Room** and the **Ferguson Room**.

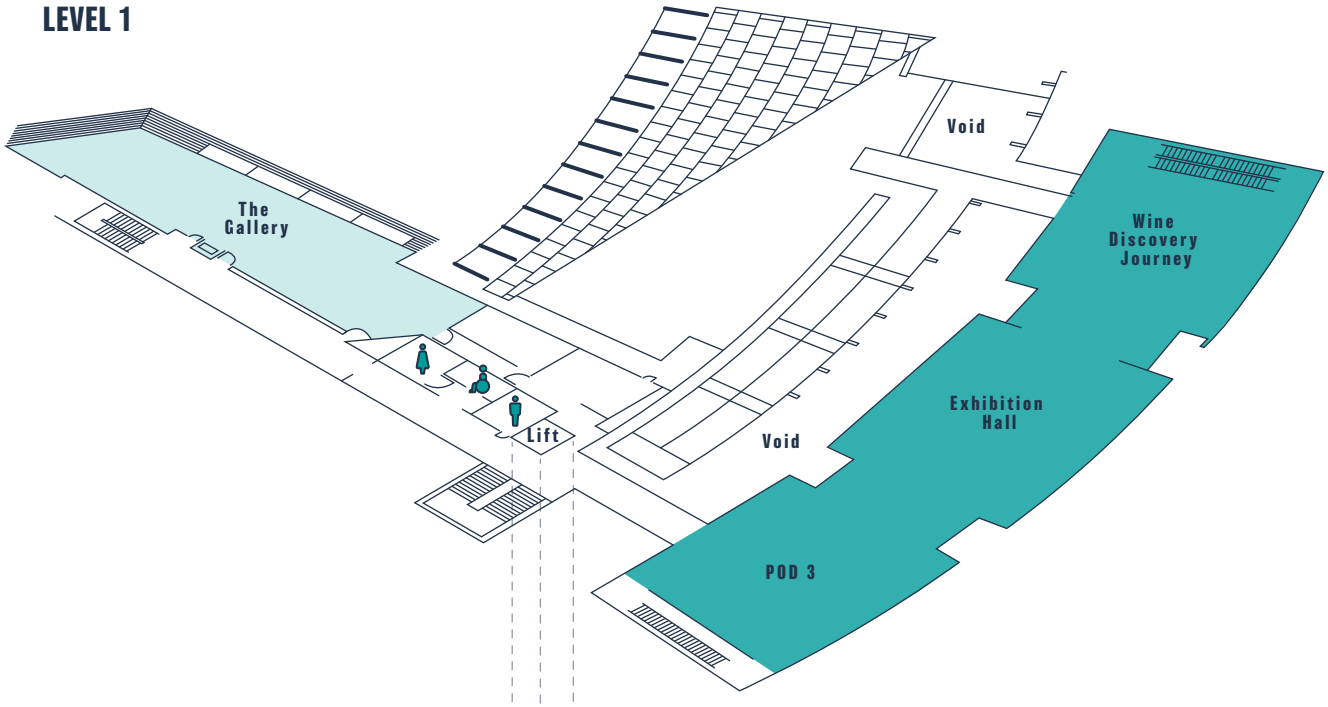
- The Vines Room will host one of the session streams for the Tuesday and Wednesday talks.
- The Annual General Meeting (AGM) will take place on Thursday in the Ferguson Room.

Level 1

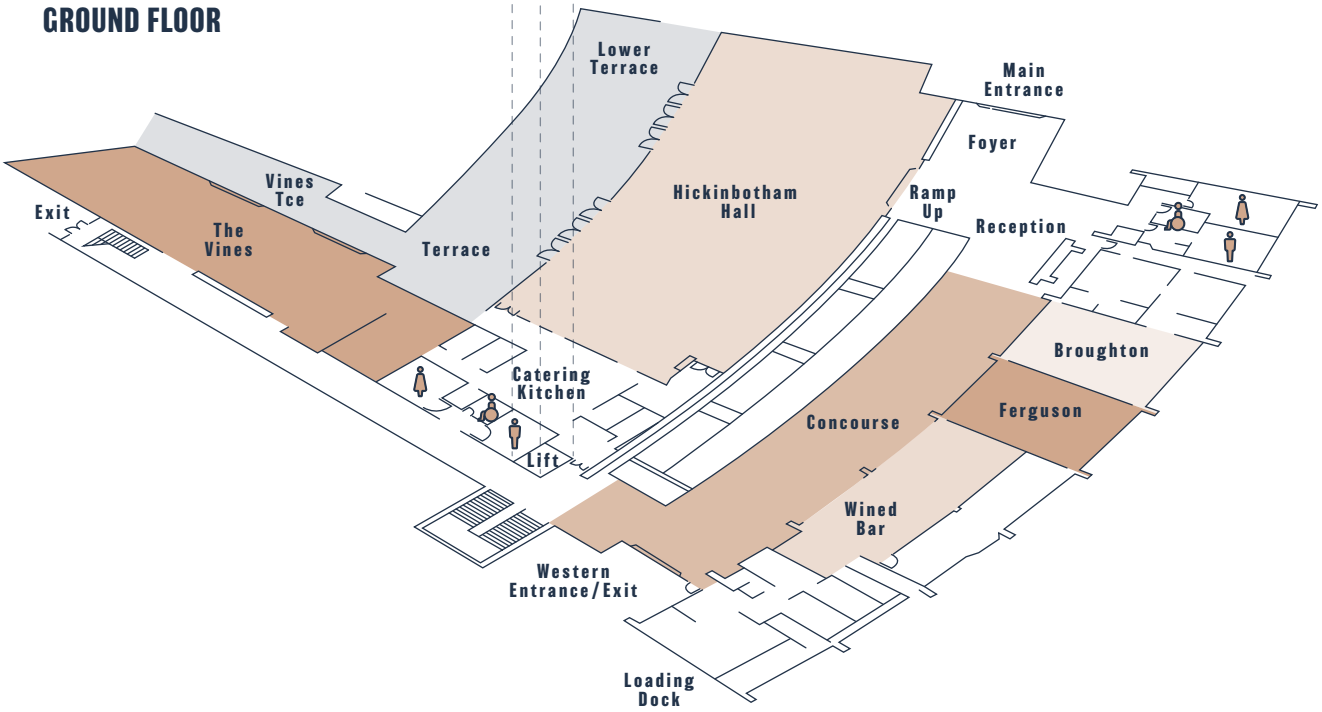
Access Level 1 by taking the lift up from the Ground floor. On Level 1 are the **Exhibition Hall** and **The Gallery**.

- The Opening address takes place in the Exhibition Hall. The Exhibition Hall will also host all of the Plenary Sessions, one of the session streams for the Tuesday, Wednesday and Thursday talks, and Morning Tea/Lunch/Afternoon Tea on each day.
- The Gallery will host one of the session streams for the Tuesday talks, and the Tuesday Poster session.

LEVEL 1



GROUND FLOOR



Conference Schedule

The conference schedule is detailed on the following pages. For convenience, the full conference schedule timetable is included in this booklet starting from page 23, followed by all of the abstracts listed in *chronological order*. The timetable and abstracts are also available on the website at `ausbiometric2019.org/outline`.

Welcome Reception

The conference welcome reception will be held from 7:00pm on Monday at the National Wine Centre. All registered conference participants are welcome to attend.

Registration

The Registration Desk opens at 6:30pm on Monday night at the National Wine Centre, preceding the Welcome Reception. On Tuesday to Friday, the Registration Desk at the National Wine Centre opens at 8:30am.

Contributed Talks

Each participant giving a contributed talk will be allotted 15 minutes for their presentation, followed by 5 minutes for audience questions and room change-over.

Poster Session

The Poster session will be held on Tuesday from 5:00–7:00pm in The Gallery. Snacks and drinks will be provided.

Young Statisticians Event

The Young Statisticians Event will take place on the Tuesday night from 7:30–9:30pm. The group will leave from the National Wine Centre at 7:00pm after the poster session and walk to the Escape Hunt Venue. The cost of attendance is free for Young Statisticians, and includes a welcome drink on arrival.

The Escape Hunts comprise two intriguing games...

Rescue a kidnapped scientist before his secret deadly formula is exploited!

It's 1941 and Australian scientist Howard Florey has been running clinical trials for penicillin. During his trials he created a penicillin-resistant bacteria. Florey has now been kidnapped and his deadly formula is at risk of falling into the wrong hands.

Enter his home, find the formula and destroy it before it's too late!

Locked in a shark cage on the ocean floor - can you escape the Great White Killer?

A dive crew has disappeared off the coast of Port Lincoln, and Captain Rocks claims they were taken by Great White Sharks. There are rumours that he and his crew were on the hunt for lost treasure, and you and your team of detectives think something smells fishy...

Under the guise of being his new crew, you are taking a cage dive with Captain Rocks to investigate. Are there really killer sharks on the loose, or is it something more sinister? Escape the shark cage and solve the case before the sharks come circling...

Contact: Mr Sam Rogers

Excursions

All Excursions are held on Wednesday afternoon, as per the timetable.

Sir Ronald Fisher in Adelaide
1:30–4:30pm, University of Adelaide + surrounds

A guided library tour, talk and afternoon tea in the Rare Collection of the Barr Smith Library to see the work by Fisher preserved in the library and learn about his last years in Adelaide, 1959–1962. This will follow by a short stroll from the

University through North Adelaide to St Peter's Cathedral and a guided tour there to visit the place of Sir Fisher ashes and hear about the significance of that for the Cathedral.

Contact: Ms Annie Conway
Cost: $20 per person

Indoor Bouldering
2:00–4:00pm, Adelaide Bouldering Club

Bouldering is a form of rock climbing that is performed on small rock formations or artificial rock walls, known as boulders, without the use of ropes or harnesses. While it can be done without any equipment, most climbers use climbing shoes to help secure footholds, chalk to keep their hands dry and provide a firmer grip, and bouldering mats to prevent injuries from falls. A 2-hour bouldering session will take place at the Adelaide Bouldering Club. No experience is required and professional instruction and demonstration will be provided.

Also, for the climbing equipment, we have:

- Bouldering shoes hire available
- Bouldering chalk available
- No harness required
- Gym clothes and water are required.

Contact: Mrs Wendy Li
Cost: $20 per person, including shoes hire and transport

Wine tasting in the Fleurieu Peninsula
12:45–6:30pm, McLaren Vale

A true journey of discovery into McLaren Vale that explores a mix of boutique producers and Australian icons, including the avant-garde Rubix Cube inspired, d'Arenberg Cube. The Cube is the most creative cellar door in the country, a combination of surrealist museum and cellar door all wrapped up in one, with

sweeping views of McLaren Vale from its top floor. At Samson Tall and Bekkers you'll meet the owners themselves, take a tour of their facilities and get intimate insights into the style of their winemaking.

Schedule:

12:45pm, Pick up CBD
Transfers in two Luxury Land Rover Discovery 4x4 vehicles.

1:30pm, Bekkers Fine Wine
Boutique, world class wines with finesse and texture, made by husband & wife team Toby & Emmanuelle Bekkers. Enjoy a private tasting with either Emmanuelle or Toby overlooking their McLaren Vale estate, and learn the secrets behind this most prestigious label.

2:30pm, d'Arenberg Cube
d'Arenberg is a founding member of Australia's First Families of Wine and the flamboyant five-story Cube complex erupts from the ground in the heart of its original McLaren Vale vineyard. Take a self-guided tour of the Alternate Realities Museum and a tasting in the sensational Cube Cellar Door.

3:45pm, Red Poles Art Gallery
Take a tour of their famous Indigenous Art Gallery, see the artworks of different aboriginal peoples and learn their stories.

4:15pm, Samson Tall
Visit Samson Tall Winery, situated in Bethany Chapel, the original vineyard site of Wirra Wirra's Church Block. Samson Tall epitomises the modern style of McLaren Vale, with a select range of single vineyard wines. Enjoy a tasting and winery tour with owner and winemaker Paul Wilson.

5:15pm, Silver Sands Beach Drive
Take a 4WD tour down Silver Sands Beach to see this famous strip of white, squeaky-fine sand, soaking up majestic views of the Fleurieu coastline and spectacular biodiversity.

6:30pm, Arrive back in CBD.

Contact: Mr Peter Kasprzak
Cost: $145 per person, includes all entrance fees, tickets and transfers

Conference Dinner and Awards Session

The conference dinner will be held on Thursday from 7:30pm until late at the Adelaide Hills Convention Centre in Hahndorf. Buses are scheduled to depart from 6:45pm from the National Wine Centre. The dinner venue is approximately a 30 minute drive into the Adelaide Hills. There will be two possible return times from the dinner venue at 10:30pm and 11:30pm, returning to the National Wine Centre.

JABES/Biometrics Showcase

On Friday morning there will be a showcase of talks for the Journal of Agricultural, Biological and Environmental Statistics (JABES) and the Biometrics journal.

More details about these journals may be found on the International Biometric Society website.

JABES: `www.biometricsociety.org/publications/jabes`
Biometrics: `www.biometricsociety.org/publications/biometrics`

<table>
<tr><th colspan="3">Tuesday 3 December</th></tr>
<tr><td>8:30–8:50am</td><td colspan="2">Registration Desk Open</td></tr>
<tr><td>8:50–9:10am</td><td colspan="2" rowspan="2">Welcome to Country
Lynette Crocker
Opening address (9:10–9:30am)
Presidential address (9:20–9:30am)
Alan Welsh
Exhibition Hall</td></tr>
<tr><td>9:10–9:30am</td></tr>
<tr><td>9:30–9:50am</td><td colspan="2" rowspan="3">Plenary Session (9:30–10:30am)
Exhibition Hall
Chair: Helena Oakey
Invited Talk
(Medical 1)
James Carpenter (p. 29)</td></tr>
<tr><td>9:50–10:10am</td></tr>
<tr><td>10:10–10:30am</td></tr>
<tr><td>10:30–11:00am</td><td colspan="2">Morning Tea (Exhibition Hall)</td></tr>
<tr><td rowspan="2"></td><td colspan="2">Contributed Session (11:00–12:40pm)</td></tr>
<tr><td>Biostatistics 1 & Imputation
(The Gallery)
Chair: Brenton Clarke</td><td>Modelling the Environment
(Exhibition Hall)
Chair: Ruth Butler</td></tr>
<tr><td>11:00–11:20am</td><td>Hans Hockey (p. 30)
(11:00–11:20am)</td><td>Rune Christiansen (p. 31)
(11:00–11:20am)</td></tr>
<tr><td>11:20–11:40am</td><td>Jessica Kasza (p. 32)
(11:20–11:40am)</td><td>Francis Hui (p. 33)
(11:20–11:40am)</td></tr>
<tr><td>11:40–12:00pm</td><td>Graham Hepworth (p. 34)
(11:40–12:00pm)</td><td>Ian Renner (p. 35)
(11:40–12:00pm)</td></tr>
<tr><td>12:00–12:20pm</td><td>Nidhi Menon (p. 36)
(12:00–12:20pm)</td><td>Emy Guilbault (p. 37)
(12:00–12:20pm)</td></tr>
<tr><td>12:20–12:40pm</td><td>Thomas Sullivan (p. 38)
(12:20–12:40pm)</td><td>Louise McMillan (p. 39)
(12:20–12:40pm)</td></tr>
<tr><td>12:40–1:30pm</td><td colspan="2">Lunch (Exhibition Hall)</td></tr>
</table>

Tuesday 3 December (continued)	
1:30–1:50pm	**Plenary Sessions** (1:30–3:30pm) *Exhibition Hall* Chair: Jessica Kasza
1:50–2:10pm	**Invited Talk** (Medical 2) Max Moldovan (p. 40) (1:30–2:30pm)
2:10–2:30pm	
2:30–2:50pm	*Exhibition Hall* **Invited Talk** (Medical 3) Richard Cook (p. 41) (2:30–3:30pm)
2:50–3:10pm	
3:10–3:30pm	
3:30–4:00pm	Afternoon Tea *(Exhibition Hall)*
	Contributed Session (4:00–5:00pm)
	Experimental Design *(Exhibition Hall)* Chair: Kathy Ruggiero
4:00–4:20pm	Emlyn Williams (p. 42) (4:00–4:20pm)
4:20–4:40pm	David Steel (p. 43) (4:20–4:40pm)
4:40–5:00pm	Richard Jarrett (p. 44) (4:40–5:00pm)
5:00–6:00pm	**Poster Session** (pp. 89–101) *The Gallery* Chair: Clayton Forknall (5:00–7:00pm)
6:00–7:00pm	
7:00–7:30pm	
7:30–9:30pm	**Young Statisticians Event** (7:30–9:30pm)

<table>
<tr><th colspan="3">Wednesday 4 December</th></tr>
<tr><td>8:30–8:50am</td><td colspan="2">Registration Desk Open</td></tr>
<tr><td rowspan="2"></td><td colspan="2">Contributed Session (8:50–10:30am)</td></tr>
<tr><td>Methods 1
(Vines Room)
Chair: Chris Triggs</td><td>Mixed Models in Agriculture
(Exhibition Hall)
Chair: Kaye Basford</td></tr>
<tr><td>8:50–9:10am</td><td>Paul Kabaila (p. 45) (8:50–9:10am)</td><td>Bethany Macdonald (p. 46) (8:50–9:10am)</td></tr>
<tr><td>9:10–9:30am</td><td>Fernando Marmolejo-Ramos (p. 47) (9:10–9:30am)</td><td>Bethany Rognoni (p. 48) (9:10–9:30am)</td></tr>
<tr><td>9:30–9:50am</td><td>Linh Nghiem (p. 49) (9:30–9:50am)</td><td>Isabel Munoz-Santa (p. 50) (9:30–9:50am)</td></tr>
<tr><td>9:50–10:10am</td><td>Brenton Clarke (p. 51) (9:50–10:10am)</td><td>Clayton Forknall (p. 52) (9:50–10:10am)</td></tr>
<tr><td>10:10–10:30am</td><td>Michael Stewart (p. 53) (10:10–10:30am)</td><td>Michael Mumford (p. 54) (10:10–10:30am)</td></tr>
<tr><td>10:30–11:00am</td><td colspan="2">Morning Tea (Exhibition Hall)</td></tr>
<tr><td>11:00–11:20am</td><td colspan="2" rowspan="3">Plenary Sessions (11:00–1:00pm)
Exhibition Hall
Chair: Petra Kuhnert
Invited Talk
(Environmental 1)
Christopher Wikle (p. 55)
(11:00–12:00pm)</td></tr>
<tr><td>11:20–11:40am</td></tr>
<tr><td>11:40–12:00pm</td></tr>
<tr><td>12:00–12:20pm</td><td colspan="2" rowspan="3">Exhibition Hall
Invited Talk
(Environmental 2)
Blair Robertson (p. 56)
(12:00–1:00pm)</td></tr>
<tr><td>12:20–12:40pm</td></tr>
<tr><td>12:40–1:00pm</td></tr>
<tr><td>1:00–1:30pm</td><td colspan="2">Packed Lunch (Collect from Exhibition Hall)</td></tr>
<tr><td>1:30–2:30pm</td><td colspan="2" rowspan="4">Excursions
(1:30–5:30pm)</td></tr>
<tr><td>2:30–3:30pm</td></tr>
<tr><td>3:30–4:30pm</td></tr>
<tr><td>4:30–5:30pm</td></tr>
</table>

<table>
<tr><th colspan="3">Thursday 5 December</th></tr>
<tr><td>8:30–8:50am</td><td colspan="2">Registration Desk Open</td></tr>
<tr><td rowspan="2"></td><td colspan="2">Contributed Session (8:50–10:30am)</td></tr>
<tr><td>Methods 2
(Vines Room)
Chair: Francis Hui</td><td>Collaboration
(Exhibition Hall)
Chair: Emi Tanaka</td></tr>
<tr><td>8:50–9:10am</td><td>Takeshi Kurosawa (p. 57)
(8:50–9:10am)</td><td>Susan Wilson (p. 58)
(8:50–9:10am)</td></tr>
<tr><td>9:10–9:30am</td><td>Zhanglong Cao (p. 59)
(9:10–9:30am)</td><td>Teresa Neeman (p. 60)
(9:10–9:30am)</td></tr>
<tr><td>9:30–9:50am</td><td>Hwan-Jin Yoon (p. 61)
(9:30–9:50am)</td><td>Sharon Nielsen (p. 62)
(9:30–9:50am)</td></tr>
<tr><td>9:50–10:10am</td><td>Warren Muller (p. 63)
(9:50–10:10am)</td><td>Esther Meenken (p. 64)
(9:50–10:10am)</td></tr>
<tr><td>10:10–10:30am</td><td></td><td>Peter Kasprzak (p. 65)
(10:10–10:30am)</td></tr>
<tr><td>10:30–11:00am</td><td colspan="2">Morning Tea (Exhibition Hall)</td></tr>
<tr><td>11:00–11:20am</td><td colspan="2" rowspan="3">Plenary Sessions (11:00–1:00pm)
Exhibition Hall
Chair: Samuel Mueller
Invited Talk
(Methods 1)
Claudia Czado (p. 66)
(11:00–12:00pm)</td></tr>
<tr><td>11:20–11:40am</td></tr>
<tr><td>11:40–12:00pm</td></tr>
<tr><td>12:00–12:20pm</td><td colspan="2" rowspan="3">Exhibition Hall
Invited Talk
(Methods 2)
Marti Anderson (p. 67)
(12:00–1:00pm)</td></tr>
<tr><td>12:20–12:40pm</td></tr>
<tr><td>12:40–1:00pm</td></tr>
<tr><td>1:00–2:10pm</td><td colspan="2">Lunch (Exhibition Hall) & AGM (Ferguson Room)</td></tr>
</table>

Thursday 5 December (continued)		
	Contributed Session (2:10–3:30pm)	
	Biostatistics 2 *(Vines Room)* Chair: Hans Hockey	**Samples & Surveys** *(Exhibition Hall)* Chair: Louise McMillan
2:10–2:30pm	Alice Richardson (p. 68) (2:10–2:30pm)	Robert Clark (p. 69) (2:10–2:30pm)
2:30–2:50pm	Tugba Akkaya-Hocagil (p. 70) (2:30–2:50pm)	Scott Foster (p. 71) (2:30–2:50pm)
2:50–3:10pm	Olivier Thas (p. 72) (2:50–3:10pm)	Omer Ozturk (p. 73) (2:50–3:10pm)
3:10–3:30pm	Tong Wang (p. 74) (3:10–3:30pm)	Dan Gladish (p. 75) (3:10–3:30pm)
3:30–4:00pm	Afternoon Tea *(Exhibition Hall)*	
	Contributed Session (4:00–5:20pm)	
	Visualisation *(Vines Room)* Chair: Alice Richardson	**Genetics & Evolution** *(Exhibition Hall)* Chair: Esther Meenken
4:00–4:20pm	Thomas Lumley (p. 76) (4:00–4:20pm)	Conrad Burden (p. 77) (4:00–4:20pm)
4:20–4:40pm	Connor James Smith (p. 78) (4:20–4:40pm)	Beata Sznajder (p. 79) (4:20–4:40pm)
4:40–5:00pm	Kevin Wang (p. 80) (4:40–5:00pm)	Anabel Forte (p. 81) (4:40–5:00pm)
5:00–5:20pm	Petra Kuhnert (p. 82) (5:00–5:20pm)	Julian Taylor (p. 83) (5:00–5:20pm)
5:20–6:30pm		
6:30–7:30pm	Buses depart at 6:45pm from the National Wine Centre	
7:30pm–late	**Conference Dinner and Awards Session** (7:30pm–late)	

Friday 6 December	
8:30–8:50am	Registration Desk Open
8:50–9:00am	**JABES/Biometrics Showcase** *Exhibition Hall* Chair: Alan Welsh
9:00–9:30am	Dan Pagendam (p. 84) (9:00–9:30am)
9:30–10:00am	Louise McMillan (p. 85) (9:30–10:00am)
10:00–10:30am	David Warton (p. 86) (10:00–10:30am)
10:30–11:00am	Morning Tea *(Exhibition Hall)*
11:00–11:20am	**Plenary Sessions** (11:00–1:00pm) *Exhibition Hall* Chair: Vanessa Cave
11:20–11:40am	**Invited Talk** (Agricultural 1) Daniela Bustos-Korts (p. 87) (11:00–12:00pm)
11:40–12:00pm	
12:00–12:20pm	*Exhibition Hall*
12:20–12:40pm	**Invited Talk** (Agricultural 2) Joanne De Faveri (p. 88) (12:00–1:00pm)
12:40–1:00pm	
1:00–1:45pm	**Conference Close**
1:45–2:30pm	Lunch *(Ferguson Room)*

Multilevel multiple imputation for health and survey data: your flexible (and robust) friend

James Carpenter[1]
Matteo Quartagno[2]
[1]London School of Hygiene & Tropical Medicine, UK, [2]MRC Clinical Trials Unit, University College London, UK
james.carpenter@lshtm.acuk

Invited Talk
Tuesday 3 December, 9:30–10:30am
Exhibition Hall

Multiple imputation is now well established as a practical and flexible method for analyzing partially observed data under the missing at random assumption. However, in large datasets there are concerns about how to preserve heterogeneity in the relationship between variables in the imputation process.

Building on recent work, we describe an imputation model (and R software) which allows the covariance matrix of the variables to vary randomly across higher level units, which may represent health districts or hospitals.

We further show how this approach allows us to (i) include weights, when the substantive model is weighted; (ii) provide a degree of robustness to misspecification of the imputation model and (iii) extends to impute data consistent with interaction and non-linear effects under investigation.

We illustrate with examples from the UK Millennium Cohort Study and the UK Clinical Practice Research Datalink.

Hockey sticks and broken sticks continued - enhancing the gold standard RCT for chronic diseases

Hans Hockey[1]
[1]Biometrics Matters Limited, New Zealand
hans@biometricsmatters.com

Contributed Talk
Tuesday 3 December, 11:00–11:20am
The Gallery

This work was motivated by a previously given chronic rare disease real data example as analysed by a longitudinal hockey stick model. The gold standard randomized controlled trial (RCT) compares active and placebo treatment arms at a post-baseline timepoint sufficiently late enough for an active effect to be apparent. For this design the analysis of covariance of final values adjusted for baseline values is standard (hopefully!).

Similar but enhanced alternative design and analysis combinations in chronic diseases will be presented which have several possible advantages: there is the possibility of assessing the size of any placebo effect; there is less ethical and recruitment need to have the active group larger than the placebo group; missing data, particularly of the final value, is less critical to the analysis.

It is also conjectured that some of the circumstances which can favour these enhanced designs can also be used to help argue for single arm studies in rare chronic diseases, despite the lack of randomization.

Towards causal inference for spatio-temporal data: adjusting for time-invariant latent confounders

Rune Christiansen[1]
Jonas Peters[1], Matthias Baumann[2], Tobias Kuemmerle[2]
[1]University of Copenhagen, Denmark, [2]Humboldt University of Berlin, Germany
krunechristiansen@math.ku.dk

Contributed Talk
Tuesday 3 December, 11:00–11:20am
Exhibition Hall

In statistical causality, we are interested not only in modeling the behaviour of a system that is passively observed, but also how the system reacts to changes in the data generating mechanism. Given knowledge of the underlying causal structure, such interventional behaviour can often be estimated from purely observational data (e.g., using covariate adjustment). Typically, the assumption is that data are generated as independent replications from the same underlying mechanism—an assertion that is often hard to justify in practice: geographically varying conditions in which the system is embedded induce spatial heterogeneity, and close-by observations (in space and time) tend to be strongly dependent. In this talk, I present causal models for spatio-temporal data that are adapted to these characteristics, and introduce a simple approach that allows for the estimation of causal effects under the influence of arbitrarily many latent confounders, as long as these confounders do not vary across time. Non-parametric hypothesis tests for the existence of causal effects are constructed based on data resampling, and do not rely on any distributional assumptions on the spatial dependence structure of the data. The method is applied to the problem of inferring the (potential) causal relationship between armed conflict and tropical forest loss, based on a spatio-temporal data set from Colombia. This talk does not require any prior knowledge in causal inference.

From the stepped wedge to the staircase: the information content of stepped wedge trials

Jessica Kasza[1]
Monica Taljaard[2], Andrew Forbes[1]
[1]Monash University, Australia, [2]Ottawa Hospital Research Institute, Canada
jessica.kasza@monash.edu

Contributed Talk
Tuesday 3 December, 11:20–11:40am
The Gallery

Stepped wedge cluster randomised trials are a type of longitudinal cluster randomised trial in which clusters, e.g. schools, hospitals, or geographical regions, are randomised to a particular set of treatment sequences. In standard stepped wedge trials, all clusters start in the control condition before switching, in a randomised order, to the intervention. The application of stepped wedge trials is increasing rapidly: a 2011 systematic review of published or registered stepped wedge trials found only 25, but as of July 2019, 210 stepped wedge trials were registered on clinicaltrials.gov. Stepped wedge trials are particularly useful in assessing interventions that will be rolled out or cannot be undone, e.g. changes in policies or cluster-wide education campaigns. However, stepped wedge trials are expensive and burdensome, requiring that all clusters contribute measurements for the entire trial duration. Recent work has shown that different cluster-period "cells" of the stepped wedge contribute different amounts of information to the estimation of the intervention effect. Such work suggests that "incomplete" stepped wedge trials in which clusters contribute measurements in a restricted set of trial periods may provide efficient alternatives to the full stepped wedge.

In this talk I will discuss the amount of information contributed by cluster-period cells of stepped wedge trials to the estimation of the effect of an intervention, where this is quantified by the increase in the variance of the intervention effect estimator when that cell is omitted. I will consider the impact of within-cluster correlation structure, treatment effect heterogeneity, implementation periods, and unequal cluster-period sizes on the pattern of information content. This work indicates that in many scenarios, "staircase" trials, a particular type of incomplete stepped wedge in which clusters provide measurements immediately before and after the treatment switch only, may prove to be efficient and less burdensome alternatives to the complete stepped wedge.

Spatial Confounding in GEEs - Why the Working Correlation Matters (well, sort of)

Francis Hui[1]
Howard Bondell[2]
[1]Australian National University, Australia, [2]University of Melbourne, Australia
francis.hui@anu.edu.au

Contributed Talk
Tuesday 3 December, 11:20–11:40am
Exhibition Hall

Generalized Estimating Equations (GEEs) are a popular tool in many scientific disciplines for investigating the effects of covariates on the mean of a response. In the context of spatial data analysis, GEEs rely on specifying a regression model for the marginal mean, a variance function, and a spatial working correlation matrix characterizing the spatial autocorrelation between observational units. One of the main advantages of GEEs is that estimation of the covariate effects is robust to misspecification of the choice of (spatial) working correlation matrix: the choice only affects the efficiency of the estimator.

In ongoing research, we investigate the impact of spatial confounding in GEEs. That is, what happens when the covariates included in a GEE, where a spatial working correlation matrix is used, are also spatially correlated. Under the conditional mixed model approach, the issue of spatial confounding is explicit and arises due to artificial multicollinearity between the spatially correlated covariates and the spatial random effect. We show that for GEEs, such multicollinearity also arises but occurs implicitly between spatially correlated covariates and the spatial working correlation matrix. Results suggests different choices of the working correlation matrix can lead to different attributions of the effect of the covariate on the mean versus on the spatial correlation i.e., on the first versus second moment. In turn, we consider using a so-called "restricted spatial working correlation matrix" that ensures all the variability in the direction of the covariates is attributed to the marginal mean, and is more in line with the underlying aim of GEEs. The issue of standard error estimation via the sandwich covariance matrix, and how it is impacted by spatial confounding, will also be discussed.

Estimation of proportions by group testing with retesting of positive groups

Graham Hepworth[1]
Stephen Walter[2]
[1]University of Melbourne, Australia, [2]McMaster University, Canada
hepworth@unimelb.edu.au

Contributed Talk
Tuesday 3 December, 11:40–12:00pm
The Gallery

In group testing (or pooled testing), material from individuals is pooled and tested in aggregate for the presence of an attribute, usually a disease. Group testing for estimation of a proportion p has been applied in a wide range of fields, including virus prevalence in flowers, blood testing (especially for HIV), and transmission of viruses by mosquitoes. Improved precision usually requires the testing of more groups, but in some situations it is difficult or expensive to obtain the required additional individuals. If the testing procedure is non-destructive, retesting of groups comprising different combinations of individuals may be a useful option.

Hepworth & Watson (2017) developed an estimator of p for the retesting of a random grouping of individuals from the positive groups at the first stage. The analytic complexity of this estimator led them to use simulation to examine its variance properties. We have developed two closed-form analytic expressions for the variance of the second-stage estimator, and compared its performance with the results from the simulation.

Our analytical solutions give acceptable approximations in a reasonable range of circumstances. They are most acceptable when the number of groups is not small and p is not large. This is a useful result for group testing, which to be of major benefit, relies on a reasonable number of groups and a small to moderate prevalence.

Hepworth G & Watson RK (2017) Revisiting retesting in the estimation of proportions by group testing. Communications in Statistics - Simulation and Computation 46:261–274.
Hepworth G & Walter SD (2019) Estimation of proportions by group testing with retesting of positive groups. Communications in Statistics - Theory and Methods DOI:10.1080/03610926.2019.1620280.

Modelling species communities using presence-only data

Ian Renner[1]
Otso Ovaskainen[2], Jarno Vanhatalo[2]
[1]University of Newcastle, Australia, [2]University of Helsinki, Finland
Ian.Renner@newcastle.edu.au

Contributed Talk
Tuesday 3 December, 11:40–12:00pm
Exhibition Hall

Often, the only available data to serve as input for species distribution models is presence-only data, which consists of observation records for the target species with no corresponding absence information. While a number of presence-only species distribution modelling methods can relate the occurrence patterns to the environment, there are limits to the types of inference that can be made from presence-only data. Indeed, there are a number of biotic and abiotic factors that influence the distribution of species and the composition of species communities, such as species co-occurrence patterns, traits, and phylogenetic relationships among species, and it is presently difficult to incorporate these factors into presence-only models.

With other data inputs such as presence-absence and count data, such inference about species and species communities is available via the hierarchical modelling of species communities (HMSC) platform. With HMSC, users can partition the variation in species occurrences to components that relate to environmental filtering, species interactions, and random processes, allowing for both species-level and community-level inference. However, the HMSC package does not currently support presence-only input.

In this talk, I will present developments of presence-only input into the HMSC framework and demonstrate the new types of inference available as a result. These developments not only enable such inference for presence-only data, but likewise allow previous analyses of species communities from HMSC to be enriched by incorporating relevant presence-only data.

The effect of number of clusters and cluster sizes on multiple imputation in multilevel models

Nidhi Menon[1]
Alice Richardson[1], Hwan-Jin Yoon[1]
[1]Australian National University, Australia
nidhi.menon@anu.edu.au

Contributed Talk
Tuesday 3 December, 12:00–12:20pm
The Gallery

Missing data are a common phenomenon in public health research. Multiple Imputation (MI) has been long recognized as an attractive approach to handle missing values. Statisticians are now advocating the use of MI as a gold standard in solving the missing data problem. Despite its early conception and its numerous advantages over the traditional ad hoc methods, there is still limited application of MI in public health research.

The theory of multiple imputation requires that imputations be made conditional on the sampling design. Not accounting for complex sample design features, such as stratification and clustering, during imputations can yield biased estimates from a design-based perspective.

Most datasets in public health research show some form of natural clustering (individuals within households, households within the same district, patients within wards, etc.). Cluster effects are often of interest in health research. In this study, we investigate through simulations different strategies for accounting for clustering when multiply imputing variables. Recent studies have identified methods to include fixed effects for clusters in imputations, however there is limited information on impact of varying number of clusters and cluster sizes on MI.

In this study, we simulate 3 level hierarchical data structures varying the number of clusters at each level. Missing values are present in covariates at each level in the data. We consider the impact of the combination of varying cluster sizes and proportion of missingness in imputation of covariates at each level in the dataset. This study implements the Gelman and Hill approach for imputation of missing data at higher levels by including aggregate forms of individual level measurements to impute for missing values at higher levels. The performance of popular methods of imputations, MICE and JoMo are compared. Performance measures include bias in estimates, mean squared errors and probability coverage of confidence intervals.

Fitting species distribution models with uncertain species labels using point process models

Emy Guilbault[1]
Ian Renner[1], Michael Mahony[1], Eric Beh[1]
[1]University of Newcastle, Australia
Emy.Guilbault@uon.edu.au

Contributed Talk
Tuesday 3 December, 11:40–12:00pm
Exhibition Hall

The popularity of species distribution models (SDMs) and the suite of tools to fit them have greatly increased over the last decade. The most common type of species data is presence-only data, which comes from citizen science. Point process models (PPMs) provide a flexible way to fit SDMs to presence-only data. Nonetheless, the quality of presence-only records (both identification and positional location) that serve as input to presence-only SDMs can be questioned. However most species distribution modelling methods applied to presence-only data assume certainty about species identity, but there are many practical situations in which this may not be the case. As an example, observers can be unable to clearly differentiate close species and taxonomists can split a species into multiple distinct species. We investigate the latter case, in which the species identities of records prior to a taxonomic change are confounded. In this talk, I will present two new tools for accommodating confounded records in PPMs. With these tools, we reclassify and incorporate records with uncertain species identities via finite mixture modelling or an iterative algorithm inspired by machine learning methods. Through simulation, we compare performance in classification and in prediction of these tools with different implementations to a standard approach which uses only records with known species labels, varying species abundance, correlation among species distributions, and the proportion of records with missing species labels. We also apply the best-performing methods to fit the distribution of 3 species of native Australian frogs that belong to the genus Myxophies. Among these species, 2 were newly described in 2006 in the northern range of the genus distribution and thus confounded previous records in the area made over the previous few decades.

Multiple imputation for missing outcome data in trials involving independent and paired observations

Thomas Sullivan[1]
[1]South Australian Health and Medical Research Institute, Australia
thomas.sullivan@sahmri.com

Contributed Talk
Tuesday 3 December, 12:20–12:40pm
The Gallery

Background: Trials involving a mixture of independent and paired data arise in many areas of health research, for example in paediatrics where outcomes can be collected on singletons and twins, and in ophthalmology, where one or both eyes may require treatment. An important consideration in these trials is the correlation in outcomes, or clustering, that occurs between observations from the same pair. When applying multiple imputation (MI) to address missing data, previous research suggests that any clustering in the data should be accounted for in the imputation and analysis models. However, most ad-hoc methods of MI for clustered data were designed with large and/or equal sized clusters in mind, and it is unclear how MI should be imeplemented in settings with independent and paired data.

Methods: Using simulated data the following MI approaches were evaluated: (1) MI ignoring clustering; (2) MI using chained equations with conditional imputation of the 2nd member of a pair; (3) MI performed separately by cluster size, and; (4) multi-level MI. Observations were allocated to one of two treatment groups using simple randomisation, with members of a pair randomised individually, to the same (cluster randomisation) or to opposite groups (opposite randomisation).

Results: When outcome data were missing at random, all MI methods produced unbiased treatment effect estimates. Although performance deficits were small, MI ignoring clustering and chained equations with conditional imputation produced confidence intervals for the treatment effect that were too narrow under cluster randomisation and too wide under opposite randomisation. MI performed separately by cluster size and multi-level MI performed well across the range of scenarios considered.

Conclusions: In trials involving a mixture of independent and paired observations, particularly those employing cluster or opposite randomisation, we recommend researchers apply multi-level MI or standard MI performed separately by cluster size to address missing outcome data.

`clustglm` and `clustord`: R packages for clustering with covariates for binary, count, and ordinal data

Louise McMillan[1]
Shirley Pledger[1], Daniel Fernández[2], Richard Arnold[1], Ivy Liu[1], Murray Efford[3]
[1]Victoria University of Wellington, New Zealand, [2]Fundació Sant Joan de Déu, Spain, [3]University of Otago, New Zealand
louise.mcmillan@vuw.ac.nz

Contributed Talk
Tuesday 3 December, 12:20–12:40pm
Exhibition Hall

We present two new R packages for model-based clustering with covariates. Both packages can perform clustering and biclustering (clustering sites and species simultaneously, for example). Both use likelihood-based methods for clustering, which enables users to compare models using AIC and BIC as measures of relative goodness of fit.

`clustglm` implements techniques from Pledger and Arnold (2014) for handling binary and count data, or data from other single-parameter exponential family distributions, such as normal distributions with constant variance. It leverages `glm` and can fit pattern detection models that include individual-level effects alongside cluster effects. For example, when applied to presence/absence data, users can cluster sites and species while also taking into account any single-species effects, and any additional covariates. `clustglm` can also be applied to capture-recapture data, to cluster individuals based on their capture patterns over multiple occasions.

`clustglm` can accommodate balanced and non-balanced designs, and numerical or categorical covariates. It provides the clustering equivalent of biplots, and also profile plots. We will illustrate the use of `clustglm` with a selection of ecological datasets.

`clustord` handles ordinal categorical data, using techniques outlined in Fernández et al. (2016). It builds on the ordered stereotype model, which accommodates flexibility in the ordinal scale used. The clustering results can reveal when two ordinal categories are effectively equivalent and can be combined to simplify the model.

Pursuing the cancer-schizophrenia disassociation paradox: genomes, phenomes and intimate conversations with inconclusive evidence

Max Moldovan[1,2]
[1]South Australian Health and Medical Research Institute, Australia, [2]Registry of Senior Australians (ROSA), Australia
max.moldovan@sahmri.com

Invited Talk
Tuesday 3 December, 1:30–2:30pm
Exhibition Hall

Positive and negative empirical findings of schizophrenia being protective against cancer remain a controversy and an active topic for debates among the intersection of oncologists, psychiatrists, epidemiologists and a diverse group of research scientists interested in the topic. I will present the vision of the paradox from different points of view, using a number of analysis and visualisation approaches from my current data science toolbox.

Firstly, I will take a position of a "genes-rule-it-all" theory proponent (while, in fact, being an opponent of the paradigm). I will share my experience of the exposure to genomics, and how my hopes and excitement were cut short by the inability of a SNP model to predict a well-defined phenotype when taken out of sample. Staying in the same role, I will introduce the Molecular Signatures Database (MSigDB) and an attempt to look at this gene expression signature information resource from a different angle.

Secondly, I will review epidemiological explanations behind the cancer-schizophrenia disassociation paradox, presenting the pair of disorders within the whole phenome network. While biases can help to justify the paradox, the doubt remains when one looks at bare numbers.

Finally, I will introduce a "geocentrism versus heliocentrism in cancer research" hypothesis and speculate about the role of dogmas in science, rates of clinical translation and prospects of moving to the brighter future.

Dependent selection and observation schemes in life history studies

Richard Cook[1]
Jerry Lawless[1]
[1]University of Waterloo, Canada
rjcook@uwaterloo.ca

Invited Talk
Tuesday 3 December, 2:30–3:30pm
Exhibition Hall

Multistate models provide a powerful framework for the analysis of life history processes when the goal is to characterize transition intensities, transition probabilities, state occupancy probabilities, and covariate effects thereon. Samples of individuals experiencing such processes are often constructed based on response-dependent selection schemes. Moreover, prospective data are often only available at random visit times realized over a finite period of follow-up. We formulate a joint multistate model for the life history, selection, visit, and loss to followup processes. This joint model is helpful when discussing the independence conditions necessary to justify the use of standard likelihoods involving the life history model alone, and provides a basis for analyses that accommodate dependence. We consider settings with disease-driven visits and routinely scheduled visits and develop likelihoods that accommodate partial information on the types of visits. Simulation studies suggest that suitably constructed joint models can yield consistent estimates of parameters of interest even under dependent visit processes providing the models are correctly specified, but identifiability and estimability issues can arise. An application is given to a cohort of individuals attending a rheumatology clinic where interest lies in progression of joint damage. This is joint work with Jerry Lawless.

Error variance bias in neighbour balance and evenness of distribution designs

Emlyn Williams[1]
Hans-Peter Piepho[2]
[1]Australian National University, Australia, [2]University of Hohenheim, Germany
emlyn@alphastat.net

Contributed Talk
Tuesday 3 December, 4:00–4:20pm
Exhibition Hall

Neighbour balance and evenness of distribution designs help to address user concerns in the two-dimensional layout of agricultural field trials. This is done by minimizing the occurrence of pairwise treatment plot neighbours and ensuring that the replications of treatments are spread out across rows and columns of a trial. Such considerations result in a restriction on the normal randomization process for a row-column design which can lead to error variance bias. In this talk results from the analysis of uniformity trial data are presented to demonstrate the degree of error variance bias for both resolvable and non-resolvable designs. Comparisons are made with linear variance spatial designs.

Williams, E.R. and Piepho, H.P. (2018). An evaluation of error variance bias in spatial designs. Journal of Agricultural, Biological, and Environmental Statistics 23, 83–91.
Piepho, H.P., Michel, V. and Williams, E.R. (2018). Neighbour balance and evenness of distribution of treatment replications in row-column designs. Biometrical Journal 60, 1172–1189.

Prof Ken Russell and the Design of Experiments for Generalized Linear Models

David Steel[1]
[1]University of Wollongong, Australia
dsteel@uow.edu.au

Contributed Talk
Tuesday 3 December, 4:20–4:40pm
Exhibition Hall

Prof Ken Russell passed away in August of this year. Ken made many contributions to the practice and theory of experimental design, through his excellent teaching, research, collaborations and work with industry. He recently published a major monograph on the Design of Experiments for Generalized Linear Models (CRC Press). To the best of my knowledge this is the first book to be written specifically on this topic. This talk will give a review of the book and its key ideas and results.

Designs with many singly replicated treatments

Richard Jarrett[1]
[1]University of Adelaide, Australia
rgjarrett@optusnet.com.au

Contributed Talk
Tuesday 3 December, 4:40–5:00pm
Exhibition Hall

This has been a topic of enduring interest, culminating in the paper on p-rep designs by Cullis et al. (2006). We will review a number of results, mainly for block designs, with and without random block effects, and extend this to the case where test lines are considered as random effects. An indication will be given of the implications for row-column designs and designs with spatial correlation.

Confidence intervals centred on bootstrap smoothed estimators

Paul Kabaila[1]
Christeen Wijethunga[1]
[1]La Trobe University, Australia
P.Kabaila@latrobe.edu.au

Contributed Talk
Wednesday 4 December, 8:50–9:10am
Vines Room

Bootstrap smoothed (or bagged; Breiman, 1996) parameter estimators have been proposed as an improvement on estimators found after preliminary data-based model selection. The key result of Efron (2014) is a formula for a delta method approximation to the standard deviation of the bootstrap smoothed estimator. This formula is valid for any exponential family of models and has the attractive feature that it simply re-uses the parametric bootstrap replications that were employed to find this estimator. It also has the attractive feature that it is applicable in the context of complicated data-based model selection. It is natural then to propose the use of a confidence interval that is centred on the bootstrap smoothed estimator and has width proportional to the estimate of this approximation to the standard deviation. We describe the results of an evaluation of the performance of this confidence interval, using a testbed that consists of two nested linear regression models and preliminary model selection using a t-test.

References

Breiman, L. (1996) Bagging predictors. Machine Learning.
Efron, B. (2014) Estimation and accuracy after model selection. Journal of the American Statistical Association.
Kabaila, P. and Wijethunga, C. (2019) Confidence intervals centred on bootstrap smoothed estimators. Australian & New Zealand Journal of Statistics.
Kabaila, P. and Wijethunga, C. (2019) On confidence intervals centred on bootstrap smoothed estimators. Stat.

Performance of factor analytic models in multi-environment trial data with small variety numbers

Bethany Macdonald[1]
Rachel King[2], Alison Kelly[1,3]
[1]Queensland Department of Agriculture and Fisheries, Australia, [2]University of Southern Queensland, Australia, [3]University of Queensland, Australia
bethany.macdonald@daf.qld.gov.au

Contributed Talk
Wednesday 4 December, 8:50–9:10am
Exhibition Hall

Much of the research undertaken in crop science involves evaluating varieties in field trials spanning a range of years and locations, where data arising from these sets of trials are known as multi-environment trial (MET) data. Analysis methods for MET data are concerned with investigating the nature of the variety by environment (V×E) effects, which describe the performance of different varieties across different environments, often with the aim of determining those varieties with superior performance. Ideally the variance of the V×E effects would be estimated using an unstructured form, assumed to be of full rank, however, this estimation is computationally difficult and can be unstable. The factor analytic (FA) model proposed by Smith et al. (2001) offers a more parsimonious option and has been shown to provide a good approximation to the unstructured form. The FA model has been shown to model the V×E effects accurately when there are large numbers of varieties, however, the accuracy of the model when variety numbers are small or varietal concurrence (the number of varieties in common between trials) is varied is unclear. The accuracy of FA models in scenarios when variety numbers are small and V×E and concurrence patterns vary was investigated through simulation. The study considered four model types for the variance of the V×E effects for MET data, with 10, 15, 25 and 50 varieties per trial. These data sets contained three types of V×E patterns, two levels of varietal concurrence and nine or 52 trials, resulting in 48 different scenarios. This study found that the accuracy of the FA model is affected when the number of varieties per trial are small, but the extent is dependent on other factors such as the number of trials and underlying V×E pattern.

Towards distributional analyses of biomarker data

Fernando Marmolejo-Ramos[1]
[1]University of South Australia, Australia
Fernando.Marmolejo-Ramos@unisa.edu.au

Contributed Talk
Wednesday 4 December, 9:10–9:30am
Vines Room

Current literature reviews of published datasets in the cognitive sciences (e.g. psychology, education, neuroscience) indicate that normally distributed data is not the norm but the exception. One such type of data is biomarker data. In a nutshell, biomarkers can be redefined as measures that index changes in neuropsychobiological states. This definition hence includes, among others, measures such as reaction times, event-related potentials, thermographic data, and blood pressure. Canonical statistical methods (e.g. t-tests, ANOVA) give biased results when dealing with non-normal data. The goal of this talk is to provide a snapshot of new statistical techniques that allow proper distributional and robust analyses. The emphasis will be on statistical graphics and conceptual definitions rather than mathematical elaborations. It is argued that adopting these techniques will ultimately lead to novel findings by revamping the way in which biomarker data are explored and analysed.

An evaluation of separable variance structures for highly genetically correlated environments

Bethany Rognoni[1]
Jason Sheedy[2], Valeria Paccapelo[1], Clayton Forknall[1], Neil Robinson[2], John Thompson[2], Alison Kelly[1]
[1]Queensland Department of Agriculture and Fisheries, Australia, [2]University of Southern Queensland, Australia
Bethany.Rognoni@daf.qld.gov.au

Contributed Talk
Wednesday 4 December, 9:10–9:30am
Exhibition Hall

In agricultural field trials, genotypes can be tested in the presence of interacting factors, resulting from either imposed treatments, repeated measurements or across multiple environments. When the aim is to select superior performing genotypes under these conditions, the factorial combination of conditions is often collapsed into one overall 'environment' factor, with a genetic variance structure fitted for this term, in a linear mixed model framework. This enables the estimation of a separate genetic variance for each of the levels (referred to as environment types) corresponding to the combinations of the original factorial structure, along with genetic covariance between the environment types.

As an alternative to this, an individual genetic variance structure for each of the factors can be fitted, resulting in a separable variance model. This may be a more parsimonious approach to modelling the genetic variance due to estimation of fewer parameters in total, and could provide a more intuitive interpretation of genotype behaviour across environment types. However, separable models have been shown to be less flexible, as they impose more restrictive variance structures.

A set of nematode resistance field trials motivated the exploration of separable variance models, where final nematode counts for each genotype were measured under three different soil depths, two previous nematode population densities resulting from prior crops and three trials. The analysis was first conducted using an overall environment model, where each environment represented a unique combination of soil depth, nematode population and trial. This model showed consistently high genetic correlations between all pairs of environments. The analysis was then performed using a separable genetic variance structure for the three-way factorial. There were substantial differences in the genetic variances and correlations that resulted from the different parameterisation of both models.

Simulation-Selection-Extrapolation: Estimation in High Dimensional Errors-in-Variables Models

Linh Nghiem[1]
[1]Australian National University, Australia
linh.nghiem@anu.edu.au

Contributed Talk
Wednesday 4 December, 9:30–9:50am
Vines Room

Errors-in-variables models in high-dimensional settings pose two challenges in application. Firstly, the number of observed covariates is larger than the sample size, while only a small number of covariates are true predictors under an assumption of model sparsity. Secondly, the presence of measurement error can result in severely biased parameter estimates, and also affects the ability of penalized methods such as the lasso to recover the true sparsity pattern. A new estimation procedure called SIMSELEX (SIMulation-SELection-EXtrapolation) is proposed. This procedure makes double use of lasso methodology. Firstly, the lasso is used to estimate sparse solutions in the simulation step, after which a group lasso is implemented to do variable selection. The SIMSELEX estimator is shown to perform well in variable selection, and has significantly lower estimation error than naive estimators that ignore measurement error. SIMSELEX can be applied in a variety of errors-in-variables settings, including linear models, generalized linear models, and Cox survival models. It is furthermore shown in the supporting information how SIMSELEX can be applied to spline-based regression models. A simulation study is conducted to compare the SIMSELEX estimators to existing methods in the linear and logistic model settings, and to evaluate performance compared to naive methods in the Cox and spline models. Finally, the method is used to analyze a microarray dataset that contains gene expression measurements of favorable histology Wilms tumors.

Multi environment trial analysis to determine the tolerance of cereal varieties to cyst nematode

Isabel Munoz-Santa[1]
[1]Universitat de València, Spain
sabela.munozsanta@adelaide.edu.au

Contributed Talk
Wednesday 4 December, 9:30–9:50am
Exhibition Hall

Cereal Cyst nematodes (CCN) are considered a major nematode pest which cause significant yield losses in wheat and barley with a high economic cost to the grain industry. A feasible solution is to breed for varieties which are tolerant; i.e. varieties which are able to yield in the presence of CCN in the field, this being an important research area for cereal nematologists.

In this study, we conducted 6 field trials from 2011 to 2018 in South Australia and Victoria with the objective of providing tolerance ratings of 92 different cereal varieties under high and low levels of pre-established nematode densities in the field. Multi environment trial analyses were used to analyse the data where the term environment refers to each year by location by nematode density combination. Spatial techniques were used to account for the spatial variability in each trial and a factor analytic model was fitted to model the genotype by environment interaction effects.

Multi environment trial analyses revealed high genetic correlation between high and low environments for each of the trials. This indicates that selecting varieties under high levels of CCN is equivalent to select varieties under low levels and thus analogous to select high yielding varieties irrespective of the nematode density. Therefore, a measure to assess the performance of varieties under high levels of nematodes independently of their performance under low levels was defined and named "tolerance index".

Multi environment trial analyses together with the tolerance index allowed us to select for high yielding varieties as well as give a more useful information to growers in relation to the comparison of varieties between high and low levels of CCN. The definition of the tolerance index and results obtained from this set of trials will be presented in this talk.

Trimmed Estimators - and a Hybrid-Censored Data Approach to Estimation

Brenton R Clarke[1]
[1]Murdoch University, Australia
B.Clarke@murdoch.edu.au

Contributed Talk
Wednesday 4 December, 9:50–10:10am
Vines Room

Using the idea of trimmed likelihood in location, we revisit the functional form of the trimmed mean and specify the functional form of the trimmed likelihood estimator as per Bednarski and Clarke (1993). This definition leads in the case of the exponential distribution to a naturally robust estimator which is highly efficient whose functional form is weakly continuous and Fréchet differentiable at the exponential model. See Clarke et al (2000). This is known as the β-trimmed mean. But what about censored data? Should we use the maximum likelihood estimator (mle) because censoring seemingly implies outliers are automatically taken care of? We explore a proposal of a hybrid estimator that combines the β-trimmed mean and the mle for some interesting results. See Clarke et al. (2017).

An across trials random regression approach to describe tolerance of wheat cultivars to disease

Clayton Forknall[1]
Alison Kelly[1], Steven Simpfendorfer[2], Ari Verbyla[3], Joanne De Faveri[3]
[1]Queensland Department of Agriculture and Fisheries, Australia, [2]NSW Department of Primary Industries, Australia, [3]CSIRO Data61, Australia
clayton.forknall@daf.qld.gov.au

Contributed Talk
Wednesday 4 December, 9:50–10:10am
Exhibition Hall

In plant pathology, tolerance to disease can be defined as the rate of change in productivity (yield), given a unit increase in pathogen burden. When applied to field crops, experiments to compare the tolerance of cultivars requires the establishment of a range of pathogen burdens over which the grain yield of a cultivar is measured, where tolerance to disease is then defined as the slope of a regression of yield against pathogen burden. A recent publication (Forknall et al. (2019), Phytopathology) describes a method for the statistically robust design and analysis of a single field experiment to quantify the rate of change in yield per unit increase in pathogen burden of five wheat cultivars for the disease crown rot. Using a random regression approach, implemented in a linear mixed model (LMM) framework, response curves describing the relationship between yield and crown rot pathogen burden were estimated for each of the cultivars in the experiment. This methodology is now extended to an across trials random regression approach, implemented in an LMM framework, which enables the estimation of different response curves for each cultivar in each experiment, where the response of each cultivar is modelled around an overall (average) yield response profile for each experiment. This modelling approach also provides a means of capturing the genetic correlation (covariance) between model parameters, both within and between experiments. The model is demonstrated by an application to 15 field experiments, estimating the yield response of the same set of five wheat cultivars to crown rot. The analysis revealed variation in the rate of change in yield, or tolerance, of cultivars across experiments, identifying that the tolerance of some cultivars was more influenced by environmental conditions than others. Also presented are graphical tools to assist in unlocking the interaction between cultivars tolerance and environmental conditions.

Robust scale estimation under small measurement errors

Michael Stewart[1]
Alan Welsh[2]
[1]University of Sydney, Australia, [2]Australian National University, Australia
michael.stewart@sydney.edu.au

Contributed Talk
Wednesday 4 December, 10:10–10:30am
Vines Room

Our motivating example is the estimation of robust scale functionals (like interquartile range or median absolute deviation) of the latent distribution in random effect models. This can be cast as a measurement error problem, where we "estimate" the random effects associated with each cluster. These estimates can in turn be regarded as the true random effects, plus some measurement errors. By "small measurement errors" we mean that we adopt a particular asymptotic scheme where both the number of clusters, and their sizes, tend to infinity. We propose to apply a "missing information principle" whereby we approximate the conditional expectation of the semiparametrically efficient score equations, given the observed data. We report some very interesting simulation study results and briefly sketch some accompanying theory.

Clustering environments in a combined trial analysis of yield response to plant population

Michael Mumford[1]
Kerry Bell[1]
[1]Queensland Department of Agriculture and Fisheries, Australia
Michael.Mumford@daf.qld.gov.au

Contributed Talk
Wednesday 4 December, 10:10–10:30am
Exhibition Hall

One of the key management practices of interest in agricultural crops is yield response to plant population and its contribution to yield in different genetic backgrounds. When field trials are run across multiple environments (i.e. years and locations), it is imperative to model the environment effects and the interaction of environment with hybrid and plant population. The challenge is then to compare the response models across environments.

A fixed effects combined trial analysis for yield response to plant population in a linear mixed model framework is presented. Clustering of the treatments (e.g. hybrid and environment combinations) is performed such that parallel curves are fitted to treatments within the same cluster to i) simplify the interaction and ii) increase the information used for fitting each response function by combining data across treatments within the same cluster. The categorisation of treatments to clusters is considered adequate when there is no significant interaction effect between plant population and treatment within each cluster. This presentation will focus on an objective method for categorising treatments into clusters. Furthermore, a sensitivity analysis will be presented to explore the adequacy of the proposed clustering methodology.

The methodology is applied to a large set of sorghum trials implemented across New South Wales, Australia in two seasons consisting of different hybrids, row spacings and trial locations. Furthermore, it provides a general framework for a fixed effects regression analysis in a linear mixed model framework that can account for i) experimental design terms, ii) separate residual variances and iii) spatial field trend at each trial.

Using Deep Neural Models to Facilitate Statistical Modeling of Complex Spatio-Temporal Dynamics

Christopher K. Wikle[1]
[1]University of Missouri, USA
wiklec@missouri.edu

Invited Talk
Wednesday 4 December, 11:00–12:00pm
Exhibition Hall

Spatio-temporal data are ubiquitous in the sciences and engineering, and their analysis is important for understanding and predicting a wide variety of processes. One of the difficulties with statistical modeling of spatial processes that change in time is the complexity of the dependence structures that must describe how such a process varies, and the presence of high-dimensional complex datasets and large prediction domains. It is particularly challenging to specify parameterizations for nonlinear dynamic spatio-temporal models (DSTMs) that are simultaneously useful scientifically, efficient computationally, and allow for proper uncertainty quantification. Alternatively, the machine learning community has developed a suite of deep neural models and learning paradigms (e.g., convolutional neural networks, recurrent neural networks, reinforcement learning) that can be combined in novel ways to predict spatio-temporal processes. However, these deep neural models are typically implemented in a deterministic framework that limits formal inference. Here we explore some recent attempts to build hybrid models in which deep neural models can be embedded within a formal statistical DSTM framework. The approaches are illustrated with examples applied to environmental and ecological data.

Quasi-random spatially balanced sampling

Blair Robertson[1]
Jennifer Brown[1], Trent McDonald[2], Chris Price[1]
[1]University of Canterbury, New Zealand, [2]Western EcoSystems Technology, Inc. USA
blair.robertson@canterbury.ac.nz

Invited Talk
Wednesday 4 December, 12:00–1:00pm
Exhibition Hall

A spatial sampling design determines where sample locations are placed in a study area. The main objective is to select sample locations in such a way that valid scientific inferences can be made to all regions of the study area. To achieve good estimates of population characteristics, the spatial pattern of the sample should be similar to the spatial pattern of the population. However, the spatial pattern of the response variable is usually not known. Fortunately, when sampling natural resources, nearby locations tend to be similar because they interact with one another and are influenced by the same set of factors. This means sample efficiency can be increased by spreading sample locations evenly over the resource. A sample that is well-spread over the resource is called a spatially balanced sample. In this talk, we show how quasi-random sequences can be used to draw spatially balanced samples from natural resources.

A new Liu estimator under a linear constraint

Takeshi Kurosawa[1]
Naoto Suzuki[1]
[1]Tokyo University of Science, Japan
tkuro@rs.tus.ac.jp

Contributed Talk
Thursday 5 December, 8:50–9:10am
Vines Room

A problem in parameter estimation for a linear model arises multi-colinearity if a data set is ill-posed. One of the solutions, Liu (1993) proposed his estimator called the Liu estimator. Furthermore, Kaciranlar et al. (1999) generalized the Liu estimator under the condition with a linear constraint among the parameters. However, their estimator is an ad-hoc method because they treated the two problems independently. In this study, we interpret the Liu estimator as the solution of a certain loss function, and then we propose a new estimator under the condition with the linear constraint. We show theoretical bias and RMSE of our estimator and give a necessary and sufficient condition for the superiority of our estimator against other estimators in terms of RMSE. We also perform a simple simulation study using empirical RMSE.

Big Biometric Data: A Biostatistical Perspective

Susan Wilson[1,2]
[1]University of New South Wales, Australia, [2]Australian National University, Australia
sue.wilson@anu.edu.au

Contributed Talk
Thursday 5 December, 8:50–9:10am
Exhibition Hall

Technological advances in biology and medicine, ranging from high-throughput sequencing to wearable electronic devices, are producing a "tsunami" of big biometric data. These data are of very widely varying types and quality. Many challenges abound on how to deal with such data, including wide-ranging deliberations concerning statistical modelling. This presentation will give an overview, including some current developments to meet the more pressing of these challenges. As well, in medicine in particular, big biometric data are giving rise to subtle and evolving ethical issues, many of which concern data analysts. Particular attention will be given to how these affect modern developments including applications such as personalised medical treatment.

Model selection and principle of parsimony in statistical modelling in agriculture

Zhanglong Cao[1]
[1]SAGI-West, Australia & Curtin University, Australia
zhanglong.cao@curtin.edu.au

Contributed Talk
Thursday 5 December, 9:10–9:30am
Vines Room

Model selection is an important issue in biostatistical, psychological and agricultural studies. Root mean squared error (RMSE), Akaike's information criterion (AIC), Bayesian information criterion (BIC) and their relatives are commonly used as selection criteria for goodness-of-fit of statistical models. However, there is no robust technique that can be applied in every aspect of parameter estimation and model selection. Sometimes, the winning model is "cursed", while the best model based on the selection criteria leads to over-fitting in practice. Goodness-of-fit must be balanced against model complexity to avoid over-fitting issues. We discuss the trap in model selection and the principle of parsimony, and present a weighted neighbouring cross-validation method. The latter will be illustrated on agricultural experimental data set.

Bridging the gap between science and statistics

Teresa Neeman[1]
[1]Australian National University, Australia
teresa.neeman@anu.edu.au

Contributed Talk
Thursday 5 December, 9:10–9:30am
Exhibition Hall

Biology has undergone a huge transformation in the last 50 years from an observational science with occasional data to an experimental science with terabytes of data. The computational challenges in visualising and analysing large biological data sets are formidable for most biologists who trained in more data-naïve environments. This is potentially a huge opportunity for statisticians, whose training can help biologists discover mechanisms, patterns of behaviour, and even new biological paradigms. But training research students often occurs in "silos" in both fields, leaving huge communication gaps between the biological sciences and statistics. Biologists miss the chance to sophisticated statistical machinery that may elucidate biological functions. Statisticians miss out on discovering the power of data to address important questions. In this talk, I explore how we can start to bridge this communication gap. As statisticians, we need to teach biologists statistical concepts that are relevant to experimental sciences. These concepts need to emphasise experimental design and mixed effects models. We need to challenge ourselves to learn the language of biologists and challenge our statistics students with real-life complex data problems. Finally, we need to connect research students with one another and encourage research collaborations.

On the effect of dependencies between regressors and random effects when analysing hierarchical structured data

Hwan-Jin Yoon[1]
Youngjo Lee[2], Alan Welsh[1]
[1]Australian National University, Australia, [2]Seoul National University, South Korea
hwan-jin.yoon@anu.edu.au

Contributed Talk
Thursday 5 December, 9:30–9:50am
Vines Room

Hierarchically structured data arises frequently in practice, in many fields of sciences including social science. Linear mixed models (LMMs) is one of the most common statistical methods to deal with the structured or clustered data.

Yoon & Welsh (2019) studied the effect of ignoring clustering in x on fitting LMM, and showed that it can be obtained misleading assessments of both the association between y and x and of the variance components. They also showed that, as the within cluster variance of x, τ_x, increases, the likelihood and the REML criterion develop two distinct local maxima and which of these is the global maximum changes at the jump point.

In LMMs for clustered or hierarchical structured data, regressors can be correlated with random effects. When the random effects and regressors independence assumptions are violated, not only regression coefficient estimators but also variance components can be severely biased.

In this study, we investigate how the violation of the random effects and regressors independence assumption could affect on the estimates of parameters and compare the results with the effect of ignoring clustering in x when fitting LMM (Yoon & Welsh). We also extend our study to the case of correlated binary data.

Statistics for agronomists: a constructive synthesis of workplace learning and community of practice

Sharon Nielsen[1]
Sam Rogers[1], Wendy Li[1], Annie Conway[1], Sabela Munoz-Santa[1,2], Olena Kravchuk[1]
[1]University of Adelaide, Australia, [2]Universitat de València, Spain
sharon.nielsen@adelaide.edu.au

Contributed Talk
Thursday 5 December, 9:30–9:50am
Exhibition Hall

Stakeholders in agronomy, and many other discipline areas, are promoting reproducible and transparent research, allowing research results to be replicated. The need for scientific rigour is paramount to the agronomic industry and can be achieved through appropriate design (with replication and randomisation) and robust and appropriate modelling techniques.

Agronomic evaluation of plants - quantity and quality, fertilisers and herbicides and pest control strategies are achieved through agronomic experiments, usually grown in the field, glasshouse or temperature control rooms. During the past three years staff from SAGI-STH training program have developed and run a series of statistical workshops, using adult learning methodologies, to improve the statistical competency of agronomists who conduct these agronomic experiments. The workshops start with an introduction to R, move through design and then analysis of agronomic experiments and end by introducing these scientists to reproducible research through R markdown. The workshops use active learning strategies and real-world examples relevant to the researchers.

On completion of the workshops, participants are then invited to join our community of practice, where we explore their data and work on solving statistical problems as a team. The community of practice is via online meetings, which are recorded so that participants can join synchronously or asynchronously. In this talk we explore the benefits of these workshops and community of practice and what it has meant in practice to the agronomists.

Modelling and Parameterization of Soil-Water Retention Curves

Warren Muller[1]
Richard Greene[2], James Noble[1]
[1]CSIRO Land and Water, Australia, [2]Australian National University, Australia
warren.muller@csiro.au

Contributed Talk
Thursday 5 December, 10:10–10:30am
Vines Room

The prediction of soil water storage and water supply to plants is essential in the investigation of vegetation response to rainfall. In particular, soil water retention curves (WRCs) which depict the relationship between the volumetric water content (VWC) and the soil water potential (h) are used by soil scientists in their studies of soil hydraulic properties.

If the shape of the water retention curve is sigmoid, then such data sets can be characterized by several models. The most commonly used model is the van Genuchten model (Soil Sci. Soc. Am. J. 44, 1980, 892–898).

Fitting the van Genuchten model to data sets relating VWC to h involves estimation of four parameters, some of which have a physical significance in understanding soil-water relationships. However typically these data sets only have 8 to 10 (VWC, h) pairs, which leads to difficulty in estimating some of the parameters and, in some cases, the van Genuchten model is difficult to fit.

We fitted the van Genuchten model to VWC vs. h data for 35 water retention curves from semi-arid rangeland soils and seven vineyard soils from the Yass Valley, NSW. We present examples of these fitted models. Some new parameters were derived from the fitted values, and relationships between the four estimated parameters and these derived parameters are also presented. The results from this study show that

(i) On some occasions the van Genuchten model fits poorly or doesn't fit, so is inappropriate;

(ii) There are strong relationships between some of the estimated and derived parameters, meaning the model is most likely over-parametrized;

(iii) Alternative models should be used when the VWC vs. h relationship is not a distinct sigmoid shape, for example curvilinear or near linear.

Uncertainty in digital agriculture: An interdisciplinary perspective

Esther Meenken[1]
Martin Espig[1], Susanna Finlay-Smits[1], Mos Sharifi[1], Mark Wever[1], Val Snow[1], David Wheeler[1]
[1]AgResearch, New Zealand
esther.meenken@agresearch.co.nz

Contributed Talk
Thursday 5 December, 9:50–10:10am
Exhibition Hall

Digitilisation in agricultural systems provides more, and increasingly real time, data to farmers, consumers, and others along the supply chain than ever before. Turning this data into actionable knowledge through eg bio-physical or statistical models should aid decisions related to activities such as farm management, policy development and product selection. A connected farm has data generating IoT systems proliferated throughout: In and on the soil, monitoring weather patterns and crop growth, and tracking the movement, interactions and welfare of animals. This burgeoning supply of data is accompanied by uncertainty that carries through and is modified as users perceive and interact with it. For example, a researcher building a crop simulation model will be subject to sensor uncertainties due to bias, lack of precision, sensor failure and incomplete calibration of the sensors. The model will exhibit direct uncertainties reflecting a) these measurement errors as well as b) lack of complete knowledge about the system being modelled and c) scenario specification. On the other hand, a farmer using information provided by the model to manage irrigation allocation may additionally experience indirect uncertainties stemming from, e.g., a lack of trust in the modellers or communicators which decrease the perceived certainty of a model, particularly in contexts where trust functions as a coping mechanism for an inability to fully assess a model's direct uncertainty. Uncertainty may also arise due to 'contextual uncertainty', defined as the bio-physical or socio-cultural environment which shapes the structure of the model or the way it is used, and determines how 'fit-for-purpose' it may be. We hypothesise a conceptual framework that attempts to visually indicate the relative uncertainties as data and biophysical/sociocultural uncertainty are incorporated into a model and as the model in turn becomes incorporated into wider, and increasingly complex, contexts once shared and used.

Partial automation of sampling and data collection to yield better estimates for faba and canola seed emergence project

Peter Kazprzak[1]
Lachlan Mitchell[1]
[1]University of Adelaide, Australia
peter.kasprzak@adelaide.edu.au

Contributed Talk
Thursday 5 December, 10:10–10:30pm
Exhibition Hall

An often overlooked part of experimental research is the data collection process. Errors in sampling propagate through the entire experiment, with protocols often comprising a significant portion of resources from limited budgets. The challenge is to automate sections of this process, while keeping the necessary human checks, to decrease time, manpower, and overall costs of this process. One important step is the selection of sampling protocol. Often a simple random sample can be improved upon in terms of smaller standard estimator error, and a more representative sample gathered from a smaller number of total samples. We attempt to show how simulation can be run on a representative data set to determine the optimal sampling protocol for the circumstance, and at the same time re-introduce a nearly forgotten sampling technique, Ranked Set Sampling. Under utilized for 60 years since its inception at the CSIRO by G. A. McIntyre in 1952, advances in modern technology now make it feasible to add this protocol to the go-to toolbox of sampling, and realize potential gains in efficiency, with the ultimate goal of wrapping these benefits in a easy to use, field ready, minimal training required piece of software.

Vine copulas and health applications

Claudia Czado[1]
[1]Technical University of Munich, Germany
cczado@ma.tum.de

Invited Talk
Thursday 5 December, 11:00–12:00pm
Exhibition Hall

Vine copulas (`vine-copula.org`) provide a wide class of multivariate dependence models. They allow for arbitrary marginal models and the dependence is characterized by a copula, which is build using only bivariate copulas. They are joined to form a valid multivariate copula using conditioning arguments, which is identified in a set of linked trees called the vine structure. Since all terms can be chosen independently they can accommodate different tail dependence both symmetric and asymmetric for groups of variables. I will introduce the construction, discuss the selection of vine structure and the step wise estimation procedure. This allows to select and fit models in very high dimensions. After this I will show how these models can be applied in some health applications.

Copula models for ecological community data

Marti J. Anderson[1]
[1]Massey University, New Zealand
m.j.anderson@massey.ac.nz

Invited Talk
Thursday 5 December, 12:00–1:00pm
Exhibition Hall

I shall describe a recently developed pathway for multivariate analysis of data consisting of counts of species abundances that includes two key components: copulas, to provide a flexible joint model of individual species, and dissimilarity-based methods, to integrate information across species and provide a holistic view of the community. Individual species are characterised using suitable (marginal) statistical distributions, with the mean, the degree of over-dispersion and/or zero-inflation being allowed to vary among a priori groups of sampling units. Associations among species are then modelled using copulas. A Gaussian copula smoothly captures changes in an index of association that excludes joint-absences in the space of the original species variables, but more flexible vine copulas might also be used. A permutation-based filter with exact family-wise error can optionally be applied a priori to reduce the dimensionality of the copula estimation problem. An MCEM algorithm provides efficient estimation of the copula correlation matrix with discrete marginals (counts). Given the resulting model, we may then simulate realistic ecological community data under fully specified null or alternative hypotheses. Distributions of community centroids derived from simulated data can be visualised in ordinations of ecologically meaningful dissimilarity spaces. Multinomial mixtures of data drawn from copula models also yield smooth power curves in dissimilarity-based settings. The proposed analysis pathway provides new opportunities to combine model-based approaches with dissimilarity-based methods, enhancing our understanding of ecological systems. I shall demonstrate implementation of the pathway with an example dataset of fish counts from a New Zealand marine reserve, the Poor Knights, and will also touch briefly on more recent extensions that embrace models of species along environmental gradients.

Determination of Indirect Reference Intervals for Immunoglobulin in an Australian Population

Alice Richardson[1]
Adain Zellner[1], Brett A. Lidbury[1], Peter Hobson[2], Tony Badrick[3]
[1]Australian National University, Australia, [2]Sulliavan Nicolaides Pathology, Australia, [3]RCPA Quality Assurance Programs, Australia
alice.richardson@anu.edu.au

Contributed Talk
Thursday 5 December, 2:10–2:30pm
Vines Room

Diagnostic pathology test results are typically supplied with reference intervals, a pair of values delineating the range within which a healthy individual's test result should lie. These reference intervals are essential for interpretation of the tests, but calculation of these intervals is controversial from both a statistical and a biological point of view.

In this presentation we introduce the indirect method for reference interval calculation, based on secondary analysis of administrative data rather than direct identification of a healthy population leading to reference interval calculation. We'll describe the statistical and biochemical issues that arise from the indirect approach.

We will also investigate the effect of two methods of outlier removal (Tukey elimination and the block method) and three methods of calculation (parametric, non-parametric and robust) on the indirect intervals obtained. The presentation will be illustrated with a large Australian data set relating to serum immunoglobulin A, G and M for males and females across the entire human age range.

The outlier elimination method was more important in the production of the indirect reference intervals than the calculation method. The Tukey elimination procedure consistently eliminated more values than the block method. If Tukey elimination was applied, variation between intervals produced by the different calculation methods was then minimal. The non-parametric intervals were actually more sensitive to outliers which, for certain assays, led to higher and wider intervals for older age groups. There were only minimal differences between robust and parametric reference intervals.

The interaction between outlier elimination and calculation method will be investigated further, and suggestions made for moving forward with indirect intervals in the diagnostic setting.

Statistical Efficiency of Distance Sampling

Robert Clark[1]
[1]Australian National University, Australia
robert.clark@anu.edu.au

Contributed Talk
Thursday 5 December, 2:10–2:30pm
Exhibition Hall

Distance sampling is a technique for estimating the abundance of animals or other objects in a region. Animals can be observed in a wide strip around the observer, while hopefully avoiding undercount by adjusting for detection rates which can decline with distance, provided that a set of strong model assumptions are justified. The impact of uncertainty about the detection model on variances of abundance estimates will be discussed. An expression for the asymptotic penalty factor is stated; it would typically be at least 2 but could be much higher if detection rates drop steeply. Various approaches are compared in simulations which incorporate clumping of animal locations. The take home message is that the significant penalty due to unknown detection parameters should be factored into decisions about the methodology and scale of abundance studies.

The Propensity Score with Semi-continuous Exposures

Tugba Akkaya-Hocagil[1]
Louise M. Ryan[2], Richard J. Cook[1], Joseph Jacobson[3], Sandra Jacobson[3]
[1]University of Waterloo, Canada, [2]University of Technology Sydney, Australia, [3]Wayne State University, Australia
takkayahocagil@uwaterloo.ca

Contributed Talk
Thursday 5 December, 2:30–2:50pm
Vines Room

Propensity score methodology has become increasingly popular in recent years as a tool for estimating causal effects of treatment or exposure using data from observational studies. For the most part, discussion has focussed on binary treatment/exposure scenarios. While some authors have discussed propensity score methodology for more general exposures, including continuous, this topic has been less well studied. In this presentation we discuss the context of environmental epidemiology where interest typically focuses on exposure variables that include zero values representing non-exposed individuals as well as long tails representing highly exposed individuals. We develop a propensity score methodology based on a two-part model and show how this can be used to more reliably estimate the causal effects of a semi-continuous exposure variable. We compare and evaluate the performance of our proposed method relative to the more standard generalized propensity score method and direct covariate adjustment through simulation studies. We find that when the outcome model satisfies linear regression model assumptions, all three methods yield unbiassed results. However, when the outcome model violates the assumptions of linear regression, our proposed method outperforms both direct covariate adjustment and the generalized propensity score method. We illustrate our method using data from the Detroit Longitudinal Cohort Study where exposure corresponds to prenatal alcohol exposure, with a long tail and many zero values.

Being Random and Efficient for Transect-Based Ecological Surveys

Scott Foster[1]
Geoffrey R. Hosack[1], Jacquomo Monk[2], Emma Lawrence[1], Neville S. Barrett[2], Alan Williams[3], Rachel Przeslawski[4]
[1]CSIRO Data61, Australia, [2]University of Tasmania, Australia, [3]CSIRO, Australia, [4]Geoscience Australia, Australia
scott.foster@data61.csiro.au

Contributed Talk
Thursday 5 December, 2:30–2:50pm
Exhibition Hall

In ecology, many sampling techniques rely on taking measurements along a transect; an example is the collection of underwater imagery from towed platforms. Despite this, methods to generate randomised survey designs have not hitherto been developed. We present an approach to generate a randomisation of transects using three steps: 1) calculate transect inclusion probabilities from user-specified cell inclusion probabilities, which allows particular environments to be sampled more often; 2) alter the cell and transect inclusion probabilities so that when transects are sampled the frequencies of sampling cells approximate the cell inclusion probabilities, and; 3) draw a spatially-balanced probability sample of transects. The resulting designs are for images constrained to lie on transects. The transect-based designs approximately respect the specified cell-inclusion probabilities whilst maintaining spatial-balance and still being straightforward to specify. We illustrate with application of the method to a towed-camera survey of deep-sea (500–2,000m depths) seamounts off Tasmania, Australia. This was a challenging area to survey due to its complex topology, and uneven inclusion probabilities for the property of interest - the presence of a stony coral that forms large areas of biogenic reef and supports elevated biodiversity.

Compositional Rank Methods for Testing for Differential Abundance in Microbiome Studies

Olivier Thas[1,2,3]
Leyla Kodalci[1], Stijn Hawinkel[2], John Rayner[4]
[1]Hasselt University, Belgium, [2]Ghent University, Belgium, [3]University of Wollongong, Australia, [4]University of Newcastle, Australia
olivier.thas@uhasselt.be

Contributed Talk
Thursday 5 December, 2:50–3:10pm
Vines Room

Microbiome count data can be considered as compositional mutivariate observations, which are characterised by a sum-constraint. Popular methods for testing for differential abundance rely on the (zero-inflated) negative binomial (NB) distribution. We have developped a new goodness-of-fit test for this distributional assumption, and we have applied the test to several public data sets. We conclude that more than 50% of the OTUs do not show a NB distribution and we demonstrate that this lack-of-fit causes poor FDR control. Robust or nonparametric methods for compositional data are needed.

Many data analysis methods have been developed for compositional data. They make use of log-ratios of counts, but these are problematic in the presence of many zero counts, as is the case for microbiome. We have developed semiparametric rank methods for compositional microbiome data. The methods do not rely on strong distributional assumptions, avoid log-ratios and account for library size variability. The methods make use of either means, rank or sign statistics. False discovery rate control happens through a new permutation method that accounts for the discreteness of the p-value null distributions. Results from a realistic simulation study suggest that the new methods perfom well and that particularly the sign-based methods perform well for overdispersed microbiome data.

Two-Stage Cluster Samples with Judgment Post-Stratification

Omer Ozturk[1]
Olena Kravchuck[2], Jennifer Brown[3]
[1]Ohio State University, USA, [2]University of Adelaide, Australia, [3]University of Canterbury, New Zealand
ozturk.4@osu.edu

Contributed Talk
Thursday 5 December, 2:50–3:10pm
Exhibition Hall

Estimation of the population mean or total in a clustered population can be done using two-stage cluster samples. Here we present a design in which cluster sample in each stage is constructed either using a judgement post-stratified (JPS) or a simle random sampling design. The JPS sampling design is implemented with or without replacement, but the SRS sample is always constructed without replacement. The paper presents design-unbiased estimators for the population mean and total, and the variances of these estimators. The efficiency improvement of the sampling designs compared with the SRS sampling design is investigated. The proposed estimator has a smaller variance than a two-stage SRS sample, but the level of improvement in efficiency depends on the intra-cluster correlation coefficient and the choices of the sampling designs in stage I and II of samplings. The paper also presents an approximate confidence interval for the population mean and total. For a fixed cost, the optimal sample sizes for stage I and stage II samples are constructed by maximizing the information content of the sample. The proposed sampling designs and estimators are applied to a two-stage sampling in an agricultural application.

Comparison of Methods for the Detection of Outlier and Associated Biomarker in Mislabeled Omics Data

Tong Wang[1]
[1]Shanxi Medical University, People's Republic of China
tongwang@sxmu.edu.cn

Contributed Talk
Thursday 5 December, 3:10–3:30pm
Vines Room

Background: Previous studies have reported that labeling errors are not uncommon in omics data. Potential outliers may severely undermine the correct classification of patients and the identification of reliable biomarkers for a particular disease. Three methods have been proposed to address the problem: sparse label-noise-robust logistic regression (Rlogreg), robust elastic net based on the least trimmed square (enetLTS), and "Ensemble". Ensemble is an ensembled classification based on distinct feature selection and modeling strategies. The accuracy of biomarker selection and outlier detection of these methods needs to be evaluated and compared so that the appropriate method can be chosen.

Results: The accuracy of variable selection, outlier identification, and prediction, and the running time of the three methods were compared in scenarios with various sample sizes, dimensions, and proportions of outliers with or without leverage points. From the results of the simulation study, Ensemble was best in terms of the variable selection accuracy. enetLTS selected the most variables with the highest positive selection rate and highest false discovery rate. In terms of the mislabeled samples detected, enetLTS performed best, with high sensitivity and a controlled false positive rate within 5%. Rlogreg detected most outliers, with the highest FPR, which exceeded 5% in some cases. Rlogreg had the lowest runtime. The three methods were applied to a triple negative breast cancer (TNBC) dataset, which included individuals with discordant labels.

Conclusions: Our study highlighted how to choose methods for feature selection and outlier detection in high-dimensional datasets with mislabeled samples. If a low FDR was required to reduce the failure of subsequent experimental validation, then Ensemble was the best choice. If screening associated genes broadly or the prediction of a response was required, then enetLTS was the best choice. In terms of the mislabeled samples detected, enetLTS performed best.

Optimal design for monitoring groundwater quality using geostatistical modelling

Dan Gladish[1]
Sreekanth Janardhanan[2], Dan Pagendam[1], Dennis Gonzalez[2]
[1]CSIRO Data61, Australia, [2]CSIRO Land and Water, Australia
dan.gladish@data61.csiro.au

Contributed Talk
Thursday 5 December, 3:10–3:30pm
Exhibition Hall

Collecting groundwater data is a critical aspect for evaluating potential impact on the regional freshwater supply from resource development. Assessing baseline water quality indicators in coal seam gas development areas is important to develop improved understanding of the groundwater system and inform managerial decisions and evaluate potential issues. However, prediction of groundwater quality indicators is difficult, and often results in high uncertainty due to the sparse nature of available data. Interest is therefore in determining future borehole locations for monitoring. However, drilling boreholes for groundwater monitoring is expensive. Therefore, determining optimal borehole locations that will reduce prediction uncertainty of groundwater quality indicators is useful. In this study, we propose a method that utilizes kriging models to interpolate groundwater quality indicators in the aquifer. We then combine the predicted values from these geostatistical models and use the Differential Evolution algorithm to determine optimal locations that would reduce spatial prediction uncertainty. We apply our method to a portion of one of the Great Artisan Basin aquifers, specifically in the Namoi region in New South Wales, Australia. Using this method, we outline 10 potential borehole locations for monitoring groundwater quality indicators that will minimize prediction uncertainty.

Multiclass Hexbin Plots

Thomas Lumley[1]
[1]University of Auckland, New Zealand
t.lumley@auckland.ac.nz

Contributed Talk
Thursday 5 December, 4:00–4:20pm
Vines Room

Hexagonal binning allows for the efficient display of very large data sets in graphs with many of the benefits of scatterplots. Hexbinning also supports weighted data. However, it does not allow the equivalent of plotting points in different colours to display a discrete third variable in a scatterplot. I will present multiclass hexbins plots that partition each hex into six triangles for the display of multiple classes. Multiclass hexbins are a form of icon plot or glyph plot and show how coarsening/stylisation of information can still be beneficial. I will also describe an R implementation in the `hextri` package.

Estimation of mutation rate matrices from genomic site frequency data

Conrad Burden[1]
Claus Vogl[2]
[1]Australian National University, Australia, [2]Vetmeduni Vienna, Austria
conrad.burden@anu.edu.au

Contributed Talk
Thursday 5 December, 4:00–4:20pm
Exhibition Hall

We describe how the full 4×4 genomic mutation rate matrix can, in principle, be estimated from population site frequency data obtained by sampling independent neutrally evolving sites within a multiple alignment of genomes. The method relies on calculated stationary sampling distributions of the diffusion limit of the mutation-drift Wright-Fisher model. These distributions can be obtained for an arbitrary general, non-reversible rate matrix, to first order in the mutation rates, either from the forward Kolmogorov equation or from a coalescent argument. The method will be applied to extracted sequence information from short autosomal introns of 196 Drosophila melanogaster individuals resulting in a site frequency spectrum of 218,942 nucleotides.

Visualisation of model stability information for better prognosis based feature extraction

Connor James Smith[1]
Samuel Mueller[1], Boris Guennewig[1,2]
[1]University of Sydney, Australia, [2]Garvan Institute of Medical Research, Australia
connor.smith@sydney.edu.au

Contributed Talk
Thursday 5 December, 4:20–4:40pm
Vines Room

Identifying key features and their regulatory relationships which underlie biological processes is the fundamental objective of much biological research; this includes the study of human diseases, with direct and important implications in the development of target therapeutics. We report new statistical approaches to identify various types of interpretable feature representations that are prognostically informative in classifying complex diseases. We present new ways to utilize information from thousands of resamples in modern selection methods and repeated subsampling to identify what features best predict disease progression. The new method VIVID utilizes feature importance measures via pairwise feature comparisons to identify significant features. We show how the selected features are repeatedly ranked higher and are more stable than other features. Taking advantage of cluster analysis, we construct a set of nested feature groups and then select an optimal group of features from the candidate models. Different methods of visualisation for this resampled information are presented. The computational speed and requirements of VIVID are discussed and how the method is able to deal with data where the number of features is continually increasing.

A multi-locus variable selection strategy for association mapping analysis

Beata Sznajder[1]
Anabel Forte[2], Julian Taylor[1]
[1]University of Adelaide, Australia, [2]Universitat de Valéncia, Spain
beata.sznajder@adelaide.edu.au

Contributed Talk
Thursday 5 December, 4:20–4:40pm
Exhibition Hall

In high dimensional gene-trait association studies often the challenge is to detect true associations of the phenotype with a small subset of genetic markers from a high dimensional set of markers. Additional challenges stem from statistical non-independence of genetic markers (linkage disequilibrium), the presence of population structure and cryptic relatedness in the population of sampled individuals. A common approach to genetic association and QTL mapping is to fit uni-variable linear mixed models testing each marker sequentially for associations with phenotype. Population structure is incorporated in the model and the false positive rate is usually controlled through adjustment of the usual 0.05 alpha level for multiple comparisons. The major disadvantage of this approach is that it does not provide an adjustment for multiple markers or potential gene-gene interactions. Testing of these additional terms can become problematic due to the substantial increase in the number of associations requiring testing compared to the number of sampled observations (the so-called $p \gg n$ scenario). We present an application of Bayesian variable selection for association mapping, implemented in the R package `BayesVarSel`. The approach of `BayesVarSel` employs model averaging to provide measures of association of phenotype with predictors (here genetic markers) across multi-variable models and includes functionality for scenarios where $p \gg n$. Additionally this approach explicitly accounts for multiple markers acting simultaneously on the phenotype. We compare the results obtained from `BayesVarSel` with the traditional uni-variable mixed model methodology through an application of QTL mapping of resistance to net-form-net-blotch (NFNB) in double haploid populations of barley.

mcvis: A new framework for collinearity discovery, diagnostic and visualization

Kevin Wang[1]
Chen Lin[2], Samuel Mueller[1]
[1]University of Sydney, Australia, [2]Fudan University, China
kevin.wang@sydney.edu.au

Contributed Talk
Thursday 5 December, 4:40–5:00pm
Vines Room

An essential step prior to any linear regression fitting is checking for collinearity. Without this step, hidden collinearity could compromise the statistical estimation, inference and interpretation of models. There exist several collinearity diagnostic statistics like the variance inflation factor and condition numbers in the literature. However, these collinearity measures have several limitations in practice. We present mcvis, a new framework that utilises conventional diagnostic statistics and sampling methods to better understand the cause of collinearity. We will present both simulated examples as well as real data examples to illustrate the novelty of mcvis. We will also introduce a bipartite graph visualisation that aids to reveal the variables causing collinearity beyond the usual diagnostics using correlation matrix.

Variable selection for genome wide association analysis in plant breeding

Anabel Forte[1]
Iker Oyanguren[1], Beata Sznajder[2], Julian Taylor[2]
[1]Universitat de València, Spain, [2]University of Adelaide, Australia
anabel.forte@uv.es

Contributed Talk
Thursday 5 December, 4:40–5:00pm
Exhibition Hall

In modern plant breeding comparative experiments, there is often a necessity to accurately determine the underlying genetic bases of industry driven phenotypic traits. With the availability of low cost high-throughput genotyping technology, this involves the mapping of quantitative trait loci (QTL) through a high dimensional genome wide association analysis pipeline. Within this pipeline, there may be several methods to reduce the dimension of the genetic marker set for further scrutiny.

In this work we present an efficient multi-step genome wide association analysis pipeline which incorporates a Bayesian variable selection strategy to identify putative QTL. In the initial step of this pipeline, a genomic prediction is conducted to obtain a complete set of marker effects. These effects are then used with a modification of a popular genome wide binning strategy to dramatically reduce the dimensionality of the marker set. With this reduced set of markers, a Bayesian variable selection is conducted to determine a small subset of markers linked to putative QTL. The complete genome wide analysis pipeline is illustrated with phenotypic and genotypic data obtained from a large Australian wheat panel. However, the genetic structure of lines may not be the only source of variability when trying to understand an output. In this sense, multienvironmental trials poses a mayor challenge adding an underlaying, complex, correlation structure.

In this work we discuss a two steps genome wide association analysis which makes use of a Bayesian variable selection strategy to determine markers significantly linked to a supposed QTL. The method is illustrated with a large Australian wheat panel.

`Vizumap`: An R package for visualizing uncertainty in spatial data

Petra Kuhnert[1]
Lydia Lucchesi[2], Christopher Wikle[3]
[1]CSIRO Data61, Australia, [2]University of Washington, USA, [3]University of Missouri, USA
Petra.Kuhnert@data61.csiro.au

Contributed Talk
Thursday 5 December, 5:00–5:20pm
Vines Room

The quantification, visualization and communication of uncertainty in spatial and spatio-temporal data is important for decision-making. It can highlight regions on a map that are poorly predicted and identify a need for further sampling. Uncertainty can also help to prioritise regions in terms of where to focus remediation efforts and allocate investment. It can also provide some assurance on where modelling efforts are working well and where it fails to trigger further investigation. Unfortunately, uncertainty is rarely included on maps that convey spatial or spatio-temporal estimates.

Approaches for visualizing uncertainty in spatial and spatio-temporal data will be presented. These include the bivariate choropleth map, map pixelation, glyph rotation and exceedance probability maps. Bivariate choropleth maps explore the "blending" of two colour schemes, one representing the estimate and a second representing the margin of error. The second approach uses map pixelation to convey uncertainty. The third approach uses a glyph to represent uncertainty and is what we refer to as glyph rotation. The final map based exploration of uncertainty is through exceedance probabilities.

`Vizumap` is an R package that has been developed within Digiscape for visualising uncertainty in spatial and spatio-temporal data. To illustrate the methods, I use an example from the Great Barrier Reef, where sediment loads were quantified from a Bayesian Hierarchical Model (BHM) that assimilated estimates of sediment concentration and flow with modelled output from a catchment model spanning 21 years of daily outputs and 411 spatial locations.

Genetic dissection of phytophthora root rot resistance in chickpea using modern statistical methods

Julian Taylor[1]
Amritha Amalraj[2], Tim Sutton[2]
[1]University of Adelaide, Australia, [2]SARDI, Australia
julian.taylor@adelaide.edu.au

Contributed Talk
Thursday 5 December, 5:00–5:20pm
Exhibition Hall

Phytophthora root rot (PRR) is a major soil borne disease that has the potential to significantly limit the yield potential of chickpea across the northern growing region of Australia. Although sources of resistance for PRR have been identified, the molecular basis of this resistance still requires quantification. This research discusses the statistical modelling approaches used to associate regions of the chickpea genome to PRR disease resistance traits obtained from multiple Recombinant Inbred (RIL) chickpea populations grown in varying field and glasshouse hydroponics experiments. Initially, the talk focusses on complex phenotypic and genetic analysis approaches used to detect quantitative trait loci (QTL) of plant survival traits across multiple field environments. The multi-environment analyses are then extended to include a plant survival trait derived from binary longitudinal measurements obtained from a high-throughput hydroponics screening system. The results from these analyses indicate PRR disease resistance in the RIL populations is strongly genetically related between field and controlled environment conditions. The results of this work have been recently published in two high impact international plant research journals.

Improving Estimates of Fried's Index from Mating Competitiveness Experiments

Dan Pagendam[1]
Nigel Snoad[2], Wen-Hsi Yang[3], Michal Segoli[4], Scott Ritchie[5], Brendan Trewin[6], Nigel Beebe[3]
[1]CSIRO Data61, Australia, [2]Verily Life Sciences, USA, [3]University of Queensland, Australia, [4]Ben-Gurion University of the Negev, Israel, [5]James Cook University, Australia, [6]CSIRO Health and Biosecurity, Australia
dan.pagendam@data61.csiro.au

Invited JABES Talk
Friday 6 December, 9:00–9:30am
Exhibition Hall

Sterile insect technique (SIT) and incompatible insect technique (IIT) are current methods for biological control of insect populations. Critical to the successful implementation of these biocontrol programs is quantifying the competitiveness of sterile/incompatible male insects for female mates relative to wildtype males. Traditionally, entomologists measure this mating competitiveness through a quantity known as Fried's Index. We establish that Fried's Index is mathematically equivalent to the mating competitiveness coefficient that features in many population models used in SIT/IIT programs. Using this insight, we propose a new approach for estimating Fried's Index from mating competitiveness experiments. We show that this approach offers greater precision and accuracy than the traditional approach that is currently used in many studies. This is demonstrated using both simulation experiments and by analysing real experimental data. To facilitate uptake of the proposed method, we also provide an R package that can be used to easily analyse data from mating competitiveness experiments.

Visualization and measures of population differentiation based on the saddlepoint approximation

Louise McMillan[1]
Rachel Fewster[2]
[1]Victoria University of Wellington, New Zealand, [2]University of Auckland, New Zealand
louise.mcmillan@vuw.ac.nz

Invited JABES Talk
Friday 6 December, 9:30–10:00am
Exhibition Hall

We propose a method for visualizing genetic assignment data by characterizing the distribution of genetic profiles for each candidate source population (McMillan & Fewster, 2017). This method enhances the assignment method of Rannala & Mountain (1997) by calculating appropriate graph positions for individuals for which some genetic data are missing. The saddlepoint method also provides a way to visualize assignment results calculated using the leave-one-out procedure.

This approach offers an advance upon assignment software such as `GeneClass2`, and is biologically more interpretable than plots provided by the widely-used software `STRUCTURE`. The visualization method makes it straightforward to detect features of population structure and to judge the discriminative power of the genetic data for assigning individuals to source populations.

We also propose new methods for quantifying population genetic structure. The measures we propose are closely related to the visualization approach, and enable visual features obvious from the plots to be expressed more formally. One measure is interloper detection probability: for two random genotypes arising from populations A and B, the probability that the one from A has the better fit to A and thus the genotype from B would be correctly identified as the 'interloper' in A. Another measure is correct assignment probability: the probability that a random genotype arising from A would be correctly assigned to A rather than B.

Using permutation tests, we can test two populations for significant population structure. These permutation tests are sensitive to subtle population structure, and are particularly useful for eliciting asymmetric features of the populations being studied, e.g. where one population has undergone extensive genetic drift but the other population has remained large enough to retain greater genetic diversity.

We illustrate these methods using microsatellite genotype data from ship rats and southern right whales, and SNP data from human populations.

Generalised latent variable models for multivariate abundances in ecology

David Warton[1]
Jenni Niku[2], Sara Taskinen[2], Francis Hui[3]
[1]University of New South Wales, Australia, [2]University of Jyväskylä, Finland, [3]Australian National University, Australia
david.warton@unsw.edu.au

Invited JABES Talk
Friday 6 December, 10:00–10:30am
Exhibition Hall

Technological advances have enabled a new class of multivariate models for ecology, with the potential to specify a statistical model for abundances jointly across many taxa, to simultaneously explore interactions across taxa and also the response of abundance to environmental variables. This talk will focus on generalised latent variable models, essentially a factor analytic extension of generalised linear models. These models can be used for a number of purposes of interest to ecologists, including: estimating patterns of residual correlation across taxa; ordination; multivariate inference about environmental associations; accounting for missing predictors; prediction of multi-taxon quantities like species richness. We discuss computation challenges fitting these models, current computational approaches and future directions.

Opportunities from combining statistical and crop growth models to predict Genotype × Environment interations over time

Daniela Bustos-Korts[1]
Scott Chapman[2,3], Karine Chenu[3], Martin P. Boer[1], Fred van Eeuwijk[1]
[1]Wageningen University and Research Centre, The Netherlands, [2]CSIRO Agriculture and Food, Australia, [3]University of Queensland, Australia
daniela.bustoskorts@wur.nl

Invited Talk
Friday 6 December, 11:00–12:00pm
Exhibition Hall

Genotype by environment interaction (G×E) for the target trait, e.g. yield, is an emerging property of agricultural systems and results from the interplay between a hierarchy of secondary traits involving the capture and allocation of environmental resources during the growing season. This hierarchy of secondary traits ranges from basic traits that correspond to response mechanisms/sensitivities, to intermediate traits that integrate a larger number of processes over time and therefore show a larger amount of G×E. Traits underlying yield differ in their contribution to adaptation across environmental conditions and have different levels of G×E. Here, we provide a framework to study the performance of genotype-to-phenotype (G2P) modelling approaches. We generated and analysed response surfaces, or adaptation landscapes, for yield and yield related traits, emphasizing trait dynamics and interactions over time. We used the crop growth model APSIM-wheat with genotype-dependent parameters as a tool to simulate non-linear trait responses over time with complex trait dependencies, and applied it to Australian environments. Such simulated data opens new opportunities to i) study the relationships between traits over time and across environments, ii) evaluate genotype-to-phenotype models for multiple traits and environments, iii) compare models with an increased integration of statistical and biological aspects, iv) develop network models that allow visualizing how causality is propagated in biological systems and v) design efficient high throughput phenotyping schedules and methods. We show applications to wheat under contrasting water deficit patterns.

Statistical Methods for data from High Throughput Phenotyping in Agriculture

Joanne De Faveri[1]
Ari Verbyla[2]
[1,2]CSIRO Data61, Australia, [1]Queensland Department of Agriculture and Fisheries, Australia
Joanne.DeFaveri@data61.csiro.au

Invited Talk
Friday 6 December, 12:00–1:00pm
Exhibition Hall

High throughput phenotyping (HTP) data is being collected in the field and laboratory by drones, helicopters, sensors and buggies, often in the form of images and spectra. In plant trials these phenotyping methods often result in large amounts of data being collected over the crop growing season. The spatial location of the plants may impact on the growth of the plants and influence the HTP traits being measured, while the temporal correlation between repeated measurements and trends over time will also impact the traits.

In plant breeding trials, modelling the genetic effects of HTP traits is of primary interest but there is also a need to account for non-genetic effects (design, spatial and temporal) to obtain accurate and unbiased estimates of the genetic effects. The size of the HTP data may cause issues with traditional spatial and temporal modelling approaches, hence alternatives are proposed based on multi-dimensional reduced rank tensor smoothing splines.

Not only can HTP data be measured over time and space but also over thousands of wavelengths for hyperspectral data and the data may also be measured at multiple sites. The interaction between environment, time, space and wavelength needs to be modelled at both genetic and non-genetic levels in suitable ways.

Often HTP traits are measured with the aim of informing other traits of primary interest, for example yield. In many cases HTP data is used together with genomic information for genomic prediction of these primary traits. Functional Data Analysis (FDA) models provide one approach for the analysis of this type of data. New functional regression models in ASReml are presented, allowing for spatio-temporal modelling and efficient modelling of Genotype by environment (G×E) effects.

This talk aims to provide a snapshot of some exciting statistical developments in the analysis of HTP data and to discuss further research topics.

Cross-validation of pasture biomass predictions from handheld NDVI sensor measures

Angela Anderson[1]
Nicole Spiegel-Janecek[1], Heather Johsson[2], Bob Shepherd[1]
[1]Queensland Department of Agriculture and Fisheries, Australia, [2]Dalrymple Landcare Committee, Australia
angela.anderson@daf.qld.gov.au

Contributed Poster
Tuesday 3 December, 5:00–7:00pm
The Gallery

Forage budgeting is a pasture management tool enabling graziers to match stocking rates with forage supply. The yields required for a forage budget are usually determined from a visual estimate. However, visual yield estimates are subject to bias. Normalised Differential Vegetation Index (NDVI) sensors offer an alternate objective method for assessing yield. This study assessed the accuracy of a handheld NDVI sensor in predicting pasture biomass. The training data, consisting of pasture measurements from 191 quadrats sampled over a 12 month period, were obtained across two properties (with a mix of pasture types) in the Burdekin grazing region in North Queensland. NDVI readings for each quadrat were made using a Trimble GreenSeekerR. Quadrats were then cut and dried to calculate various measures, including biomass and crude protein. Linear and curvilinear regression models were fitted to the training data to gain prediction equations for green biomass from NDVI or NDVI x pasture height, for the wet and dry seasons. The best models were chosen by assessing the adjusted R2 and normal probability plots. The validation process involved using the best model obtained from the training data, to predict the biomass response for the NDVI observations from an independent validation data set. The predictive ability of the models was assessed by the mean square error of prediction (MSEP), the root mean square error (RMSE - to quantify the error of prediction on the original scale of the data) and the prediction bias. Cross-validation indicated the use of NDVI technology for predicting biomass of mixed pastures showed promise. Future work will involve refining the method to enable more precise biomass predictions (by using a modified NDVI value) and assessing the prediction accuracy of other pasture measures (e.g. crude protein). This will improve the practicality of the sensors' use for graziers.

Bayesian Approaches to Carbon Cycle Modelling

Mohammad Javad Davoudabadi[1]
Gentry White[1], Christopher Drovandi[1], Daniel Pagendam[2]
[1]Queensland University of Technology, Australia, [2]CSIRO Data61, Australia
mohammadjavad.davoudabadi@hdr.qut.edu.au

Contributed Poster
Tuesday 3 December, 5:00–7:00pm
The Gallery

The exchange of carbon between and within four main pools (or reservoirs) is called the global carbon cycle. These pools include the oceans, the atmosphere, fossil fuels, and land. Increasing greenhouse gases especially CO_2 which is considered as a response for global climate change, reduction of global soil quality and consequently reduction of crop productivity caused scientists try to modelling the soil carbon process and forecasting the change of soil carbon in the future to deal with these problems. In addition to that financial gains by selling carbon credits is another aspect that encourage to study in this area. Soil carbon storage and temperature have a negative correlation throughout the world, so that, it generally increases when mean annual temperature decreases. Sequestration which alludes a combination of both storage and capture, has many benefits includes declining greenhouse gases, improved crop productivity, and potential financial gains. The challenge is that the sequestration process and the carbon cycle are poorly understood. Several deterministic models of soil carbon dynamics have been introduced to describe the sequestration process with more precision, but these models do not account for uncertainty in a comprehensive manner. Uncertainty affects parameters, model inputs, and consequently predictions from the model. Some researches have been done for quantifying uncertainty in model outputs especially either under simulated climate change or via a sensitivity analysis by running models for different sets of parameter values. In this context, Bayesian models are fit using Markov chain Monte Carlo (MCMC) methods and applied to soil carbon modelling to quantify uncertainty in soil carbon cycle. This research is expected to provide new and extended methods under a Bayesian framework to evaluate more complex models. This research will lead to efficiency gains in computational processes and the discovery of important factors that affect the global carbon cycle.

Analysis of continuous water use data for wheat plants grown on a Droughtspotter platform

Nathaniel Jewell[1]
Chris Brien[1], Abdeljalil El Habti[1], Penny Tricker[1], Trevor Garnett[1]
[1]University of Adelaide, Australia
nathaniel.jewell@adelaide.edu.au

Contributed Poster
Tuesday 3 December, 5:00–7:00pm
The Gallery

Gravimetric watering platforms such as the Droughtspotter enable potted plants to be automatically weighed and watered multiple times per day. The benefits for controlled-environment experiments include reduced manual labour, improved quality control, and the ability to continuously measure water loss from individual pots. A Droughtspotter experiment at The Plant Accelerator shows that transpiration rate in wheat plants is strongly correlated with Vapour Pressure Deficit (VPD), a simple measure of the drying power of air. The experiment encompassed several wheat varieties and watering regimes, including simulation of drought or heatwave conditions during early spring. Linear mixed models are presented to relate water use and water use efficiency to VPD, variety, watering regime and final grain yield.

Comparative study of probability models for agriculture field index data

Olena Kravchuk[1]
John van der Hoek[2]
[1]University of Adelaide, Australia, [2]University of South Australia, Australia
olena.kravchuk@adelaide.edu.au

Contributed Poster
Tuesday 3 December, 5:00–7:00pm
The Gallery

This simulation study is motivated by the challenges of the analysis of field indices, like Harvest Index and Nutrient Use Efficiencies, in agriculture research. The index data we are interested in represents the quotient of two positively correlated random variables, $X/(X+Y)$, but is only observed as a single index variate, Z. Based on the analysis of that index, important decisions are made about the ranking of agronomic factors and conditions. In this study, we are concerned with probability models for such indices, considering and contrasting several choices for X and Y. The choices include: normal, Beta and log-normal distributions, as well as mixtures of either Beta or log-normal. We demonstrate that the mixture of log-normal provides an identifiable and flexible model.

Fitting a nugget variance to the analysis of National Variety Trials

Chris Lisle[1]
David Hughes[1], Ky Mathews[1]
[1]University of Wollongong, Australia
clisle@uow.edu.au

Contributed Poster
Tuesday 3 December, 5:00–7:00pm
The Gallery

It has been over 20 years since the Gilmour et al. (1997) landmark paper, which developed much of the recent statistical methodology for the analysis of variety trials. These methods are still used widely today in many crop improvement programs around the world, including the analysis of variety trials in the GRDC funded National Variety Trial (NVT) program. This program has been running since 2005 with over 10,000 trials assessed across eleven crops.

Gilmour et al. (1997) recommends using a separable first-order autoregressive model in two dimensions (denoted as AR1xAR1). Gilmour et al. (1997) also suggested the inclusion of the so-called 'nugget' variance (i.e. measurement error). The use and requirement of this term has caused conjecture, as well as issues with model convergence. These have created debate over the appropriateness of this method, and in particular, the use of the AR1xAR1 spatial process for the analysis of variety trials. However, the nugget variance is not routinely fitted to NVT and many other crop improvement programs.

This poster summarises the analysis of 170 NVT wheat trials, each with and without a nugget variance fitted. Fitting this term caused slight model convergence issues, however, there was negligible changes to the predicted variety effects, with a correlation of 0.99 across all trials with and without nugget effects. The estimated size of the nugget effect has also been shown to contradict values presented in simulation studies within several papers.

Yield response curves using fixed effects vs random coefficient regression across environments

Yao Lu[1]
Kerry Bell[1], Michael Mumford[1], David Lester[1]
[1]Queensland Department of Agriculture and Fisheries, Australia
yao.lu@daf.qld.gov.au

Contributed Poster
Tuesday 3 December, 5:00–7:00pm
The Gallery

Understanding crop yield response to fertiliser nutrients is crucial to agronomists and growers. To investigate the impact of differing conditions, it is important to implement field trials across multiple environments. The measured yield response to fertiliser rates for each environment is influenced by background nutrient levels and the 'noise' in the trial. The comparison of estimating response curves using fixed effects or random coefficient regression approach in a linear mixed model framework will be explored.

Estimating response curves using fixed effects has the advantage of fitting individual curves to each environment without biasing the fit. When there are limited and noisy data the disadvantages include fits that are readily influenced by points with high leverage and outliers. Clustering environments with similar shaped response curves can help identify meaningful groups with similar characteristics.

Estimating response curves via random coefficient regression has the advantage of 'borrowing' information from other environments through modelling covariance matrices. A disadvantage is that variance components related to the formation of the response curves across environments can bound to zero and force each environment to have the same response curve shape. This would be unrealistic where it was expected that some environments would have no yield response while others should exhibit a response due to the background nutrient levels of each environment.

Examples of deep phosphorus application to field trials conducted over the last few years in several locations in Queensland and New South Wales will be given and analyses fitting response curves with fixed effects and random coefficient regression will be compared. Each trial has replicated levels of the rate of fertiliser and this will be taken into account during the modelling process. The results of this comparison will provide clarity for when random coefficient regression should be implemented over a fixed effect approach and vice versa.

Multivariate analysis for Greenhouse Gas Emissions from New Zealand Sheep and Beef Farm Systems

Esther Meenken[1]
Ronaldo Vibart[1], Kathryn Hutchinson[1], Grant Rennie[1], Andrew Burtt[2], Jane Chrystal[2], Robyn Dynes[1]
[1]AgResearch, New Zealand, [2]Beef + Lamb New Zealand, New Zealand
esther.meenken@agresearch.co.nz

Contributed Poster
Tuesday 3 December, 5:00–7:00pm
The Gallery

Research investigating greenhouse gas (GHG) emissions from pastoral farming in New Zealand has, until recently, been focused on the impact of dairy. The New Zealand Agricultural Greenhouse Gas Research Centre and Beef + Lamb New Zealand have identified that an understanding of the impact of less intensive pastoral practice on GHG emissions is needed. The drivers of GHG emissions on largely sheep and beef mixed farm systems are unlikely to follow those of dairy farms because of substantial differences in farm management practices, climate and landscape characteristics. The biological feasibility (e.g. matching feed supply with feed demand) of the various stock classes on the farm was determined using Farmax. The whole-farm nutrient model Overseer® was used to examine the environmental outputs, including GHG emissions of farm blocks. These models characterise farms in relatively high detail and can provide hundreds of descriptive variables, and many non-independent representations of GHG emissions. The objective of this analysis was to identify farm management practices that resulted in fewer GHG emissions under various landscape and management conditions. Feature selection and dimension reduction techniques were compared to help describe, visualise and understand the key drivers of GHG emissions and how to reduce them in the sheep and beef sector in New Zealand.

Is the SpATS model as good as we would like it to be for the spatial analysis of field trials?

Lucas Peitton[1,2]
Luciana Magnano[2], Valeria Paccapelo[1], M. Gabriela Borgognone[1]
[1]Queensland Department of Agriculture and Fisheries, Australia, [2]Universidad Nacional de Rosario, Argentina
Lucas.Peitton@daf.qld.gov.au

Contributed Poster
Tuesday 3 December, 5:00–7:00pm
The Gallery

Variation in agricultural field experiments must be accounted for to ensure accurate comparisons of tested breeding lines. The most generally used statistical method for the spatial analysis of field experiments is the linear mixed model approach introduced by Gilmour et al. (1997), which consists of a multi-step procedure that uses graphical diagnostics and formal statistical tests to obtain the best spatial model. This model is based on an underlying separable variance structure for an auto-regressive process across the row and column dimensions in the field. Recently, a novel one-step-procedure based on penalized splines was developed in the mixed model framework (Rodríguez-Álvarez et al., 2016a). This modelling approach, called Spatial Analysis of field Trials with Splines (SpATS), was tested in large sorghum breeding trials that followed partially replicated designs. In this context, the SpATS model performed similarly to the model obtained under the traditional approach. However, it has not yet been evaluated for other conditions such as other crops, smaller trials with less breeding lines, or trials following different experimental designs. The aim of the present study is to compare the SpATS and the traditional spatial models in large and small mungbean and chickpea trials that follow fully and partially replicated designs. The analysis of each trial is performed utilising the `SpATS` package (Rodríguez-Álvarez et al., 2016b) and the `ASReml-R` package (Butler et al., 2009) for the traditional approach. The comparison is focussed on the variance components estimates and on the correlations between predicted values for the tested lines obtained using each approach in the same trial.

Computation of the expected value of a function of a chi-distributed random variable

Nishika Ranathunga[1]
Paul Kabaila[1]
[1]La Trobe University, Australia
n.kapuruge@latrobe.edu.au

Contributed Poster
Tuesday 3 December, 5:00–7:00pm
The Gallery

We consider the problem of numerically evaluating the expected value of a smooth bounded function of a chi-distributed random variable, divided by the square root of the number of degrees of freedom. This problem arises in the contexts of simultaneous inference, the selection and ranking of populations and in the evaluation of multivariate t probabilities. It also arises in the assessment of the coverage probability and expected volume properties of some non-standard confidence regions. We propose the application of the "Mixed Rule" transformation, followed by the application of the trapezoidal rule. This rule has the remarkable property that, for suitable integrands, it is exponentially convergent. We provide some numerical illustrations of the advantages of this method.

A comparison of linear mixed model packages in R for analysis of plant breeding experiments

Sam Rogers[1]
Julian Taylor[1]
[1]University of Adelaide, Australia
s.rogers@adelaide.edu.au

Contributed Poster
Tuesday 3 December, 5:00–7:00pm
The Gallery

There is an abundance of packages in R that provide functionality for data analysis using linear mixed models. In this research we investigated three mature linear mixed model centric packages (`ASReml-R V4`, `LME4` and `sommer`) and compared their functionality and computational efficiency in the context of plant breeding genomic selection models. The ultimate aim of this work was to assess the viability of each of the packages for use in future plant breeding research involving very large numbers of individuals. To motivate this assessment, we analysed data from subsets of a large plant breeding field trial conducted by Australian Grains Technologies (AGT) containing 10375 lines and 17171 genetic markers. In this poster we will present a theoretical overview of the linear mixed models required for genomic selection analysis of the AGT data as well as the computational syntax required to fit these models using the functionality in each of these packages. We will also include some of the challenges we encountered during this work and concludes with a comparative discussion on the computational efficiency of the linear mixed model algorithm in each package.

Exploration of trait relationships in mungbean using a multivariate linear mixed model

Eugenia Settecase[1,2]
Cristina Cuesta[2], M.Gabriela Borgognone[1], Valeria Paccapelo[1]
[1]Queensland Department of Agriculture and Fisheries, Australia, [2]Universidad Nacional de Rosario, Argentina
eugenia.settecase@daf.qld.gov.au

Contributed Poster
Tuesday 3 December, 5:00–7:00pm
The Gallery

Plant improvement programs involve the evaluation of potential new varieties in designed experiments. The selection of the best performing genotypes is usually based on multiple traits measured on the same experimental unit (plot) from field trials. The study of the association between these traits can help with the selection process and the development of new genotypes. When two traits are involved, Ganesalingam et al. (2013) proposed a bivariate linear mixed model that incorporates different error and genotypic variances per trait as well as genotypic and residual correlations between traits, resulting in more accurate predicted values for each genotype.

The National Mungbean Improvement Program develops new varieties with combined characteristics or traits specifications according to different market needs. A field experiment was carried out to evaluate the diversity of a set of genotypes in terms of two different groups of seed traits: seed coat colour related traits (lightness level and two chromaticity coordinates) and yield related traits (pod length, number of seeds per pod, and seed weight).

To analyse the three traits corresponding to the seed coat colour and the yield related traits, the modelling approach proposed by Ganesalingam et al. will be extended, retaining all the advantages of the bivariate analysis. Results from this modelling approach will provide insights to help understand the genotypic relationships between pairs of traits within each group. Furthermore, the genotype predicted values for each trait will support genotype selection as well as parent identification for the development of new mungbean genotypes.

Random thoughts on p-values

Alan Welsh[1]
[1]Australian National University, Australia
Alan.Welsh@anu.edu.au

Contributed Poster
Tuesday 3 December, 5:00–7:00pm
The Gallery

Much of the recent commentary on p-values has been negative with some journals and societies taking strong stands against their use. They are even blamed for the reproducibility crisis in some fields of applied statistics. Some of this may be due to misunderstandings of what p-values are. We consider a very simple example for which we can do various calculations exactly and use the results of these calculations to gain insight into the properties of p-values. These insights then help to clarify the nature and interpretation of p-values.

NIRS Classification

Carole Wright[1]
[1]Queensland Department of Agriculture and Fisheries, Australia
carole.wright@daf.qld.gov.au

Contributed Poster
Tuesday 3 December, 5:00–7:00pm
The Gallery

There are numerous supervised classification techniques which have the purpose of assigning objects into different groups. Three of the most commonly used techniques for near-infrared spectroscopy (NIRS) data include principal components discriminate analysis (PC-DA), support vector machines (SVM) and partial least squares discriminate analysis (PLS-DA). PC-DA produces a number of orthogonal linear discriminant functions that minimise the variance within groups, while maximising the separation between the groups. Often the number of wavelengths contributing to each spectra is greater than the number of samples. To overcome the requirement that there are more samples than variables, the data dimensionality is reduced using PCA prior to applying the DA. SVM is a technique which has the ability to model linear and non-linear classification problems by identifying a hyperplane that maximises group separation. The third technique, PLS-DA, uses PLS regression to discriminate between the groups by assigning a binary coding to identify group association. Although these three techniques are commonly used, it is not usually clear which technique is the most appropriate. This study compares the application of these three classification methods on three NIRS data sets. The correct classification rate for the individual groups and the overall correct classification rate is used as the criterion to compare the three methods.

Abstracts Index

Akkaya-Hocagil, Tugba, 70
Anderson, Angela, 89
Anderson, Marti J., 67

Burden, Conrad, 77
Bustos-Korts, Daniela, 87

Cao, Zhanglong, 59
Carpenter, James, 29
Christiansen, Rune, 31
Clark, Robert, 69
Clarke, Brenton R, 51
Cook, Richard, 41
Czado, Claudia, 66

Davoudabadi, Mohammad Javad, 90

Faveri, Joanne De, 88
Forknall, Clayton, 52
Forte, Anabel, 81
Foster, Scott, 71

Gladish, Dan, 75
Guilbault, Emy, 37

Hepworth, Graham, 34
Hockey, Hans, 30
Hui, Francis, 33

Jarrett, Richard, 44
Jewell, Nathaniel, 91

Kabaila, Paul, 45
Kasprzak, Peter, 65
Kasza, Jessica, 32
Kravchuk, Olena, 92
Kuhnert, Petra, 82
Kurosawa, Takeshi, 57

Lisle, Chris, 93
Lu, Yao, 94
Lumley, Thomas, 76

Macdonald, Bethany, 46
Marmolejo-Ramos, Fernando, 47
McMillan, Louise, 39, 85
Meenken, Esther, 64, 95
Menon, Nidhi, 36
Moldovan, Max, 40
Muller, Warren, 63
Mumford, Michael, 54
Munoz-Santa, Isabel, 50

Neeman, Teresa, 60
Nghiem, Linh, 49
Nielsen, Sharon, 62

Ozturk, Omer, 73

Pagendam, Dan, 84
Peitton, Lucas, 96

Ranathunga, Nishika, 97
Renner, Ian, 35
Richardson, Alice, 68
Robertson, Blair, 56
Rogers, Sam, 98
Rognoni, Bethany, 48

Settecase, Eugenia, 99
Smith, Connor James, 78
Steel, David, 43
Stewart, Michael, 53
Sullivan, Thomas, 38
Sznajder, Beata, 79

Taylor, Julian, 83
Thas, Olivier, 72

Wang, Kevin, 80
Wang, Tong, 74

Warton, David, 86
Welsh, Alan, 100
Wikle, Christopher K., 55
Williams, Emlyn, 42
Wilson, Susan, 58
Wright, Carole, 101

Yoon, Hwan-Jin, 61