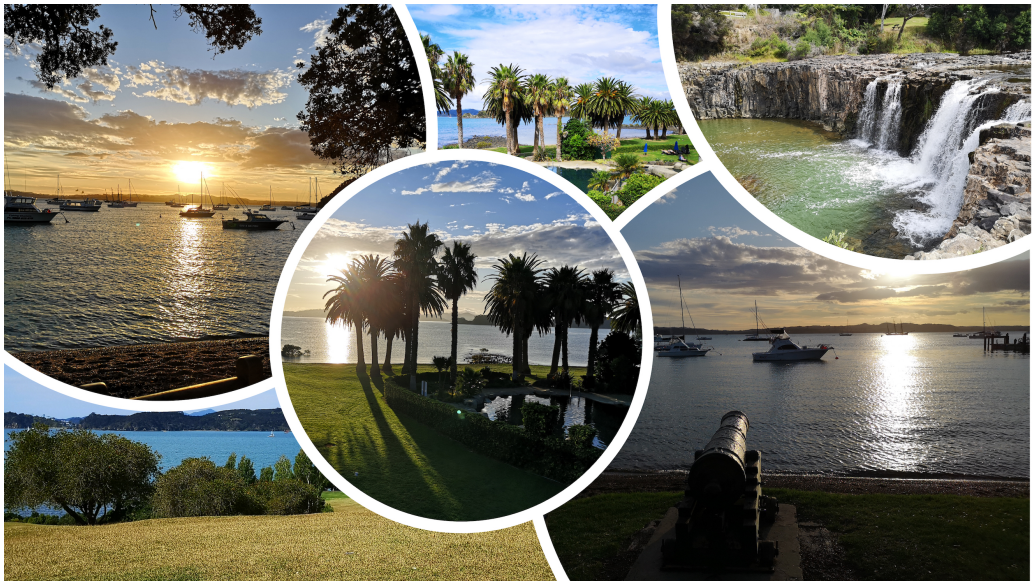
IBS-AR/SEEM Conference



Bay of Islands
Aotearoa
New Zealand
27 November–1 December 2023

Scan me for an online copy!

Programme

Tuesday and Wednesday:

Time	Tuesday 28/11		Time	Wednesday 29/11	
855	Welcome James Curran, President		855	Housekeeping	
900	Keynote 1: Hans-Pieter Piepho (Chair: Emlyn Williams)		900	Keynote 2: Lisa Warbrick (Chair: Esther Meenken)	
950	Morning Tea (30 minutes)		950	Morning Tea (30 minutes)	
	Ecology 1 (Chair: Jing Liu)	Biostatistics (Chair: John Pearson)		New Methods in Regression (Chair: Francis Hui)	Industry and Ethics (Chair: Lisa Thomasen)
1020	Shirley Pledger	Insha Ullah	1020	Alan Welsh	Alba Cervantes Loreto
1045	Godrick Oketch	Chuyao Xu	1045	Thomas Yee	Esther Meenken
1110	Ellen Cieraad	Alain Vandal	1110	James Curran	Garth Tarr
1135	Kathy Ruggiero	Alastair Lamont	1135	Samuel Muller	Peter Green
1200	Jing Liu	John Pearson	1200	Francis Hui	Lisa Thomasen
1225	Lunch (1 hour 10 minutes)		1225	Lunch (1hr 10 minutes)	
	'Omics (Chair: Lindy Guo)	Applied Statistics 1 (Chair: Andrew Balemi)	1300	IBS-AR AGM in Treaty Room 1 Bring your lunch	
1335	Min Zhang	Siwei Zhai	1335	Conference Excursions	
1400	Louise McMillan	Simon Urbanek			
1425	Lindy Guo	Andrew Balemi			
1450	Afternoon tea (30 minutes)				
	Bayesian (Chair: Beatrix Jones)	Environment (Chair: Alasdair Noble)			
1520	Lizbeth Naranjo Albarrán	Warren Müller			
1545	Renate Meyer	Dongwen Luo			
1610	Beatrix Jones	Alasdair Noble			
1700	Cultural experience				

Thursday and Friday:

Time	Thursday 30/11		Time	Friday 1/12	
855	Housekeeping		855	Housekeeping	
900	Keynote 3: David Warton (Chair: James Curran)		900	Keynote 5: Jo Potts (Chair: Vanessa Cave)	
950	Morning Tea (30 minutes)		950	Morning Tea (30 minutes)	
	Design of Experiments (Chair: Emi Tanaka)	Spatial (Chair: Leslie New)		Biometrics (Chair: David Baird)	Ecology 3 (Chair: David Chan)
1020	Emlyn Williams	Charlotte Jones-Todd	1020	Brian Cullis	Matthew Schofield
1045	Chris Brien	Xu Ning	1045	Chanatda Somchit	David Chan
1110	Brian Cullis	Sam Mason	1110	Vanessa Cave	Russell Millar
1135	Alison Smith	Monique Jordan	1135	Rina Hannaford	Gordana Popovic
1200	Emi Tanaka	Leslie New	1200	David Baird	Rachel Fewster
1225	Lunch (1 hour 10 minutes)		1225	Lunch and conference close (1 hour 10 minutes)	
1335	Keynote 4: Adrian Baddeley (Chair: Ben Stevenson)				
	Multi-environment trials (Chair: Alison Kelly)	Statistical Methods (Chair: Michael Stewart)			
1425	Aidan McGarty	Kate Lee			
1450	Lu Wang	Thomas Lumley			
1515	David Hughes	Rishika Chopara			
1540	Alison Kelly	Michael Stewart			
1605	Afternoon tea (15 minutes)				
	Ecology 2 (Chair: Alec van Helsdingen)	GLMM's (Chair: Matt Wand)			
1620	Taylor Hamlin	Quan Vu			
1645	Ben Stevenson	Priya Parmar			
1710	Alec van Helsdingen	Matt Wand			
1800	Conference Dinner Garden Restaurant 1800-2200				

Haere mai, welcome to the IBS-AR/SEEM conference!

Here are a few further details about some of the items in our programme.

Monday night

Esther Meenken has kindly organised, in conjunction with the local iwi, a pōwhiri (Māori welcome) to take part before the Welcome Barbecue. We will be welcomed by Ngati Kawa Taituha, Chairman of the Waitangi Marae. The welcome is formal, and as such, delegates are asked to observe a smart casual dress code. Those interested in attending should gather by the palm trees near the pool around 1720 for a 1730 start. The Welcome Barbecue will start after the conclusion of the welcome at approximately 1800.

Tuesday night

Matt Schofield has organised a 30-minute cultural event at the Waitangi Treaty Grounds Museum at 1700. The talk sessions finish at 1635, allowing for us to take the approximately 5–10 minute walk to the museum and be ready for the event to start on time at 1700. The event will be followed by canapés and drinks for about an hour in the Courtyard at the Copthorne.

Wednesday lunch

We have moved the IBS-AR AGM to coincide with the lunch period on Wednesday. IBS-AR members are asked to attend. Please get some lunch and then assemble in Treaty Room 1 by 1300. We do not anticipate a long AGM, and so 30–40 minutes should be sufficient.

Wednesday excursions

There are plenty of things to do in the Bay of Islands (for example taking a boat trip to see the Hole in the Rock, or swimming with the dolphins). However, these are best arranged on an individual basis.

Excursion options:

1. **Haruru Falls walk** (free): The start of this trail is just opposite the entrance to the Copthorne Hotel.

2. **Waitangi Treaty Grounds**: The Waitangi Treaty Grounds (`https://www.waitangi.org.nz/`) are a combination of two museums and a park. The Treaty Gounds are located right next-door to the Copthorne Hotel. Admission, which includes a guided tour and cultural experience, starts at NZ$30 for NZ residents and NZ$60 for international visitors.

3. **Waitangi Golf Club**: For the keen golfers, Waitangi Golf Club (`http://www.waitangigolf.co.nz/bookingsfees/`) is a short 1.5-km (20-minute) walk from the hotel. The club also has equipment for hire at very reasonable rates for those who just want to give it a go.

4. **Historic Russell**: Those interested in the short ferry ride across the harbour followed by an unorganised exploration of historic Russell should meet at in the hotel lobby at 1345 for a short walk to Paihia where ferry tickets can be purchased to travel to Russell (NZ$14). The ferry leaves every 15 minutes. Russell has a collection of small shops, bars, restaurants, as well as some historical attractions. Some of us may get no further than the Duke of Marlborough Hotel.

Dining options

We would also remind you that there is no evening meal planned for either Tuesday night or Wednesday night. The hotel has advised that guests are welcome to eat in the hotel restaurant, but they do need to book in advance. We also advise booking in advance if you are planning to eat in Paihia. There is a reasonable number of restaurants in Paihia, which is an easy 2-km walk from the hotel, but it is worth remembering that it is midweek in a small town, in a shoulder tourist season.

Enjoy the conference!

Keynote 1, Tuesday 0900

Chair: Emlyn Williams

An adjusted coefficient of determination (R^2) for generalized linear mixed models in one go

Hans-Pieter Piepho

University of Hohenheim

The coefficient of determination (R^2) is a common measure of goodness of fit for linear models. Various proposals have been made for extension of this measure to generalized linear and mixed models. When the model has random effects or correlated residual effects, the observed responses are correlated. This talk proposes a new coefficient of determination for this setting that accounts for any such correlation. A key advantage of the proposed method is that it only requires the fit of the model under consideration, with no need to also fit a null model. Also, the approach entails a bias correction in the estimator assessing the variance explained by fixed effects. I will use several examples to illustrate this new measure. A simulation will be presented showing that the proposed estimator of the new coefficient of determination has only minimal bias.

Keynote 2, Wednesday 0900

Chair: Esther Meenken

Māori, data and sovereignty: Weaving the threads

Lisa Warbrick

Te Pū Oranga Whenua

Lisa is Director for the Indigenous Genomics Institute (IGI) and Pou Arataki for Te Pū Oranga Whenua (TPOW), a national collective of Māori Agribusinesses with wāhine leadership. With over 25 years' experience in Māori economic development and Kaupapa Māori, whanau wellbeing is at the heart of Lisa's 25 years management, community and enterprise experience. Lisa's leadership roles with the IGI and TPOW have had her working at the interface between Mātauranga Māori and western eResearch data infrastructure, governance and ethics as New Zealand and the world grapples with what sovereignty of data looks like in practice. She will share her broad, practical experiences with understanding what

makes data tika (accurate, fair) not just from government, research and industry perspective, but from a grass-roots Māoridom point of view (te ao Māori).

Keynote 3, Thursday 0900

Chair: James Curran

Multivariate spatial models for community ecology using a basis function approach

David Warton*, Elliot Dovers, and Jakub Stoklosa

*University of New South Wales

Basis function expansions are a fundamental tool for doing spatial statistics on large datasets or complicated problems. Essentially, a basis function approximation to a random field lets us treat a spatial model like a generalised additive model. Here we will look at three problems from ecology that require the fitting of spatial models to point event or multivariate data and show how we can develop powerful new tools using a basis function approximation, extended (where needed) to the multivariate setting by combining with factor analytical techniques. We show how we can then: fit a log-Gaussian Cox process using standard GAM software; partition biodiversity along spatial gradients to identify regions of high turnover and its main drivers; fit multivariate point process models to account for observer bias and study co-occurrence from citizen science data.

Keynote 4, Thursday 1335

Chair: Ben Stevenson

ROC curves for spatial point patterns and presence-absence data

Adrian Baddeley*, Ege Rubak, Suman Rakshit, and Gopalan Nair

*Curtin University

Receiver Operating Characteristic (ROC) curves have recently been used to evaluate the performance of models for spatial presence-absence or presence-only data.

Applications include species distribution modelling in spatial ecology, and mineral prospectivity analysis in geoscience. We clarify the interpretation of the ROC curve in this context. Contrary to statements in the literature, ROC does not measure goodness-of-fit of a spatial model, and its interpretation as a measure of predictive ability is weak. To gain insight we draw connections between ROC and other statistical techniques for spatial data, including point process modelling of mapped spatial point patterns. The area under the ROC curve (AUC) is related to hypothesis tests of the null hypothesis that the explanatory variables have no effect. This suggests several new techniques, which extend the scope of application of ROC curves for spatial data, to support variable selection and model selection, analysis of segregation between different types of points, adjustment for a baseline, and analysis of spatial case-control data. The new techniques are illustrated with several real example datasets.

Keynote 5, Friday 0900

Chair: Vanessa Cave

Statistical consulting: Bridging the gap between numbers and advice

Joanne Potts

The Analytical Edge

To be candid, I fell into statistical consulting back in 2012 by virtue of personal circumstances rather than an inherent ambition to run a business, but here we are, turning 11 years young this December and still going strong! Over the past decade I have found statistical consulting to be a highly rewarding profession. I have had the privilege of collaborating with some remarkable colleagues on a variety of interesting projects. I will take the opportunity in this presentation to highlight the enjoyable aspects of my experience as a statistical consultant, teaching professional development workshops (which I love!), and going on the odd field trip to far flung places like Barrow Island to catch burrowing bettongs and participating in detector dog training in Kosciuszko National Park.

In the spirit of honesty and transparency, I will also share the challenges I've encountered as well. By doing so, I hope to provide support and guidance to others who may be facing similar obstacles. These challenges include scope-creep, contending with messy datasets (and clients who aren't exactly sure what they need), managing feelings of professional isolation, handling overwhelming workloads and addressing demanding and occasionally troublesome clients.

Session: Ecology 1, Tuesday 0950–1225

Chair: Jing Liu

Estimating frog growth curves when both age and sex are unknown

Shirley Pledger* and Ben Bell

*Victoria University of Wellington

The Hamilton's frog, a New Zealand native, is unusually long-lived. A 40-year capture-recapture study on Maud Island has intermittent length data for a few hundred frogs, but in most cases both age and sex are unknown. Two growth curves are needed as females ultimately grow larger than males. We use finite mixtures and the generalized EM algorithm to fit the two curves while allowing for the missing data.

A new framework for saddlepoint methods with an application to capture-recapture

Godrick Oketch

University of Auckland

The saddlepoint approximation transforms a random variable's moment generating function (MGF) into an approximation of the probability density function. Interpreting the saddlepoint approximation as a likelihood function offers an alternative to maximum likelihood estimation when the true likelihood function is computationally intractable. However, to maximise these saddlepoint-based likelihoods, one must access the MGF and its derivatives as a function of the relevant parameters. Our study introduces a framework that automates the computation of these MGFs, even for complicated/composite random variables. We provide user-friendly R functions to construct MGFs and utilise automatic differentiation tools for accurate gradient-based optimisation. These functions do not require any expertise in saddlepoint methods. Finally, to demonstrate the utility of our framework, we apply it in fitting certain capture-recapture models, whose true likelihood functions are unknown.

Networks in aquatic communities collapse upon neonicotinoid-induced stress

Ellen Cieraad*, Henrik Barmentlo, Maarten Schrama, Geert de Snoo, Martina Vijver, Kees Musters, and Peter van Bodegom

*Department of Conservation

The risk of insecticides, such as neonicotinoids, to freshwater ecosystems remains unclear, since translation from single species bioassays to actual community or ecosystem responses has proven notoriously difficult. We investigated the effects of the neonicotinoid thiacloprid on invertebrate species alpha and beta-diversity, species co-occurrence and ecosystem functioning using an outdoor mesocosm facility consisting of 36 individual ditches. Using a combination of structural equation models and network-based analysis of species co-occurrence, we found alterations in ecosystem functioning at increasing levels of insecticides, which appeared to have occurred through losses in invertebrate food web functionality (in the absence of changes in taxon richness).

Predicting albumen gland length of cereal crop pest snails

Kathy Ruggiero

University of Auckland

Terrestrial snails introduced from the western Mediterranean have become pests in southern Australia, causing crop damage and grain contamination. Effective molluscicidal baiting hinges on timing, typically just before adult snails initiate reproduction in mid- to late autumn. In this talk, I will discuss how Gaussian Mixture Models are valuable in classifying these pest snails into reproductive and non-reproductive states. I will also explore how this information can help estimate the albumen gland length of a standard-sized snail over time, and its potential in correlating environmental cues with reproductive state to advance baiting programs.

Closed-form likelihood functions for spatial capture-recapture

Jing Liu* and Ben Stevenson

*University of Auckland

Spatial capture-recapture models are commonly used to estimate animal population density. These models allow the probability that an animal is detected to vary amongst spatially separated detectors based on the location of its activity centre. However, activity centres are not observed on a spatial capture-recapture survey and are treated as random effects. Existing methods require sampling or numerical integration over activity centres to obtain the marginal likelihood function.

In this talk, we show that the marginal likelihood function for a popular spatial capture-recapture model is available in closed form and some consequences of this closed form.

Session: Biostatistics, Tuesday 0950–1225

Chair: John Pearson

Impact of eplet mismatches on the risk of donor-specific antibodies and antibody-mediated rejection in simultaneous pancreas-kidney transplant recipients

Insha Ullah

Australian National University

While increased human leukocyte antigen (HLA) eplet mismatches are known to elevate the risk of donor-specific antibodies (dnDSA) after solid organ transplantation, the specific eplets that predict dnDSA and acute rejection have been unidentified. We analyzed eplet mismatches at HLA class I and II loci in 202 recipients of a first simultaneous pancreas-kidney (SPK) transplant recipients using Cox regression models. In the cohort followed for a median of 3.9 years, 37% developed dnDSA. While no class I eplet mismatches correlated with class I dnDSA, specific class II mismatches at locations 55PP, 182N and 57S, 16Y, 74L doubled the risk of developing class II dnDSA (adjusted HR=2.1). For those who developed dnDSA against a specific eplet mismatch, the risk of antibody-mediated rejection (AMR) tripled (HR=3.0). Our study identifies specific class II eplet mismatches that significantly increase the risk of dnDSA and subsequent AMR. Avoiding mismatches at these specific locations may lower the risk of adverse post-transplant outcomes.

Response adaptive randomization

Chuyao Xu*, Thomas Lumley, and Alain Vandal

*University of Auckland

Response adaptive randomization assigns future patients with higher probability to a more effective treatment and lower probability to an inferior treatment. However, when considering patients' benefit, we also care about patients in the population who can take the superior treatment. So we analyse the impact of response-adaptive randomisation on treatment allocation in the population (as well as the trial) and identify advantages and disadvantages of applying response adaptive randomization in real clinical trials compared to the group sequential design and equal randomisation.

Trials of a statistician: The ROBUST RCT

Alain Vandal*, Lata Jayaram, Conroy Wong, and the ROBUST research team

*University of Auckland

The ROBUST trial (Conroy Wong, PI) was an HRC-funded two-period randomised crossover trial of an inhaled steroid, tiotropium, in participants with bronchiectasis. I will talk about the timeline of the trial from conception to publication, with brief descriptions of activities involving the statistician related to study design, monitoring, data management and wrangling, analytical design, analysis, interpretation and manuscript writing. The title is a bad pun having to do with my involvement as trial statistician in 4 HRC-funded RCTs in the same year. These studies were not actually sent to try me.

Development of a population simulator to optimise study designs and estimate polygenic disease risk/resilience in Aotearoa New Zealand Māori populations

Alastair Lamont*, Phillip Wilcox, and Mik Black

*University of Otago

Disease risk/resilience (DR) prediction requires statistical models that are typically generated from empirical studies. For commonly occurring polygenically inherited conditions such as gout, type 2 diabetes, and cardiovascular conditions, risk/resilience estimates have most often been derived from GWAS (genome-wide association studies). Such studies require large sample sizes ($n > 10^4$ participants) genotyped with 10^4–10^7 DNA markers.

However, such datasets often do not include indigenous peoples, who can have important genetic differences from more commonly represented populations of predominantly European descent. Moreover, existing datasets from Māori (and Pasifika) domiciled in New Zealand are few, and those that could be utilised, consist of fewer than two thousand individuals—typically from case-control studies (e.g. Flynn, Phipps-Green et al. (2013))—thus are underpowered for clinically accurate DR prediction. In addition, establishing sufficiently large GWAS is highly unlikely in Aotearoa/NZ because of substantive costs associated with generating genotypic data and reluctance of many Māori to participate in such studies.

In order to offset further health inequities arising from lack of Māori-specific DR prediction models, new studies are required. Such studies require both (a) optimal designs that incorporate known genetic relationships on non-genotyped as well as genotyped individuals, and (b) analytical methods that more accurately predict phenotype than GWAS-based methods such as polygenic risk score (PRS).

We have used a population simulator (SLiM) to model genetic structures of Māori communities (i.e., whānau/hapū/iwi), incorporating estimates of effective population sizes prior to European admixture, as well as post-colonisation admixture with Europeans. We are using these simulations to explore what features of study design and analytical methods lead to optimal DR prediction. I will illustrate and present current results on this.

The incidence of early-onset colorectal cancer in Aotearoa New Zealand: 2000 to 2020

John Pearson*, Olliver Waddell, Andrew McCombie, Harriet Marshall, Rachel Purcell, Jacqui Keenan, Tamara Glyn, and Frank Frizelle

*University of Otago, Christchurch

Colorectal cancer is the second most common cancer in Aotearoa New Zealand resulting in over 1,200 deaths each year. While age standardised incidence rates in the total population have been steadily decreasing against a background of static incidence rates and an aging population, incidence rates for early onset colorectal cancer (EOCRC) in those under 50 years old have been increasing in Aotearoa New Zealand, as has been found elsewhere. The New Zealand Cancer Registry data from 2000-2020 was examined by Poisson regression to quantify trends in incidence and age standardised incidence by age group, sex and cancer location for total and Māori ethnicity and projected to 2040 using population projections from Statistics New Zealand. A critical aspect of this research is the production of meaningful statistics in the New Zealand context that are sympathetic to Māori world views, Māori health advancement and indigenous research methodology against a background of global efforts that minimise such considerations.

Session: 'Omics, Tuesday 1335–1450

Chair: Lindy Guo

Using cellular composition estimated from bulk RNA-seq data to suggest cellular level markers for melanoma survival

Min Zhang*, Kaye Basford, Vivi Arief, Geoff McLachlan, and Quan Nguyen

*University of Queensland

Past studies have shown cancer tissues can have very different cellular compositions (i.e., the abundance of different cell types in the tissues), and there are studies suggesting the potential role of cellular composition in cancer patient progression. However, these studies usually based on small sample sizes and/or only focused on one or two cell types. Here, we used the combination of a few statistical methods, including support-vector-regression-based tool "Cell-type Identification By Estimating Relative Subsets Of RNA Transcripts" (CIBERSORT) and survival analysis, to extract cellular composition information of eight cell types from The Cancer Genome Atlas (TCGA) melanoma bulk transcriptomic (RNA-seq) data ($n = 457$), and to suggest cellular level markers for melanoma patient survival.

Clustering of mixed-type data with nuclear and mitochondrial genetic data as a case study

Louise McMillan

Victoria University of Wellington

Clustering data of mixed type (a mixture of continuous, count, binary, ordinal, nominal) has proved to be a surprisingly difficult challenge, yet such a method would be useful in a huge range of fields. Existing methods often treat numbered categorical levels as numerical data with equally spaced levels, or model categorical variables as if they have underlying Gaussian distributions; in many cases, neither of these assumptions hold.

Population genetics offers a different but related problem: that of clustering data that are all categorical, but where some variables (nuclear DNA) have two or more responses per variable (diploid or polyploid data) and others (mitochondrial DNA) have one response per variable. Clustering these types of data together is a different form of mixed-data clustering. I will discuss my latest work on this topic, intended to enable biologists to combine their different genetic data types into a single analysis, and show how the solution also provides a possible avenue of research into the more general challenge of clustering mixed-type data.

I will also discuss surrogate residuals, which are akin to randomised-quantile residuals, and how they offer an alternative solution to mixed-type clustering.

Multi-omics integration pipeline: A test framework to analyze multiple high-dimensional biological data

Lindy Guo*, Olivia Angelin-Bonnet, Blake List, Leonardo Salgado, Nigel Joyce, Susan Thomson, and Roy Storey

*Plant and Food Research

In recent decades, the explosion of new technologies that measure cell contents or states across different omics layers, combined with a reduction in processing costs, has enabled a shift towards a holistic approach to the study of biological systems. This multi-omics integration pipeline is a test framework designed to integrate data from multiple sources, such as genomics, transcriptomics, and metabolomics, aiming to systematically evaluate and compare existing tools for multi-omics integration, to inform future integration analyses, and to promote best practices. In this talk, we will discuss our ongoing effort and user cases in constructing this test framework.

Session: Applied Statistics 1, Tuesday 1335–1450

Chair: Andrew Balemi

Respiratory health of Pacific youth: Nutrition resilience and risk in childhood

Siwei Zhai*, Alain Vandal, Shabnam Jalili-Moghaddam, Catherine Byrnes, Conroy Wong, Leon Iusitini, and El-Shadan Tautolo

*University of Auckland

In New Zealand, 7% of deaths are related to respiratory diseases and Pacific people are at higher risk. This work investigated causal effects of early-life nutritional factors on early-adulthood lung function amongst Pacific Islands Families Study cohort members, who consist of the 1398 individuals born from Pacific Island families in Middlemore Hospital between March and December 2000. 466 from the cohort participated in the respiratory study. Primary outcome was forced expiratory volume in 1 second (FEV1) z-score at age 18 years. FEV1 and healthy lung function (HLF), defined as the z-score being larger than -1.64, were secondary outcomes. Nutrition and other information were previously collected in 4 measurement waves at ages 4, 6, 9 and 14 years. Food portions consumed daily were totaled within each of 12 food categories at each measurement wave. Exploratory and multi-group confirmatory factor analyses identified 4 eating patterns represented by nutritional factor scores (NFS), identified as "Occasional", "Seafood", "Fruit and vegetables", and "Meat". NFS were scaled to portions per day. Confounders were identified using a causal directed acyclical graph. Semi-parametric linear and relative risk regression models were fitted to estimate causal effects of NFS on respiratory outcomes, using estimated weights compensating for attrition-induced selection bias. The population attributable fractions of HLF of each NFS were estimated for each measurement wave. HLF cohort prevalence was estimated at 90% (95% confidence interval [CI] [0.86,1.00]), smaller than the expected 95%. Only the "Fruit and vegetables" eating pattern at 9 years was found to have a statistically significant causal effect on the FEV1 z-score in early adulthood (change in FEV1 z-score: +0.25, 95%CI [0.00,0.43]). The proportion of HLF attributable to "Fruit and vegetables" eating pattern at 9 years was estimated at 11% (95%CI [0.00,0.19]). Results suggest a positive impact of consuming more fruit and vegetables during childhood on respiratory health later in life. There is a need to support healthier food environments for Pacific children and access to healthier food choices.

Estimating human mobility: Computing commute graph across Aotearoa

Simon Urbanek* and Kathlyn Ycong

*University of Auckland

In this talk we will show a way to construct a spatial network representing commutes to/from work, illustrated on the Aotearoa New Zealand 2018 Census recording the usual residence and work location of New Zealanders. Based on the spatial information we can estimate the commute flows between and through regions. The resulting spatial network can be combined with other available data to answer various questions such as how dangerous are commutes for people living in certain areas. The main focus is to construct a spatial network that can be used independently of the generation process to link existing data such as demographics or tagged locations with the human mobility aspect captured by the transition network.

Teaching GLMs to undergraduates and graduates: Challenges and successes

Andrew Balemi

University of Auckland

GLMs have been around for about 40+ years now. In this talk I will discuss the approach we have undertaken at the University of Auckland to impart these ideas to undergraduate and post-graduate students. I will discuss our pedagogy and discuss what we have learnt in this process—the challenges we have encountered and the successes. Clearly, this is a work in progress. This may be useful for any other teachers of this wonderful regression technique.

Session: Bayesian, Tuesday 1520–1635

Chair: Beatrix Jones

A Bayesian latent class linear mixed model for monotonic processes subject to measurement error

Lizbeth Naranjo Albarrán*, Osvaldo Espin-Garcia, and Ruth Fuentes-García

*Universidad Nacional Autónoma de México

Motivated by a longitudinal study on radiographic diagnosis of osteoarthritis, for which a biometric index of interest has been possibly measured with error, we propose a Bayesian approach to identify latent classes in a model with continuous response naturally subject to a monotonic constraint, i.e. non-decreasing or non-increasing process. A latent class linear mixed model has been defined to consider measurement error where the monotonic process under study is restricted by using truncated normal distributions. The main purpose is to classify the trajectories of the response variable through the latent classes, i.e. to get homogeneous sub-populations from a heterogeneous population, which describe in a better way the disease progression.

Spectral analysis of multivariate time series

Renate Meyer*, Yixuan Liu, Kate Lee, and Claudia Kirch

*University of Auckland

The analysis of multivariate time series can give important insights into periodicities and coherencies. We present a novel approach to Bayesian nonparametric spectral analysis of stationary multivariate time series which is based on Whittle's likelihood. Starting with a parametric vector-autoregressive model, the parametric likelihood is nonparametrically adjusted in the frequency domain to account for potential deviations from parametric assumptions. The nonparametric prior used is a multivariate extension of the nonparametric Bernstein-Dirichlet process prior for univariate spectral densities to the space of Hermitian positive definite spectral density matrices. We demonstrate that the nonparametrically corrected likelihood combines the efficiencies of a parametric with the robustness of a nonparametric model. We demonstrate the practical benefits through a spectral examination of a medical time series containing cardiorespiratory measurements, as well as two environmental datasets: one comprising a bivariate time series of the Southern Oscillation Index and fish recruitment, and another consisting of time series data for windspeed at six different locations in California.

Using Bayesian networks for hypothesis generation

Beatrix Jones* and Innocenter Amima

*University of Auckland

We consider a case study from the Vineyard Ecosystem programme. A Bayesian Network encompassing 131 variables including weather, soil attributes, vineyard management and vineyard performance has been inferred from observational data collected as part of this programme. We outline how we discussed this large, complex network with vineyard scientists. Because of the observational nature of the data, and the presence of substantial uncertainty about the network, we viewed this as a hypothesis generation exercise. We describe our structured conversations, points where there were misunderstandings or the vineyard scientists needed more information, and instances where interesting hypotheses were generated.

Session: Environment, Tuesday 1520–1635

Chair: Alasdair Noble

Limitations of using the van Genuchten model to fit soil water retention curves

Warren Müller*, Richard Greene, Vilim Filipović, and James Noble

*CSIRO

The soil water retention curve (SWRC) represents a relationship between soil water content and soil water (matric) potential. Determining this relationship is crucial to understanding processes such as soil water storage, water flow and solute transport, and plant water uptake. In order to fit SWRC, soil scientists have routinely used the van Genuchten model

$$\theta(h) = \theta_r + (\theta_s - \theta_r)[1 + (\alpha h)^N]^{\frac{1\breve{}N}{N}},$$

where $\theta(h)$ is water content, h is soil water potential, and $\theta_r, \theta_s, \alpha$, and N are fitted parameters determining the shape of the curve.

Although widely accepted and used, there are a few concerning issues with using this model both due to its parametrization and its inappropriateness for particular soil types. We demonstrate the limitations in estimating the parameters when applying this model using datasets derived from 42 Australian soil samples with a limited soil water potential range (pF(=$\log_{10} h$) 0 to 4.2). Further datasets, obtained from published studies, which covered a wider soil water potential range (pF 0 to 7) were examined to compare parameter estimates with those obtained by restricting the same datasets to the pF range 0 to 4.2. Our results indicate the need to have measurements over broader soil water potential ranges to properly fit SWRC data. This could result in more precise determination of soil hydraulic parameters and consequently more accurate descriptions of soil water dynamics in the soil.

Deriving global reference concentrations of water quality contaminants

Dongwen Luo

AgResearch Ltd

Water stands as a fundamental necessity for all human life. However, mounting consumption, climate change, and pollution are exerting significant pressures on water supply systems. In response to this challenge, companies and urban areas urgently require a well-defined strategy to reduce their ecological footprint

on water resources. Reference conditions offer a vital benchmark, enabling us to assess the extent of river disturbances and identify opportunities for enhancement. This approach empowers companies and municipalities to safeguard their growth prospects while safeguarding water resources from depletion and pollution, ultimately steering us toward a water- secure future.

During the presentation, we will elucidate the methodology for establishing global reference concentrations utilizing available global databases, incorporating techniques such as random forest modeling, data collection, feature selection, hyperparameter tuning, model validation, and model deployment.

Estimating power of detection of mitigation of contaminants in waterways

Alasdair Noble*, Richard McDowell, Oliver Ausseil, David Hamilton, and Mike Kittridge

*AgResearch Ltd

Estimating the power of detecting a change in the level of contaminants under various sampling strategies in a river poses many challenges and assumptions. We discuss a programme of work to estimate the standard deviation of the distribution of the residuals of sampled data for detecting change over time using currently sampled sites. We extend our analysis to predict the residuals in unsampled river reaches. Hence an estimate of the power, given a level of change, the sampling frequency, and the time-period of sampling, can be estimated. Models are required for the data for each river reach over time, for the standard deviations using meta data for the river reaches and for the future scenarios. Each of these will be discussed and various options presented. Results for a range of indicators for New Zealand river reaches of stream order 3 and above as examples to assess the usefulness of the tool will be presented.

Session: Methods in Regression, Wednesday 1020–1225

Chair: Francis Hui

Insights into small area estimation using the nested error regression model

Alan Welsh* and Ziyang Lyu

*Australian National University

Estimating characteristics of domains (referred to as small areas) within a population from sample surveys of the population is an important problem in survey statistics. We consider using model-based small area prediction intervals under the nested error regression model. We present model-based simulations that show the performance of our asymptotic prediction intervals in quite small, finite samples. We also carry out a design-based simulation using data on consumer expenditure on fresh milk products to explore the design-based properties of the mixed model-based inferences. We explain and interpret some surprising simulation results through analysis of the population and further design-based simulations. The simulations highlight important differences between the model- and design-based properties of mixed model prediction intervals in small area estimation.

Generally-altered, -inflated, -truncated and deflated regression

Thomas Yee

University of Auckland

Models such as the zero-inflated and zero-altered Poisson and zero-truncated binomial are now well-established, especially in biometrics. Along with deflation and modification, we review important ideas behind the five operators and subsequent incremental extensions, and propose a super mixture model that jointly and maximally unifies alteration, inflation, truncation and deflation for counts, given a 1- or 2-parameter parent or base distribution. Seven disjoint sets of special value types are accommodated because all but truncation have parametric and nonparametric variants. Some highlights include: (i) the mixture distribution is exceeding flexible, e.g., up to seven modes; (ii) under-, equi- and over-dispersion can be handled using a negative binomial (NB) parent, with underdispersion handled by a novel Generally-Truncated-Expansion method; (iii) under- and over-dispersion is studied holistically in terms of the operators; (iv) while GA regression explains why observations are there, GI regression accounts for why they are there in excess, and GD regression explains why observations are not there. We illustrate GAITD

regression with some biometric examples using the VGAM R package. The important application of heaped and seeped data from retrospective self-reported surveys is briefly mentioned. The GAITD-NB has potential to become a Swiss army knife for count responses.

What the zeta?

James Curran

University of Auckland

Many people are familiar with forensic evidence such as DNA, fingerprints, fibres and so on. Many fewer people are familiar with forensic glass evidence. Glass evidence arises when glass is broken during the commission of a crime. The statistical interpretation of glass evidence preceded the statistical interpretation of DNA evidence by almost a decade. In this talk I will discuss an estimation problem that arises when considering activity level propositions for glass evidence—i.e. propositions that consider how the glass might have been deposited on a person of interest as well as the physical characteristics of the glass that link it to the crime scene. This talk will involve an obscure discrete distribution, as well as some extensions to said distribution. Some might liken it to Morris Dancing.

CR-Lasso: Robust cellwise regularized sparse regression with `regcell`

Samuel Muller*, Peng Su, Garth Tarr, and Suojin Wang

*Macquarie University

Robust variable selection currently has some focus on dealing with cellwise contamination in the design matrix where only some but not all elements of an observation vector are contaminated. The problem is particularly challenging when the number of variables is large. Traditional robust methods can fail when the data is high-dimensional and too many observation rows experience some cellwise contamination. We explore how using initial robust empirical covariance matrix estimators together with regularization approaches, helps in robustly selecting variables by simultaneously shrinking regression coefficients and identifying outlying cells in the data matrix. Specifically, we highlight the performance of CR-Lasso, a new approach which incorporates a constraint on the deviation of each cell in the loss function to detect outliers based on regression residuals and cell deviations, by combining L1 and cellwise outlier regularization.

Homogeneity pursuit and variable selection in regression models for multivariate abundance data

Francis Hui*, Luca Maestrini, and Alan Welsh

*Australian National University

We propose a generalized estimating equation (GEE) approach for simultaneous homogeneity pursuit (i.e., grouping coefficients so that they share the same values in their unknown clusters) and variable selection in regression models for multivariate abundance data in ecology. Using GEEs allows us to straightforwardly account for between-species correlations, while we augment the GEE with both adaptive fused lasso and adaptive lasso-type penalties to cluster the species-specific coefficients within each covariate and encourage differing levels of sparsity across the covariates, respectively. We apply the proposed method to presence-absence records collected along the Great Barrier Reef in Australia, revealing both a substantial degree of homogeneity and sparsity in species-environmental relationships, while the estimated model produces stronger out-of-sample predictive performance compared to methods that do not accommodate such features.

Session: Industry and Ethics, Wednesday 1020–1225

Chair: Lisa Thomasen

Response propensity and nonresponse bias for the 2022 Agricultural Production Census

Alba Cervantes Loreto

Stats NZ

This investigation aimed to explore whether the 2022 Agricultural Production Census (APC) in New Zealand was affected by low response rates and a boycott organized by farmers in protest against government policies. It used a Bayesian multilevel models to predict the likelihood of farms responding to the census based on past responses and other farm characteristics. The research focused on three key variables; the total number of dairy cows, total number of beef cows, and the total number of sheep. This investigation showed that the boycott did not introduce systematic nonresponse bias since the patterns of response propensity remained similar to those observed in 2017, when the boycott was not in effect. In other words, the variables that influenced farms to respond in 2017 were still the most important factors in 2022, despite the lower overall response rate. This indicates that the boycott did not systematically influence certain types of farms to respond more or less than they would have otherwise, which would have led to biased estimates. Additionally, the results of the donor imputation and response propensity weighting methods were consistent, indicating that either method could be used without introducing significant bias.

Toward cross-CRI Māori data governance principles

Esther Meenken

AgResearch Ltd

CRIs hold considerable amounts of data. Much of this data is about Māori, about the environments Māori have relationships with or about species considered taonga. Sovereignty of these taonga is guaranteed by Te Tiriti, but it is only recently that CRIs have started to consider the implications of what this means. Recently Māori thought-leaders have published aspirations and recommendations for the management and sovereignty of taonga data. In this talk I will give examples of initiatives by work across the CRIs to better manage data that we hold on behalf of kaitiaki and ask what would need to happen within CRIs for Māori data sovereignty to be achieved.

Statistical challenges for the red meat industry

Garth Tarr

University of Sydney

This talk will discuss some elements of the work I do with Meat and Livestock Australia. A key focus has been predicting eating quality in beef and sheep based on vast amounts of consumer data collected over decades of research trials. In recent years, there has been a desire to extract more insights from routinely collected data, improve the feedback given to producers and processors, and incorporate new technologies in an effort to make carcase grading more objective.

How does uncertainty influence potential to make decisions when integrating complex data sets?

Peter Green*, Esther Meenken, and Delphine Rapp

*AgResearch Ltd

Data has increasing influence on decisions, but the majority of state-of-the-art use-cases that utilise big and/or disparate data and ML in agriculture are still relatively modest when compared to the aspirations of the agri-tech sector.

When integrating data, an important consideration is sensitivity, where the relative importance of one factor on the outcome may be more than or less than that of another factor. The value of sensitivity analysis lies in understanding the relative importance of various pieces of data to the outcome, which in turn helps to understand where to focus attention in describing the types and sources of uncertainty.

We will present a case study applying sensitivity analysis to pilot data linking drought duration to food safety risks.

A taste of variability

Lisa Thomasen

Fonterra

Sensory data has a reputation for being variable. This can make it challenging to distinguish between products or explore relationships with functional properties. This is often compounded by slapdash data management practices and surprising stakeholder interactions. I will share a case study involving sensory data and how I teased out the story from amongst the noise.

Session: Design of Experiments, Thursday 1020–1225

Chair: Emi Tanaka

Spatial design and analysis of tree improvement trials

Emlyn Williams

Australian National University

Field trials are used extensively in many disciplines to develop and evaluate new germplasm. Tree improvement trials can differ from, say, cereal breeding trials in that they can be designed with single or multiple tree plots. This talk will discuss some spatial design and analysis options. Results from the spatial analysis of examples will be compared at both the plots and trees levels.

Piepho, H-P., Williams, E.R. and 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, 745-53.

Piepho, H-P., Boer, M.P. and Williams, E.R. (2022). Two-dimensional P-spline smoothing for spatial analysis of plant breeding trials. *Biometrical Journal*, 64, 835-57.

The anatomy of a two-phase experiment involving human subjects using `dae`

Chris Brien

University of Adelaide, University of South Australia

Brien (2020, `https://doi.org//10.1177/09622802211031612`) and Brien, Sermarini, and Demetrio (2023, `https://doi.org/10.1002/bimj.202200284`) describe using the anatomy of a design for understanding the confounding relationships between the terms in a linear mixed model. In spite of the links of an anatomy to anova, they argue that the anatomy of a design is relevant irrespective of whether the data analysis is to be carried out using anova or linear mixed model analyses. In this talk, obtaining and interpreting the anatomy corresponding to a linear mixed model will be illustrated for the two-phase pain-rating experiment described in Brien (2020). Also touched upon will be the inclusion of block-treatment, or the more general intertier, interactions, pseudoreplication, the comparison of an anatomy with a traditional, skeleton anova, and the production of anatomies both manually and with `dae` (Brien, 2023, `https://cran.at.r-project.org/package=dae/`).

A model-based design approach for the design of selection experiments using ODW

Brian Cullis*, Alison Smith, and David Butler

*University of Wollongong

The success of a plant improvement program is based on its ability to maintain high levels of genetic gain. From a purely statistical perspective, maintenance of genetic gain relies on the use of near optimal experiment designs and appropriate methods of analyses. The natural phenotyping instrument for selection in the advanced evaluation phase is a multi-environment trial (MET). The analysis of METs has received wide attention for many years, culminating in the use of a single-step factor analytic mixed model by many plant breeding programs in Australia and elsewhere. There is, however, a dearth of literature concerned with the design of single and multiple selection experiments. There is a long history concerned with the design of experiments for fixed treatment effects, but these designs are not appropriate for selection experiments. In this talk we present a model-based approach for the design of single and multiple selection experiments. Our approach is model-based and hence allows for the use of genetic relatedness for each stage in the design process. The construction of optimal or near-optimal designs is achieved using the R package, ODW. We illustrate our ideas using two MET datasets.

Novel statistical design that enables valid comparisons of canola varieties across herbicide technology groups

Alison Smith* and Brian Cullis

*University of Wollongong

Historically, the evaluation of commercial or near-to-release canola varieties has been conducted using field trials with separate sub-experiments for different herbicide technologies. This has precluded valid comparisons of varieties across technologies, despite the distinct need of such information by Australian growers and agronomists. In this presentation we discuss a new approach for experimental design that addresses this problem. It has been led by a private company, Pacific Seeds, and enhanced by statisticians at the University of Wollongong. We show the evolution of the collaborative scientific process and highlight key statistical issues of false replication and aliasing.

A tool to easily simulate valid experimental data

Emi Tanaka

Australian National University

Simulation is often employed in statistics to empirically quantify and understand the performance of various statistical methods. Simulating data is advisable as another diagnostic tool for experimental design to ensure the collected experimental data can be effectively analysed in accordance with their intended goals. Nevertheless, simulating data requires effort, particularly for novices, thus presenting a challenge in routine adoption. In this talk, I present a tool, implemented in the edibble R-package, that easily simulates experimental data that respects the experimental structure and other encodings. The automated simulation scheme can make use of information that the user already prespecified as part of their experimental design, thus requiring little extra effort for the user to simulate experimental data.

Session: Spatial, Thursday 1020–1225

Chair: Leslie New

`stelfi`: An R package for fitting Hawkes and log-Gaussian Cox point process models

Charlotte Jones-Todd

University of Auckland

Events cluster in time and space: bees swarm, whales click, spores disperse infecting trees. Temporal and spatial proximity are major factors in the chain reaction of events. Yet, the nature of how and why these patterns of events propagate is much more complex. In this talk, I will present the R package `stelfi`, which fits a range of spatiotemporal point process models that include self-exciting mechanisms.

A range of models for spatial, temporal, and spatiotemporal point pattern data are implemented in the package and are fitted using automatic differentiation via the R package `TMB`. This talk will cover the implementation of `stelfi` models and illustrate their use with real-world examples. To date, the development of spatiotemporal self-exciting point process models has been restricted to the fields of seismology and criminology, where a process is developed for a specific application. This talk will illustrate how such models can be used to model self-exciting mechanisms inherent in many environmental and ecological settings.

A double fixed rank kriging approach to spatial regression models with covariate measurement error

Xu Ning*, Francis Hui, and Alan Welsh

*Australian National University

In many applications of spatial regression modeling, the spatially-indexed covariates are measured with error, and it is known that ignoring this measurement error can lead to attenuation of the estimated regression coefficients. Classical measurement error techniques may not be appropriate in the spatial setting, due to the lack of validation data and the presence of (residual) spatial correlation among the responses. We propose a double fixed rank kriging (FRK) approach to obtain bias-corrected estimates of and inference on coefficients in spatial regression models, where the covariates are spatially indexed and subject to measurement error. Assuming they vary smoothly in space, the proposed method first fits an FRK model regressing the covariates against spatial basis functions to obtain predictions of the error-free covariates. These are then passed into a second FRK model,

where the response is regressed against the predicted covariates plus another set of spatial basis functions to account for spatial correlation. A simulation study and an application to presence-absence records of Carolina wren from the North American Breeding Bird Survey demonstrate that the proposed double FRK approach can be effective in adjusting for measurement error in spatially correlated data.

Spatio-temporal species distribution modelling

Sam Mason* and David Warton

*University of New South Wales

Detecting species response to climate change is a critical concern in ecology requiring relevant climatic predictors over long temporal windows and large spatial extents. However, it is currently typical to see species distributions modelled in a temporally static fashion, as a function of 30-year averaged climate datasets. In this talk we will move beyond this and use spatio-temporal data to look directly at the question of whether species responses are changing as the climate changes.

Long-term datasets of species records are now available in many areas, and recent advances in data acquisition technology make it relatively easy to obtain a wide range of environmental predictors at appropriate spatio-temporal scales and at consistent resolutions needed to build a spatio-temporal species distribution model. We will show how to obtain such data and use it to construct such a model.

By using dynamic predictors and anomalies from their mean values, together with a basis function approach to handle large datasets with spatio-temporal structure, we look at the question of whether we can detect species response to climate change and construct models for species distribution with better predictive performance than current practice.

Formal diagnostics for modelling spatial processes in field trials

Monique Jordan*, Alison Smith, and Brian Cullis

*University of Wollongong

Each year in Australia field trials are conducted to compare the yield performance of crop varieties across different environments. The trials typically comprise rectangular arrays of plots indexed by rows and columns. Current methods of analyses for individual trials follow those of Gilmour et al. (1997) where a separable (row $\times$ column) autoregressive process of order one (often denoted AR1 $\times$ AR1) is used as a baseline model for modelling smooth local trend and a graphic of the sample

variogram is used as a diagnostic for detecting non-stationarity such as in the form of extraneous variation. This diagnostic is informal and is open to interpretation leading to large disparities in the final models fitted by different practitioners. We investigate formal diagnostics for spatial modelling in field trials to provide a more vigorous framework for such analyses.

Species distribution models for eagle use in the continental United States

Leslie New*, Justine Fretz, Annaliese Chen, and Eugene Montforte

*Ursinus College

The United States Fish and Wildlife Services uses statistical models to estimate and predict bald and golden eagle fatalities at operating wind facilities. These models require data on eagle use of the landscape as an input, but this information is often missing, particularly for older facilities. One approach to filling this gap is the prediction of eagle use based on environmental covariates. In this study we used species distribution models, implemented through generalized additive mixed models, to predict the distribution of bald and golden eagle use throughout the continental U.S., along with the uncertainty in these predictions.

Session: Multi-Environment Trials, Thursday 1425–1605

Chair: Alison Kelly

Assessing disease resistance in chickpeas through the bivariate analysis of normal and binomial traits

Aidan McGarty*, Brian Cullis, Kristy Hobson, and Ahsan Asif

*University of Wollongong

Assessing disease resistance in chickpea is a complex problem which often necessitates the measurement of both normally and non-normally distributed traits. This issue has motivated the statistical methods discussed in this talk, namely the implementation of a bivariate generalised linear mixed model. This model was formulated to produce a valid bivariate analysis of two sets of phenotypic data. The first of which involved a glasshouse experiment where the recorded response was assumed to follow a binomial distribution and the second, a field experiment with response variable assumed to follow a normal distribution. The analysis required novel use of the software package `ASReml-R` which is able to model complex multivariate data, allowing estimation of genetic correlations between the two sets.

Variety selection using interaction classes derived from factor analytic linear mixed models in a single step multi-environment trial analysis with information on genetic relatedness

Lu Wang*, Chris Proud, Alison Smith, and Brian Cullis

*University of Wollongong

The key objective of plant breeding programs is to increase genetic gain by selecting superior individuals in the analysis of multi-environment trial (MET) data. The use of appropriate statistical methods has a key role in improving selection accuracy. The current recommended approach of analysis involves a fully efficient one stage factor analytic linear mixed model (FALMM) analysis that incorporates genetic relatedness through ancestral (pedigree) information or genomic (marker) data, as well as proper modelling of all sources of variation. The use of the factor analytic structure for modelling the genetic effects in the MET data allows for the investigation of variety by environment interaction (VEI), which represents the differential response of varieties to a change in environment (Smith et al., 2023). Selection requires meaningful and concise summaries of variety performance across

the environments in the MET. This becomes complex with the presence of VEI, in particular cross-over VEI which indicates changes in variety rankings. The motivating example considered in this talk was the MET analysis conducted for Rice Breeding Australia. We will demonstrate the use of interaction classes (iClasses) derived from FALMMs to aid variety selection with the presence of VEI in a single step MET analysis with the inclusion of pedigree information.

Factor analytic mixed models for multi-phase multi-environment trial data

David Hughes*, William Fairlie, Marijka Batterham, Alison Smith, and Brian Cullis

*University of Wollongong

The Hagberg-Perten falling number (FN) test is the industry standard to measure starch degradation cause by late maturity α-amylase enzyme activity in flour. The measurement of FN is a so-called multi-phase trial involving two phases, namely a field phase and a laboratory phase. In this talk, we extend the concepts presented in Smith et al. (2006) and Smith et al. (2015) to encompass multi-environment trial data. We present the analysis of a multi-phase multi-environment trial dataset in which the trait of interest is FN. This dataset spans 6 years from 2014-2019 and contains more than 230 environments and 124 genotypes. Using the Design Tableau approach of Smith and Cullis (2019), an appropriate mixed model is specified which accommodates the block structure for each phase as well as allowing for additional sources of variation and correlation. The extent of the genotype by environment interaction in the Australian wheat growing regions is explored using a factor analytic linear mixed model.

Smith, A. B., Lim, P., and Cullis, B. R. (2006). The design and analysis of multi-phase plant breeding experiments. *Journal of Agricultural Science*, 144(5), 393-409.

Smith, A. B., Butler, D. G., Cavanagh, C. R., and Cullis, B. R. (2015). Multi-phase variety trials using both composite and individual replicate samples: a model-based design approach. *Journal of Agricultural Science*, 153(6), 1017-1029.

Smith, A. B., and Cullis, B. R. (2019). Design Tableau: an aid to specifying the linear mixed model for a comparative experiment. Fisher Memorial Lecture.

Digging up the dirt on competition in multi-environment agricultural field trials

Alison Kelly*, Tolera Keno, Emma Mace, Ian Godwin, and David Jordan

*University of Queensland

Experimental techniques can be hampered by limited resources when conducting agricultural research, particularly when cost is a major constraint, as is often the case in developing countries. It is well known that spatial dependence exists between small plots in agricultural field trials, but it is less common to find interference due to the treatments applied to each small plot. The motivating example for our work is from a Sub-Saharan African maize breeding program growing maize hybrids in single-row plots in multi- environment field trials (METs). We extended the linear mixed model (LMM) for competition effects in a single trial to include random treatment terms for both direct and neighbour genetic effects across environments. The LMM also incorporated a residual covariance model estimating positive spatial correlation due to field trend, and negative correlation due to inter-plot interference, through a conditional auto-regressive structure across the spatial dimensions in each trial. We found that both direct and neighbour genetic effects differed across the trials confirming the presence of genotype by environment interaction for these genetic components. The results demonstrated that inter-plot competition biases grain yield predictions from single-row plots in hybrid maize METs and thus reduces the rate of genetic gain in the breeding program. A statistical model provides an efficient solution for maintaining selection accuracy when testing maize hybrids in single-row plots. This analysis approach ensures that the current resource-efficient field-testing approach can be maintained, requiring less land and seed for testing, in addition to further cost reductions for trial management and measurement of data.

Session: Statistical Methods, Thursday 1425–1605

Chair: Michael Stewart

Partially ordered data analysis

Kate Lee

University of Auckland

Partial orders are a natural model for partially ordered data in which it often seen in social hierarchies, genetics. In this presentation, we will present how to estimate the unknown partial orders in the Bayesian framework with a marginally consistent prior. I will also present some recent work in this area, a new class of models for time series data and computationally efficient method for many actors.

Optimal sampling design is sensitive to model misspecification

Thomas Lumley* and Tong Chen

*University of Auckland

For a linear regression model under two-phase sampling, optimal designs when the model is assumed correct involves sampling only from extreme values of predictor or outcome, but when the model is not assumed correct, the optimal design involves sampling the entire range of predictor and outcome. Previous work and the local asymptotic minimax theorem suggest that for some form of model misspecification the transition between these two optima must happen over a family of contiguous alternatives to the true model. We explore what this transition looks like.

Dealing with the badness of goodness-of-fit

Rishika Chopara*, Ben Stevenson, and Rachel Fewster

*University of Auckland

Goodness-of-fit (GOF) testing is vital to statistical analysis, as it allows us to validate the reliability of any statistical inference we make.

In many statistical models, the deviance is used to assess GOF by comparing it against a Chi-squared distribution. However, in some situations (e.g. when

dealing with sparse counts) the deviance does not have a Chi-squared distribution, even approximately, yielding such tests unusable. In principle, the true distribution for the deviance is computable, however in practice it is often intractable. We show that generally, we can accurately approximate the true underlying distribution of the deviance using a Gamma distribution. Using this approximation, we enhance the usability and power of GOF tests while retaining the familiarity and convenience of the deviance statistic.

Using a range of capture-recapture models for illustration, we show how our method can be used to accurately approximate the distribution of the deviance when dealing with various levels of data sparsity. With this approach, we aim to provide an accessible and effective GOF testing framework for complex modelling scenarios.

Robust estimation under small measurement errors

Michael Stewart* and Alan Welsh

*University of Sydney

Certain robust location and scale functionals cannot be estimated at the usual parametric rate under fixed measurement errors. We therefore allow the spread of the errors to decrease as the sample size increases. We report on some novel methods and some interesting phenomena including one which at first glance appears to be something of a free lunch.

Session: Ecology 2, Thursday 1620–1735

Chair: Alec van Helsdingen

Linking foraging movements with reproductive success in Adélie penguins

Taylor Hamlin*, Dean Anderson, Phil Seddon, and Matthew Schofield

*University of Otago

Linking movement decisions to external processes, particularly demographic outcomes, has been limited within the field of movement ecology. We have developed statistical models to link movement strategies made during foraging periods by Adélie penguins to the health and survival of their chicks. These techniques utilize flexible multi-state random walk models (state-space models) informed by satellite location data and time-depth recordings, as well as survival models informed by chick mass and daily observation data. Models were fitted in a Bayesian framework utilizing Stan and JAGS.

Estimating animal population density from passive acoustic surveys: Current methods and new challenges

Ben Stevenson*, Jing Liu, Melissa Bather, Sarah McGrath, and Lily Martin

*University of Auckland

Every year, acoustic recording devices become cheaper and cheaper. Every year, ecologists who are armed with such equipment head out into the field in greater numbers than ever before. The primary goal of many of these ecologists is straightforward: to estimate the population size of their study species.

Acoustic spatial capture-recapture is one particularly useful way to estimate animal population density from passive acoustic surveys, but deciding how you should analyse the data is complicated and depends on a variety of factors. Can you tell when the same sound is captured by multiple different devices, rather than the detections being of different sounds? Can you tell when you detect two different vocalisations by the same individual, rather than detections being of different individuals? Do individuals move during a single recording session? How do you know when you have recorded a vocalisation of your target species rather than a gust of wind, or a vocalisation of some other species? The list of considerations goes on.

In this talk, I will introduce existing spatial capture-recapture methods for acoustic surveys using two examples: the northern yellow-cheeked crested gibbon

Nomascus annamensis in Cambodia, and the great horned owl *Bubo virginianus* in Georgia, USA, highlighting the challenges faced by a practitioner on their journey from collecting acoustic recordings through to obtaining population density estimates. I will discuss ongoing work on the design of user-friendly software to alleviate these difficulties, and new challenges for statisticians to overcome as passive acoustic monitoring increases in popularity.

Extending spatial capture-recapture with the Hawkes process

Alec van Helsdingen

University of Auckland

Spatial capture-recapture (SCR) is a well-established method used to estimate animal population size from animal sighting or trapping data. Standard SCR methods assume animal movements are temporally independent and consequently cannot incorporate site fidelity (attachment to a particular region) and the temporal correlation of an animal's location. Recent work seeks to solve these issues by explicitly modelling animal movement. In this talk we propose an alternative solution based on a multivariate self-exciting Hawkes process. Here the rates of detection of a given animal at a given camera are a function of not only the location and its proximity to the animal's activity center, but also on where and when the animal has been previously detected. This allows us to account for both site fidelity and the inherent temporal correlation in detections that have not previously been accounted for in SCR-type models. In this talk, I will 1) give a conceptual overview of Self-Exciting Spatial Capture-Recapture (SESCR) models, 2) outline the challenges of fitting these models and potential solutions and 3) use case studies to compare traditional SCR and SESCR.

Session: GLMMs, Thursday 1620–1735

Chair: Matt Wand

Prediction and prediction uncertainty with generalized linear mixed models

Quan Vu

Australian National University

Generalized linear mixed model (GLMM) is a widely used tool in biology, ecology, and environmental sciences, because of its capability to model non-Gaussian clustered or correlated data. One important aspect when dealing with GLMMs is prediction of random effects and mean responses. In this talk, we investigate the approach using best predictor, which minimizes the mean squared prediction error, and discuss a few different methods to quantify the uncertainty associated with this point predictor. Particularly, we look into expanding the mean squared prediction error using Taylor series, and approximating the mean squared prediction error by bootstrapping. The methods are illustrated by an example with data simulated from a Poisson GLMM.

Results from the 10-year traumatic brain injury study

Priya Parmar*, Alice Theadom, and Patria Hume

*University of Auckland

Analysis from a New Zealand population-based Traumatic Brain Injury (TBI) incidence cohort study is shown here. The impact of TBI on executive functioning as a measure of cognitive decline in adults is assessed using data collected at 12-months and 10-years post event. A generalised linear mixed model assessing cognitive functioning over 10-years will account for age, gender, socioeconomic status, TBI severity type and recurrent TBI events. Comparisons of covariance-variance structures accounting for the repeated measures collected per individual will be shown and contrasted against the typically requested simplistic approach discussed under "forgoing statistical complexity for clinicians' comfort—where to draw the line".

The generalized linear mixed model leading terms

Matt Wand*, Aishwarya Bhaskaran, Jiming Jiang, and Luca Maestrini

*University of Technology Sydney

Generalized linear mixed models were born in the early 1990s as the love child of linear mixed models (1950s) and generalized linear models (1970s). Now, in the 2020s, every day ends with the publication of around 3 new papers on the topic. Despite their ever-increasing ubiquity, there has been very little in the way of asymptotic theory for the maximum likelihood estimators of generalized linear mixed model parameters. Apart from simple conveyance of estimator behaviour, there are the usual payoffs concerning statistical inference, sample size calculations and optimal design. This talk will describe new results concerning the generalized linear mixed model leading terms. Ramifications concerning variational approximation will also be mentioned.

Session: Biometrics, Friday 1020–1225

Chair: David Baird

`DWreml`: An R package for fitting the linear mixed model

Brian Cullis* and David Butler

*University of Wollongong

The average information (AI) algorithm for efficient residual maximum likelihood (REML) estimation of variance components has proven highly effective in the analysis of data arising from the biological sciences. However, as problem size and model complexity have increased, a key computing step, the assembly and solution of the mixed model equations (MME) using traditional tools, has become a significant time constraint. With falling memory costs, these methods have become compute-bound and are limited in their extent to exploit parallel processing opportunities.

The `DWreml` package implements the AI algorithm for REML estimation, and leverages advances in equation solver technology that take advantage of sparsity and parallelism. In particular, `DWreml` uses the MUMPS (MUltifrontal Massively Parallel Sparse direct solver) library to solve the MME. The `DWreml` package offers a familiar R user-interface based on ASReml-R and we illustrate its use in crop breeding applications, with an emphasis on highlighting elapesd times for key steps in fitting the linear mixed model. It is intended that `DWreml` become available in the public domain.

Generalised additive mixed models for large datasets: Modeling ruminal temperature sensor data from dairy cows

Chanatda Somchit

AgResearch Ltd

Generalised additive models (GAMs) are a nonparametric extension of generalised linear models (GLMs). Wood (2000) represented GAMs as penalised generalised linear models (GLMs), where each smooth term is represented using an appropriate set of basis functions and has an associated penalty measuring the wiggliness, where the smoothing parameters are given to each penalty in the penalised likelihood to control the wiggliness. More recently, Wood, Goude and Shaw (2015) developed a method for estimating GAMs for large data sets by using an iterative

update scheme on a model matrix factorisation to avoid formation of the whole model matrix in the GAM context.

The GAM approach for very large datasets is introduced here as a powerful exploratory tool in the analysis of on-farm sensor data. On-farm sensor technologies such as boluses and collars are now widely used on commercial farms in New Zealand. They are used to monitor cow health and to improve the productivity and efficiency of modern agriculture. However, the use of GAMs in on-farm sensor application is sparse. This talk begins by summarising GAM fitting methods for large data sets. It next represents generalised additive mixed models incorporating a tensor product of two or more numerical predictors for the analysis of ruminal temperature sensor data with respect to climate. The talk concludes with a discussion of the potential of the use of the more modern GAMs for large datasets in a modern agriculture application.

Analysis of Censored Cannabidiol Data

Vanessa Cave*, Roger Payne, and David Baird

*VSNi Ltd

Censoring occurs during data collection 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 detection/quantification. Lifetime data is often right-censored, in that not all experimental units will fail before the end of a study. In some circumstances, data may even be subjected to both left- and right- censoring, where the quantification method has both a lower and upper limit of Detection.

Censored data can be analysed by the Tobit method, using an E-M (expectation-maximization) algorithm to estimate values for the censored observations. In this talk, we demonstrate the Tobit method using an example from medical cannabis research. Here, an experiment was conducted to compare mean cannabidiol (CBD) concentration between different nutrient treatments and cultivars. The method used to quantify CBD resulted in a data set with both left- and right-censored values. Genstat was used to analyse the data using a Tobit linear mixed model.

Working with a different kind of AI

Rina Hannaford*, Juliana Yeung, Sofia Khanum, Peter Janssen, and Neil Wedlock

*AgResearch Ltd

We use a method that goes by the lovely name ELISA to quantify the production of antibodies, where the data is generally presented as a plot of optical density

versus logged dilution. (What is this density, what is being diluted? Come to the talk to find out!) But you don't need to understand the details to see that, typically, the data follows a sigmoidal curve.

In our project we have two such curves, one showing the immune response where the animal was treated versus the curve from its control sample.

Now, while ELISA is a well established and trusted procedure, there is less consensus at present on how to compare the response curves. A parameter that has proven popular is our different kind of AI, the so-called avidity index. We will discuss some of the commonly used methods for computing this index in the literature, and present what we believe to be an improved approach.

Spatial distribution of fertilizer spreading

David Baird* and Allister Holmes

*VSN NZ

A large-scale experiment was performed to collect the spatial distribution for a range of fertilizer types (pure and blended components) for different trucks and regions of New Zealand. This was analyzed with a generalized linear model and the results input into a simulation program to estimate the spatial distribution over a whole paddock. The physics of spreading fertilizer from a turning truck had to be estimated to allow for corners in the truck's track. The simulation can then be used to optimize the distance between tracks to give the optimal economy for the spreading operation.

Session: Ecology 3, Friday 1020–1225

Chair: David Chan

Estimating population size: The importance of model and estimator choice

Matthew Schofield*, Richard Barker, William Link, and Heloise Pavanato

*University of Otago

This work is motivated by a mark-recapture distance sampling analysis where we found unexpectedly large differences between Bayesian and frequentist estimates of abundance despite a moderately large number of observations (approx 600). Further exploration revealed similar sensitivity to estimator choice when focusing on frequentist estimation. To understand these differences, we consider abundance estimation from general mark-recapture models with three estimation strategies (maximum likelihood estimation, conditional maximum likelihood estimation, and Bayesian estimation) for both binomial and Poisson capture-recapture models. We find that assuming the data have a binomial or multinomial distribution introduces implicit and unnoticed assumptions that are not addressed when fitting with maximum likelihood estimation. This can have an important effect, particularly if our data arise from multiple populations. We compare our results to those of restricted maximum likelihood in linear mixed effects models.

Are acoustic spatial capture-recapture models also robust to misspecified detection functions?

David Chan* and Ben Stevenson

*University of Waikato

Animal density estimators from spatial capture-recapture (SCR) are commonly advertised as being remarkably robust to model misspecifications, particularly to the choice of detection function. The literature since its inception supports this with sampling designs involving many detectors distributed across the study area. However, the typical number of detectors used in an acoustic survey for density estimation is noticeably smaller than other sampling methods, such as camera traps and DNA sampling surveys. Currently, there is no known guidance on whether the density estimators from SCR remain robust to the choice of detection function for a typical acoustic survey design. Our work looks at this particular scenario via simulation studies, focusing on whether model selection methods, such as AIC, assist with selecting a detection function that results in more robust density estimates in practice.

Tweedie vs $\log(y+1)$ for analysis of zero-inflated non-negative continuous data

Russell Millar

University of Auckland

Measurements of rainfall amount, or catch weight of fish are two examples of zero-inflated non-negative continuous data. When a quick and dirty analysis of such data is sufficient then the transformed $\log(y+1)$ values are often modelled as normally distributed, and this is very common in the published literature. The Tweedie distribution provides a quick and convenient alternative. This talk will compare these two approaches, and provide recommendations for their use (or not).

Four principles for improved statistical ecology

Gordana Popovic*, Tanya Mason, Szymon Drobniak, Tiago Marques, Joanne Potts, Rocío Joo, Res Altwegg, Carolyn Burns, Michael McCarthy, Alison Johnston, Shinichi Nakagawa, Louise McMillan, Kadambari Devarajan, Alison Wunderlich, Magdalena Mair, Juan Andrés Martínez-Lanfranco, Malgorzata Lagisz, and Patrice Pottier

*University of New South Wales Sydney

Increasing attention has been drawn to the misuse of statistical methods over recent years, with particular concern about the prevalence of practices such as poor experimental design, cherry-picking and inadequate reporting. These failures are largely unintentional and no more common in ecology than in other scientific disciplines, with many of them easily remedied given the right guidance.

Originating from a discussion at the 2020 International Statistical Ecology Conference, we show how ecologists can build their research following four guiding principles for impactful statistical research practices: 1) Define a focused research question, then plan sampling and analysis to answer it; 2) Develop a model that accounts for the distribution and dependence of your data; 3) Emphasise effect sizes to replace statistical significance with ecological relevance; and 4) Report your methods and findings in sufficient detail so that your research is valid and reproducible.

These principles provide a framework for experimental design and reporting that guards against unsound practices. Starting with a well-defined research question allows researchers to create an efficient study to answer it, and guards against poor research practices that lead to poor estimation of the direction, magnitude, and uncertainty of ecological relationships, and to poor replicability. Correct and appropriate statistical models give sound conclusions. Good reporting practices and a focus on ecological relevance make results impactful and replicable.

Illustrated with two examples—an experiment to study the impact of disturbance on upland wetlands, and an observational study on Blue Tit colouring—this paper explains the rationale for the selection and use of effective statistical practices and provides practical guidance for ecologists seeking to improve their use of statistical methods.

Curiosity killed the rat—but saved the whale: How the saddlepoint approximation is saving NZ's biodiversity

Rachel Fewster*, Jesse Goodman, Godrick Oketch, and Louise McMillan

*University of Auckland

The saddlepoint approximation is a simple formula to convert a moment generating function into a probability density function or likelihood. It is ideal for scenarios where it is hard to compute the probability density, but easy to compute the moment generating function. This situation arises in various contexts, for example when the random variables we can observe arise as a linear transformation of random variables that can be more readily modelled. However, despite its advantages, the saddlepoint approximation is not widely known and has seen relatively little use. I will tell the story of an idea that spent many decades as little more than a curiosity, but is now beginning to gain some traction—not least in the realms of NZ conservation where it has been put to work in deleting unwanted rats and totting up more agreeable creatures such as whales.