NZSA Conference 2025

Schedule & Book of Abstracts





University of Otago
ŌTĀKOU WHAKAIHU WAKA

Monday 8th			
Time	Lecture Theatre - LT01	Seminar Room C - G.03	Seminar Room A - G.02
12:00pm	Registration		
1:00pm	Welcome and Housekeeping		
1:10pm	**Keynote - Kerrie Mengersen**		
2:00pm	Address from Minister of Statistics, Hon Dr Shane Reti		
	Session 111	**Session 112**	**Session 113**
2:20pm	**Christiana Akua Akyaa Agyei:** *Predicting Flowering Timing in Pinot Noir*	**Alec van Helsdingen:** *Modelling Under-Dispersed but Clustered Whale Cues*	**Siwei Zhai:** *Childhood Risk and Resilience Factors for Pasifika Youth Respiratory Health: Accounting for Attrition and Missingness*
2:40pm	**John Holmes:** *Variational Bayes in the GLM context: Can Bayesian(ish) inference be made efficient?*	**Joseph Reps:** *Spatial Capture-Recapture for Acoustic Data: Models and Implementation in the acre R Package*	**Ting Wang:** *Better forecasting of extreme geomagnetic storms using non-stationary extreme value models*
3:00pm	**Kate Lee:** *Bayesian inference for the learning rate in Generalized Bayesian inference*	**Tilman Davies:** *Chi-Squared Cox Processes: Is There Any Point?*	**Richard Arnold:** *Trust in science: sheep farmers, Flat Earthers and the rest of us*
3:20pm	Afternoon Tea		
	Session 121		**Session 122**
3:40pm	**Ruiting Mao:** *Rapid Gravitational Wave Parameter Estimation on LISA Signals with Gaps using Flow Matching*	**Women in Data Science**: *Data Science in Climate Finance – A Reflection* presented by Quyen Nguyen, followed by a panel discussion on the support needed for women in data science across both academia and industry.	**Laia Egea Cortés:** *Evaluating the Impact of Measurement Cost on Predicting Salmon Health with a Novel Ordinal Regression Model*
4:00pm	**Tarin Eccleston:** *Rapid Bayesian Inference for Galactic Supernova Gravitational Waves*		**Sarah Croft:** *Estimating abundance in small populations using pedigree reconstruction*
4:20pm	**Yuan Li:** *The Time Dilation of High Redshift Quasars: Insights from Bayesian Inference*		**Osamu Komori:** *Cumulant-based approximation for fast and efficient prediction for species distribution*
4:40pm	**Matt Edwards:** *Detecting the next galactic supernova using variational autoencoders*		**Ben Stevenson:** *Measurement error models for morphometric data*
5:30pm	Welcome Drinks at the Staff Club		

Tuesday 9th

Time	Lecture Theatre - LT01	Seminar Room C - G.03	Seminar Room A - G.02
9:00am	Housekeeping		
9:10am	**Keynote - Darryl MacKenzie**		
	Session 211	**Biostatistics Session**	**Session 212**
10:00am	**Unnati Nigam:** *Statistical Modeling of Repetitive Dynamics via Quasi-Periodic Gaussian Processes*	10:00am - **Peter Herbison:** *Reflections on 50 years in biostatistics*	**Miaotian Li:** *Fractional Cepstral Techniques for Real-World Time Series Clustering*
10:20am	**Janet Stacey:** *Automating the Evaluation of Forensic Evidence*	10:20am - **Priya Parmar:** *Real-world experiential biostatistics – examples, reflections and extensions to practical teaching applications*	**Hannah Yun:** *Condition Monitoring of Wind Turbine Bearings Using Dictionary Learning in the Frequency Domain*
10:40am	**Zhijian Wen:** *Detecting body fluids on fibres using hyperspectral imaging*	10:45am - **Seong Hoon Yoon:** *Weight Calibration in the Joint Modelling of Medical Cost and Mortality*	**Charlotte Jones-Todd:** *Modelling crime event data as a time-evolving bipartite network*
11:00pm	Morning Tea		Morning Tea
	Session 221	11:00am - Morning Tea	**Session 222**
11:30am	**Hao Xie:** *A Statistical Framework for Partial Multi-Label Feature Selection with Graph Regularization*	11:30am - **Alain Vandal:** *The ROBUST story told by the trial statistician*	**Qingyu Meng:** *A Reproducible and unified Benchmark of Deep Learning Feature Selection Across Simulations and Multi-Omics Cohorts*
11:50am	**Kaisi Chen:** *Phase Retrieval and Reconstruction in X-ray Computed Tomography: A Review*	11:45am - **Alice Kim:** *Estimating enteric hospitalisation rates using the National Collections*	**Xiaoyan Sun:** *Multi-head Attention Feature Selection for High-Dimensional Biomedical Data via Deep Fusion of Filter and Embedded Methods*
12:10pm	**Qian Chang:** *Understanding topological heterogeneity within social networks for cross-domain disinformation detection*	12:00pm - **Nisa Widyastuti:** *Older adults' driving self-regulation patterns over time* 12:15pm - Panel discussion with **Claire Cameron, Andrew Gray, Yannan Jiang, Keith Muller, Robin Turner**	**James Bristow:** *Joint hyperparameter optimisation and neural architecture search for simulation-based inference: Applications in phylodynamics*
12:30pm	**Simon Urbanek:** *Applying LLM Embeddings to R Journal Articles*		**James Curran:** *Data fusion in forensic paint analysis*
1:00pm	Lunch - Poster Session		
		Session 231	
2:00pm		**Rasika Dilhani:** *Entropy–Complexity Based Clustering of ARMA Simulated Time Series Models*	
2:20pm		**Guoping Hu:** *Joint Forecasting and Reconciliation for Hierarchical Time Series Using GNNs*	
2:40pm		**Nasim Sadra:** *Statistical Analysis of Landslide-Induced Vegetation Disturbance Following Cyclone Gabrielle in Hawke's Bay, New Zealand*	
3:00pm		**Ciprian Doru Giurcaneanu:** *Fractional Integral Transforms*	
	Census Panel with Len Cook, Gary Dunnet, Thomas Lumley		**Pipeline Development Workshop** : development and model deployment using the tidymodels framework in R. Note: Contrary to the indicated room and end-time, this is a three hour session taking place in the **room Math 124 in the Science III building.**
3:20pm	Afternoon Tea		

		Session 241	**Session 242**
3:40pm	**AI Panel**: perspectives on the opportunities and challenges statisticians face in relation to all things AI. Hosted by Lisa Thomasen includes Gary Dunnet, Priya Parmar, Alana Docherty and Andrew Gray.	**Linus Fromm:** *Using Fibre Extensions To Build Efficient Fibre Samplers*	**Xinyue Wan:** *Making Predictions for Model-based Clustering*
4:00pm		**Bo Peng:** *Weighted Pairwise Likelihood for GLMMs in Complex Survey Data*	**Tori Diamond:** *Advancing Official Health Statistics: Bayesian approaches for making Māori health data relevant to local decision-making*
4:20pm		**Martin Hazelton:** *Statistics Research in Aotearoa New Zealand: An International Comparison*	**Tieta Vesty:** *Iwi affiliation in linked admin data and historical censuses*
4:40pm	NZSA AGM		
6:00pm	Conference Dinner at Otago Museum		

Wednesday 10th

Time	Lecture Theatre - LT01	Seminar Room C - G.03	Seminar Room A - G.02
9:00am	Housekeeping		
9:10am	**Keynote - Thomas Lumley**		
10:00am	Aitken Documentary		
	Session 311	**Session 312**	**Session 313**
10:10am	**Andrew Balemi:** *Deviancy meets Deviance.*	**Tom Elliott:** *Widgets for Rserve and Typescript*	**Alexandra Ferguson:** *Using longitudinal surveys to discuss labour force stability*
10:30am	**Anna Fergusson:** *Branching out data science education: Developing task and computational environment design principles for teaching data science at the high school level through an international research collaboration*	**Thomas Yee:** *Variance Factorized Loglinear Models for Overdispersed and Underdispersed Counts*	**Alicia Crabtree:** *Evaluating alternative methods to derive ethnicity estimates from information available in administrative data*
10:50am	**Lisa Chen:** *AI without data sense is like fish without bicycle*	**Alain Vandal:** *Extending diagnostic validity meta-analysis to several diagnostic guidelines*	**Takafumi Kubota:** *Social Isolation among Japanese Adults Based on Microdata from the Basic Survey on Social Life*
11:10am	Morning Tea		
	Session 321		**Session 322**
11:40am	**Andrew Martin:** *Bayesian linkage error estimation for linked administrative data.*		**Priya Parmar:** *Real-time, real-world longitudinal analyses . . . all in one semester*
12:00pm	**Lafaele Mapusua:** *Data Formation: A Renaissance - Building an Inclusive Data System, One Portfolio at a Time*	**Essential Skills Workshop**: interactive discussion about how we can build and maintain essential skills throughout a career in statistics. Organised by Lisa Thomasen.	**Liza Bolton:** *Talk data to me? Supporting English Language Learners in Statistics and Data Science*
12:20pm	**Jamas Enright, Chen Chen:** *Linking People to Places of Work: Methodological Advances in the Integrated Statistical Data System (ISDS)*		**James O'Malley:** *Causal Inference for Intervention Spillover in a Cluster-randomized Stepped Wedge Trial: Application to Physician Networks*
12:40pm	**Gary Dunnet:** *Modernising the NZ Population Census*		**Vanessa Cave:** *Analysis of Censored Cannabidiol Data*
1:00pm	Lunch		

Keynote Abstracts

Kerrie Mengersen, Queensland University of Technology

1:10pm Monday 8th. *Decoding the SDGs: A Bayesian Perspective.*

The United Nations (UN) Sustainable Development Goals are a set of 17 interconnected global goals adopted by the United Nations in 2015 to create a more peaceful and prosperous world by 2030. They are a "universal call to action" to end poverty, protect the planet, and ensure that all people have peace and prosperity.

A critical activity common to each of the SDGs is trustworthy acquisition, analysis and interpretation of data – for rigorous monitoring, evidence-based decision-making, transparent communication and demonstrable impact. Statisticians, therefore, play a key role in achieving the SDGs.

In this presentation, I will describe some of the small efforts of our team to employ Bayesian modelling and analysis to advance the SDGs. Five case studies will be presented in the context of SDG 3 (Good health and wellbeing), SDG 10 (Reduced inequalities), SDG 6 (Clean water and sanitation), SDG 8 (Decent work and economic growth), SDG 16 (Peace, justice and strong institutions), SDG 13 (Climate action) and SDG 14 (Life below water). The case studies will spotlight our work on developing the Australian Cancer Atlas, high-dimensional anomaly detection, AI to combat modern slavery, predicting terrorism, and protecting the Great Barrier Reef. For each study, I will also touch on challenges in modelling and computation, and our efforts to address to address these issues.

This work has been undertaken by a large number of researchers in our Centre for Data Science at QUT. I will acknowledge key contributors in the presentation.

Darryl MacKenzie, Proteus

9:10am Tuesday 9th. *A statistical path less travelled.*

I was first exposed to environmental and ecological applications of statistics during the final year of my undergraduate degree, through the work of the late Prof Bryan Manly, who also introduced me to statistical consulting. A career as a consultant ecological statistician intrigued me from that point on, and I thought that once I had gained some experience working in an academic or government organisation, it might be something I could realistically contemplate. My opportunity for doing so came much sooner than I expected.

In this talk I shall share some of the experiences, challenges and lucky breaks I have had in my career to date as a consulting statistician; from being a student at the University of Otago, helping to develop a suite of modelling approaches that have become widely used in ecology and wildlife applications, involvement in a range of interesting consulting projects, to building a consulting business.

Thomas Lumley, University of Auckland

9:10am Wednesday 10th. *Doing more with fewer.*

In health research it is common to have a large existing database and to want to measure new variables. Sometimes these come from newly-invented biological assays; sometimes from expert audits of administrative data variables. Measuring everything on everyone is expensive, so we want to use the data we have to choose good subsets of people for the new measurements. I will talk about the efficient design and analysis of these two-phase medical studies, and about software, collaboration, and reading the literature.

Special Sessions

Women in Data Science New Zealand

Chair: Ivy Liu, Victoria University of Wellington.
3:40pm Monday 8th.

The session features a talk titled "Data Science in Climate Finance – A Reflection" presented by Quyen Nguyen (Earth Sciences New Zealand), followed by a panel discussion on the support needed for women in data science across both academia and industry. In Quyen's talk, she will share insights from her research and professional experience at the intersection of data science, climate change, and finance. It will provide an overview of how data-driven approaches are being applied to measure, model, and manage climate-related financial risks. She will discuss recent advances in machine learning for corporate carbon footprint prediction, evaluating climate risk disclosure, and assessing climate risks in the real estate and financial market. The talk will conclude with reflections on the challenges and opportunities for integrating data science into financial decision-making and climate policy.

Note: The mission of Women in Data Science New Zealand is to increase participation of women in data science and to feature outstanding women doing outstanding work. All genders are invited to attend.

Biostatistics Session

10:00am Tuesday 9th.

Statisticians are increasingly called upon to lead and collaborate in the development of innovative methods for health research. This year, we are excited to introduce a dedicated Biostatistics stream, featuring statisticians who work across a range of organisational settings. The sessions will highlight strategies for building a sustainable and fulfilling career in Biostatistics, and will celebrate the contributions of statisticians in medical and health research. The stream is designed to engage statisticians at all career stages – from early career professionals to established experts.

Peter Herbison, University of Otago

10:00am. *Reflections on 50 years in biostatistics.*

In this talk I will first describe the situation when I started work in August 1975. I will mention the computing resources available and what work involved. Then I will talk about the changes in statistics and computing over my working life. I will cover some of the projects I have been involved in over the years finishing with a study on Network Meta-analysis that we published in June.

Priya Parmar, University of Auckland

10:20am. *Real-world experiential biostatistics – examples, reflections and extensions to practical teaching applications.*

I showcase several examples of research project partnering from (1) traditional 15-point post-graduate longitudinal data analysis teaching courses with Growing Up in NZ cohort; (2) 30-point Master of Data Science dissertations with global TATA industries; (3) summer internships for stage 2 applied statistics students; (4) cross-faculty Master of Biostatistics degree (60 or 90-point thesis) research projects collaborating with clinical students, co-supervised by clinicians and biostatisticians; (5) globally networking the experientially learnt analyst as mentee then mentor and contributing to better understanding of experiences through Professor Jackie Carter's qualitative research with resource generation for future iterations.

Designed to develop the real-world applied analyst in real-time. We link to our "client" via our collaborative analytic teaching experiential model with an embedded "soft skills communication" framework.

Seong Hoon Yoon, University of Waikato

10:45am. *Weight Calibration in the Joint Modelling of Medical Cost and Mortality.*

Joint modelling of longitudinal and time-to-event data is a method that recognises the dependency between the two data types, and combines the two outcomes into a single model, which leads to more efficient estimates. These models are applicable when individuals are followed over a period of time, generally to monitor the progression of a disease or a medical condition, and also when longitudinal covariates related to the time-to-event variable are available. However, some longitudinal datasets (e.g. medical cost) are usually obtained using a complex sampling design rather than simple random sampling, which needs to be recognised in the statistical analysis. To address this, we combine survey calibration with standard joint modelling. The proposed method is applied to the survival and cost data on patients with diagnosed dementia in New Zealand.

Alain Vandal, University of Auckland

11:30am. *The ROBUST story told by the trial statistician.*

The ROBUST trial (Conroy Wong, PI) was an HRC-funded two-period randomised crossover trial of tiotropium vs placebo in participants with bronchiectasis. The main hypotheses were that anti-inflammatory properties of tiotropium, well-evidenced in people with COPD, would port to a bronchiectatic population by reducing pulmonary exacerbation frequency. I will talk about the timeline of the trial from conception to publication, with brief descriptions of statistical activities related to study design, grant application, data management, data monitoring, analytical design, analysis, interpretation, and manuscript writing.

Alice Hyun Min Kim, University of Otago

11:45am. *Estimating enteric hospitalisation rates using the National Collections.*

We conducted a population-based study of hospitalisation rates for Acute gastroenteric infection (AGI) from 2010 to 2019. Age-standardisation helps to compare rates for populations with different age structures. We discuss how using weights from different standard populations and disease age-distributions impact on the estimated rates and rate ratios. We outline the filtering process for the sequence of discharge events. To address co-infection, the total counting method was used for estimating the hospitalisation rates for specific disease codes. Age-standardised rates showed dose-response disparity by neighbourhood deprivation. The age-standardised rates linked to key specific pathogens such as Campylobacter, Salmonella and Rotavirus showed distinct annual trends over 2010-2019.

Nisa Widyastuti, University of Otago

12:00pm. *Older adults' driving self-regulation patterns over time.*

Aim: This study aims to 1) quantify the prevalence of specific self-regulation behaviours among older drivers at 4 time points; and 2) describe longitudinal patterns (trajectories) of driving self-regulation between 2016 and 2023. Method: Self-reported data from four waves of NZPATHS longitudinal cohort study of drivers (65y+) was analysed. The prevalence of each self-regulation behaviour was treated as repeated cross-sectional data. Change in behaviours across time was estimated using group-based trajectory models (GBTM). Results: A total of 3280 observations from 934 participants were included in the analysis. The most commonly reported avoidance behaviours were avoiding driving in bad weather at night, avoiding driving at night, and avoiding driving in bad weather (in general). Each of these was reported by over 30% of participants in most waves, though the prevalence of avoiding driving in bad weather (in general) was lower in 2023 than in preceding waves. The behaviour of leaving greater following distances remained relatively stable, ranging between 21–24% across survey waves. GBTM showed three distinct driver groups, Group1 (23.6% participants) had a zero self-regulation score on average; Group2 (49.3%) had at least one driving avoidance behaviour on average over the period, increasing over time; Group3 (27.1%) had higher total driving avoidance than Group2 over time. Conclusions: Self-regulation may assist drivers to maintain independent mobility and improve safety by reducing exposure to situations that can increase crash risk. These findings show the development of self-regulation over time in a cohort of ageing drivers. The next step is to ascertain factors associated with the different patterns.

Panel Discussion

Claire Cameron, Andrew Gray, Yannan Jiang, Keith Muller, Robin Turner.
12:15pm.

Biostatistics Session 3. *Statisticians thriving in health research: building successful careers through collaboration, resilience and adaptability.*

Statisticians work across a wide range of different organisational structures and disciplines. This panel explores strategies for building a sustainable and fulfilling career in biostatistics. The panellists bring decades of experience across the public and research sectors, in both academic and administrative leadership roles. Drawing on real-world experiences, they will discuss how statisticians can thrive by networking across disciplines, seeking and offering mentorship at all career stages, and adapting to the unique cultures of academia, government, and clinical environments. The panellists will share their insight into how to remain relevant and impactful in a field that is both technically demanding and highly collaborative. Whether early-career or established, participants will leave with practical ideas for fostering long-term growth, visibility, and career satisfaction in health research.

Pipeline Development with R

James Bristow, Bioeconomy Science Institute.
2:00pm Tuesday 9th.

This workshop provides an overview of pipeline development and model deployment using the tidymodels framework in R. We incorporate tidymodels-compatible packages such as stacks, DALEX, and workflowsets. We include a light introduction to MLOps principles for model deployment using the vetiver and plumber R packages. We demonstrate containerisation of models using Docker for deployment onto Kubernetes.

This is a three hour session taking place in the room Math 124 in the Science III building.

The Significance of AI for Statisticians – Panel Discussion

3:40pm Tuesday 9th.

The use of AI tools and LLMs has become ubiquitous in many workplaces. Irrespective of our opinions and personal biases, these tools are having a significant impact on statisticians and their collaborators. This panel discussion will be hosted by the Committee for Engagement with Statistics in Industry and will feature different perspectives on the opportunities and challenges statisticians face in relation to all things AI.

Lisa Thomasen will host the discussion and will be joined by Gary Dunnet, Priya Parmar, Alana Docherty and Andrew Gray as panellists.

Essential Skills for Statisticians: Interactive Workshop

Lisa Thomasen, Fonterra.
11:40am Wednesday 10th.

Statisticians leave university with an arsenal of technical knowledge only to find that some of the essential skills for being a successful statistician are ones we're left to learn on the job. The purpose of this interactive workshop is to bring attention to some of these essential skills with a focus on how we can build and maintain them throughout a career in statistics.

The essential skills discussed in this session may include:

- Tips for successful statistical consulting
- Communicating statistics to non-statisticians
- Managing stakeholder relationships
- Developing leadership skills and the ability to influence

Poster Abtracts

Tarin Eccleston, University of Auckland

Rapid Bayesian Inference for Galactic Supernova Gravitational Waves.

Gravitational waves—ripples in spacetime—let us listen to how matter behaves in one of the Universe's most extreme environments: the collapsing cores of dying stars. Yet waves from galactic core-collapse supernovae remain undetected, and we need to be ready for the next one.

We are developing a Bayesian inference pipeline based on neural posterior estimation, using variational autoencoders and normalising flows. Because this approach relies on amortised inference, once the model is trained it can produce posterior estimates almost instantly, avoiding the need to run computationally expensive procedures like MCMC at detection time.

Charlotte Jones-Todd, University of Auckland

Principles in practice: Activities for developing students' ethical data dispositions.

Statistics educators are uniquely positioned to support the development of students' ethical data dispositions and practices. While professional frameworks central to ethical statistical practice could provide direction for educators, there remains a need for practical guidance in designing learning activities and assessments. Supporting educators with achievable and accessible activities enables them to support students in engaging with ethical considerations throughout any data-related learning. We present classroom activities that draw on established principles, directly connect to employment-readiness, and are designed to develop students' ethical dispositions to achieve learning outcomes at different levels. We aim to encourage educators, in statistics and beyond, to embed data ethics across their teaching by peppering activities grounded in established principles across their course.

Alastair Lamont, University of Otago

Population Simulation to Optimise Study Designs and Estimate Polygenic Disease Risk/Resilience in Aotearoa Māori Populations.

For commonly occurring polygenically inherited conditions, disease risk/resilience estimates have most often been derived from GWAS (genome-wide association studies), which require large genotyped datasets. These datasets often do not include indigenous peoples, who can have important genetic differences from more commonly represented populations of predominantly European descent. Existing datasets from Māori domiciled in New Zealand are few and are underpowered for prediction due to their small size. In addition, establishing sufficiently large GWAS is unlikely in Aotearoa/NZ because of 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 a 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 scores. We have used a population simulator to model genetic structures of Māori communities, incorporating estimates historical population sizes, as well as post-colonisation admixture with Europeans. Our research uses these simulations to explore what features of study design and analytical methods lead to optimal disease risk/resilience prediction.

Yvonne Li, University of Auckland

Choosing pathways through hospital care for older persons.

Hospitals aim to allocate limited bed resources fairly while ensuring timely care for urgent patients. Our work was motivated by challenges at Waikato Hospital in Aotearoa New Zealand, which provides integrated care for urgent patients, primarily older adults, in its Older Persons and Rehabilitation (OPR) unit. Non-urgent patients may also be admitted to the OPR when beds are available; however, if the unit reaches capacity, urgent patients will be admitted to general wards, where service times are longer.

We have developed a compartmental model for patient admissions to OPR and general wards. Our model considers heterogeneous arrival and service rates across three different patient priority classes: urgent, non-urgent, and elective, and includes readmissions. We use a simulation model to test various thresholds for non-urgent admissions to OPR, to evaluate their impact on bed occupancy and the probability of urgent patients being diverted to general wards. Simulation results can assist in setting safe admission thresholds for non-urgent patients while ensuring that urgent care is not compromised.

Maryann Staincliffe, AgResearch

Adaptive Smoothing of Irregular Pasture Growth Data Using Flexi within R.

Pasture growth datasets often exhibit irregular temporal sampling, with dense observations during peak growing seasons and sparse data in cooler months. This uneven structure, compounded by multi-year and multi-seasonal variation, presents challenges for estimating daily growth rates and making inferences during unsampled periods. To address these complexities, I have applied Flexi, an adaptive smoothing algorithm developed by Martin Upsdell, which is particularly effective for handling non-uniform time series data. Through ongoing collaboration with the developer, I have explored Flexi's capacity to model biologically plausible growth curves while accommodating the assumptions necessary for interpolation. To enhance accessibility and reproducibility, I have developed an R package that integrates Flexi into the R environment, allowing for streamlined analysis and visualization. This presentation will discuss the statistical underpinnings of Flexi, demonstrate its application to pasture growth data, and highlight its utility in agricultural research settings where traditional smoothing methods may fall short.

Speaker Abstracts

Christiana Akua Akyaa Agyei, University of Canterbury

Session 111. *Predicting Flowering Timing in Pinot Noir.*

Flowering timing in grapevines influences yield and fruit quality, critical to New Zealand's wine industry. Predicting flowering enables growers to manage vines for optimal outcomes. Timing differences arise from vine structure, management practices, and environmental conditions. This study models flowering dynamics in Pinot Noir using Bayesian hierarchical logistic regression, quantifying the effects of cane type, shoot position, and node insertion point. Data were collected from a Waipara vineyard. The flowering proportion (open flowers / total) was modelled with fixed effects for day, cane, shoot, and node position, and random effects for block and vine. Models were run in WinBUGS using MCMC. Population-level predictions captured the sigmoidal flowering pattern, allowing estimation of the day when 50% of flowers were open. Lower nodes reached 50% flowering earlier, showing a linear trend across node positions. Shoot and node positions strongly influence flowering timing, providing insight for vineyard management..

Richard Arnold, Victoria University of Wellington

Session 113. *Trust in science: sheep farmers, Flat Earthers and the rest of us.*

When we as statisticians and scientists communicate our findings to the public, we're asking them to trust us, and for them to have confidence in the way that we work. When the subject doesn't have consequences for the way people live, such trust is often easily given. However if we're talking about climate change, and its consequences for policy and regulation, then that trust can readily evaporate. But distrust is driven by more than that: the bewildering growth of interest in the Flat Earth seems utterly disconnected from policy.

In this presentation I'll discuss a project investigating trust in science in New Zealand – I'll talk about two local case studies: acceptance of climate science among high country sheep farmers, and the nature of Flat Earth belief in New Zealand. I'll also present some early results from a population survey about levels of acceptance of settled science.

Andrew Balemi, University of Auckland

Session 311. *Deviancy meets Deviance.*

This is a talk for people who wish to teach GLMs to undergraduates. It will use an interesting example (NZ's notorious Peter Lumley-Walker trial) to illustrate fundamental ideas in undergraduate classes from hypothesis testing, proof of guilt, linear models, logistic regression, multicollinearity, and discriminant analysis... all from one interesting data set.

Liza Bolton, University of Auckland

Session 322. *Talk data to me? Supporting English Language Learners in Statistics and Data Science.*

Statistics and data science courses are evolving to meet modern academic, workforce and societal needs by incorporating ethics, communication, and authentic assessments. For many courses, this means an increased need for skill development in academic reading, speaking, listening and writing. At English-medium universities with large English Language Learner (ELL) cohorts, this is also often fundamental for these students' success and well-being. Additionally, all students can benefit from improved support for these skills. In this talk, I will explore how support for ELL students can be integrated at various scales — from small modifications to teaching practices to large changes in assessment and design. This is illustrated with examples from my and my collaborators' teaching across junior and senior undergraduate statistics and data science courses in Canada, Aotearoa New Zealand and China.

James Bristow, Massey University

Session 222. *Joint hyperparameter optimisation and neural architecture search for simulation-based inference: Applications in phylodynamics.*

The field of phylodynamics is crucial for understanding the evolutionary drivers and processes of epidemics. Traditional Bayesian phylodynamical models may produce biased and unstable parameter estimates when fitted to large and uncertain phylogenetic trees. Recent advances in simulation-based inference via deep generative models have demonstrated promise in inferring more stable and less biased parameter estimates. However, the incorporation of deep learning within phylodynamics introduces further complexity, requiring both model hyperparameters and architecture to be carefully calibrated. Here, we demonstrate the application of Bayesian optimisation to jointly tune the hyperparameters and neural architecture of a deep generative model. We perform a showcase study on a phylogenetic tree reconstructed from 200 HIV-1 sequences using a Birth-Death Skyline Serial (BDSS) model, noting that our approach can scale to larger and more complex phylodynamic models and phylogenetic data.

Vanessa Cave, University of Auckland

Session 322. *Analysis of Censored Cannabidiol Data.*

Censoring arises when measurements cannot be taken above or below a bound. For example, chemical concentrations may be left-censored if they fall below a detection limit, while lifetime data are right-censored when some units survive beyond the study period. In some situations, both left- and right-censoring occur when the quantification method has both lower and upper detection limits.

Censored data can be analysed using the Tobit method, which employs an E-M algorithm to estimate values for the censored observations. In this talk, I'll demonstrate the Tobit method using an example from medical cannabis research. Here, an experiment was conducted to compare mean cannabidiol (CBD) concentrations across nutrient treatments and cultivars, but the quantification method produced both left- and right-censored values. Genstat was used to analyse the data using a Tobit linear mixed model.

Qian Chang, University of Auckland

Session 221. *Understanding topological heterogeneity within social networks for cross-domain disinformation detection.*

Misinformation on social media spreads across diverse domains, where linguistic styles and propagation behaviors vary widely. Detecting misinformation in this cross-domain setting remains challenging due to three factors: semantic-structural heterogeneity, representation shifts across domains, and data scarcity in emerging domains. Existing methods largely ignore structural variability, enforce overly rigid domain-invariant spaces, and fail to exploit complementary information from other domains. In this work, we explore a graph-based perspective for cross-domain misinformation detection. This talk will represent online discussions as structured graphs and introduce a framework that leverages prompting and adaptive expert modeling to capture both transferable patterns and domain-specific signals.

Kaisi Chen, University of Canterbury

Session 221. *Phase Retrieval and Reconstruction in X-ray Computed Tomography: A Review.*

This paper presents a comprehensive review of X-ray computed tomography (CT) with emphasis on phase retrieval and reconstruction methods. The physical and mathematical foundations of X-ray imaging are summarised, covering wave propagation, attenuation, and phase contrast mechanisms. Common CT artefacts such as beam hardening, ring effects, and misalignment are examined together with current correction strategies. Phase retrieval models, including Single Material Phase Retrieval (SMPR), Data Constrained Phase Retrieval (DCPR), and Linear Iterative Phase Retrieval (LIPR), are discussed within the framework of the Transport of Intensity Equation and the Alvarez–Macovski model. Simulation studies show that LIPR provides high accuracy and adaptability in recovering Compton and photoelectric signals, and a preliminary test on experimental datasets confirms its robustness under realistic noise and geometric variations. The review highlights key challenges in achieving reliable material differentiation and artefact reduction in modern X-ray imaging.

Lisa Chen, University of Auckland

Session 311. *AI without data sense is like fish without bicycle.*

Today, students and academics are quickly turning to AI in their studies, while industry is also making fast use of these tools to help with modelling and decision-making. But in this rush, one important piece is often missed: data sense. Often overlooked, data sense – the simple (and somewhat basic) awareness and judgement we bring when working with data – is what keeps AI useful and stops results from becoming misleading.

In this talk, I'll share some stories and reflections from both teaching and industry work on what "data sense" looks like, why it matters for data scientists, and how missing it can trip us up.

Alicia Crabtree, Stats NZ

Session 313. *Evaluating alternative methods to derive ethnicity estimates from information available in administrative data.*

Ethnicity is a measure of cultural identity and a key attribute used to describe the population and inform policy. Individuals may identify with more than one ethnic group, and people may respond with different ethnicities in different contexts.

Sarah Croft, University of Otago

Session 122. *Estimating abundance in small populations using pedigree reconstruction.*

Accurate measures of abundance are essential for successful monitoring of animal populations, and for assessing the efficacy of conservation interventions. In previous work, genotypic information has been incorporated into Mark-Recapture models to enable the identification of individuals, as well as determination of kinship between observed individuals in Close-Kin Mark-Recapture (CKMR) models. Generally, CKMR models make large sample assumptions, limiting their application to many endangered and at-risk species.

We have developed Bayesian methodology to estimate population abundance and dynamics for small, isolated populations using pedigree reconstruction. The true underlying pedigree completely describes the abundance and population structure over time, however the true relationships between individuals in wild populations are rarely known. Given a set of observed genotypes, along with supplementary data, our methodology is able to successfully reconstruct the pedigree without the need for large sample assumptions. Prior knowledge of the mating structure and reproductive dynamics of the population can also be incorporated in the model. In this talk I will present our pedigree reconstruction approach for population estimation using dead recovery data, and will discuss the challenges associated with the full pedigree approach.

James Curran, University of Auckland

Session 222. *Data fusion in forensic paint analysis.*

Paint is frequently encountered in criminal casework, particularly in burglary, hit-and-run incidents, and vandalism. The ability to associate paint recovered from a person of interest with an evidential item is useful in the judicial process.

Tilman Davies, University of Otago

Session 112. *Chi-Squared Cox Processes: Is There Any Point?.*

The trusty old log-Gaussian Cox process (LGCP) is one of the most popular tools in a spatial statistician's kit. Tractable and flexible, it is based upon exponentiation of a Gaussian random field. Another option, which also results in a valid, positively-valued spatial driving intensity, is to *square* said field, or sum a set of several squared fields — yielding a Chi-squared Cox process (CSCP). A CSCP has different properties to an LGCP and hence describes different styles of spatial dependence. But the CSCP has essentially only ever received basic, in-passing mentions in the literature, quite possibly because much smarter people have ignored it with good reason. Let's investigate.

Tori Diamond, University of Auckland

Session 242. *Advancing Official Health Statistics: Bayesian approaches for making Māori health data relevant to local decision-making.*

Inequitable health outcomes persist for Māori in Aotearoa, yet monitoring these disparities faces significant methodological challenges. The 'small n problem' creates a critical paradox: populations most requiring targeted healthcare often have the least reliable statistical estimates to inform local decision-making.

This project builds on Bayesian methods to address small count challenges in official statistics, focusing on Māori health outcomes. Using Stats NZ's Integrated Data Infrastructure (IDI), this talk informs two case studies: rheumatic fever (a rare event with small counts) and COVID-19 vaccinations (a common outcome with spatial variation). We propose Bayesian hierarchical models to address the fundamental challenge of balancing local-level precision, statistical accuracy and the reality of small count sizes.

This research contributes to methodological advances and novel implementation of linked administrative data, while addressing critical gaps in official health statistics with real implications for small populations.

Rasika Dilhani, Victoria University of Wellington

Session 231. *Entropy–Complexity Based Clustering of ARMA Simulated Time Series Models.*

Ordinal pattern analysis provides a robust and computationally efficient approach for time series analysis. This study applies the method to simulated Autoregressive (AR), Moving Average (MA), and Autoregressive Moving Average (ARMA) models. Time series are transformed into sequences of symbols that represent ordering relationships among data points within specific time windows, from which Shannon entropy and statistical complexity are computed. These features serve as input for clustering and model discrimination.

Simulated series of length n=1000 were generated under stationarity and invertibility conditions, with several parameter variants and 100 replications per configuration. Analysis in the entropy–complexity plane revealed distinct structural patterns: ARMA(1,1) formed two clusters determined by AR coefficients, whereas ARMA(2,2) exhibited overlapping but sample-size-dependent groupings.

Results demonstrate that entropy and complexity effectively reveal intrinsic model differences. The proposed framework supports unsupervised time series classification using ordinal pattern analysis, with applications in finance, engineering, and bioinformatics.

Gary Dunnet, Stats NZ

Session 321. *Modernising the NZ Population Census.*

There is no doubt that Census data is critical to the way we gain insights about people and communities that support Aotearoa New Zealand's economic growth and social wellbeing. Census data is essential for policy making, resource allocation, long-term workforce planning, and for targeting investment in health, education, and infrastructure; and enables many communities and organisations around the country, helping to guide their planning and decision-making.

In June 2025, following Cabinet's endorsement, it was announced that Aotearoa New Zealand will move toward a new approach that prioritises the use of admin data, supported by a new annual sample survey of the population, to produce New Zealand's official census data and statistics. The new approach builds on the combined models used in 2018 Census and 2023 Census, which both used admin data to support survey information. While admin data will cover most of the population, the new annual sample survey will play an important role in testing accuracy and gathering more information. As part of the new approach, Stats NZ will also develop tailored solutions, such as targeted surveys, to help meet the information needs of iwi Māori and smaller population groups. This new approach will deliver a range of important benefits for New Zealand: it will allow data and statistics to be produced in a way that is sustainable now and into the future and meet growing customer demand for more frequent and responsive data.

This paper will present some of the methodological feasibility work we have done, is in progress, and future areas of development so to deliver a solution comprising the combination of admin data, annual sample survey and tailored solutions.

Tarin Eccleston, University of Auckland

Session 121. *Rapid Bayesian Inference for Galactic Supernova Gravitational Waves.*

Gravitational waves—ripples in spacetime—let us listen to how matter behaves in one of the Universe's most extreme environments: the collapsing cores of dying stars. Yet waves from galactic core-collapse supernovae remain undetected, and we need to be ready for the next one.

We are developing a Bayesian inference pipeline based on neural posterior estimation, using variational autoencoders and normalising flows. Because this approach relies on amortised inference, once the model is trained it can produce posterior estimates almost instantly, avoiding the need to run computationally expensive procedures like MCMC at detection time.

[warning: please skip the first half of the talk if you're sensitive to flashing lights!].

Matt Edwards, University of Auckland

Session 121. *Detecting the next galactic supernova using variational autoencoders.*

We introduce a novel methodology that employs a variational autoencoder (VAE) to compute Bayesian odds for use as a ranking statistic for LIGO/Virgo/KAGRA gravitational wave (GW) data analysis. Analysis with simulated data shows that this VAE-based framework offers robust signal-noise discrimination, presenting a promising alternative to conventional signal-to-noise ratio ranking statistics. While we focus on rotating stellar core-collapse and bounce GW signals, our methodology is generalisable and could potentially enhance sensitivity in searches for various types of GW signals. This work serves as a proof-of-concept, illustrating how deep learning surrogates combined with Bayesian analysis can improve signal-noise discrimination in GW astronomy.

Laia Egea Cortés, Victoria University of Wellington

Session 122. *Evaluating the Impact of Measurement Cost on Predicting Salmon Health with a Novel Ordinal Regression Model.*

Salmon health can be assessed through a wide range of measurements that differ substantially in monetary cost. Some measurements are very cheap, such as those obtained from photographs or videos, while others require specific analysis techniques, which are expensive, such as histology analysis. For the industry, reducing costs while maintaining performance remains a constant challenge. In this analysis, we asked ourselves if we could analyse the impact of the cost of the measurements on their ability to predict salmon health using regression models.

To address this question, we used the novel and flexible Partial Ordered Stereotype Model (POSM) as a regression model to predict the salmon health status, measured as an ordinal variable with three levels (healthy, sub-clinical, and sick). This model allows us to account for the varying discriminant ability of different groups of covariates across the categories of the ordinal response. To evaluate whether including more expensive covariates is worthwhile for predicting salmon health, we selected covariates according to their cost using a Forward–Backward Stepwise model selection method, considering three sets: (1) all measurements (expensive, moderate, and cheap); (2) only moderate and cheap measurements; and (3) only cheap measurements.

Tom Elliott, University of Auckland

Session 312. *Widgets for Rserve and Typescript.*

Web front-ends offer seamless deployment across desktop and mobile platforms, and with an ever growing variety of libraries and tools, there's almost no limit to what web developers can create.

Jamas Enright, Chen Chen, Stats NZ

Session 321. *Linking People to Places of Work: Methodological Advances in the Integrated Statistical Data System (ISDS).*

Stats NZ is developing the Integrated Statistical Data System (ISDS) to provide a robust, high quality, consistent, and connected set of statistical registers across three core data domains that enables integrated data for People, Places and Business. It is a critical enabler of transformation projects within Stats NZ, including Census, Monthly Consumers Price Index, Measuring a Modern Economy and Integrated Data for Social Investment.

A key ISDS capability is linking employees to business locations, which is essential for use cases such as Census industry and workplace statistics, and for analysing job and worker flows. This talk will provide an overview of the ISDS and highlight the development of the Employee Allocation Methodology and Process that enables these integrations.

Alexandra Ferguson, Stats NZ

Session 313. *Using longitudinal surveys to discuss labour force stability.*

Labour market stability is an underexamined characteristic in traditional labour market statistics, despite its importance.

Anna Fergusson, University of Auckland

Session 311. *Branching out data science education: Developing task and computational environment design principles for teaching data science at the high school level through an international research collaboration.*

As an emerging field, data science demands new pedagogical approaches that adapt to the evolving data and technology, preparing students for an ever-changing future. Research is needed to develop generalised theoretical frameworks or pedagogical principles that can effectively guide data science education across different countries and educational systems. Our Design-Based Research (DBR) project involves researchers from New Zealand and Germany. We used relevant literature and our previous research findings to identify principles for designing high school data science tasks and computational learning environments that could be applied across different educational systems. We propose three key design "branches": (1) emphasising a humanistic focus through data construction and exploration; (2) using decision trees to develop algorithmic modelling concepts; (3) using computer programming for supporting creativity and model tinkering. In this talk, I will describe and illustrate these design "branches" using specific features of research tasks or computational environments.

Linus Fromm, University of Otago

Session 241. *Using Fibre Extensions To Build Efficient Fibre Samplers.*

Discrete linear statistical inverse problems appear across fields such as ecology, genetics and network tomography. The aim is to conduct inference on the distribution of some latent discrete variable of interest conditional on corrupted or aggregated observations. This involves the evaluation of a sum over all values of the latent vector consistent with the observations, which is infeasible in most circumstances. We turn to Markov chain Monte-Carlo (MCMC) methods. This requires the use of sets of moves called Markov bases. A fundamental theorem by Diaconis and Sturmfels from 1998 shows how we can find Markov bases algebraically.

Unfortunately, the convergence of MCMC with full Markov bases can be very poor. Additionally, the size of full Markov bases has no upper bound even on medium-sized problems.

As a solution we propose extending the fibre naturally. In some cases, this allows for smaller bases with faster convergence. Existing results indicate that MCMC over certain extensions defined by column partition lattice bases have rapid mixing.

Ciprian Doru Giurcaneanu, University of Auckland

Session 231. *Fractional Integral Transforms.*

Fractional integral transforms have been extensively used in physics and engineering, but they do not appear to be as popular in statistics. In this talk, we will present some well-known results on the Fractional Fourier Transform, a key member of this family of transforms, and briefly discuss some newer findings related to other transforms in the same family.

Martin Hazelton, University of Otago

Session 241. *Statistics Research in Aotearoa New Zealand: An International Comparison.*

New Zealand has some excellent researchers in statistics. But how does the country stack up in statistical research against international comparison? This talk will describe various bibliometric attempts to answer that question.

John Holmes, University of Canterbury

Session 111. *Variational Bayes in the GLM context: Can Bayesian(ish) inference be made efficient?.*

Statisticians have historically relied on MCMC to perform Bayesian inference. The associated computational cost has motivated research into approximate Bayesian methods, such as Variational Bayes. In this talk, we will demonstrate how a Variational Bayes approach, assuming a multivariate normal posterior, not the more typical mean field methods, with a flat prior for coefficients, for GLM corresponds to a modified version of Newton's method used to find Maximum Likelihood estimates. We will show our approach is much closer to MLE estimation than MCMC in terms of computational efficiency, while providing a close approximation to the true posterior. We will show that Bayesian GLM under a flat prior is not going to replicate maximum likelihood, particularly for logistic regression. Lastly, we will discuss how Variational Bayes can be used to construct proxy sufficient statistics that let us fit more complicated models using more efficient MCMC samplers than currently available.

Guoping Hu, University of Auckland

Session 231. *Joint Forecasting and Reconciliation for Hierarchical Time Series Using GNNs.*

Multiple time series forecasting with hierarchical structure plays a vital role in real-world applications, such as sales prediction for supply chain decision-making. Traditionally, hierarchical time series forecasting follows a two-stage paradigm: generating base forecasts independently, followed by reconciliation strategies to ensure coherence across levels.

In this talk, we present a novel end-to-end framework that jointly optimizes forecasting and reconciliation using graph neural networks. We evaluate our approach on real-world retail data under scenarios both with and without promotional information.

Charlotte Jones-Todd, University of Auckland

Session 212. *Modelling crime event data as a time-evolving bipartite network.*

Crime incidents are often linked through shared or co-offending perpetrators; this structure naturally creates a bipartite network that evolves over time as perpetrators commit new crimes. We present a framework for modeling such crime event data assuming a marked Hawkes process that incorporates a time-evolving network structure. The node arrival times of the bipartite network are modelled as a temporal Hawkes process, which allows the network to update conditional upon its history. By parameterising the conditional intensity of the marked Hawkes process accordingly our method models the time of updates concurrently with the network updates. This approach captures both the self-exciting nature of crime occurrence and the dynamic growth of the underlying perpetrator-crime network.

Osamu Komori, Seikei University

Session 122. *Cumulant-based approximation for fast and efficient prediction for species distribution.*

Species distribution modeling is widely used in ecology to estimate the relationship between target species and environmental variables associated with the species' habitat. To estimate the habitat map, Maxent is recognized as one of the most powerful tools, in which the model's parameters are estimated by sequentially maximizing the likelihood. However, computing the likelihood becomes extremely expensive when the number of locations in the study area is large. Therefore, we propose a new species distribution modeling approach based on gamma-divergence, in which the normalization term in the gamma-loss function is approximated using cumulant coefficients for the study area. This results in a computationally efficient estimation of the model's parameters, as well as high estimation accuracy. Moreover, by using the cumulant-based approximation, we show that the estimation of Maxent is equivalent to Fisher linear analysis when the normality assumption holds. We demonstrate the proposed method using simulation studies and a dataset of Japanese vascular plants.

Takafumi Kubota, Tama University

Session 313. *Social Isolation among Japanese Adults Based on Microdata from the Basic Survey on Social Life.*

Using microdata from Japan's Basic Survey on Social Life, this study investigates the determinants of social isolation among adults aged 20 to 59. Social isolation is defined in two forms: spending the entire day alone and spending the entire day only with family. The analysis first focuses on family-type isolation and applies logistic regression to identify the factors associated with this condition. Based on the findings, a multinomial logistic model is then employed to comprehensively examine the predictors of different isolation types. This sequential approach allows for a detailed understanding of how demographic, socioeconomic, and regional factors contribute to isolation patterns in contemporary Japan.

Kate Lee, University of Auckland

Session 111. *Bayesian inference for the learning rate in Generalized Bayesian inference .*

Bayesian evidence combination often involves models with multiple parametric modules. When one module is misspecified, its influence can be reduced by cutting information flow—this leads to Cut Models. Semi-modular inference (SMI) generalizes this by introducing a learning rate, loss function, and Gibbs posterior to allow partial information flow between modules. Estimating the learning rate and loss hyperparameters is challenging, as they can't be learned jointly with other parameters due to model misspecification. We consider settings with a pseudo-optimal learning rate and use held-out data to infer its posterior. Under certain conditions, this posterior concentrates on the pseudo-optimal value, enabling principled hyperparameter estimation. Experiments show that SMI posteriors outperform both Cut-model and standard Bayesian approaches on test data.

Miaotian Li, University of Auckland

Session 212. *Fractional Cepstral Techniques for Real-World Time Series Clustering.*

A class of methods for measuring dissimilarity between time series relies on cepstral coefficients. Typically, these coefficients are estimated by applying the Fourier transform (FT) to obtain the log-spectrum, followed by the inverse FT. Various techniques are then used to zero out selected cepstral coefficients. In this talk, we explore the impact of replacing the FT with the fractional Fourier transform in the coefficient computation process. Real- world data are used to illustrate the resulting effects on time series clustering.

Yuan Li, University of Auckland

Session 121. *The Time Dilation of High Redshift Quasars: Insights from Bayesian Inference.*

After decades of investigation, cosmological time dilation has recently been confirmed in the variability timescales of distant quasars. This study builds upon previous analyses by separating the cosmological signal from the effects of source population properties, specifically quasar bolometric luminosity and the rest-frame emission wavelength of observed variability. Using Bayesian modelling, the variability timescales in 190 quasars were examined. The findings indicate that the model incorporating all explanatory factors provided the best fit as indicated by marginal likelihoods. If the cosmological dependence is assumed to be of the form $(1+z)^n$, the inference yields $[n = 1.13]^{+0.35}_{-0.26}$, aligning with the results of Lewis and Brewer (2023). This supports the hypothesis that quasar variability is driven by cosmological time dilation, challenging earlier conclusions by Hawkins (1993, 2001, 2010).

Ruiting Mao, University of Auckland

Session 121. *Rapid Gravitational Wave Parameter Estimation on LISA Signals with Gaps using Flow Matching.*

The Laser Interferometer Space Antenna (LISA) will revolutionize gravitational wave astronomy, but its data will inevitably have gaps and instrumental noise. These imperfections challenge traditional analysis methods, which can be slow and unreliable when applied to incomplete signals. To address this, we present a novel deep learning framework that performs parameter estimation directly on noisy, gapped data using conditional flow matching.

Our model extracts essential information by processing the signal in two ways: directly in the time domain and in a time-frequency representation using wavelets. Validated on simulated Lisa-like signals, our approach yields improved results compared to normalizing flows, which are currently widely applied in Simulation-Based Inference on gravitational wave signals, while being more flexible and efficient to train. This work offers a computationally efficient solution for analyzing realistic gravitational wave signals, helping to maximize the scientific return from the LISA mission.

Lafaele Mapusua, Stats NZ

Session 321. *Data Formation: A Renaissance - Building an Inclusive Data System, One Portfolio at a Time.*

Stats NZ is investing in coordinated approaches across Māori, Pacific, Rainbow, and Accessibility data portfolios. Each area faces distinct evidence gaps, yet many challenges are shared – from inclusive classifications to ethical data governance and equitable survey design. This presentation will share how cross-portfolio collaboration reduces duplication, improves consistency, and accelerates learning across the system. Case studies will illustrate how insights from one portfolio can inform another, supporting better decisions for all priority populations. We will also explore risks of siloed approaches, the role of inclusive data advisory models, and opportunities to scale innovations system-wide. The session will highlight practical lessons, benefits, and challenges in building equity-driven data capability that works across diverse communities.

Andrew Martin, Stats NZ

Session 321. *Bayesian linkage error estimation for linked administrative data.*

New Zealand has run a Census every five years since 1877, with four exceptions. Rising costs, declining response rates, and vulnerability to natural disasters have led Stats NZ to develop an alternative census model using government administrative data. One goal of the programme is to produce high-quality, timely population estimates from linked data. Our experimental approach applies dual-system estimation to two administrative lists, linked through a semi-automated process. Linkage errors — missed and false links — can bias estimates. False links are minimized through conservative matching. We estimate missed links using dual-system estimation, requiring either two independent linkage methods or one method plus a set of records known to appear in both lists. Within a Bayesian framework, our approach estimates the number of missed links and associated uncertainty for small domains. We demonstrate the method using both simulated and real administrative data.

Qingyu Meng, University of Auckland

Session 222. *A Reproducible and unified Benchmark of Deep Learning Feature Selection Across Simulations and Multi-Omics Cohorts.*

Reliable feature selection (FS) is critical for extracting small, interpretable biomarker panels from high-dimensional omics, yet deep learning–based FS methods have rarely been compared systematically. We present a unified and reproducible benchmark of post-hoc explainers—attribution (GradientSHAP, DeepLIFT, and LRP), perturbation (LIME, Feature Ablation, and Occlusion) and embedded selectors (CancelOut, EAR-FS, and GRACES). Using standardized preprocessing, a shared neural network, and Optuna tuning, we evaluate FS performance and computational efficiency in simulations and downstream predictive accuracy on diverse real datasets, including gene expression datasets with binary outcomes, five TCGA projects spanning mRNA, methylation, and SNP modalities with multi-class outcomes, and ADNI gene expression with continuous neuroimaging phenotypes. EAR-FS and GRACES consistently perform best: GRACES is most robust but computationally intensive, whereas EAR-FS achieves similar accuracy with much lower computational cost. Classical methods remain competitive when signals are sparse and near-linear. Post-hoc explainers contribute most as interpretability tools and model auditors rather than as primary subset selectors. To enable reproducibility and broad adoption, we provide a software hub implementing all these methods with standardized pipelines and evaluation routines, facilitating efficient FS under typical constraints of budget, sample size, and turnaround time.

Unnati Nigam, Monash University

Session 211. *Statistical Modeling of Repetitive Dynamics via Quasi-Periodic Gaussian Processes.*

Repetitive processes arise in many scientific and engineering applications, where system performance may decline due to environmental changes or system variability. Iterative Learning Control (ILC) improves performance by exploiting information from previous repetitions to adjust for expected errors in future ones. We introduce Quasi-Periodic Gaussian Processes (QPGPs) into a predictive ILC framework to model and forecast upcoming errors. Leveraging a recent structural equation formulation, our approach enables efficient likelihood-based inference and parameter estimation, making continual GP learning feasible within sequential decision-making. By predicting error trajectories in future repetition, rather than relying only on past observations, the framework achieves faster convergence and robustness under time-varying disturbances. We compare QPGP-based predictive ILC with standard ILC and conventional GP-based predictive ILC in simulations and real-world experiments, including autonomous racing, a robotic manipulator, and robot trials. The results demonstrate that combining the QPGP model with ILC yields statistically efficient, computationally scalable, and robust methods for sequential prediction.

James O'Malley, Dartmouth

Session 322. *Causal Inference for Intervention Spillover in a Cluster-randomized Stepped Wedge Trial: Application to Physician Networks.*

Often motivated by concerns about contamination of control subjects, cluster-randomized stepped wedge designs assign interventions to distinct clusters (e.g., hospitals) and protect against confounding by randomizing the timing of intervention delivery. However, when trial units are embedded in professional networks that span clusters, such designs are vulnerable to spillover from intervention subjects to control subjects. In this study, we use a longitudinal cluster-randomized trial of an advance care planning (ACP) intervention and contemporaneous physician networks to model and estimate the direct effect of the intervention in a counterfactual world absent contamination, the indirect or spillover effect from intervention to control subjects and the direct-indirect interaction effect of whether spillover modifies the intervention's effect. Our findings suggest that the diffusion of intervention behaviors through the network were primarily driven by a large physician spillover effect that modified the direct effect at the time of intervention and subsequently over follow-up.

Priya Parmar, University of Auckland

Session 322. *Real-time, real-world longitudinal analyses ... all in one semester.*

Extending on 2024's exploration of the learnings from real-time real-world teaching applications of longitudinal analyses we use the example from the Growing Up in New Zealand study to explore and examine the longitudinal trajectories in adolescents using self-reported data from ages 8 to 13, focusing on the impact of adverse events such as COVID-19 and extreme weather events.

We reflect upon two semesters of developing the real-world applied analyst [student] linked to our "client" [clinical researcher] via the collaborative analytic teaching experiential model with an embedded "soft skills communication" framework.

Students were tasked with comparing sub-group analyses and assessment of key modifiable factors that influence mental health outcomes for different repeated measures analyses. We then present results that would provide evidence-based recommendations for parents, schools, and policymakers as well as targets for interventions aimed at mitigating the impact of adverse events on adolescent mental health and support mental health resilience.

Bo Peng, University of Auckland

Session 241. *Weighted Pairwise Likelihood for GLMMs in Complex Survey Data.*

Complex survey data often involve hierarchical structures and non-Gaussian outcomes, making likelihood-based inference challenging. Recent weighted pairwise likelihood methods address this for linear mixed models, but extensions to generalized linear mixed models (GLMMs) remain limited. This work develops weighted pairwise likelihood estimators for GLMMs that incorporate survey design weights and link-specific approximations to handle binary and count outcomes efficiently. Implemented in R and evaluated through simulation studies, the proposed estimators are compared with glmmML and WeMix to assess bias, variance, and computational performance. The results show that the weighted pairwise approach achieves design-consistent estimation but with some loss of efficiency across outcome types and cluster sizes. The method provides a practical framework for valid design-based inference in large-scale surveys, contributing to applications in social science, health, and education research.

Joseph Reps, University of Auckland

Session 112. *Spatial Capture-Recapture for Acoustic Data: Models and Implementation in the acre R Package.*

Passive acoustic monitoring provides an efficient, non-invasive way to survey wildlife populations. By modelling detections from clusters of acoustic detectors, spatial capture-recapture (SCR) methods can be used to estimate population density and spatial distribution. The acre R package offers a flexible framework for fitting such models to acoustic survey data, extending traditional SCR approaches to incorporate additional information such as signal strength, bearing, time of arrival, and relevant covariates. In this presentation, I will outline the modelling framework implemented in acre, demonstrate its functionality, and illustrate how it facilitates practical, reproducible analysis of acoustic SCR data. The package is actively maintained and continues to expand in scope. This presentation highlights acre as a versatile, open source tool for acoustic spatial capture-recapture modelling and wildlife density estimation.

Nasim Sadra, Massey University

Session 231. *Statistical Analysis of Landslide-Induced Vegetation Disturbance Following Cyclone Gabrielle in Hawke's Bay, New Zealand.*

This research proposes a statistical framework to evaluate the spatiotemporal dependencies between extreme weather events and coastal sediment dynamics. Sentinel-2 NDVI (Normalised Difference Vegetation Index) timeseries from Hawke's Bay contain possible anomalies associated with Cyclone Gabrielle, serving as a landslide proxy. A Bayesian change point is proposed to detect anomalies, which will also estimate the timing, magnitude, as well as recovery rate after the event. A multivariate distributed lag model (DLM) quantifies temporal dependencies between landslides and coastal turbidity, while Generalised Extreme Value (GEV) distributions can characterise extreme sediment probabilities. The preliminary results demonstrate clear temporal anomalies after the cyclone, which can determine a starting point for the analysis of the delayed response assumed to be associated with the cyclone. Careful consideration will be required to define causal relationships based on the patterns and associations identified in this study.

Janet Stacey, University of Auckland

Session 211. *Automating the Evaluation of Forensic Evidence.*

Forensic science is undergoing a paradigm shift with a growing emphasis on activity-level propositions, which focus not just on the source of evidence, but on the mechanisms and timing of its deposition. Bayesian Networks (BNs) offer a robust framework for evaluating these propositions, integrating experimental results and expert knowledge to support scenario-based reasoning. However, there is still reticence to their use for the evaluation of activity-level propositions. Some very vocal commentators go as far as suggesting that activity level propositions should not be considered at all. This presentation introduces a project that aims to address these issues using responsibly developed generative AI and statistical methods. To embrace robust and reliable automation of Activity Level Evaluations, an understanding of the available data is needed. The first objective of the project will be covered where topic modelling of collated published materials has been carried out using a Sentence Transformer based approach.

Ben Stevenson, University of Auckland

Session 122. *Measurement error models for morphometric data.*

Authors: Ben C Stevenson, Elizabeth Smit, Edy Setyawan.

Understanding the body size of individuals and the relationships between different dimensions is critical for monitoring the status and health of wildlife populations. Recent technological advancements allow researchers to use affordable tools like drones and camera traps to photograph animals and extract morphometric data, but these can be less accurate than direct measurements.

In this talk, I will describe a new linear mixed-effects modelling approach that uses multivariate normal distributions to describe latent true measurements and account for measurement error. Our method estimates relationships between dimensions, predicts one dimension from any subset of others, and enables inference about allometry, including tests for allometric growth. We illustrate the use of our method with an application to morphometric data of the reef manta ray collected by drones.

Xiaoyan Sun, University of Auckland

Session 222. *Multi-head Attention Feature Selection for High-Dimensional Biomedical Data via Deep Fusion of Filter and Embedded Methods.*

Feature selection in biomedical data analysis confronts significant challenges: existing approaches typically employ singular strategies and exhibit inherent limitations. Filter methods demonstrate computational efficiency but fail to identify complex feature interaction patterns, while embedded methods can capture nonlinear relationships yet lack statistical prior guidance and incur high computational costs. Most existing methods cannot simultaneously leverage the interpretability of statistical approaches and the sophisticated modeling advantages of deep learning, a problem particularly pronounced in high-dimensional biomedical data. We propose a Multi-Head Attention-based Feature Selection Framework (MAFS), a unified model that integrates statistical priors and deep learning capabilities. This framework initially employs filter methods to compute initial correlation weights, then utilizes multi-head attention mechanisms with Lasso adaptive regularization to learn complex feature interactions from multiple perspectives, ensuring statistically guided deep modeling results. Through a multi-stage dimensionality reduction strategy, the most discriminative features are selected from each attention head to construct optimal subsets. Comprehensive experiments on multiple high-dimensional biomedical datasets, including cancer gene expression and Alzheimer's disease data, demonstrate the performance of this approach. Downstream analyses validate the potential of our method in terms of feature selection effectiveness and classification performance.

Simon Urbanek, University of Auckland

Session 221. *Applying LLM Embeddings to R Journal Articles.*

In this talk we take some ideas from the AI world in the form of LLM embeddings and see how they could aid journal editors. We use published articles of the R Journal to explore the possibilities and its properties with the aim to perform tasks such as keyword matching, similarity assessment or finding reviewers.

Alec van Helsdingen, University of Auckland

Session 112. *Modelling Under-Dispersed but Clustered Whale Cues.*

Many whale species emit cues, whose times form a temporal point pattern. The event times are usually strongly clustered, with long periods of silence followed by a rapid burst of cues. Conversely, at shorter time scales the cues are more evenly spaced (under-dispersed) than expected by an inhomogeneous Poisson process. The Hawkes process can be used to model the clustering by assuming that cues self-excite future cues, but this does not account for underdispersion.

In this talk, we develop a framework for incorporating under-dispersion and over-dispersion into the Hawkes process. We use the Weibull distribution and mixtures of two Weibulls to model the interarrival times between cues rather than the exponential distribution. This gives more flexibility and a better fit.

We demonstrate our model on the cues of an individual sperm whale, confirm our intuition that the cues are under-dispersed, and quantify the relationship between the cue rate and covariates.

Alain Vandal, University of Auckland

Session 312. *Extending diagnostic validity meta-analysis to several diagnostic guidelines.*

Optimal management of acute biliary disease should include an assessment for possible choledocholithiasis. Uncertainties remain regarding the performance of existing guidelines in practice. Thirty-one studies were identified for a meta-analysis of diagnostic validity. All focused on one or more of three international guidelines. Each study reported on the diagnostic validity of one or more of these guidelines at one or more score cutpoints. The analysis was carried out using the diagmeta R package [Rücker G, Steinhauser S, Kolampally S, Schwarzer G (2022)]. Their approach allows the meta-analytical estimation of the AUROC, and consists in fitting transformed values of sensitivity and specificity at various cutpoints using linear mixed effects regression. While the diagmeta approach applies to the meta-analysis of a single diagnostic procedure, we show how its framework can be extended to test for performance similarity between guidelines. We also show how it can be extended to identify outlier studies.

Tieta Vesty, Stats NZ

Session 242. *Iwi affiliation in linked admin data and historical censuses .*

This work presents a current snapshot of availability and quality of admin data sources for iwi affiliation in the Integrated Data Infrastructure (IDI). The data sources are from the Ministry of Education for primary and secondary schooling and tertiary levels, Student Loans and Allowances, and historical census data from 2013 and 2018 Censuses. Quality measurement is done by measuring coverage and consistency rate with 2023 Census for each data source along with metadata and collection context analysis. The combined coverage rate of admin data sources is 69.2 percent, while the combined coverage rate also including historical census is 90.6 percent. The consistency results for all data sources range from 67.9 to 77.0 percent. Stats NZ are currently modernising the census, and possible future solutions for iwi affiliation will be designed in collaboration with iwi Māori, and could include a combination of admin data, survey data and tailored solutions.

Xinyue Wan, Victoria University of Wellington

Session 242. *Making Predictions for Model-based Clustering.*

This proposal develops statistical methods for modeling clustered ordinal data. In the current work, we focus on a two-step approach using Dirichlet regression combined with proportional odds models to estimate latent cluster memberships and predict category probabilities, especially when only partial information is available. For future work, a one-step estimation approach is proposed to incorporate covariates directly into the clustering framework. Finally, the method is evaluated for estimating performance across various simulation scenarios, including continuous and discrete covariates, and is further applied to a real dataset to demonstrate its practical utility.

Ting Wang, University of Otago

Session 113. *Better forecasting of extreme geomagnetic storms using non-stationary extreme value models.*

An assessment of the risk of extreme geomagnetic storms is critically important for modern society. However, current methods mainly focus on using stationary statistical models to analyze extreme geomagnetic events. These models ignore the non-stationary nature of the data, caused by effects of the solar cycle and the seasons, and thus could provide unreliable estimates of return levels. We propose use of hidden Markov models and generalized additive models, both involving time-varying parameters, in order to capture these features of the data. We use these models to analyze extreme values of the magnitude of the derivative in the horizontal component R1(t) of geomagnetic observations from the Eyrewell geomagnetic observatory in New Zealand. We demonstrate that they can successfully model the effects of both the solar cycle and the seasons and use the best-fitting models to provide more reliable estimates of return levels.

Zhijian Wen, University of Auckland

Session 211. *Detecting body fluids on fibres using hyperspectral imaging.*

Detection of latent body fluids on fabrics is an important step in forensic investigations. Conventional approaches rely on presumptive chemical tests or alternative light sources (ALS) to visualise fluids; however, ALS methods—such as ultraviolet (UV) illumination, may damage DNA present. This study looks at an alternative instrument – hyperspectral imaging (HSI), for non-destructive detection and localisation of body fluids deposited on various fibre types. We collected a hyperspectral images dataset using a Specim IQ camera. A CNN-transformer Unet model have been employed to segment fluid regions from high-dimensional spectral data. The model achieved an intersection-over-union (IoU) score of 0.66, compared with 0.31 for a principal component analysis (PCA)-based baseline. The results demonstrate that spectral–spatial feature learning using neural network methods can enhance the localisation of latent body fluids on various fibre backgrounds.

Hao Xie, Victoria University of Wellington

Session 221. *A Statistical Framework for Partial Multi-Label Feature Selection with Graph Regularization.*

Partial Multi-Label Learning (PML) tackles the challenge of developing accurate models based on candidate label sets that include both ground-truth labels and noisy ones. This uncertainty, compounded by high-dimensional features, leads to biased estimation, overfitting, and inflated variance. Existing feature selection methods generally assume fully reliable labels and thus perform poorly under label noise. We propose a statistical framework for feature selection in partial multi-label data. The approach employs a low-rank representation to recover the ground-truth label distribution and a sparsity penalty to suppress spurious labels. To preserve structural information, we incorporate a dual-graph regularization that captures non-linear geometry in both the feature and label spaces. Feature selection is further promoted through an L2,1-norm penalty. An alternating iterative optimization algorithm is developed for parameter estimation, with guaranteed convergence. Experimental results on both synthetic and real partial multi-label datasets demonstrate that the proposed method consistently outperforms existing approaches.

Thomas Yee, University of Auckland

Session 312. *Variance Factorized Loglinear Models for Overdispersed and Underdispersed Counts.*

A more structured approach to modelling over- and under-dispersed counts is proposed: variance factorized loglinear (VFL) models are a parameterization of an (enlarged) distribution whereby the overdispersed variance is expressed as a product of the nominal variance and a remainder term involving a highly interpretable loglinear effect with a potentially different set of covariates. This facilitates mean–variance analysis: the variance is modelled more separately from the mean, especially for GLMs. They offer advantages over other methods, such as a unifying framework and greater interpretability. Examples of enlarged distributions that supply the equidispersion as a special case include the extended beta-binomial (with a newly proposed "clog" link and special offset), and negative binomial and generalized Poisson. VFL models are constructed as a vector generalized linear model with a judicious combination of constraint matrices, offsets and links. VFL models are implemented in the VGAM package on CRAN.

Hannah Yun, University of Auckland

Session 212. *Condition Monitoring of Wind Turbine Bearings Using Dictionary Learning in the Frequency Domain.*

Effective condition monitoring (CM) of wind turbines can significantly reduce maintenance costs and enhance the efficient use of wind energy by enabling early fault detection. In this talk, we present a novel CM approach based on spectral analysis of bearing vibration signals. We demonstrate the capabilities of the novel method in experiments conducted with laboratory and industrial data.

Siwei Zhai, University of Auckland

Session 113. *Childhood Risk and Resilience Factors for Pasifika Youth Respiratory Health: Accounting for Attrition and Missingness.*

In New Zealand, 7% of deaths are related to respiratory diseases, with Pacific people at higher risk. Our work aims to investigate the causal effects of early-life risk and resilience factors on early-adulthood lung function amongst Pacific Islands Families Study (PIFS) cohort members (n=1,398), of whom 466 participated in a respiratory study. Primary outcome was forced expiratory volume in 1 second (FEV1) z-score at age 18 years. FEV1 and healthy lung function, defined as the z-score ¿ -1.64, were secondary outcomes. Our study will explore the effects of factors from relevant domains based on the PIFS cohort by applying an integrated model. Simulations were conducted to validate the integrated model, from which we found that the causal estimate is generally biased. The direction and magnitude of this bias are affected by several factors: 1. Sample size; 2. Link function non-collapsibility; 3. Link function boundary.

Monday 8th			
Time	Lecture Theatre - LT01	Seminar Room C - G.03	Seminar Room A - G.02
12:00pm	Registration		
1:00pm	Welcome and Housekeeping		
1:10pm	Keynote - Kerrie Mengersen		
2:00pm	Address from Minister of Statistics, Hon Dr Shane Reti		
2:20pm	**Session 111** Christiana Akua Akyaa Agyei, John Holmes, Kate Lee	**Session 112** Alec van Helsdingen, Joseph Reps, Tilman Davies	**Session 113** Siwei Zhai, Ting Wang, Richard Arnold
3:20pm	Afternoon Tea		
3:40pm	**Session 121** Ruiting Mao, Tarin Eccleston, Yuan Li, Matt Edwards	**Women in Data Science**	**Session 122** Laia Egea Cortés, Sarah Croft, Osamu Komori, Ben Stevenson
5:30pm	Welcome Drinks at the Staff Club		

Tuesday 9th			
9:00am	Housekeeping		
9:10am	Keynote - Darryl MacKenzie		
10:00am	**Session 211** Unnati Nigam, Janet Stacey, Zhijian Wen	**Biostatistics Session**	**Session 212** Miaotian Li, Hannah Yun, Charlotte Jones-Todd
11:00pm	Morning Tea		Morning Tea
11:30am	**Session 221** Hao Xie, Kaisi Chen, Qian Chang, Simon Urbanek		**Session 222** Qingyu Meng, Xiaoyan Sun, James Bristow, James Curran
1:00pm	Lunch - Poster Session		
2:00pm	**Census Panel**	**Session 231** Rasika Dilhani, Guoping Hu, Nasim Sadra, Ciprian Giurcaneanu	**Pipeline Development Workshop** Three hour duration in **Math 124**.
3:20pm	Afternoon Tea		
3:40pm	**AI Panel**	**Session 241** Linus Fromm, Bo Peng, Martin Hazelton	**Session 242** Xinyue Wan, Tori Diamond, Tieta Vesty
4:40pm	NZSA AGM		
6:00pm	Conference Dinner at Otago Museum		

Wednesday 10th			
9:00am	Housekeeping		
9:10am	Keynote - Thomas Lumley		
10:00am	Aitken Documentary		
10:10am	**Session 311** Andrew Balemi, Anna Fergusson, Lisa Chen	**Session 312** Tom Elliott, Thomas Yee, Alain Vandal	**Session 313** Alexandra Ferguson, Alicia Crabtree, Takafumi Kubota
11:10am	Morning Tea		
11:40am	**Session 321** Andrew Martin, Lafaele Mapusua, Jamas Enright & Chen Chen, Gary Dunnet	**Essential Skills Workshop**	**Session 322** Priya Parmar, Liza Bolton, James O'Malley, Vanessa Cave
1:00pm	Lunch		

1 Code of Conduct

We value your and everyone's participation at the NZSA Conference 2025. Your support will help keep our community and our conference a safe, welcoming, and friendly space for all fellow participants.

Our conference is dedicated to providing an harassment-free conference experience for everyone, regardless of gender, gender identity and expression, age, sexual orientation, disability, physical appearance, body size, ethnicity, religion (or lack thereof), or technology choices.

It is the responsibility of the community as a whole to promote a kind, inclusive and positive environment for our activities. All attendees, speakers and sponsors at our conference are required to agree with the following code of conduct.

We do not tolerate harassment in any form. Harassment includes offensive or demeaning verbal comments; sexual images in public spaces; deliberate intimidation, stalking, harassing photography or recording; sustained disruption of talks or other events; inappropriate physical contact; unwelcome sexual attention.

We expect participants to follow these rules at conference and workshop venues, conference-related social events and online media.

Participants asked to stop any harassing behaviour are expected to comply immediately.

If you are being harassed, notice that someone else is being harassed, or have any other concerns, please contact the following representatives:

- Conor Kresin (Organiser) – `conor.kresin@otago.ac.nz`
- Bethany Macdonald (Organiser) – `bethany.macdonald@otago.ac.nz`

These representatives will be available in person and via e-mail. Be assured that if you approach them, your concerns will be kept in strict confidence and they will consult with you on any actions taken. In case of a formal complaint, they will speak to all parties involved to try to resolve the issue – without presupposition of guilt.