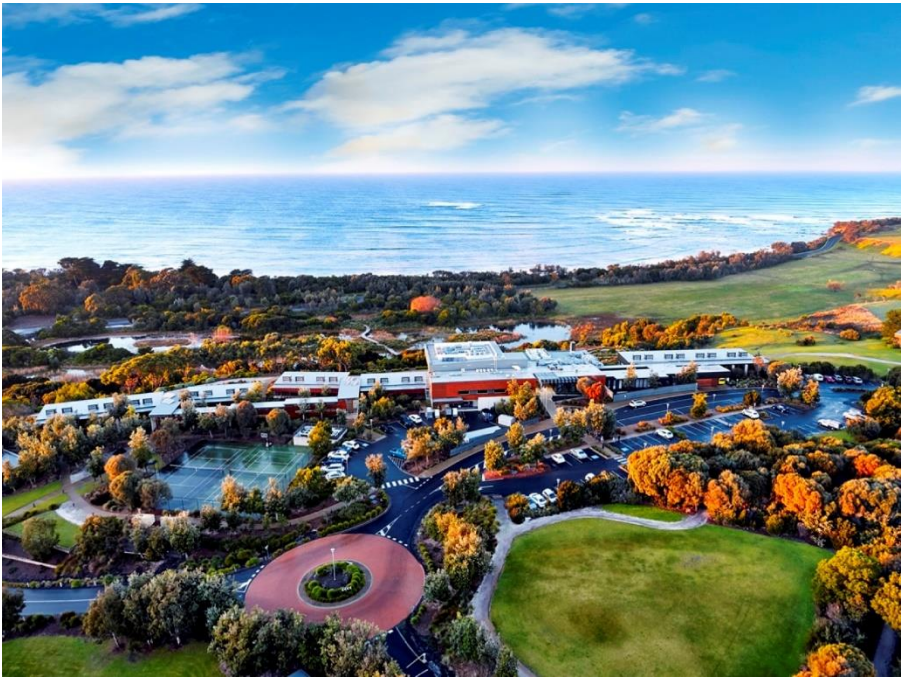
AASC 2022

28 November – 2 December 2022

RACV Inverloch Resort, Victoria, Australia



AUSTRALASIAN APPLIED STATISTICS CONFERENCE

ABSTRACT AND PROGRAMME BOOK

Sponsor:

VSN international limited

Contents

About AASC and AGUAi

This conference is the most recent in a series of conferences originating with the initial Australasian Genstat conference in Canberra in 1979. At the 2008 conference in Victoria it was voted to broaden the focus of the conference to encompass the wider community of applied statisticians. The conference was therefore renamed as the Australasian Applied Statistics Conference. The conference attracts biometricians and statisticians working mainly in primary industries and environmental science research in Australia and New Zealand. The first AASC was in 2011 at Palm Cove (Aus), followed by conferences every two years in Queenstown (2012 NZ), Port Lincoln (2014 Aus), Bermagui (2016 Aus) and Rotorua (2018 NZ). The conference due in 2020 was postponed to 2022 due to the Covid epidemic.

The Australasian Genstat Users' Association incorporated (AGUAi) exists to organise the biennial AASC. The AGUAi is an independent, non-commercial, not-for profit, association. It maintains an informal relationship with the Genstat software and its developers VSN international. Anyone can become a member of the AGUAi by simply attending an AASC and indicating their agreement to membership – there are no membership fees or other requirements. Being a legally incorporated association offers some on-going structure and a degree of legal protection for the committee organising conferences. The AGUAi AGM will be held on Thursday 1 Dec, where the main item is to select a new AGUAi committee.

AGUAi Committee and AASC organizing sub-committees

Position	Name	Affiliation	Sub-committees
Officers:			
Chair	Murray Hannah	Retired	Local
Vice-Chair	Andrew van Burgel	DPIRD Western Australia	
Secretary	Khageswor Giri	Ag Victoria	Local
Secretary	David Baird	VSNi, NZ	Program
Treasurer	Debra Partington	Ag Victoria	Local
Members:			
	Roger Payne	VSNi, UK	
	Angela Anderson	DAFF, QLD	Program
	Ruth Butler	StatsWork 2022 Ltd	Program
	Vanessa Cave	VSNi NZ	Program
	Benoit Liquet-Welland	Macquarie University	
	Chris Triggs	University of Auckland	Program
	Emi Tanaka	Monash University	Website
	Peter Kasprzak	University of Adelaide	Website
	Pierre Lafaye de Micheaux	University of New South Wales	
	Pauline O'Shaughnessy	University of Wollongong	Program
	Sam Rogers	University of Adelaide	Website
	Kym Butler	University of Melbourne	Local

AASC 2022: Workshops

Monday 28th November: 9am-4:30pm

Genstat Master Class – Data Mining and Advanced Modelling

Cape Liptrap room. Presenters: Dr. Vanessa Cave and Dr. David Baird, VSNi New Zealand

Data Visualisation with R

The Board Room. Presenter: Dr. Emi Tanaka, Monash University

AASC 2022 Conference Social Functions:

The cost of attendance for each of these are included in the conference registration fee. Extra tickets may be available for purchase: see the registration desk.

Welcome Buffet: 7pm on Monday 28 November (Panorama Room)

Poster-session Canapes and Drinks: 6:45pm on Tuesday 29 November (Panorama Room)

Conference Dinner: 7pm on Thursday 1 December (Panorama Room)

Registration Desk

The registration desk will be outside of the main conference space in Reception.

Desk Opening Times:

Sunday	27th	6:30-8:00 pm
Monday	28th	8:30-9:00 am and 4:00-5:30 pm
Tuesday	29th	8:00-8:30 am

Buses To/From Melbourne:

	To Inverloch		**To Melbourne**
	Sunday 27th Nov:	Monday 28th Nov:	Friday 2nd Dec
Airport	4:00	1:00	4:30
Southern Cross Station	4:45	1:45	5:15
Inverloch	7:15	5:00	1:30

AASC 2022: Conference Trips

Wed 30th Nov

There are five basic trips, and a further seven combinations of these trips. For full details see the AASC2022 website. https://aasc2022.netlify.app/

A: In-Cider Story: **CANCELLED**

~~Gurney's Cidery and Promontory Restaurant & Winery: this is a chance to see how cider is made rom start to finish. https://gurneyscider.com.au/~~

B: Churchill Island:

Churchill Island is a 50.7-hectare island reachable by road bridges. It contains many historical features, and is adjacent to the Churchill Island Marine National Park. Enjoy the farm demonstrations or go for an easy stroll around the island amongst the remnant flora, and/or relax at the café.

C: Seal Rocks Boat Cruise:

Experience an up-close encounter with thousands of seals at Australia's largest fur seal colony at Seal Rocks, on a 2-hour cruise

D: Phillip Island Penguin Parade:

The Penguin Parade is Victoria's most popular tourist attraction. Phillip Island is home to the largest Little Penguin colony in the world. You can experience the magic of watching these amazing seabirds waddle home from the ocean to their burrows.

Warm clothes and wind protection are essential

H-K: George Bass Coastal Walk:

This is a picturesque 7km (2+ hr) walk along headlands from Kilcunda to Punchbowl. Suitable footwear and all-weather clothing are essential. Because of transport constraints people taking this walk are obliged to join either the Churchill Island or Seal Rocks excursion.

Conference Trips Outline Timetable

Option	Trip	Time	Detail
A	~~Cidery & Winery~~		CANCELLED
B	Churchill Island	11:00	Depart RACV Inverloch
		11:45	San Remo, Pelicans, Lunch
		13:00	Bus departs San Remo
		13:15	Churchill Island
		16:30	Bus pickup at Churchill Island
		17:30	Arrive back at RACV Inverloch
C	Seal Rocks	11:00	Depart RACV Inverloch
		11:45	San Remo, Pelicans, Lunch
		13:15	Bus departs San Remo
		13:45	Cowes jetty and boat to Seal Rocks
		16:00	Bus pickup at Cowes jetty
		17:15	Arrive back at RACV Inverloch
D	Penguin Parade	15:45	Depart RACV Inverloch
		17:00	Dinner at Cowes
		18:15	Bus departs from Cowes
		18:30	Arrive at The Nobbies
		19:45	Penguin Parade
		21:30	Depart Penguin Parade
		22:30	Arrive back at RACV Inverloch
E	~~Cidery & Winery + Penguins~~		CANCELLED
F	Churchill Island + Penguins	11:00	Option B until 16:30
		16:30	Penguin bus pick-up at Churchill Island
		17:00	Dinner at Cowes, follow Option D.
G	Seal Rocks + Penguins	11:00	Option C until 16:00
		16:00	Remain at Cowes, 1hr free time
		17:00	Join in with Option D.
H	Coastal Walk + Churchill Island	8:30	Depart RACV Inverloch
		9:15	Walk from Kilcunda
		11:15	Bus pickup at Punchbowl
		11:45	Join in with Option B
I	Coastal Walk + Seal Rocks	8:30	Same as Option H, but …
		11:45	join in Option C
J	Coastal Walk + Churchill Island + Penguins	8:30	Same as Option H, but …
		11:45	join in Option B until 16:30
		16:30	join in Option F
K	Coastal Walk + Seal Rocks + Penguins	8:30	Same as Option H, but …
		11:45	join in Option C until 16:00 …

Session guidance for presenters and chairs

Given the large number of talks, we strongly request that everyone keep to the times given.

Invited Talks: 45 min including 5 min for questions.

Talks: 15min: Questions to be left to the end of the session. Sessions of six 15min talks will run with a short break after the first three talks. Up to 10 min will be available for questions after all talks have been completed.

QuickFire Talks: 10min. Questions will be left to after all of the talks have been completed, with up to 10 min available.

Posters: The poster session will start with 1min talks for each poster, in the order given on page 54. You may wish to present your poster as a single slide as a background to this talk (one slide only please). Please email the slide to aascconf2022@gmail.com with subject 'Poster {your name}' or take it to the registration desk. Poster presenters are invited to also use this method to make their poster available on the website (https://aasc2022.netlify.app/).

Session Chairs will be given two 'time left' signs and a 'stop' sign to hold up at the appropriate moment. Presenters should expect to stop talking when the 'stop' sign is shown.

Uploading talk slides: Talks need to be uploaded to the conference computer by the start of the day on which the talk is to be given. Speakers can either email their slides to aascconf2022@gmail.com with subject 'Talk {your name}' or take a file with their slides to the registration desk.

Invited Speakers

Carolyn Huston: Tuesday, 9:00 – 9:45

Emulation of the Spark bushfire solver, and building trust in uncertainty

Carolyn Huston is currently a Senior Research Consultant at CSIRO's Data61, where she has been involved in numerous commercial and strategic research projects related to bushfire risk and resilience, including estimation of fire consequence and loss. She completed a BSc. in Ecology at the University of Alberta, Canada; an MSc. in Biostatistics at the University of Alberta, Canada; and a PhD in Statistics at Simon Fraser University, where she researched West Coast of British Columbia Sockeye fisheries estimation, openings, and related problems using Bayesian hierarchical methods.

Francis Hui: Tuesday 1:00 – 1:45

All Under One Roof – The Rise of Joint Species Distribution Modeling in Ecology

Francis Hui is a Senior Lecturer in Statistics at the Australian National University. Having completed his PhD at the University of New South Wales in 2014, Francis moved to Canberra to undertake a postdoctoral fellowship at the ANU, and has been willingly stuck there since. His research spans a mixture of methodological, computational, and applied statistics, including longitudinal and correlated data analysis using mixed and/or marginal models, dimension reduction and variable selection, and approximate statistical estimation and inference. Much of his applications are motivated by joint modeling in community ecology, and temporal analysis of social and environmental drivers for mental health. All of his research is complemented by copious amounts of tea drinking and unhealthy amounts of anime watching.

Ed Cripps: Tuesday 3:30 – 4:15

Uncertainty quantification and communication for the earth sciences

Edward Cripps, deputy director of DARE centre (Data Analytics for Resources and Environments), is an Associate Professor in the Department of Mathematics and Statistics at the University of WA. His research interests are in Bayesian longitudinal analysis and spatio-temporal models, and the integration of statistical and physical models. His primary applications are in the statistical modelling of environmental, meteorological and oceanographic processes and their interaction with engineering decision making and asset management. Edward has extensive experience in industry collaboration and translating academic research output into commercially industrial applications. He is currently working in collaboration with Shell, Woodside, Lloyds and the Alan Turing Institute, the UK's national data science institute.

Vanessa Cave and David Baird: Thursday, 8:45-9:30

New Developments in Genstat 22

David Baird is a consultant statistician with 35-years' experience and has been a Genstat developer for 25 years. He was a biometrician at AgResearch for 25 years before starting his own company VSN NZ. He has worked in a wide range of disciplines including biosecurity, entomology, agriculture, ecology, soil science, plant breeding and microarrays. His statistical interests include experimental design, spatial analysis, data mining and statistical modelling. For the last 9 years he has been the NZ Earthquakes Commission's statistical consultant. In 2019 he was awarded an ALF Cornish award for contributions to biometrics in Australasia. David has a MSc in Applied Statistics from the University of Reading and a PhD in Statistics from the University of Otago.

Vanessa Cave is an applied statistician interested in the application of statistics to the biosciences, in particular agriculture and ecology, and is a developer of

the Genstat statistical software package. She has many years' experience collaborating with scientists in the agricultural and environmental sciences, using statistics to solve real-world problems. As a biometrician, Vanessa provides expertise on experiment and survey design, data collection and management, statistical analysis, and the interpretation of statistical findings. Vanessa is also an active member of the Australasian statistical community, serving on the New Zealand Statistical Association committee and president-elect of the Australasian Region of the International Biometric Society. She is also an editorial board member for New Zealand Veterinary Journal, an associate editor for Agronomy Journal and an honorary academic at the University of Auckland. Vanessa has an honours degree in Statistics from the University of Otago and a PhD in Statistics from the University of St Andrews.

Patricia Menéndez: Thursday 1:00 – 1:45

Practical steps toward reproducibility

Patricia Menéndez is a senior lecturer at the Department of Econometrics and Business Statistics at Monash University Business School. Patricia's training is in mathematics and statistics and she received her PhD from ETH Zurich in Switzerland. Since completing her PhD she has held academic positions at Wageningen University, University of New South Wales and University of Queensland. Before joining the department she has also worked outside academia as statistician/ecological statistician for the NSW Bureau of Crime Statistics and Research, and for the Australian Institute of Marine Science.

Patricia has expertise developing and applying statistical methodology and computational methods as well as providing statistical training both in academic and non-academic environments. During her time outside of academia, she has worked on multidisciplinary projects to answer research and policy-making questions in the fields of climate change, environmental and marine sciences besides criminology.

Her research interests include statistical inference, functional data analysis, methods for time series, computational statistics, data visualisation and data science tools.

Emlyn Williams: Friday, 8:45 – 9:30

The application of CycDesigN in designing experiments for use in agriculture

Emlyn worked as a statistical consultant with CSIRO for 35 years before becoming Director of the Statistical Consulting Unit at ANU (2006-2013). He has wide experience in the application of statistical methods and has published around 150 journal articles and 2 books. In 2019 he received the E.A. Cornish Award from the Australasian Region of the International Biometric Society in recognition of his long-time service to the Biometric Society and to the advancement of Biometry.

Emlyn's main areas of research interest are in the design and analysis of experiments and the use of mixed models in practice. He is also actively involved in the development and maintenance of the design generation package CycDesigN.

Emlyn has given many short courses in Australia and overseas with a focus on statistical training in conjunction with computer packages. He has served on the Central Council of the International Biometric Society and has been an Associate Editor for the Australian and New Zealand Journal of Statistics and Biometrics and was a co-editor of the JABES 2020 special issue "Recent advances in design and analysis of experiments and observational studies in agriculture".

Arthur Gilmour: Friday, 11:45 - 12:30

ASReML in Practice

Arthur Gilmour obtained his BSc (Agr) from Sydney University majoring in Biometry in 1970 with a NSW Government traineeship. He then served as a biometrician until his retirement from NSW Agriculture as a Principal Research Scientist in 2009. From the outset, he was involved in software development to meet the current statistical analysis needs of his clients and colleagues. He obtained his PhD in animal breeding from Massey University in 1983 during which time he came into contact with Robin Thompson. This led to a collaboration, also involving Brian Cullis, resulting in the development of ASReml in 1996. He continues to support and develop ASReml and is in regular contact with Robin.

Program

Sunday

Monday

Tuesday

8:30-9:00: Welcome to AASC 2022: Murray Hannah, Roger Payne (by Video)

Invited Talk

TEA

MultiPhase experiments.

Short Break

On the Farm.

11:50-12:00 Questions/Discussion

LUNCH

Invited Talk

Tuesday Quick Fire 10 min talks.

Short Break

2:50-3:00 Questions/Discussion

TEA

3:30-5:10 DARE special session.

Invited Talk

5:00-5:10 Questions/Discussion
6:45 Canapes and Drinks
7pm-8:30 Poster session 54

Wednesday

Thursday

Friday

8:30-8:45 Welcome/Housekeeping

Invited Talk.

TEA

Mixed Topics 1.

Short Break

Mixed Topics 2.

11:35-11:45 Questions/Discussion

Invited Talk.

12:30-12:40 Conference Close: Murray Hannah, Andrew van Burgel

LUNCH

1: 30 Bus to Melbourne Airport & Southern Cross Station…… 3

Invited Talk *Chair: David Baird*

Tues 9:00-9:45

Emulation of the Spark bushfire solver, and building trust in uncertainty

Carolyn Huston
CSIRO, carolyn.huston@csiro.au

Emulations is a potentially useful tool in the context of estimating and predicting potential fire perimeters and associated areas burnt. A key advantage of interest is a speedup in estimation of fire progression, allowing faster exploration of scenarios and more efficient incorporation of uncertainty in reported results. Equally important to the development of the emulator is building trust in the technology such that decision makers feel empowered to make decisions using this tool. This work considers both the construction of an emulator, as well as develops a vocabulary and framework to facilitate discussions to build trust better allowing users to understand, accept, and utilize features enabled by the emulator.

MultiPhase experiments. *Chair: Debra Partington*

Tues 10:15-10:30

A linear mixed model framework with clustering for analysis of multi-phase proteomics experiments

Clayton Forknall

Department of Agriculture and Fisheries, Clayton.Forknall@daf.qld.gov.au

Authors: Clayton Forknall, Arunas Verbyla, Yoni Nazarathy, Adel Yousif, Sarah Osama, Shirley Jones, Edward Kerr, Benjamin L. Schulz, Glen Fox and Alison Kelly

The malting process is fundamental to beer brewing. This process involves the controlled and limited germination of raw barley grains before the grain is kilned, becoming malt. While much is known about the chemical and physical changes that the grain undergoes during malting, there is limited knowledge of how the many proteins composing the grain are affected.

Mass spectrometry (MS) based proteomics techniques enable the identification and quantification of proteins composing biological samples. The application of such techniques results in high dimensional data sets, with the abundance of upwards of hundreds of proteins quantified from the processing of a single sample.

We describe the conduct of an MS based proteomics experiment, in which barley grain samples, taken at twelve time points during a malting process, are processed using MS techniques to quantify changes in protein composition. In what we consider to be a first for this type of experiment, a multi-phase experimental design is implemented to facilitate an exploration of the magnitude of potential sources of variation arising in both phases of the experiment; i) the malt sample collection phase, and ii) the MS processing phase.

We present a two-stage linear mixed model framework for the analysis of protein abundance. A two-stage approach is necessitated due to the dimensionality of the data arising from MS experiments.

To commence the first stage of the analysis, a ‘baseline’ model is fitted, whereby effects for each protein by sampling time combination are fitted as fixed, while effects describing the experimental design structure are fitted as random. The residual variance structure of the ‘baseline’ model is defined using a separable variance model, which assumes heterogenous variance for each protein, scaled independently for each sampling time.

Complex relationships can exist between proteins at the elementary level, and modelling such relationships often proves computationally challenging. As such, to approximate these relationships, the residuals from the ‘baseline’ model are used to inform the clustering of like protein groups. Through the introduction of these clusters into the residual variance structure, a parsimonious modelling of the covariance between proteins is achieved.

From the first stage analysis, the predictions (and associated weights) for protein by sampling time combinations are carried forward. In the second stage, the linear mixed model representation of the cubic smoothing spline is used to estimate differential responses to time in the malting process for each protein. Furthermore, a second clustering procedure is performed to group like responses in protein abundance to malting time and identify a set of differential 'typical' protein responses to the malting process.

This modelling approach provides a parsimonious and computationally feasible means of analysing structured, high dimensional data in a linear mixed model framework. When applied to an MS based proteomics experiment, non-negligible variation is revealed in both phases of experimentation. Furthermore, significant variation in the responses of proteins to time in the malting process is evident. Altogether, the application of this modelling approach provides novel insights to both the brewing industry and MS practitioners alike.

Tues 10:30-10:45

Multi-phase design and analysis using a single step multi-experiment approach with factor analytic models to improve accuracy of late maturity α-amylase classification in wheat

Lu Wang
Centre for Biometrics and Data Science for Sustainable Primary Industries (CBADS-SPI), University of Wollongong, luw@uow.edu.au
Authors: Lu Wang, Jess Meza, Brian Cullis and Bettina Berger

Late maturity alpha-amylase (LMA) is an important trait routinely assessed through LMA expression experiments (LMAEEs) that provide LMA classification of Australian wheat varieties. Australian wheat breeding lines with high levels of LMA are deemed unsuitable for high value end products regardless of their yield capacities. Two sets of experiments are conducted annually, one in winter and one in the following summer, which form a pair of LMAEEs with a high proportion of lines in common. The current pair of experiments is analysed together with previous experiments in a multi-environment trial (MET) analysis.

The aim of the MET analysis for the LMAEEs is to classify the current set of test lines against the benchmark variety RAC655, which is a known LMA expressor. The use of appropriate statistical methods has a key role in providing accurate predictions. The current recommended method of analysis follows Smith and Cullis (2018) and involves a linear mixed model (LMM) with factor analytic (FA) variance structure for the line by experiment effects, along with appropriate modelling of the non-genetic effects that reflects both the experimental design and physical blocking factors that are not a part of the experimental design (e.g. columns and rows on the glasshouse benches and ELISA slides) and a residual model that accommodates spatial dependence where appropriate. In addition to utilising appropriate statistical methods for analysis, it is critical to ensure that the experiment has been designed in an appropriate and cost-efficient manner. Change of trial management and access

to improved facilities since 2019 has enabled the development of a more appropriate trial design which better captures the aims of the experiment.

The LMAEE is an example of a multi-phase experiment that comprises a glasshouse (GH) phase, a deep well slide phase and an enzyme-linked immunosorbent assay (ELISA) laboratory phase. It involves several time periods and has observational units which are completely different from the preceding phases (Butler et al., 2009). We consider a model-based design approach for constructing a multi-phase design that seeks to find an optimal or near optimal design under a pre-specified LMM (Butler, 2022). Importantly, this approach allows design information from the glasshouse phase to be carried on to the ELISA phase. That is, the design of the ELISA experiment accommodates sources of variations from both the glasshouse and ELISA phases.

In this presentation we will demonstrate the design process for the most recent pair of LMAEEs using the ODW package (Butler, 2022) in R, and present results from the most recent MET analysis using the ASReml package (Butler et al., 2019) in R.

References: Butler, D., Tan, M.K., and Cullis, B. R. Improving the accuracy of selection for late maturity alpha-amylase in wheat using multi-phase designs.Crops & Pasture Science, 60(12), 1202-1208, 2009.
Smith, A. B., and Cullis, B. R., (2018). Plant breeding selection tools built on factor analytic mixed models for multi-environment trial data. Euphytica, 214 (143), 2018.
Butler, David., Smith, A. B., Cullis, B.R., Gogel, B., Gilmour, A.R., and Thompson, R. ASReml-R Reference Manual Version 4, 2019.
Butler, David. Optimal experimental design under the linear mixed model, odw package manual, mmade.org, 2022.

Tues 10:45-11:00

The analysis of a two-phase experiment involving human subjects using ASReml-R and asremlPlus

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

A pain-rating experiment was reported in a 1997 paper by Solomon, Prkachin, & Farewell. A subset of the data was analyzed by Farewell and Herberg (2003) and this was reanalyzed to include block-treatment interactions by Jarrett, Farewell and Herzberg (2020), who identified it as being a two-phase experiment. The analysis was extended to include heterogeneous variances and to consider how the design might be improved for future experiments by Brien (2022) Designing, understanding and modelling two-phase experiments with human subjects. *Statistical Methods in Medical Research*, **31(4)**, 626-645.

A description will be given of the use of ASReml-R (Butler et al., 2020) to fit the linear mixed models and asremlPlus (Brien, 2022a) to select, employing Akaike Information Criteria, between models that differed in their variance parameters.

Heterogeneous variances complicate the deployment of multiple comparison procedures, such as the LSD, in that multiple values of the test criteria occur for comparing pairs of a set of predictions. An outline will be given of the tools available in asremlPlus for specifying the comparisons for which LSDs are calculated and to explore summary statistics of the LSDs with a view to choosing a statistic that minimizes, or even eliminates, the numbers of false conclusions arising from the use of the chosen statistic in comparing pairs of predictions.

References:

Brien, C. J. (2022). Designing, understanding and modelling two-phase experiments with human subjects. *Statistical Methods in Medical Research*, **31(4)**, 626-645.

Brien, C. J. (2022). *asremlPlus: augments ASReml-R in fitting mixed models and packages generally in exploring prediction differences*. R package version 4.3.36. Retrieved from https://cran.at.r-project.org/package=asremlPlus/

Butler, D. G., Cullis, B. R., Gilmour, A. R., Gogel, B. J., & Thompson, R. (2020). *ASReml-R reference manual*. Version 4.1.0.130. Retrieved from http://asreml.org

Farewell, V. T., & Herzberg, A. M. (2003). Plaid designs for the evaluation of training for medical practitioners. *Journal of Applied Statistics,* **30(9)**, 957-965.

Jarrett, R. G., Farewell, V. T., & Herzberg, A. M. (2020). Random effects models for complex designs. *Statistical Methods in Medical Research,* **29(12)**, 3695-3706.

Solomon, P. E., Prkachin, K. M., & Farewell, V. (1997). Enhancing sensitivity to facial expression of pain. *Pain,* **71(3)**, 279-284.

On the Farm. *Chair: Debra Partington*

Tues 11:05-11:20

When two's a crowd: Using experimental design to explore soilborne disease interactions

Bethany Rognoni
Department of Agriculture and Fisheries, Bethany.Rognoni@daf.qld.gov.au
Authors: Bethany Rognoni and Clayton Forknall

Plant diseases caused by soilborne pathogens can be devastating for cereal crop growth, resulting in adverse impacts on crop productivity. Typically, experiments focus on measuring the impact of a single pathogen on crop response variables such as grain yield and grain quality. However, in practice, growers are faced with disease complexes, where multiple soilborne pathogens can be present in the farming system.

To explore the impact and potential interaction between pairs of soilborne pathogens in wheat, experiments were designed such that a range of disease pressures would be established, for both pathogens in combination. These experimental designs included a factorial treatment structure, with a minimum of four applied levels of each pathogen. This ensured that quantitative measurements of pathogen burdens in each plot formed robust continuums for both pathogens, and a chosen response variable was able to be modelled against both explanatory pathogen continuums using a 3D response surface.

In the case of some pathogens, namely plant parasitic root-lesion nematodes (RLNs), the only reliable way to establish differential pathogen levels is by growing host wheat varieties of differing susceptibility in the year prior to experimentation. These varieties influence the RLN population densities in each plot, serving to establish a range of RLN pathogen burdens in the year of experimentation. However, the realised RLN population densities can vary between plots and between setup varieties, so it was important to quantify RLN population densities in each plot prior to the application of other treatments in the second year of experimentation. This enabled the use of a conditioned randomisation approach, which allowed for the randomisation of treatments over a measured covariate (in this case, RLN population densities) in such a way as to ensure that each treatment was exposed to as consistent a range of the covariate as possible (Reeves et al., 2020).

Prior to conducting these soilborne disease interaction experiments, the researchers identified concerns with being able to detect significant grain yield differences or interactions, and so wanted to bolster accuracy of predictions at the extreme ends of pathogen burden where the greatest yield differences were expected to occur. This resulted in the inclusion of extra replicates of the lowest and highest pathogen level treatments amongst the original four pathogen levels, to provide greater power for testing while still enabling the development of a continuum of disease pressures.

Our design solution to this problem will be presented via the step-by-step development of an optimal design for a soilborne disease interaction experiment involving RLNs, highlighting the flexibility and power of the *odw* package in R when implementing conditioned randomisation, and when dealing with an unequal number of replications of factor treatments. This work showcases the fun and creative side of puzzle-solving with experimental design to develop statistically sound, yet practically feasible experimental design solutions to answer the often-times complex research questions arising in applied agricultural research.

Tues 11:20-11:35

Statistical analysis of large plot on-farm trials with qualitative treatments

Jordan Brown
Curtin University, jordan.brown@curtin.edu.au
Authors: Jordan Brown, Andrew Grose, Katia Stefanova, Zhanglong Cao, Mark Gibberd, Joseph Sigar and Suman Rakshit

Statistical methods used for small plot analyses are unsuitable for analysing large-scale on-farm experiments (OFE) because these methods fail to take into account the spatial variability in treatment effects within large paddocks. Several new methods have been proposed that are inspired by geo-statistical analyses of spatially-varying treatment effects, which are typical of site-specific crop management trials with quantitative treatments. However, most applied biometricians, who routinely analyse data from field trials, are either unfamiliar with the new methods or reluctant to include these in their regular analytical toolkit. This reluctance often comes from the unavailability of easy-to-use software implementations of these methods. One of the widely used methods for analysing field trials is the linear mixed model because it is extremely versatile in modelling spatial and extraneous variability and is accessible through user-friendly software implementation. Motivated by the analysis of comparative experiments conducted in large strip trials with categorical treatment factors, we propose a linear mixed effects model for determining the best treatment at both local and global spatial scales within a paddock, based on yield and profit estimates. To account for the large spatial variation present in on-farm experiments, we divide the trial into smaller regions or management blocks, each containing at least two replicates. We propose two approaches for creating these management blocks. In the presence of appropriate spatial covariates, a clustering method is proposed; otherwise, the trial region is stratified into equal-sized rectangular blocks using a systematic partitioning scheme. Management blocks are used to estimate the varying treatment effects by incorporating treatment-by-block interactions in linear mixed effects models. The optimum treatment of each block is found by either comparing the best linear unbiased prediction (BLUP) solely or incorporating profit and comparing economic performance. To illustrate the applicability of our method, we have analysed two large on-farm trials conducted in Western Australia.

Tues 11:35-11:50

Analysis of Farm Scale Strip Trials

Andrew van Burgel
Department of Primary Industries and Regional Development, andrew.vanburgel@dpird.wa.gov.au

Many farmers are keen to see research done on a large scale. One common trial design is long strip plots where treatments are applied in strips across a paddock that are hundreds of meters long. While many of these trials maintain replication of treatments, the total number of strips is typically low, leading to statistical challenges in analysis. Recent papers in this area have applied Geographically Weighted Regression1. This talk comments on this methodology and seeks feedback on an alternative approach based more closely on traditional statistical analysis methods for agricultural designed experiments.

1 S Rakshit et al. (2020) Novel approach to the analysis of spatially-varying treatment effects in onfarm experiments. Field Crops Research 255, doi.org/10.1016/j.fcr.2020.107783

and

FH Evans et al. (2020) Assessment of the Use of Geographically Weighted Regression for Analysis of Large On-Farm Experiments and Implications for Practical Application. Agronomy 10(11), 1720, doi.org/10.3390/agronomy10111720

Invited Talk *Chair: Graham Hepworth*

Tues 1:00-1:45

All Under One Roof – The Rise of Joint Species Distribution Modeling in Ecology

Francis Hui
ANU.EDU, francis.hui@anu.edu.au

Prompted by improvements in computing power and an increasing number of scientific questions that are multi-response in nature, the landscape of statistical analysis in community ecology has undergone a major shift in the last five years with the explosion of joint species distribution modeling. In this talk, I will provide an overview of how such models (which are largely based around the use of latent variables or some variation thereof) have come to dominate the discipline, and how they have been adapted to solve questions by ecologists regarding the environmental and biotic processes driving species assemblages. I will then offer an (opinionated) view of where-to-next for joint species distribution models, including the use of “modern” statistical approaches such as covariance/correlation regression and spatio-temporal methods, and the growing software market. Finally, I will discuss related research opportunities across other disciplines where latent variable models are applied, such as in the analysis of multi-environmental field trials in plant breeding.

Tuesday Quick Fire 10 min talks. *Chair: Andrew van Burgel*

Tues 1:45-1:55

Automated visualisation of experimental designs

Emi Tanaka
Monash Univeristy, emi.tanaka@monash.edu

As those in the forefront of statistical consulting would well know, communication with the domain expert is of paramount importance in the construction of a valid experimental design. This communication can be better enhanced by the use of visualisation of the experimental design layout that allows all parties to assess understanding and clarify structure for the desired experimental design. There is, however, a friction to the manual construction of this visual representation. In this talk, I extend the "the grammar of experimental designs" system in the edibble R-package to automate construction of the visualisation object; this system is implemented as the deggust R-package.

Tues 1:55-2:05

Exploratory data analysis of TCGA skin cutaneous melanoma RNA-seq data

Min Zhang
The University of Queensland, m.zhang1@uq.net.au
Authors: Min Zhang, Vivi Arief, Geoffrey McLachlan, Quan Nguyen and Kaye Basford

RNA-sequencing is a popular tool in quantitative cancer research to provide insights into tumour progression. However, the high-dimensional nature of the RNA-seq gene expression data, which is usually in the format of thousands of genes by tens/hundreds of samples, makes it not a straightforward task to capture important features from the data. This is when exploratory data analysis (EDA) becomes useful. By applying EDA techniques including scatter plot matrices, clustering, and principal component analysis (PCA) on skin cutaneous melanoma RNA-seq data from The Cancer Genome Atlas (TCGA), we prioritized a list of genes that could be associated with the progression of skin cutaneous melanoma. Many genes from our gene list have been linked with epidermis and/or cancer related biological activities in the literature, suggesting the power of EDA in quantitative cancer research. Together, our study provided insights into understanding the variations between melanoma patients and added knowledge into melanoma progression.

Tues 2:05-2:15

Model-based design of QTL, GP and GWAS phenotyping experiments using genetic relatedness

Aidan McGarty
University of Wollongong, amcgarty@uow.edu.au
Authors: Aidan McGarty, Brian Cullis, Ahsan Asif and Kristy Hobson

Model-based designs have become a popular component of experimental design in Australian plant breeding programs. More recently, information on genetic relatedness has been included in models for design generation through pedigree information. Including such ancestral information allows the model used for design to more closely match the model used for analysis. This study examines the inclusion of genetic relatedness information in model-based designs through marker data and the inclusion of the genomic relationship matrix. The ideas are illustrated using an experiment conducted in a growth chamber which aims to identify the genetic architecture of Phytophthora root rot (PRR) in chickpeas (Cicer arietinum) using a recombinant inbred line (RIL) population involving 185 lines.

Tues 2:20-2:30

Flexible Regression for Count Data using the Poisson-Beta model

Alan Herschtal
Monash University, Alan.Herschtal@monash.edu

Historically, the standard method used to analyse count data has been the Poisson model. In the early years of count data analysis, it's elegant simplicity and interpretability have made it the default option for count data analysis in many fields. However, in recent decades increasing attention has been brought to its shortcomings. It assumes a constant underlying event rate across observational units, and thus does not account for the commonly occurring real-world phenomena of overdispersion or zero inflation. It has thus given way to the two-parameter negative binomial and zero inflated Poisson models, which extend the one-parameter Poisson model in each of these directions respectively. In recent years, however, there has been increasing disquiet regarding use of these Poisson extensions, and recognition that even these models fail to fully cater for the full diversity of count data observed in the real world. This has spawned a wide variety of yet more generic count data models, including the zero-inflated negative binomial, Delaporte, Sichel, and others. One of these newer models is the three-parameter Poisson-Beta model, which is a Poisson-Beta mixture in which the underlying Beta mixing density is scaled by a scaling parameter, which acts as a third parameter for the Beta model. The Poisson-Beta mixture model has various attractive properties. It is a suitable fit for a variety of real-world datasets, and in particular has been shown to have utility in the model of genomic sequencing data. It contains both the zero-inflated Poisson and the negative binomial distributions as special cases, but is more generic, allowing for the shape to be controlled independently of the location

and scale. However, the fact that its density function does not have a closed form expression, being expressible only as an infinite sum or as an integral, leads to computational tractability issues when developing a regression method. This work discusses how these hurdles can been overcome and describes the construction of such a regression method.

Tues 2:30-2:40

A change of perspective: The importance of evaluating domain specific assumptions for robust model development

Vihanga Gunadasa
University of Sydney, vihanga.gunadasa@sydney.edu.au
Authors: Vihanga Gunadasa, Prof. Glenda M Wardle, Prof. Robert Kohn and Dr. Aaron Greenville

Data science modelling approaches rely on both statistical and domain specific assumptions. In applications of data science in ecology, the statistical assumptions are generally addressed while the domain assumptions are often overlooked. However, the ecological assumptions can either introduce a bias or assist in minimising the bias based on the incorporated subjective opinion and the simplifications of the model. Even though the initial ecological assumptions represent our subjective expectation of the ecosystem behaviour, testing alternative ecological assumptions can provide insights into the inner workings of the ecosystem. In ecology, modelling the species population abundance is one of the most prominent areas that builds upon domain assumptions. State-space models (SSM) which are a class of forecasting models built upon statistical assumptions have been widely used to model animal abundance as uncertainties related to both measurement and environment are captured. In this study, we performed an exploratory analysis on Bayesian SSMs by 1) investigating ecological assumptions through changes in state-process parameters and 2) testing validity of statistical assumptions. Our experiment was conducted on a foundational study with a Multivariate SSM used on a long-term abundance data set of five small mammal species across nine sites in arid central, Australia. The mammals were categorized based on the existence or lack of cross correlation of population abundances across the nine geographical locations. For the two groups the ecological assumptions in the foundational study subjectively defined the dynamics of the ecosystem. In our study, under domain assumptions we tested if the abundance of species depends on the density, if both predator-prey interactions affect the abundance of the species and finally if the animal count depends on their habitat. These ecological assumptions were investigated by considering a three-way factorial combination of model specifications that assume scenarios between density dependence, predator-prey interactions and spatial independence of abundances. The parameter combination of the best model was selected by comparing the SSMs using Akaike Information Criterion corrected (AICc). From the examination of the ecological assumptions, all species showed density dependence in abundance and predator-prey interactions. Species with growth abundance patterns synchronized across spatial sites assumed different measurement errors while asynchronous species assumed different environment variability. In

line with the ecological assumptions, normality and auto-correlation of residuals in the model were tested under statistical assumptions using Shapiro-Wilk and Ljung-box tests respectively. Validating statistical assumptions revealed that 1) the residuals from the observation model violated the normality assumption in 5 sites and 2) the residuals were dependent in 3 sites. Sparse zero-inflated short-term time series with missing values due to non-survey violated these assumptions. In conclusion, data science methods should be backed by testing both the statistical and domain specific assumptions for the outcomes to be valid and robust to domain shifts in the future.

Tues 2:40-2:50

Semi-parametric regression in 'predictmeans'

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

R package 'predictmeans' is used to calculate predicted means for linear models. It provides functions to visualize, diagnose and make inferences such as predicted means and standard errors, contrasts, multiple comparisons and permutation test. Recently, three new functions have been added in the package, named 'smZ', 'semireg' and 'semipred', for fitting, visualizing (2D or 3D) and inferencing a semi-parametric regression using 'lme4' framework, especially, with O'Sullivan-Type spline. In the talk, we will illustrate the usage of those functions with various examples.

Tues 3:30-5:10 DARE special session. *Chair: Willem Vervoort*

Over the last decade there has been a rush to use Data Science applications in areas such as Geology, Ecology and Hydrology. However, in many cases this has either been a relatively limited application of Data Science tools within a complex Domain Science problem, or a more theoretical Data Science approach on a very limited Domain Science problem. In other words, in most cases the Domains simply borrow techniques or problems. This can be seen as a relatively weak integration of the two domains. This session seeks contributions of examples of strong integration, where problems of Data Science and Domain Science are co-developed or are working towards co-developed problems. Contributions highlighting issues and barriers related to such co-development are also encouraged.

Invited Talk Tues 3:30-4:15

Uncertainty quantification and communication for the earth sciences

Ed Cripps
University of Western Australia, edward.cripps@uwa.edu.au

Advances in technology and the availability of data-acquisition devices have increasingly centralised the role of the data analytics in the earth sciences, which in turn inform data driven decision making across science, industry and government. Still, despite the preponderance of data, empirical based decision making continues to be made under conditions of uncertainty: data is messy; statistical model selection/estimation is complex; underlying physics that discretised numeric methods attempt to resolve are mis-specified. This recognition implies that, when the consequences of decisions are substantial, robust uncertainty quantification ought to accompany the fusion of domain knowledge and empirical evidence. This talk is based on a series of recent papers, providing an overview on: recent applications/methods developed with earth scientists and industry partners for probabilistic models of meteorological, oceanographical and geophysical processes; experiences on conveying to non-statistical colleagues the meaning of uncertainty and its consequences for decision making; deployment of software for private (industry) and public use.

Tues 4:15-4:30

Application of Bayesian model selection to Lower Namoi aquifer water balance models

Katherine L. Silversides

ARC Training Centre in Data Analytics for Resources and Environments (DARE), The University of Sydney, katherine.silversides@sydney.edu.au

Authors: Katherine L. Silversides and R. Willem Vervoort

Groundwater systems are complex and unobservable, being underground. Water balance models can be used to constrain the overall inputs and outputs. However, in many cases there are several competing models that can fit the overall water balance. Bayesian model selection offers a way to probabilistically evaluate the different models. Enemark et al (2019) highlighted the use of Bayesian model selection to test different hypotheses about a groundwater system using competing water balance models. We use this method to test the likelihood of different hypothesised components contributing the water balance in the Lower Namoi representing different inflows and outflows. We investigate the impact the magnitude of the components and how different reported ranges of pumping values can affect the likelihoods of other components.

The certain (from well-established literature sources) input components were Great Artesian Basin (GAB), river recharge, rainfall and lateral flow in. The certain output components were pumping, river discharge and lateral flow out. The uncertain components were a decrease in storage, additional lateral flow out, additional pumping, recharge from irrigation water, additional GAB and evapotranspiration. A range of values was assigned to each component. The range for pumping varied based on different estimates and real measurements. These included the minimum and maximum of actual values pumped, mean ± stdev, mean ± stdev/2, and the planned pumping values.

A set of model structures including every possibly combination of the uncertain components was developed. Multiple runs generated random values for each component (within its defined range). Using Metropolis-Hastings sampling the model with the smallest water balance error in each run was identified. The number of times a specific model structure was identified indicates the posterior likelihood of those hypotheses being true.

When the largest range of initial values was allowed for the pumping, the range was multiple times the range of the other components, representing essentially a large uncertainty. As a result, there was no support for any hypothesis. Restricting the pumping range increased the support for different models. Particularly, increased support was observed for a decrease in storage in the groundwater system and against additional pumping or evapotranspiration. Using the predicted pumping values changed the interactions of the components, making them more conditional on other included components. Additionally, when a component has a relatively small value compared to the other components, it is hard to get support for it.

This demonstrates that when model selection is used, the values used need to be carefully considered so that they are realistic, but the scale between the different components is considered. One potential method of addressing this is to model individual years, where values such as the known pumping can be tightly constrained. This may also provide information on different behaviours of the aquifer during wet and dry years.

Enemark, T., Peeters, L.J.M., Mallants, D., Batelaan, O., Valentine, A.P., Sambridge, M. 2019. Hydrogeological Bayesian Hypothesis Testing through Trans-Dimensional Sampling of a Stochastic Water Balance Model. *Water*, 11. doi:10.3390/w11071463

Tues 4:30-4:45

A Comparative study of the Priors used in Bayesian Variable Selection for Quantile Regression

Dilani Kaveendri
ARC centre in Data Analytics for Resources and Environments(DARE) and University of Sydney, dilani.damburehewage@sydney.edu.au
Authors: Dilani Kaveendri, Nandini Ramesh and Sally Cripps

Bayesian quantile regression is a logically consistent framework with which to deal with both parameter estimation and uncertainty quantification. Variable selection in such a paradigm plays an important role in quantile regression because it allows one to make inferences on the dependency between potential predictive factors on the tail, or extreme, of the distribution and decide on the regressors of the final model to enable predictive accuracy. Our objective is to evaluate the impact of various priors on Bayesian variable selection in quantile regression under the assumption of linearity. The horseshoe prior which is a global-local shrinkage prior, Laplacian priors such as LASSO and Adaptive LASSO, and spike and slab prior which is a discrete mixture prior considered as the gold standard for sparse problems are the priors that are used in our study. The frequentist properties of different prior settings are examined via a simulated data set. We also apply these methods to a real data example considering rainfall at a single meteorological station (Sydney Observatory Hill) as the response variable and a set of meteorological variables as covariates. Rainfall is chosen here as an example to demonstrate these techniques as it is an aggregate of different climate phenomena, which means that the different parts of the distribution of rainfall may be affected by various combinations of factors. The most significant variables selected under each scenario can be used to identify which factors influence rainfall for different quantiles. Furthermore, the use of posterior estimates of the selected variables via Gibbs sampling can be used to derive the marginalized predictive density under each method. Thereafter, predictive density can be used to compare with actual observed values at a future point in time to draw conclusions in terms of predictive accuracy. This will provide new insights into quantile regression models starting from variable selection up to making predictions while quantifying uncertainty.

Tues 4:45-5:00

Application of feedback particle filter with diffusion-map-based gain to state-space models

Yiyi Ma
UNSW and DARE, yiyi.ma1@unsw.edu.au
Authors: Yiyi Ma, Sahani Pathiraja and Lucy Marshall

A general starting approach for investigating a complicated environmental system involving various parameters is summarising it using process-based models. On the other side, state-space models can be regarded as a further simplified version of process-based models, a set of differential equations consisting of observations and hidden states to describe a system's dynamic. A critical task for understanding such models is approximating the posterior distribution of the model at each time state, given initial conditions and historical information. Ensemble methods (such as ensemble Kalman filters and particle filters) have achieved promising results for such tasks. A common feature of these methods is to rely on Monte Carlo samples to empirically approximate prior and posterior distributions. Although ensemble Kalman methods have shown great success in many real-world problems, they are inaccurate for filtering problems where the state space model or observation operators are nonlinear. Importance sampling-based particle filters are theoretically an elegant solution, but suffer from the curse of dimensionality. Feedback particle filter (Yang, P. G. Mehta, S. P. Meyn, 2013) provides an alternative solution for the non-linear filtering estimation task by introducing a control term to move the particles in state space, much like ensemble Kalman methods. This control term contains two main components: an innovation term and a gain function. The latter element is the solution to the Euler-Lagrange boundary value problem (E-L BVP). (Taghvaei, P. Mehta, S. Meyn, 2020) provides a diffusion-map-based algorithm to estimate this solution. It also shows that a feedback particle filter with this diffusion-map-based gain approximation achieves better results for one-dimensional stochastic processes than adopting a constant gain or using a sequential importance sampling (SIR) particle filter. We will explore the method's performance for higher dimensional state-space models, starting with the toy model Lorenz 63 to more complex ones related to environmental systems.

Yang, T., Mehta, P. G., Meyn, S. P. (2013). ÒFeedback Particle FilterÓ. IEEE Transactions on Automatic Control 58.10, pp. 2465Ð2480. doi: 10.1109/TAC.2013.2258825.
Taghvaei, A., Mehta, P., Meyn, S. (2020). ÒDiffusion Map-based Algorithm for Gain Function Approximation in the Feedback Particle FilterÓ. SIAM/ASA Journal on Uncertainty Quantification 8.3, pp. 1090Ð1117. doi: 10.1137/19M124513X.

Invited Talk *Chair: Ruth Butler*

Thurs 9:00-9:45

New Developments in Genstat 22

Vanessa Cave, David Baird
VSNI NZ, vanessa@vsni.co.uk, david@vsni.co.uk

The 22nd Edition of Genstat was released earlier this year. In this presentation, we will showcase some of the new enhancements and features in Genstat 22 and provide a glimpse into the developments underway in Genstat 23.

A key statistical enhancement in Genstat 22 is provided by the new menus and procedures for generalised linear mixed model (GLMM) analysis. Genstat 22 provides considerably greater functionality for displaying and saving output from a GLMM, producing predictions, plotting residuals, and visualising the fit of a GLMM in a separation plot. Furthermore, you can now assess the significance of the fixed terms in a GLMM using permutation tests.

Another development is the addition of equivalence, non-inferiority and non-superiority tests. These are extremely useful tests when the aim of the study is to demonstrate that two treatment means are effectively the same, or that one treatment mean is effectively no smaller or no larger than another. For example, in a medical trial when the aim is to prove that a new drug treatment is just as effective as the standard drug, or in the plant breeding context, when the aim is to show that a new cultivar is at least as resistant to disease as the industry standard. The new menus for t-tests and ANOVA make it easy for you perform equivalence, non-inferiority and non-superiority tests.

Yet another innovation is the ability to analyse data with either fixed-threshold left- or right-censoring using a linear mixed model. During data collection, censoring occurs when measurements cannot be taken above or below a bound. For example, chemical concentrations may be left-censored when they fall below a minimum level of quantification. The new Linear Mixed Models with Censoring menu enables users to easily and quickly fit a linear mixed model to censored data. In the Genstat 23 these facilities are being extended to include left- and right-censored Poisson data.

With Genstat 22 the more flexible RLM web-based licensing system was rolled out. This makes accessing, amending and renewing a license a much smoother process, and moreover it allows us to deliver and manage your license entitlements via a cloud-based license server without installing license server software on your individual device or network.

Genstat 22 also offers many other new menus and commands to help you perform your desired statistical task. For example, there are new menus and procedures for analysing rainfall data, plotting confidence, prediction and equal-frequency ellipses for bivariate data, assessing the importance of fixed effects

in a REML analysis using random permutation tests, importing Excel file cell formulae and formatting information into Genstat, editing command windows and lots more! Unfortunately, we cannot showcase all these new features here, so to learn more please do visit: https://genstat.kb.vsni.co.uk/22/whats-new22nd/.

Development of Genstat 23 is well underway, including new features for displaying large bivariate data sets with observations classified into groups and also for exploring multi-dimensional data.

The Human Dimension *Chair: Pauline O'Shaughnessy*

Thurs 10:15-10:30

Measuring and modelling surgical outcomes with Days Alive and Out of Hospital

Luke Boyle
University of Auckland, lboy505@aucklanduni.ac.nz

Days alive and out of hospital (DAOH) is an emerging metric for measuring healthcare outcomes or more specifically, surgical outcomes. DAOH is objective and is reduced by any complication within a defined period sufficiently serious to cause death, prolong a hospital admission or lead to a readmission. It is a continuous composite variable, making analysis of DOAH differences more informative than the traditional focus on point estimates such as average one month mortality. In New Zealand the information required to compute DAOH is collected regularly and stored in the New Zealand National Minimum Data Set. We can also easily track patients across the health system via their National Health Index Number. This makes New Zealand the ideal place to investigate how to use DAOH as a measure of surgical outcomes.

DAOH data has a complex non-linear distribution with two peaks, whose size depends on the risk level of the operation. Through my PhD I have been developing methods to model, analyse and risk adjust DAOH to allow for more complex analyses. Using these techniques, we have performed analysis of hospital differences using DAOH through a comparison of DHBs within New Zealand. We demonstrated that we can identify outlier hospitals through DAOH scores and we showed that DAOH had face validity when compared against mortality rates or overall operation risk. We have also applied DAOH to investigate equity in outcomes after Coronary Artery Bypass Graft (CABG) operations. We found that for higher risk patients, there is a large equity gap which shrinks as the overall health of the patient pre-operation improves.

Through this work we found that DAOH is more effective at assessing outcomes than some common metrics such as mortality and we have shown that a focus on the median or mean values of DAOH can often hide important information. This talk will illustrate some of our results and discuss the best methods for risk-adjustment and best models for prediction of DAOH data in a healthcare context.

Thurs 10:30-10:45

Constructing multivariate disease progression curves for Alzheimer's disease

Timothy Cox
CSIRO, timothy.cox@csiro.au
Authors: Timothy Cox, Rosita Shishegar, Pierrick Bourgeat, James Doecke, Jurgen Fripp, Christopher Rowe, Colin Masters, Victor Villemagne and Samantha Burnham

Understanding the timing of changes in different biomarkers during the progression of Alzheimer's will allow us to build a "map" of the disease progression. This has the potential to provide crucial knowledge for the design and timing of effective clinical therapeutic trials. Most longitudinal data available on Alzheimer's disease progression is collected on over a shorter timespan than overall course of the disease. We present a multi-modal phase plane (MMPP) method to construct long-term multivariate disease trajectory curves from short term longitudinal data for degenerative diseases like Alzheimer's and evaluate its effectiveness. The MMPP method is an extension to previously presented four-step methods for constructing single variable disease trajectories. A novel anchoring step is added, which uses study participants' multivariate data to infer the staging of the different single variable progression trajectories allows multivariate disease trajectory curves to be generated. Further, the anchoring step provides disease staging at the individual level. We demonstrate that the MMPP method is able to accurately reconstruct multivariate disease trajectory curves and individuals' disease stage from simulated noisy short term longitudinal data.

Thurs 10:45-11:00

Estimating the number of undetected COVID-19 cases in Australia through Capture-Recapture methods

Joanne Thandrayen
Australian National University, Joanne.Thandrayen@anu.edu.au
Authors: Joanne Thandrayen and Bernard Baffour

Worldwide, most countries have publicly reported their daily counts of confirmed COVID-19 cases and related deaths. Nevertheless, there remains a substantial number of undetected infections due to various factors including difficulties in identification of asymptomatic cases, delayed symptoms, rapidly evolving variants and strains, efficacy of contract tracing, amongst others. For these reasons, the total number of COVID-19 cases is unknown. Capture-recapture methods, which originated in ecology with the aim of estimating the size of an unknown (possibly elusive) population, have now been extensively used in epidemiology and public health. By definition, since the number of undetected COVID-19 cases is unknown, capture-recapture methods have increasingly been applied to estimate the true number of cases (taking into account both the observed and unobserved infections).

Australia has been one of the few developed countries in the world that has successfully slowed down the spread of COVID-19 by introducing a number of public health measures including Australian borders closure, quarantine for returning residents, lockdowns in individual states and territories, social distancing rules, rollout of vaccinations, in an attempt to curb the spread of the disease. As a result, Australia has experienced lower infection and death rates than many comparable developed countries. However, these figures are only those reported and collected by the state/territory official health systems, and as such can under-estimate the true magnitude of the pandemic. To our knowledge, estimation of the true number of COVID-19 cases using capture-recapture methods has not been undertaken in Australia.

Our study provides a lower bound (best case scenario) and upper bound (worst case scenario) estimates of the true number of COVID-19 cases that includes the number of undetected cases during the first wave of the pandemic (March-April 2020). In addition, those same estimates were provided for the states of New South Wales, Victoria, Queensland, Western Australia, and Tasmania. Data comprised records from the “COVID-19 in Australia” public database and verified with federal and state/territory health departments. Our results showed that, in general, the total number of cases were at least twice and at most four times the observed number of cases that has been reported in Australia. A ratio (total/observed) of 2 and 4 implied that for every confirmed case of COVID-19 there were 2 to 4 people with COVID-19 respectively but the latter have not been identified by the official health systems. There were differences at the state levels with New South Wales having the highest ratio and Victoria the lowest ratio. Overall, the results confirmed that the COVID-19 outbreak was more prevalent than what has been officially recorded by the state/territory health departments. These results are important to better appreciate the spread and impact of the COVID-19 pandemic for health monitoring, planning and evaluation.

A little bit of Bayes. *Chair: Pauline O'Shaughnessy*

Thurs 11:05-11:20

clusterBMA: Combine insights from multiple clustering algorithms with Bayesian model averaging

Owen Forbes

Queensland University of Technology, owen.forbes@hdr.qut.edu.au

Authors: Owen Forbes, Paul E. Schwenn, Paul Pao-Yen Wu, Edgar Santos-Fernandez, Hong-Bo Xie, Jim Lagopoulos, Lia Mills, Dashiell D. Sacks, Daniel F. Hermens and Kerrie Mengersen

Clustering is one of the most common tasks for applied statisticians across a wide variety of industry, government and research settings. When an analyst reports results from one 'best' model out of several candidate clustering models, this ignores the uncertainty that arises from model selection, and results in inferences that are sensitive to the particular model and parameters chosen.

In this work we introduce clusterBMA, extending Bayesian Model Averaging (BMA) methodology to combine inference across multiple algorithms for unsupervised clustering of a given dataset, using a combination of clustering internal validation criteria to weight results from each model. BMA offers some attractive benefits over other existing approaches. Benefits include intuitive probabilistic interpretation of an overall cluster structure integrated across multiple sets of clustering results, flexibility to accommodate various input algorithms, and quantification of model-based uncertainty. These features enable improved communication of uncertainty and variability across models for clustering applications, allowing clients to gain clearer understanding of the insights offered by different clustering methods and uncertainty in cluster structure across models.

We present results from a simulation study to explore the utility of this technique for identifying robust integrated clusters with model-based uncertainty, under varying conditions of separation between simulated clusters. We then implement this method in a substantive real world case study, clustering young people based on electrical brain activity and relating these clusters to measures of mental health and cognitive function. Our method offers extra insight compared to clustering results from individual algorithms, particularly regarding consistency or ambiguity in cluster allocations between multiple clustering algorithms. This case study demonstrates the utility of clusterBMA in health applications where model-based uncertainty is relevant for communication of risk to clinicians and patients.

The method is implemented in the freely available R package “clusterBMA”, and this session will include a practical demonstration to facilitate understanding of how this tool may be useful for audience members in their work.

Thurs 11:20-11:35

Bayesian framework to quantify uncertainty in water accounting in NSW catchments

R. Willem Vervoort
DARE ARC ITTC & The University of Sydney, willem.vervoort@sydney.edu.au
Authors: R. Willem Vervoort, Joshua Simmons, Gilad Francis and Richard Scalzo

To quantify and protect our water resources around Australia, analysis of the total inflows and outflows (water balance) to our catchments are routinely undertaken. Analyses by the NSW governments have, however, identified large-scale "unaccounted differences" in the water balance across numerous NSW catchments. These volumes are typically around 20% of inflows but may occasionally reach ~50% of inflows in some catchments. Given the public concern about water sharing and environmental flows in the Murray Darling Basin catchments, uncertainty in the water balance is a major source of risk for decision-making in water management.

As part of the NSW Smart Sensing Network coordinated "Where Is All The Water" (WIATW) multidisciplinary research project a probabilistic modelling framework was developed to explain and quantify unaccounted differences in the water balance for major rivers up to catchment scales. As an initial step, simplified proof of concept calculations were undertaken to quantify uncertainty. These calculations for a subset of the water balance components form part of a probabilistic modelling framework based on the application of well-established Bayesian inference techniques. This paper will highlight a pilot project, where the uncertainty quantification of three components of the system is demonstrated focusing on the years 2019 - 2020. This specifically incorporates: 1) the uncertainty in estimating the evaporation and transpiration from river reaches and riparian zones; 2) the uncertainty in estimating the surface water groundwater connection and groundwater flow given river reach and groundwater observations; and 3) the uncertainty in the estimated surface water inflows/outflows stemming from the fitting of rating curves at gauges in the river.

The paper demonstrates a Gaussian Processes (GP) approach to quantify groundwater-surface water interactions, evapotranspiration losses and to fit rating curves. While a Bayesian approach to many of these processes is not new, incorporating this as part of an overall probabilistic framework is a novel approach. The current pilot only demonstrates three components of the overall system to highlight how the GPs can be used. Future work will include a full river reach system integrated into a complete water balance, which can be constrained to sum to zero using the likelihood.

Thurs 11:35-11:50

Bayesian Zero-Inflated Negative Binomial Model to Correct Bias in Crowdsourced Cycling Data

Alaa Amri
The University of Edinburgh, alaa.amri@ed.ac.uk
Authors: Alaa Amri, Amy Wilson, Chris Dent and Gail Robertson

There is a plethora of methods intended to collect data about cycling activities in cities such as count stations. However, such data suffer from a lack in both temporal and spatial details in addition to being expensive and possibly protracted. Alternatively, crowdsourced data like Strava Metro data could be depicted as incredibly informative in terms of spatio-temporal details.On the other hand, the degree of representativeness and potential bias in the data could be problematic for analyses in urban planning.

We compared three different regression models namely Poisson, Negative Binomial and Zero-Inflated Negative Binomial (ZINB) using the Bayesian framework to find the relationship between data from count stations and Strava along with weather-related information. We have shown that ZINB demonstrates good performance with the used data, compared to the two other methods which are frequently used in the literature.

Finally, a careful prior specification must be considered to improve the efficiency of the model.

Invited Talk. *Chair: Kerry Bell*

Thurs 1:00-1:45

Practical steps toward reproducibility

Patricia Menéndez
Monash University , Patricia.Menendez@monash.edu

Government organisations, industry players, and universities are embracing the notion of reproducibility around projects involving statistical and data analyses as a cornerstone in their working practices. It has become a basic requirement that projects be built in a reproducible manner to ensure that consistent computational results are obtained when using the same input data, computational methods, code and conditions for analysis. In this talk, I will discuss practical steps that can help managers and data science practitioners to set up their statistics and data analysis projects in a reproducible way. Meanwhile, I will also address some of the fundamental challenges in making work reproducible. I will draw on a few examples from my experience working outside of academia and discuss how reproducibility can be incorporated in university curriculums.

Thursday Quick Fire 10 min talks. *Chair: Clayton Forknall*

Thurs 1:45-1:55

'Where the creeks run dry or ten feet high': A probabilistic approach to streamflow forecasting in Australia

Rajitha Athukorala
DARE, The University of Sydney, rajitha.athukorala@sydney.edu.au
Authors: Rajitha Athukorala, Sally Cripps and R. Willem Vervoort

Streamflow prediction is a challenging but an important task because of its impact on all forms of life on Earth. In Australia, this task is made more complex because more than 70% of the rivers are non-perennial (i.e., rivers which have no flow for at least a part of the year) Furthermore, the transition from zero flows to high flows can be very rapid causing flash floods which is typical in Australia as experienced first hand in the recent past. Therefore, any model for Australian streamflow prediction and variability need to account for these unique features.

In this paper we propose a fully Bayesian mixture model where the distribution of streamflow has two components; the first component is a point mass at zero and the second is a Gamma distribution. The probability of belonging to one of the two components is modelled as a logistic regression. The logistic regression model as well as the parameters which prescribe the Gamma distribution are parameterized to depend upon adjacent streamflow and rainfall (where adjacent refers to both in space and time).

The model is estimated in a Bayesian framework using Hamiltonian Monte Carlo to perform the required multidimensional integration and generate samples from the posterior distributions of the quantities of interest.

The methodology was applied to the Avoca river catchment from Northern Victoria, which is a non-perennial river catchment. The preliminary results show potential in predicting the occurrence of peak flows but not quite the magnitude. The current areas of research include: 1) the use of other covariates such as temperature, soil moisture and accumulated rainfall, 2) the use of a mixture of gamma distributions for the non-zero flow days.

Thurs 1:55-2:05

Comparison of spatial methods: a newcomer's perspective

Sevvandi Kandanaarachchi
CSIRO, sevvandi.kandanaarachchi@data61.csiro.au
Authors: Sevvandi Kandanaarachchi, Jennifer Flegg and Nick Golding

With so many spatial/spatio-temporal modelling techniques to choose from, which model would you select to model the data you're interested in? In this talk, we discuss a newcomer's perspective of spatial modelling and compare four modelling techniques on Malaria prevalence in Kenya. The four models of interest are 1. Integrated Nested Laplace Approximations, commonly known as

INLA, 2. Spatial Random Forests, an extension of random forests to the spatial domain, 3. GPBoost, a tree boosting technique with Gaussian Processes and 4. Fixed Rank Kriging. We will discuss the challenges associated with a comparison of such diverse models and share some results.

Thurs 2:05-2:15

Strategies for Bayesian network modelling when some variables are more important than others

Innocenter Amima
University of Auckland, iami742@aucklanduni.ac.nz
Authors: Innocenter Amima and Beatrix Jones

Bayesian networks (BNs) are widely used for knowledge representation, reasoning, handling missing data and uncertainty. This study assesses the development of robust Bayesian network (BN) models to investigate the impact of two distinct management styles on the vineyard's longevity and sustainability in New Zealand. We have used a greedy search algorithm with bootstrap sampled data to learn multiple network structures from data with missing observations. We used a penalised likelihood (Schwarz's BIC) for model selection of these network structures. In addition, we propose using a cross-validation (CV) approach to build a consensus network by assessing the model's validity and predictive performance on a set of priority outcome variables. The final BN model provides suggestions that scientists can use for modelling (in)direct influences, predicting and generating potential hypotheses for further investigation.

Thurs 2:20-2:30

Changepoints methods for animal behaviour classification

Peter Green
AgResearch, peter.green@agresearch.co.nz
Authors: Peter Green and Bryan Thompson

Fitting sensors to animals allows for data collection on larger spatial and temporal scales than labour-intensive direct observation. However, that increase in data comes with the challenge of converting from sensor readings (e.g. GPS locations or accelerometer data) to meaningful animal behaviours. Current methods generally break the sensor data into fixed intervals (e.g. 30s or 1min), resulting in a discrete series of predicted behaviours. Animal behaviours though have a variety of durations. We might want the ability to recognise both a resting bout lasting for several hours and a short frolic lasting a few seconds. A long foraging bout might be composed of a series of short movements and feeding stations.

We explore the use of changepoint algorithms to translate high frequency sensor data into continuous intervals of predicted behaviours.

Thurs 2:30-2:40

Optimal design for on-farm strip trials --- systematic or randomised?

Zhanglong Cao
Curtin University, Zhanglong.Cao@curtin.edu.au
Authors: Zhanglong Cao, Andrew Grose, Jordan Brown and Suman Rakshit

There is no doubt on the importance of randomisation in agricultural experiments by agronomists and biometricians. Even when agronomists extend the experimentation from small trials to large on-farm trials, randomised designs predominate over systematic designs. However, the situation may change depending on the objective of the on-farm experiments (OFE). If the goal of OFE is obtaining a smooth map showing the optimal level of a controllable input across a grid made by rows and columns covering the whole field, a systematic design should be preferred over a randomised design in terms of robustness and reliability. With the novel geographically weighted regression (GWR) for OFE and simulation studies, we conclude that, for large OFE strip trials, the difference between randomised designs and systematic designs are not significant if a linear model of treatments is fitted or if spatial variation is not taken into account. But for a quadratic model, systematic designs are superior to randomised designs.

Thurs 2:40-2:50

Correcting for bias in the estimation of proportions by pooled testing

Graham Hepworth
Statistical Consulting Centre, The University of Melbourne, hepworth@unimelb.edu.au
Authors: Graham Hepworth and Brad Biggerstaff

Pooled testing (or group testing) arises when units are pooled together and tested as a group for the presence of an attribute, such as a disease. It originated in blood testing, but has been applied in many fields, including plant disease assessment and prevalence estimation of mosquito-borne viruses – the two fields in which we have encountered the technique.

In the estimation of proportions by pooled testing, the MLE is biased, and several methods of correcting the bias have been presented in previous studies. We have developed an estimator based on the bias correction method introduced by Firth (1993), which uses a modification of the score function. Our estimator is almost unbiased across a range of pooled testing problems, whether or not there are equal pool sizes.

In more recent work we have relaxed the assumption of perfect testing. This adds considerable complication to the derivation of the estimator, but it still results in an almost unbiased estimator in most situations of practical interest.

GxE and Plant Breeding. *Chair: Kaye Basford*

Thurs 3:30-3:45

Incorporating environmental covariates into a MET analysis with GxExM interaction effects

Michael Mumford
Department of Agriculture and Fisheries, Michael.Mumford@daf.qld.gov.au
Authors: Michael Mumford, Clayton Forknall, Daniel Rodriguez, Joseph Eyre and Alison Kelly

Incorporating a large number of explanatory variables in statistical models can lead to issues with model fitting such as variable selection, feature extraction and multi-collinearity. Inferential challenges can also arise, with the model fitting process needing to address overfitting and false discovery rates. Additionally, when data are collected from designed experiments, the design strata defined by the randomisation process need to be accounted for in the model.

In multi-environment trial (MET) analyses, genotype by environment (G×E) interactions are ubiquitous. The development of statistical methods to model the G×E interaction using environmental covariates (ECs) has almost exclusively occurred within the context of plant improvement programs. In agronomic research, measuring the impact of management practice (M) on genotype performance is also a key objective, giving rise to the genotype by environment by management practice (G×E×M) interaction.

A one-stage analysis approach is presented, including ECs to untangle the G×E×M interaction. This analysis approach uses a linear mixed model (LMM) framework, allowing for design effects and spatial field trend, along with heterogeneous residual variance across environments. The motivating dataset is composed of 17 environments across the northern grains region of Australia, each consisting of a common, continuously varying M treatment. The non-linear trait response to M, including interactions between G and M, are modelled via the LMM representation of the natural cubic smoothing spline. Similarly, the non-linear trait response to an EC is also captured using cubic smoothing splines, while the non-linear interaction between M and an EC is described using tensor cubic smoothing splines.

Nineteen ECs representing known eco-physiological drivers of crop growth, development and stress were derived from weather records and crop observation within each environment. Covariate data were obtained at different stratum, with some covariates measured at the G×E level, and others at the environment level. Covariates were incorporated into the model via a combination of forward and backwards selection procedures, to identify the most important predictors, avoid multi-collinearity, and ensure a parsimonious model. Cross validation was used to identify important ECs in the backwards selection procedure, and to assess the predictive performance of the final model in an unobserved (new) environment. The statistical software package *ASReml-R* enabled the fitting of the model and the calculation of predictions in ‘observed’ and ‘unobserved’ environments.

Initial results identified a subset of ECs determined to be the key drivers of G×E×M interaction in the motivating dataset, allowing researchers to identify the combination of G, M and ECs that optimised grain yield. The final model was capable of differentiating between high and low yielding environments when the subset of key ECs were used as a surrogate for the 'environment' model terms. Cross validation resulted in minimal decreases in predictive performance, implying that the final model has accounted for over-fitting, and can provide reasonable predictions in an unobserved environment.

This study is the first step in developing one-stage statistical models to identify the key eco-physiological drivers of G×E×M interaction, allowing for the development of more robust and targeted agronomic recommendations.

Thurs 3:45-4:00

On the use of factor analysis and iClasses to assess genotype by environment interactions in falling number across Australia.

David Hughes
University of Wollongong, hughesd@uow.edu.au
Authors: David Hughes, William Fairlie, Alison Smith and Brian Cullis

The Hagberg-Perten falling number (FN) test is the industry standard to measure starch degradation cause by late maturity α-amylase (LMA) enzyme activity in flour. In a recent study Sjoberg et al (2020) investigated the utility of a factor analytic mixed model to assess overall performance and stability of FN for a set of 129 soft wheat genotypes grown in 35 environments. They concluded that examination of factor loadings enabled the separation of environments and genotypes which were responsive to either pre-harvest sprouting or late maturity alpha amylase.

In this talk we use a multi-environment trial dataset spanning 6 years from 2014-2019, with more than 230 environments and 124 genotypes to examine the extent of genotype by environment interaction (GEI) in the Australian wheat growing regions. We use the so-called iClass approach of Smith et al (2021) to examine the overall performance within environment types with the same or similar GEI. We will also present some preliminary results on the relationship between overall performance within iClasses for FN and the propensity to express LMA based on the laboratory assay of Mrva and Mares (2001).

Thurs 4:00-4:15

Spatial Models for Co-located Trials

Monique Jordan
University of Wollongong, mj633@uowmail.edu.au
Authors: Monique Jordan, Alison Smith, Brian Cullis and Daniel Mullan

Breeding programs evaluate varieties using trials conducted at various locations and years with the combination of the two often being termed an environment. An early stage plant breeding trial conducted in a particular year and location is

often very large phenotypically evaluating a large number of cultivars. These large trials are often split into several smaller trials or management blocks and may even evaluate varieties from different stages. These trials are usually conducted in the same field with; (i) very similar management practices, (ii) sowing dates and (ii) harvesting dates. In this instance they are referred to as colocated trials and are considered to be in one environment. In contrast if trials in a particular year and location do not satisfy (i), (ii), (iii) they will not be considered to be colocated and would be considered to be in different environments. In the literature there is limited information on appropriate statistical models for the analysis of such colocated trials. Thus we aim to address this by investigating several statistical models used to analyse such colocated trials and recommend which one should be used in different scenarios. Each method used assumes a different spatial process is occurring on the field and are extensions of the baseline model used in Gilmour et al. (1997) for single trials to colocated trials. Of course there are limitations to which spatial process can be assumed if information is not known about the layout or configuration of trials and thus information pertaining to the distances between plots in different trials. We conducted an in silico experiment designed to compare and assess the performance of the approaches over various scenarios. We present the results of that simulation study and final recommendations on which models should be used to analyse colocated trials.

Gilmour, A. R., Cullis, B. R. & Verbyla, A. (1997). Accounting for Natural and Extraneous Variation in the Analysis of Field Experiments, vol. 2.

Thurs 4:15-4:30

Empirical comparison of times series models and tensor product penalised splines for modelling spatial dependence in plant breeding field trials

Bev Gogel
University of Wollongong, bgogel@uow.edu.au
Authors: Bev Gogel, Sue Welham and Brian Cullis

Plant breeding field trials have been widely analysed using linear mixed models in which low order autoregressive integrated moving average (ARIMA) time series models, and the subclass of separable lattice processes, are used to account for two-dimensional spatial dependence between the plot errors. A separable first order autoregressive model has been shown to be particularly useful in the analysis of plant breeding field trials. Recently, tensor product penalised splines (TPS) have been proposed to model two-dimensional random smooth variation in field trial data. This represents a non-stochastic smoothing approach and is in contrast to the separable autoregressive approach that models a stochastic covariance structure between the lattice of errors.

In this talk we will present the results of an empirical study in which we have compared the autoregressive (AR) and TPS models for a set of early generation plant breeding field trials. The fitted models include genetic relatedness through

ancestral (pedigree) information. This provides a more relevant framework for comparison than the assumption of independent genetic effects.

Judged by Akaike Information Criteria, the AR models resulted in a better fit than the TPS model for more than 80% of trials. In the cases where the TPS model provided a better fit it did so by only a small amount whereas the AR models made a substantial improvement across a range of trials. When the AR and TPS models differ, there can be marked differences in the ranking of genotypes between the two methods of analysis based on their predicted genetic effects. Using the best fitting model for a trial as the benchmark, the rate of mis-classification of entries for selection was greater for the TPS model than the AR models. Practically, this has implications for breeder selection decisions.

Invited Talk. *Chair: Chris Triggs*

Fri 8:45-9:30

The application of CycDesigN in designing experiments for use in agriculture

Emlyn Williams
ALPHASTAT, emlyn@alphastat.net

CycDesigN version 8.0 (CD8) has extensive features for the design of field and glasshouse experiments, including plant phenotyping. This talk will focus on major upgrades in the construction of partially replicated (p-rep) designs for one or more locations. Examples of neighbour balance and evenness of distribution (NB & ED) designs as applied to resolvable, non-resolvable and p-rep designs will also be presented.

Piepho, H.P., Michel, V. & Williams, E.R. (2018). Neighbour balance and evenness of distribution of treatment replications in row-column designs. Biometrical Journal 60, 1172-1189.

Piepho, H.P., Williams, E.R. & Michel, V. (2021). Generating row-column field experimental designs with good neighbour balance and even distribution of treatment replications. Journal of Agronomy and Crop Science 207, 745-753.

Mixed Topics 1. *Chair: Emi Tanaka*

Fri 10:00-10:15

The number of distinct alleles in mixed DNA profiles when contributors are related

James Curran
University of Auckland, j.curran@auckland.ac.nz
Authors: Maarten Kruijver and James Curran

The total number of alleles appearing in a DNA profile developed from a crime scene stain is sometimes used as a quick estimate of how many people may have contributed to this stain. The number of contributors is an important quantity when assessing the strength of DNA evidence.

DNA mixtures occur when more than one person contributes to a stain. The resulting mixed DNA profile may appear to have fewer contributors than actually contributed due to allele sharing and allelic dropout. Previous work [1,2] has estimated the risk of underestimating the number of contributors based on the maximum allele count across loci for mixtures of unrelated persons without dropout.

In this talk I will talk about a method we have developed for predicting the total number of distinct alleles in a mixed DNA profile. This method differs from previous work in that it allows contributors may be related according to a pedigree, and it allows for the incorporation of allelic dropout modelling.

[1] Coble M.D. et al., Uncertainty in the number of contributors in the proposed new CODIS set, Forensic Sci. Int.: Genet., 19 (2015), pp. 207-211

[2] Tvedebrink T., On the exact distribution of the numbers of alleles in DNA mixtures, International Journal of Legal Medicine, 128(3) (2014), pp. 427-437

Fri 10:15-10:30

Enabling statistical analysis without disclosing data values

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

In the age of data, data mining provides feasible tools to handle large datasets, which can consist of data from multiple sources. When data contain confidential information, there is limited research on how to retrieve statistical information from protected data when the intended analysis is undefined. It is of interest to develop feasible ways to allow the application of existing data mining techniques to confidential data. In the talk, we propose a new framework for simultaneously protecting data privacy at publishing and retrieving statistical information accurately. In particular, we introduce a method for determining the sample size of the resampled data, which is unique to data mining. The implementation of

the proposed framework is illustrated in an example of data clustering analysis using real data from Australian soybean seed data.

Fri 10:30-10:45

Predicting effective dose levels from exponential growth curves for residual herbicides on crop seedlings

Kerry Bell
Department of Agriculture and Fisheries, Kerry.Bell@daf.qld.gov.au
Authors: Kerry Bell, Emily Plant, Annemieke Ruttledge, Jesse Muller and Michael Widderick

Logistic curves were used to describe the detrimental rates of residual herbicides on crop seedlings grown in a controlled environment. These curves were used to estimate the concentration of herbicide at different 'effective doses' (ED).

In field crop research there is a need to understand at what level crops can tolerate residual herbicides applied to the field for weed control prior sowing a crop. For example, to be able to estimate the maximum concentration of herbicide that induces less than 30% plant damage. The measurements used to assess plant damage are often continuous and normally distributed (e. g. plant height, root length, plant dry weight). The logistic curve is appropriate for normally distributed data showing an 'S-shaped' trend and fitted well for this data. Often the survivorship curves based on binomial data produce effective doses, and a definition of an equivalent ED was needed for the logistic curve.

To determine an equivalent ED, the required ED was redefined as the percentage of the distance between the upper and lower asymptote. Using this value, an inverse-prediction was performed to determine the rate of herbicide required to obtain this ED, with each prediction being accompanied by an adjusted bootstrap percential (BCa) intervals.

To obtain a suitable data set to estimate the logistic curve parameters from the experiment, a series of preliminary experiments were run to determine the range of herbicide rates and intervals between rates, needed to accurately estimate the inflection point and the lower and upper asymptotes.

These curves and ED rates were supplied to the field crop industry to provide information about crop safety implications of herbicide concentrations detected in soil samples. This information will help industry to identify and minimise crop damage following the use of herbicides.

Mixed Topics 2. *Chair: Emi Tanaka*

Fri 10:50-11:05

Weighted estimation of linear mixed models under two-phase sampling for kākāpō genomics data

Pei Luo
University of Auckland, pluo244@aucklanduni.ac.nz
Authors: Pei Luo, Thomas Lumley and Ben Stevenson

Whole-genome sequencing of the critically endangered kākāpō has been completed, allowing genome-wide association studies for the entire population. However, this sort of effort is not feasible in most situations, and only some individuals can be sequenced. Despite the decreasing cost of DNA sequencing, the budget remains a substantial problem for most research funders. A cost-saving strategy is to select a subsample for whole-genome sequencing and then use the subsample data to estimate the same parameters as would be estimated with the complete data, and such a design is called a two-phase sampling. In many genome-wide association studies, the polygenic model is considered the founding principle because it allows the possibility that thousands of variants could contribute to the phenotypic variation in the population. Under the polygenic model, one can fit mixed models to measure the genetic effect of a particular variant while accounting the other variants as correlations between related individuals. I proposed a weighted maximum likelihood approach that takes advantage of the fact that the kākāpō population kinship structure is known, and I will talk about fitting linear mixed models to the kākāpō sample data and the model inference under two-phase sampling designs.

Fri 11:05-11:20

Managing and modelling multiple-response data

Thomas Lumley
Department of Statistics, University of Auckland, t.lumley@auckland.ac.nz

Multiple-response data arise from "Choose all that apply" or "Choose up to N" survey questions. They extend factor/class/category data, where exactly one option must be chosen. I will present tools in R for managing, displaying, and modelling multiple-response data. The 'rimu' package provides a class for representing multiple responses as single observations and functions for managing and visualising them. The 'rata' package provides regression modelling tools based on marginal generalised linear models, marginal loglinear models, and marginal multinomial regression models.

Fri 11:20-11:35

Sensitivity - a flexible criterion for comparison of measurement methods.

Murray Hannah
Agriculture Victoria Research, gchrismuz@gmail.com
Authors: Murray Hannah, Khageswor Giri and Rodrigo Albornoz

The problem of comparing measurement methods is addressed when none of the available methods can be considered a priori to be a “gold-standard” method of measurement. In this case, analysis of agreement between measurement methods is not enough and cannot establish which method is best. For this, an objective performance criterion is needed. In this paper, a sensitivity criterion introduced by Mandel and Stiehler (J Res Natl Bur Stand., 1954) but not well known outside the discipline of analytical chemistry is expanded in terms of components of variance. Sensitivity, so defined, is invariant to linear transformation and can be used to compare measurement methods with differing units, whilst encapsulating fundamental properties of the measurement method. The relative sensitivity provides an objective criterion upon which the merits of the competing measurement methods can be judged. Reformulation in terms of components of variance allows the sensitivity criterion to be more flexibly tailored to context, applied to structured data and fit for purpose. As an example, two methods of measuring body condition score in dairy cattle are considered.

Invited Talk. *Chair: Bev Gogel*

Fri 11:45-12:30

ASReML in Practice

Arthur Gilmour
CARGOVALE.COM, arthur.gilmour@cargovale.com.au

The linear mixed model has proved a very useful tool for data analysis, especially in the contexts of plant and animal breeding, agricultural experimentation and data exploration. The design of ASReml allows it to fit a large range of linear mixed models in a variety of contexts. The key features are the use of various correlation structures, use of direct product variance structures and sparse computing methods facilitating the fitting of large models. The linear mixed model has been extended to generalised linear mixed models and hierarchal generalised linear mixed models. The paper will outline the joint model of mean and dispersion.

Poster Presenters in order of Presentation

The poster session will start with 1min talks for each poster, in the order given above. You may wish to present your poster as a single slide as a background to this talk (one slide only please). Please email the slide to aascconf2022@gmail.com with subject 'Poster {your name}' or take it to the registration desk. Poster presenters are invited to also use this method to make their poster available on the website (https://aasc2022.netlify.app/).

Simulation Meta-Model Variance Reduction for Latin Hypercube Designs

Jessica Penfold
Human and Decision Sciences Division, Defence Science and Technology Group,
jessica.penfold@defence.gov.au
Authors: Stelios Georgiou, Andrew Gill, Stella Stylianou and Haydar Demirhan

Constructing a meta-model of a computationally expensive computer simulation is an important analytical task, as the meta-model can then be used to serve various analytical objectives such as sensitivity analysis, prediction or optimisation. A fitted meta-model requires generation of a sample of input-output simulation data, and the quality of the fit - and thus of the experimental design - is often with regard to the size of the generalised variance of the fitted model parameters. While a larger design will clearly help, the additional computational cost has led researchers to investigate alternative variance reduction techniques.

For stochastic simulations, which require a stream of pseudo-random numbers (PRNs) in order to execute a given design point, one such technique involves assigning these PRN streams to the set of design points in such a way that the resultant correlations between simulation outputs has a net beneficial effect on the size of the generalised variance of the fitted model parameters. Initial research considered a main-effects first-order meta-model and the designs analysed were factorial-based, which allowed an assignment strategy to be associated with a confounding contrast that achieves orthogonal blocking. Since then, various extensions have been made, including second-order meta-models, multiple blocks, and multiple classes of effects. However, the designs explored in each case remained within the factorial-based class.

Latin Hypercube Designs (LHDs) are a popular alternative for fitting meta-models, including second-order meta-models, but there appears to be no analytic means (PRN assignment strategy) of achieving orthogonal blocking. There are numerous ways of constructing LHDs and of particular interest in this paper is the class of LHDs constructed by rotating an underlying factorial-based design. The intent of this paper is to investigate whether the confounding contrast of the underlying factorial-based design can be used to assign PRN streams to the design points of these rotated LHDs as a variance reduction technique.

While exhaustive enumeration of all blockings did confirm the optimality of the assignment strategy, it also revealed multiple other blockings with the same variances - thus the optimal blocking was not unique. Furthermore, while the trace of the covariance matrix of the fitted parameters of both a first-order linear

meta-model in two factors, and a second-order linear meta-model in three factors was shown analytically to be smaller than when using independent PRNs, the relative improvement of using two blocks and inducing negative correlations between blocks by using ARNs was observed to be minor compared with simply using common random numbers (CRNs).

The results suggest that if optimal blocking is to be pursued, then some form of intelligent computer search (as opposed to exhaustive enumeration) will be required. Indeed, if one broadens the class of meta-models to those within generalised linear models (to enable the modelling of discrete (or binary) response variables), then this is almost a surety, as even the notion of an optimal design is computationally-based. In this spirit, and perhaps as a remedy to unlocking the potential of blocking, research into jointly optimising the PRN assignment and design could be followed.

Constructing Large Nonstationary Spatio-Temporal Covariance Models via Compositional Warpings

Quan Vu
University of Wollongong, quanv@uow.edu.au
Authors: Quan Vu, Andrew Zammit Mangion and Stephen Chuter

Understanding and predicting environmental phenomena often requires the construction of spatio-temporal statistical models, which are typically Gaussian processes. A common assumption made on Gausian processes is that of covariance stationarity, which is unrealistic in many geophysical applications. In this talk, we introduce a new approach to construct descriptive nonstationary spatio-temporal models by modeling stationary processes on warped spatio-temporal domains. The warping functions we use are constructed using several simple injective warping units which, when combined through composition, can induce complex warpings. A stationary spatio-temporal covariance function on the warped domain induces covariance nonstationarity on the original domain. Sparse linear algebraic methods are used to reduce the computational complexity when fitting the model in a big data setting. We show that our proposed nonstationary spatio-temporal model can capture covariance nonstationarity in both space and time, and provide better probabilistic predictions than conventional stationary models in both simulation studies and on a real-world data set.

Cause-specific hazard, or cause-specific incidence, that is the question.

Tomomi Yamada
Osaka University, tomomi.yamada@dmi.med.osaka-u.ac.jp
Authors: Tomomi Yamada, Tsuyoshi Nakamura, Hiroyuki Mori, Todd Saunders and Yoshiaki Nose

The Fine-Gray subdistribution hazard model has been widely used and is regarded as one of the most popular models to estimate the incidence of outcomes over time in the presence of competing risks. Recently, however, some studies have reported unexpected effects of censors on the Fine-Gray

model analysis. The most basic requirement in failure time, or survival data analysis, is for the results of statistical analysis to be independent of the distribution of independent censoring. For instance, the Cox proportional hazards model is not affected by the distribution of independent censoring.

This consideration prompted us to explore any root causes of these unexpected censor effects on the Fine-Gray model analysis.

The objective of this paper is to show that the Fine-Gray hazard depends on the distribution of independent censoring; and to propose countermeasures to avoid problems with the Fine-Gray hazard in applications. As far as we can determine, the dependency of the Fine-Gray on the distribution of independent censoring has not been described in literature. This paper points out a gap in the mathematical proof described by Fine and Gray[1], and then provides a condition under which the Fine-Gray hazard does not depend on the distribution of independent censoring. The results are numerically demonstrated by a simple simulation. We also discuss the peculiarities of the example that Gray[2] presented as a drawback of the cause-specific hazard that motivated the development of the Fine-Gray hazard. Finally, we discuss the advantages and disadvantages of the cause-specific hazard and the cumulative incidence function in competitive risk analysis of clinical data.

Fine JP, Gray RJ. A proportional hazards model for the subdistribution of a competing risk. J Am Stat Assoc 94: 496 - 509, 1999.

Gray, R. J. A class of k-sample tests for comparing the cumulative incidence of a competing risk. Annals of Statistics 16: 1141- 1154, 1988.

ATLA & the forward search; for robust outlier detection and clustering in multivariate data

Andrew Grose
Curtin University, andrew.grose@curtin.edu.au
Authors: Andrew Grose and Brenton Clarke

Comparisons of procedures in multivariate outlier detection are sparse. Yet those that exist often utilise disparate procedures with varying measures of performance. This combined with the wide variety of classification techniques made under various assumptions make comparisons difficult. To mitigate this we look at one such class of outlier detection procedures with comparison achieved out of simulation study through several generating functions and measures of performance. In particular, automated procedures that use the forward search along with Mahalanobis distances to identify and classify multivariate outliers subject to predefined criteria are examined. Procedures utilizing a parametric model criterion based on a χ^2-distribution are among these, whereas the multivariate Adaptive Trimmed Likelihood Algorithm (ATLA) identifies outliers based on an objective function that is derived from the asymptotics of the location estimator assuming a multivariate normal

distribution. Criterion including size (false positive rate), sensitivity and relative efficiency are canvassed. To illustrate a novel measure; variability of multivariate location estimation are used through simulation out of Tukey-Huber ϵ-contamination models. Mean slippage models are also entertained in order to ensure the results are not limited to the generating model. The simulation results here are illuminating and demonstrate there is no broadly accepted procedure that outperforms in all situations, albeit one may ascertain circumstances for which a particular method may be best if implemented. Finally the research explores graphical monitoring for existence of clusters and the potential of classification through the objective function using ATLA. That is, we demonstrate the possibility to identify and categorise clusters as well as outliers generated out of each these contaminating distributions. This is illustrated through simulation out of multiple predefined cluster distributions as well as by extension to real world data.

Optimal design of two-phase sampling for generalized raking estimations of regression modelling

Tong Chen
The University of Auckland, tche929@aucklanduni.ac.nz
Authors: Tong Chen and Thomas Lumley

Two-phase designs measure variables of interest on a subcohort where the outcome and covariates are readily available or cheap to collect on all individuals in the cohort. Given limited resource availability, it is of interest to find an optimal design that includes more informative individuals in the final sample. We explore the optimal designs and efficiencies for analyses by design-based estimators.

Generalized raking is an efficient class of design-based estimators, and they improve on the inverse-probability weighted (IPW) estimator by adjusting weights based on the auxiliary information. We derive a closed-form solution of the optimal design for estimating regression coefficients from generalized raking estimators.

We compare it with the optimal design for analysis via the IPW estimator and other two-phase designs in measurement-error settings. Our results show that the optimal designs for analyses by the two classes of design-based estimators can be very different. The optimal design for analysis via the IPW estimator is optimal for IPW estimation and typically gives near-optimal efficiency for generalized raking estimation, though we show there is potential improvement in some settings.

Post-processing approach for estimating valid standard errors for the Bayesian Lasso

Mohammad Javad Davoudabadi
University of Sydney, mohammad.davoudabadi@sydney.edu.au
Authors: Mohammad Javad Davoudabadi, John Ormerod, Samuel Muller and Garth Tarr

Penalised regression methods are widely used for simultaneous parameter estimation and variable selection. In the Bayesian paradigm, Bayesian penalised regression models can be fit using computationally expensive Markov chain Monte Carlo sampling. Alternative faster methods such as variational Bayes (VB) are commonly employed as computationally simpler alternatives. However, VB can be shown to underestimate the posterior variance and not produce valid standard errors. In this study, we find improved approximations to posterior variances for the Bayesian Lasso through post-processing of the VB estimated model parameters. We will demonstrate these findings on real data including on a melanoma dataset.

The Application of Supernodal Methods to the Fitting of Factor Analytic Linear Mixed Models for Plant Variety Trials

Luke Mazur
University of Wollongong, lmazur@uow.edu.au

Factor analytic linear mixed models (FALMM, Smith et al. 2001) are commonly used to produce single stage analyses of multi-environment trials (METs) that accommodate non-genetic sources of variation (including variety by environment interaction) as well as using spatial methods for the residuals at each trial. The Average Information algorithm (Gilmour et al. 1995) as implemented in ASReml is commonly used to fit these models, and each iteration requires factorization of the coefficient matrix of the mixed model equations, followed by a sparse inversion. This matrix is sparse when the variety effects are treated as independent, but modern analyses tend to incorporate relatedness between varieties in the form of pedigree or genomic information, making the associated matrices far denser. This results in these computational steps taking an unreasonable amount of time, and so the software required to perform the analyses must evolve to meet these challenges.

One possible solution is the addition of supernodal methods (Duff & Reid, 1983), which allow for processing dense blocks within the coefficient matrix using highly optimized dense linear algebra libraries. These methods were compared to other computing methods in a series of two in-silico experiments, featuring a coefficient matrix that has a large dense block at the top and a number of supernodes below it, respectively corresponding to the factors and the environments in a typical coefficient matrix arising from an FALMM fitted to a MET. The first uses a simplified structure in order to partition the runtime in terms of the the time taken for completing the floating point operations (FLOPs) and the time taken to manage the sparsity pattern of the matrix.

The second considered a more realistic structure, and different computing methods were compared across various sizes of root and supernodes, as well as with a number of different computing threads.

The best supernodal methods are shown to be up to a factor of 10 times faster than ASReml.

Duff, Iain S & Reid, John K (1983). The multifrontal solution of indefinite sparse symmetric linear. ACM Transactions on Mathematical Software (TOMS) 9(3), 302–325.
Gilmour, Arthur R, Thompson, Robin & Cullis, Brian R (1995). Average information REML: An efficient algorithm for variance parameter estimation in linear mixed models. Biometrics 51(4), 1440–1450.
Smith, Alison B, Cullis, Brian R & Thompson, Robin (2001). Analyzing variety by environment data using multiplicative mixed models and adjustments for spatial field trend. Biometrics
4971 57(4), 1138–1147.

Survey design for rank order reactive questions using R Shiny

Maryann Staincliffe
AgResearch Limited, maryann.staincliffe@agresearch.co.nz
Authors: Maryann Staincliffe and Chris Buddenhagen

Our aim was to create a survey to help identify potentially stretchy impactful science in the field of pasture and weed science and management. Workshops with New Zealand experts were held to identify emerging issues that will drive change in pasture weed science and management. We asked the responders to rank themes based on three criteria: “Is it likely to become more important in 10 to 20 years”, “Work in this area will stretch knowledge boundaries”, “Successful research on this could result in significant changes to weed management”. We had six major themes (parent) that were ranked, and then in each of these parent themes we had 3 to 5 subordinate themes to be ranked. Having each responder rank subordinate themes for all the major themes would be too time consuming, so it was decided that each responder would only be shown the subordinate themes for their three highest ranked major themes. Thus, the survey questionnaire was reactive to their responses. This presentation will show how shiny can be used to create reactive survey design.

Evaluation of telehealth usage in child and family allied health services during the COVID-19 pandemic

Alex Shaw
The University of Sydney, alexander.shaw@sydney.edu.au
Authors: Alex Shaw, Eva Litherland, Donna Thomas, Sarah Masso, Melanie Keep and Vijay Kapadiya

Telehealth refers to the delivery of healthcare services via information and communication technologies. There was a dramatic uptake of telehealth as a mode of clinical engagement during the onset of the COVID pandemic in early 2020. In the child and family allied health setting, many clinical appointments

(including both assessments and interventions) can effectively be conducted either face-to-face or by telehealth. The choice of engaging in a face-to-face or telehealth appointment during the COVID pandemic is likely driven by multiple patient, clinician, and appointment-level clinical factors. With the exponential growth and subsequent persistence of telehealth as a mode of clinical engagement, the factors that drive telehealth usage in the pandemic and post-pandemic era need to be clearly understood.

The routine collection of patient demographic information and appointment-level clinical data (including mode of engagement) across child and family health clinics in the Western Sydney Local Health District during 2020 provided an opportunity to study patterns of telehealth usage at the onset of COVID. The study period encompassed the lockdown period accompanying the first wave of COVID, and subsequent months of eased restrictions and local COVID suppression. Data was collected on 5264 appointments across 466 patients and 66 clinicians.

We examined whether several known patient demographic and clinical factors were associated with choice of telehealth. We used a mixed-effects logistic regression approach, with the choice of telehealth for each appointment as the outcome, and both clinician and patient IDs as random factors. The base model included fixed effects for patient gender, patient age range, home centre location, service type (speech therapy, occupational therapy or counselling) and time period of appointment. Models with additional fixed effects were considered beyond the base model were evaluated and the fit compared using AIC. The final model chosen included additional fixed factors of sequential appointment number (for individual patient), and the interaction between appointment number and appointment time period. Of the fixed factors there was strong support ($p < 0.001$ for LRT) for the service type, patient age, appointment time period main effect and interaction with visit number as being associated with choice of telehealth. Patient ID and clinician ID had similar effect sizes (ICC=0.247 and ICC=0.239 respectively), and both factors collectively appeared important in model prediction (conditional model pseudo R^2=64.5%; marginal model=30.8%). Alternate modelling approaches incorporating time-series analysis will also be presented.

Patient age and service type was clearly associated with telehealth usage, which is not surprising given differing clinical need for face-to-face appointments based on these factors. There was little evidence for other known patient factors. Unknown patient and clinician level factors clearly had a strong influence. In future, these could be characterised through surveys of attitudes and preferences towards telehealth among patients and clinicians. Telehealth usage appeared most strongly driven by the time period of appointment, consistent with average telehealth engagement varying across the year in line with changing health advice, knowledge and risk perception throughout 2020.

Applications of machine learning models in making operational decisions: a case study in blood transshipment problem

Babak Abbasi
RMIT University, babak.abbasi@rmit.edu.au
Authors: Mahdi Abolghasemia, Babak Abbasi and Zahra Hossenifard

Using machine learning in solving constraint optimization problems is becoming an active research area due its capabilities in speeding up the computations, lowering the costs, and providing accurate solutions. This is in particular important for organisations that needs to make fast tactical decisions on a regular basis but do not have the resources to develop and implement expensive models. This paper studies a problem in a blood supply chain and explores how we can predict the solutions of constraint optimization problems using machine learning techniques. We develop an algorithm where we use multi-output machine learning models that not only provide near-optimal solutions but also organically take into account the constraints of the models to a large extent. We investigate the importance of loss functions and measure the performance of the predictive models in optimising various decisions and their associated costs. Our results show that while using LightGBM model with the mean absolute deviation criterion as loss function provides solutions with only 2% higher total cost than a mathematical model, it reduces transshipment and shortage costs by 27% and 6%, respectively.

Exploring the performance of Inductive Linearisation for simulation of the Van der Pol oscillator

Sepideh Sharif
University of Otago, sepi.sharif@postgrad.otago.ac.nz
Authors: Sepideh Sharif, Chihiro Hasegawa and Stephen Duffull

Background: A dynamical system is one in which a function defines the temporal dependence of a point in an arbitrary space set expressed mathematically by differential equations. This means that a dynamic system is anything that depends on variations of values *x(t)* with time *t*. Dynamical systems are pervasive across disciplines including finance, meteorology and pharmacology. Although exact solutions for nonlinear dynamical systems cannot generally be established using algebraic methods, speedy and accurate solutions are desirable. Consequently, solving systems of nonlinear differential equations is of considerable importance. In this work we explore a new method (*Inductive Linearisation*) for solving nonlinear systems when applied to stiff systems. This method approximates solutions to nonlinear systems utilising iterative linearisation to create a linear time-varying (LTV) system of differential equations that spans the entire time interval of interest. The resultant LTV is then solved using *eigenvalue decomposition* (EVD). This method has been optimised and compared favourably to an inbuilt differential equation solver in MATLAB (*ode45*) for solving non-stiff nonlinear systems in pharmacology (1). Here we apply the method to the Van der Pol oscillator as a common example of a nonlinear dynamical system with arbitrary stiffness. The performance of the *Inductive Linearisation* solver was investigated in this study.

Objectives: To investigate the performance of *Inductive Linearisation* for simulation of the Van der pol oscillator.

Methods: Simulations were performed in MATLAB. The Van der Pol system was examined for two values of the damping parameter (μ=1,10), denoting non-stiff and very stiff. Execution time and graphical precision were examined for Inductive Linearisation where ode45 and ode23s were chosen as the reference solvers for the non-stiff system and ode23s was selected for the stiff system. The time span for evaluation was 30 units allowing for 2 cycles of the oscillator.

Results: For the non-stiff Van der Pol system (μ=1), the Inductive Linearisation solution solved the system in 0.57 seconds. The reference ode45 solver took 0.58 seconds whereas the ode23s solver required 1.20 seconds. There was excellent graphical agreement between the two solutions. For the stiff Van der pol system (μ=10), Inductive Linearisation took 165 seconds and the reference solver, ode23s, took 2.25 seconds. There was excellent graphical agreement between the two solutions.

Conclusion: Our results indicate that the *Inductive Linearisation* solver performed well for both the non-stiff and stiff systems but was considerably less efficient than the reference solver for the stiff system. Importantly, *Inductive*

Linearisation worked regardless of stiffness, hence there is no need for users to know this beforehand in order to guess which ode solver they need to employ, while this may not be the case for numerical methods (e.g., choice of either *ode45* or *ode23s*). There is a need for further optimisation of the *Inductive Linearisation* for stiff systems.

References: 1.Sharif S, Hasegawa C, Duffull SB. Exploring Inductive Linearization for simulation and estimation with an application to the Michaelis–Menten model. J Pharmacokinet Pharmacodyn. 2022:1-9.

Blood units substitution decisions: a stochastic optimisation approach

Zahra Hosseinifard
University of Melbourne, zahra.h@unimelb.edu.au
Authors: Zahra Hosseinifard, Mostafa Khatami and Babak Abbasi

Abstract: Substitution is known as an efficient strategy to mitigate the supply chain risk in dealing with demand uncertainty. Substitution is generally considered in designing effective inventory replenishment policies, as it can reduce the constraints imposed by limited storage capacity and alleviate the negative impact of demand uncertainty. If efficiently designed, it can reduce shortage and holding costs. A well-known example of substitution practice is in blood transfusion of different compatible blood types at hospitals or emergency departments. Among different blood types, O-negative is most commonly used for substitution, due to its compatibility property, which means it can be given to any patient without knowing their blood type. A recent study by Australian Red Cross Lifeblood (Hirani et al. 2017) revealed the issue of O-negative over-ordering at hospitals. They found that O-negative red cells units were mainly transfused to prevent wastage and in-close-to-expiry situations whereas identical blood groups may have been more suitable.

This research focuses on ordering policy with consideration of effective substitution decisions for red blood cells at hospitals and emergency requisition from the blood service with consideration of the optimal substitution policy that has not been explored in the blood supply chain literature. We consider demand and supply as stochastic. The mathematical modelling approach to the problem is by considering a stochastic optimisation model under substitutions, uncertain demand, stochastic supply, perishable items with fixed shelf life and the age of items in inventory. To improve the performance of blood supply chain, the outdates and shortages and the age of transfused items should be minimised.

References: Hirani, R., Wong, J., Diaz, P., Mondy, P., Hogan, C., Dennington, P.M., Pink, J., Irving, D.O., 2017. Anational review of the clinical use of group od–red blood cell units. Transfusion 57, 1254–1261

'Data Stories' : Illustrating and de-mystifying agricultural statistics

Annie Conway
The University of Adelaide, annie.conway@adelaide.edu.au
Authors: Annie Conway, Sam Rogers and Olena Kravchuk

Our project, titled 'Data Stories', is an online resource consisting of a curated collection of real datasets, showcasing the variety of research undertaken in agricultural science in South Australia. Accompanying each dataset is a story - a short document that includes the background of the research project, and a tutorial on statistical methods that can be used to analyse the data. The goal of the resource is to promote and publicise agricultural research, while at the same time providing an educational resource to teach statistical applications at various levels. The stories and the data in the online repository may be used directly as teaching tools, or as an option for self-study. In this poster presentation we will demonstrate the resource and its uses.

The AIBL Toolbox: An R Shiny application to aid research and clinical decision making

Rodrigo Canovas
CSIRO Health and Biosecurity, Australian e-Health Research Centre, Parkville 3052, Victoria, Australia, rodrigo.canovas@csiro.au
Authors: Rodrigo Canovas, Angelina Duan and James Doecke

Background: In the study of Alzheimer's disease (AD) the ability to both predict and detect a participant's beta-amyloid (Aβ) status (Aβ+ or Aβ-), as well as the ability to estimate the prevalence of Aβ+ in a cohort, is indispensable when designing research projects.

Methods & Results: In this work we present the RShiny "AIBL Toolbox" application, which allows the user to explore the graphical user interface showcasing blood-based biomarker data across different disease groups and then explore results from statistical modelling using different biomarkers to predict Aβ status. By modifying various experimental parameters, such as threshold cut-offs, treatment groups, and selected biomarkers, researchers are able to see how these measures change in different scenarios. Within the toolbox, users can statically and dynamically select appropriate relevant parameters to perform statistical analysis for calculating cohort requirements during the planning phase of an experimental set-up. That is, the tool provides statistical information that help to layout and test possible scenarios which allows the use to define future cohort requirements before expending time and money into it.

Application: The application is divided into the following: (1) applications that provide the user with an approximation on how statistics (such as positive predictive value, negative predictive value, area under the curve, age range, predictive prevalence, risk prediction and so on) perform in the selected reference cohort; and (2) applications that allow the user to create a simple

linear model and check the relevant statistics associated with the parameters in question. Additionally, the user has the option between inputting their own data for the reference cohort or using the default subset of Australian Imaging, Biomarkers and Lifestyle (AIBL) study of ageing data, which contains follow-up data on biomarkers, cognitive characteristics, and health and lifestyle factors from an Australia cohort (approximate sample size = 600, aged between 46 and 96 years old, and equally gender distributed).

Highlights of "AIBL Toolbox" functionalities include: (1) provide assistance to envision the number of expected positive predicted Aβ+ participants in a new cohort, based on the expected size of the new cohort, the reference cohort and biomarker univariable model selected, the model sensitivity desired, and the cognitive status of the samples on the reference cohort; (2) approximate the sample size required depending on the number of predicted Aβ+ desired, assuming that the new cohort behaves similarly to the reference cohort; (3) dynamically calculate the individual predictive risk of being Aβ+ within the pre-computed generalised linear model given their personal parameters.

Summary: This app presents an AIBL based statistical summary tool, allowing researchers to access a range of statistical analyses through numerical and graphical representation. This provides a fast way to estimate the effect of the most relevant Aβ+ biomarker predictors, giving information on sample size calculations, appropriate Aβ+ group assignment and definitions, cut-off biomarker thresholds, and other logistical and design considerations needed in setting up new research projects.

Using imputation to harmonize measures of cognition across three large cohorts: AIBL, ADNI and OASIS

Rosita Shishegar
The Australian e-Health Research Centre, CSIRO, Melbourne, Australia, rosita.shishegar@data61.csiro.au
Authors: Rosita Shishegar, Christopher Fowler, Fiona Lamb, Joanne Robertson, Simon Laws, Paul Maruff, Hamid R Sohrabi, Stephanie Rainey-smith, Greg Savage, Jason Hassenstab, Christopher Rowe, Colin L. Masters, Michael Weiner, Victor Villemagne and James D. Doecke

Background: To improve understanding of Alzheimer's disease (AD), large observational studies are needed to increase power for more nuanced analyses. Combining data from different but similar study cohorts is one solution. However, the disparity of these datasets, e.g. using different tests to assess specific cognitive domains, makes this a challenging task. Here, we propose a harmonisation solution using imputation strategies for cognitive memory performance in 3 large Alzheimer's datasets (AIBL, ADNI and OASIS).

Methods: AIBL participants (N=2936 [925 normal control (NC), 1214 mild cognitive impairment (MCI) and 797 AD; aged 75.54±7.48; 1360 females]), ADNI participants (N=2327 [847 NC, 1061 MCI, 419 AD; aged 73.12±7.4; 1091 females]) and OASIS participants (N=780 [539 NC, 149 MCI, 92 AD; aged

73.48±7.25; 403 females]) were included in this study. Scores for cognitive tests administered in one cohort but not the other were imputed in the cohort for which they were missing (with the Rey's Auditory Verbal Learning Test (RAVLT) being used in ADNI, the California Verbal Learning Test (CVLT-II) in AIBL and Selective Reminding Test in OASIS). Non-parametric multivariate imputation using random forests (missForest)[8] was employed to impute the 'missing' data and test scores across the cohorts from a selection of similar cognitive test scores from the neuropsychological testing batteries, and age, gender, years of education and the major genetic risk factor for AD (*APOE*-ε4 status) were used to inform the models. The stability of the method was assessed by comparing imputed test scores imputed using 2 datasets and the scores imputed using all 3 datasets. To further validate the method, statistical differences between AIBL CVLT-II delayed recall scores imputed using a model trained with ADNI and a secondary model trained with OASIS were compared using two-tailed T-tests.

Results: Comparing imputed and harmonised AIBL and ADNI data using 3 datasets (AIBL, OASIS and ADNI) with harmonised data using only AIBL and ADNI data resulted in correlation coefficients of ≥96% and the best mean absolute error (MSE) of 2.2 for the RAVLT immediate recall test (range 0-60) and the worst MAE of 1.54 for the Boston naming test (range 0-30). No significant statistical differences ($p<.001$) between imputed CVLT-II test scores were observed using the model trained with ADNI and the one trained with OASIS.

Conclusion: The need to combine large datasets will increase as we aim to improve the generalisability of findings and test more complex hypotheses. The results here suggest it is possible to use data imputation, capitalising on underlying structures and relationships, to impute specific test scores in a cohort for which that test was not administered. In turn, providing a practical solution for data harmonization across large datasets.

References:

1 doi:10.1038/srep21689.
2 doi:10.1002/sim.5894.
3 doi:10.1017/S1041610209009405.
4 doi:10.1212/WNL.0b013e3181cb3e25
5 doi:10.1076/clin.14.4.526.7204.
6 doi:10.1002/1097-4679(199311)49:6<883::AID-JCLP2270490616>3.0.CO;2-6
7 doi:10.1212/wnl.43.8.1467. 8doi:10.1093/bioinformatics/btr597.

Development of a thyroid cancer recurrence prediction calculator: A regression approach

Jiaxu (Jimmy) Zeng
Department of Preventive and Social Medicine, Dunedin School of Medicine, University of Otago, Dunedin, jimmy.zeng@otago.ac.nz
Authors: Jiaxu (Jimmy) Zeng, Jean Li, Mark Sywak, Carolina Nylén, Robin Turner

Background: The thyroid cancer staging calculator has been recognised as one of the most efficient tools for assisting clinicians in making clinical treatment decisions. However, the current calculator is missing patients' serum Thyroglobulin information, which is crucial for staging cancer patients in practice. The primary aim of this study is to update current calculator with serum thyroglobulin included based on the tertiary thyroid cancer service database from Australia.

Methods: Records from 3962 thyroid patients were analysed for training a logistic model for predicting recurrence. Twelve predictive variables were chosen under close guidance of thyroid cancer specialists, which includes age at operation, sex, number of carcinomas presented in the operation, size of the greatest tumour, histologic type of carcinoma, extrathyroidal extension status of tumours, pathologic staging of the primary tumour, presence of venous invasion of the primary tumour, immunohistochemistry for the primary tumour, presence of extranodal spread, number of lymph nodes and serum thyroglobulin level presented in the scans.

Results: The strongest predictors were number of lymph nodes, histologic type of carcinoma and most importantly, the serum thyroglobulin level.

Conclusions: This study has addressed an important concern that serum thyroglobulin information was not used to predict thyroid cancer recurrent in practice.

On the Conway-Maxwell-Poisson point process

Yan Wang
RMIT University, yan.wang@rmit.edu.au
Authors: Yan Wang, Ian Flint and Aihua Xia

The Poisson point process plays a pivotal role in modelling spatial point patterns. One of its key features is that the variance and the mean of the total number of points in a given region are equal, making it unsuitable for modelling point patterns that exhibit significantly different mean and variance. To tackle such point patterns, we introduce the class of Conway-Maxwell-Poisson point processes. Our model can easily be fitted with a logistic regression, its point counts in different regions are correlated and its log-likelihood in any subregion can be easily extracted. Both simulations and real data analyses have been carried out to demonstrate the performance of the proposed model.

A model-robust design approach for optimally sub-sampling Big data

Amalan Mahendran
School of Mathematical Sciences, Queensland University of Technology, mahendr3@qut.edu.au
Authors: Amalan Mahendran, Helen Thompson and James McGree

For the analysis of Big data, computationally efficient and scalable methods are needed to support timely insights and informed decision making. One such method is known as sub-sampling where an informative subset of the Big data is analysed and used as the basis for inference, rather than considering the whole data set. A key question when implementing such a method is how to obtain an informative subset based on the questions being asked of the Big data. For this purpose, random sampling has been proposed based on sub-sampling probabilities determined via methods from optimal experimental design. However, a major drawback of this approach is that the sub-sampling probabilities can rely on an assumed model for the Big data. To address this, we propose a model robust approach, where a set of models is instead considered, and the sub-sampling probabilities are evaluated based on the weighted average of the probabilities that would be obtained if each model was considered singularly. Theoretical support for this approach is provided, and it is then applied in Generalised Linear Model settings in simulation and in two real world applications. The results from this show that our model robust approach outperforms current sub-sampling methods through providing informative data for estimating a range of potential models for the Big data.

List of Delegates

Alphabetical by first name

Name	Email
Aidan McGarty	amcgarty@uow.edu.au
Alaa Amri	alaa.amri@ed.ac.uk
Alan Herschtal	alan.herschtal@monash.edu
Alex Shaw	alexander.shaw@sydney.edu.au
Amalan Mahendran	n10678905@qut.edu.au
Andrew Grose	andrew.grose@curtin.edu.au
Andrew van Burgel	andrew.vanburgel@dpird.wa.gov.au
Anil Dolgun	Anil.Dolgun@csl.com.au
Anne Harris	anne.harris@dpi.nsw.gov.au
Annie Conway	annie.conway@adelaide.edu.au
Arjun Sekhar	arjun.sekhar177@gmail.com
Arthur Gilmour	arthur.gilmour@cargovale.com.au
Babak Abbasi	babak.abbasi@rmit.edu.au
Bethany Rognoni	Bethany.Rognoni@daf.qld.gov.au
Bev Gogel	bgogel@uow.edu.au
Bradley Wakefield	bnw722@uowmail.edu.au
Carolyn Huston	carolyn.huston@csiro.au
Chris Brien	chris.brien@iinet.net.au
Chris Howden	chris@trickysolutions.com.au
Chris Triggs	triggs@stat.auckland.ac.nz
Christopher Chung	chungc@uow.edu.au
Clayton Forknall	clayton.forknall@daf.qld.gov.au
Damian Collins	damian.collins@dpi.nsw.gov.au
Daniel Pesu	Daniel.Pesu@environment.nsw.gov.au
David Baird	david@vsn.nz
David Hughes	Hughesd@uow.edu.au
Debra Partington	debra.partington@agriculture.vic.gov.au
Dilani Kaveendri Dambure Hewage	dilanikaveendri@gmail.com
Dongwen Luo	dongwen.luo@agresearch.co.nz
Edward Cripps	edward.cripps@uwa.edu.au
Emi Tanaka	emi.tanaka@monash.edu
Emlyn Williams	emlyn@alphastat.net
Francis Hui	francis.hui@anu.edu.au
Giri Khageswor	khageswor.giri@agriculture.vic.gov.au
Graham Hepworth	hepworth@unimelb.edu.au
Innocenter Amima	iami742@aucklanduni.ac.nz
James Curran	j.curran@auckland.ac.nz
James Doecke	james.doecke@csiro.au
Jessica Penfold	jessicapenfold123@gmail.com
Jiaxu Zeng	jimmy.zeng@otago.ac.nz
Joanne Thandrayen	Joanne.Thandrayen@anu.edu.au
Jordan Brown	jordan.brown@curtin.edu.au
Kai Bagley	kaibagley98@gmail.com
Kathrin Schemann	kathrin.schemann@sydney.edu.au
Katie Silversides	katherine.silversides@sydney.edu.au
Kaye Basford	k.e.basford@uq.edu.au
Kerry Bell	Kerry.Bell@daf.qld.gov.au

Kerry de Vent	kerry@de-vent.net
Kym Butler	kym.butler@unimelb.edu.au
Lu Wang	luw@uow.edu.au
Luke Boyle	lboy505@aucklanduni.ac.nz
Luke Mazur	lmazur@uow.edu.au
Maheswaran Rohan	maheswaran.rohan@dpi.nsw.gov.au
Maryann Staincliffe	sarah.hornell@agresearch.co.nz
Michael Mumford	Michael.Mumford@daf.qld.gov.au
Michael Vacher	michael.vacher@csiro.au
Min Zhang	m.zhang1@uq.net.au
Mohammad Javad Davoudabadi	mohammad.davoudabadi@sydney.edu.au
Monique Jordan	mj633@uowmail.edu.au
Murray Hannah	gchrismuz@gmail.com
Owen Forbes	chantel.thoms@qut.edu.au
Patricia Menendez	patricia.menendezhb4@gmail.com
Pauline O'Shaughnessy	poshaugh@uow.edu.au
Pei Luo	pluo244@aucklanduni.ac.nz
Peter Green	peter.green@agresearch.co.nz
Quan Vu	quanv@uow.edu.au
Rajitha Athukorala	dare.centre@sydney.edu.au
Reinier de Vos	reinier.devos@health.nsw.gov.au
Rodrigo Canovas	rodrigo.canovas@csiro.au
Rosita Shishegar	rosita.shishegar@csiro.au
Ruth Butler	rcb962@gmail.com
Sepideh Sharif	sepi.sharif@postgrad.otago.ac.nz
Sevvandi Kandanaarachchi	sevvandi.kandanaarachchi@data61.csiro.au
Stelios Georgiou	stelios.georgiou@rmit.edu.au
Stella Stylianou	stella.stylianou@rmit.edu.au
Steven Harden	steven.harden@dpi.nsw.gov.au
Subhash Chandra	subhash.chandra@agriculture.vic.gov.au
Thomas Lumley	t.lumley@auckland.ac.nz
Timothy Cox	timothy.cox@csiro.au
Tomomi Yamada	tomomi.yamada@dmi.med.osaka-u.ac.jp
Tong Chen	tche929@aucklanduni.ac.nz
Tsuyoshi Nakamura	naka@nagasaki-u.ac.jp
Vanessa Cave	vanessa@vsni.co.uk
Vihanga Gunadasa	vihanga.gunadasa@sydney.edu.au
Willem Vervoort	willem.vervoort@sydney.edu.au
Yan Wang	yan.wang@rmit.edu.au
Yiyi Ma	yiyi.ma1@unsw.edu.au
Zahra HosseiniFard	zahra.h@unimelb.edu.au
Zhanglong Cao	Zhanglong.Cao@curtin.edu.au

Presenter Index

Outline Time Table

Sunday		Page	Monday		Page	Tuesday	Page
			8:30-9:00 Registration		3		
					3	8:30-9:00 Welcome to AASC 2022: Murray Hannah, Roger Payne (by Video)	
						9:00-9:45 Carolyn Huston	15
						9:45-10:15 TEA	
						MultiPhase experiments.	
						10:15-10:30 Clayton Forknall	16
						10:30-10:45 Lu Wang	17
						10:45-11:00 Chris Brien	18
						11:00-11:05 Short Break	
						On the Farm.	
						11:05-11:20 Bethany Rognoni	20
						11:20-11:35 Jordan Brown	21
						11:35-11:50 Andrew van Burgel	22
						11:50-12:00 Questions/Discussion	
			9-4:30 Workshops			12:00-1:00 LUNCH	
				1:00 Bus from airport	3	**1:00-1:45 Francis Hui**	23
						Tuesday Quick Fire 10 min talks.	
				1:45 Bus from City		1:45-1:55 Emi Tanaka	24
						1:55-2:05 Min Zhang	24
						2:05-2:15 Aidan McGarty	25
						2:15-2:20 Short Break	
						2:20-2:30 Alan Herschtal	25
						2:30-2:40 Vihanga Gunadasa	26
						2:40-2:50 Dongwen Luo	27
						2:50-3:00 Questions/Discussion	
						3:00-3:30 TEA	
						3:30-5:15 DARE special session.	
						3:30-4:15 Ed Cripps	28
4:00 Bus from airport		3				4:15-4:30 Katherine L. Silversides	
						4:30-4:45 Dilani Kaveendri	29
						4:45-5:00 Yiyi Ma	30
			4-5:30 Registration	5:00 Inverloch	3	5:00-5:15 Questions/Discussion	31
4:45 Bus from City							
7:15 Inverloch	7-8:15 Registration	3				6:45 Canapes and Drinks	
						7-8:30 pm Poster-session	55

Wednesday		Thursday		Friday	
	Page		*Page*		*Page*
	4			8:30-8:45 Welcome/ Houskeepin	
		8:45-9:00: Welcome/ Housekeeping		**8:45-9:30 Emlyn Williams**	49
		9:00-9:45 Vanessa Cave, David Baird	32		
				9:30-10:00 TEA	
		9:45-10:15 TEA		Mixed Topics 1.	
		The Human Dimension		10:00-10:15 James Curran	50
		10:15-10:30 Luke Boyle		10:15-10:30 Pauline O'Shaughnessy	50
		10:30-10:45 Timothy Cox	34	10:30-10:45 Kerry Bell	51
		10:45-11:00 Joanne Thandrayen	35	10:45-10:50 Short Break	
		11:00-11:05 Short Break	35	Mixed Topics 2.	
		A little bit of Bayes.		10:50-11:05 Pei Luo	52
		11:05-11:20 Owen Forbes		11:05-11:20 Thomas Lumley	52
		11:20-11:35 R. Willem Vervoort	37	11:20-11:35 Murray Hannah	53
		11:35-11:50 Alaa Amri	38	11:35-11:45 Questions/Discussion	
		11:50-12:00 Questions/Discussion	39	**11:45-12:30 Arthur Gilmour**	54
		12:00-1:00 LUNCH		12:30-12:40 Conference Close	
		1:00-1:45 Patricia Menéndez	40	12:40-1:30 LUNCH	
				1: 30 Bus to Melbourne Airport & Southern C	3
		Thursday Quick Fire 10 min talks.			
		1:45-1:55 Rajitha Athukorala	41	Key:	
		1:55-2:05 Sewandi Kandanaarachchi	41	Bus	
		2:05-2:15 Innocenter Amima	42	Registration	
		2:15-2:20 Short Break		Admin etc	
		2:20-2:30 Peter Green	42	Workshops	
		2:30-2:40 Zhanglong Cao	43	Invited Session	
		2:40-2:50 Graham Hepworth	43	15 min talks	
		2:50-3:00 Questions/Discussion		10 min talks	
		3:00-3:30 TEA		Breaks/ Functions	
		GxE and Plant Breeding.		Posters	
		3:30-3:45 Michael Mumford	44		
		3:45-4:00 David Hughes	45		
		4:00-4:15 Monique Jordan	46		
		4:15-4:30 Bev Gogel	46		
		4:30-4:40 Questions/Discussion			
		4:45-5:15 AGUAi AGM			
			2		
		7 pm Conference Dinner			

Conference Trips