New Zealand Statistical Association 2024 Conference

2–4 December
Victoria University of Wellington
Wellington, New Zealand

Presenters and Abstracts

Plenary, Sponsor's and ANZJS Read Paper sessions are listed first, in temporal order

Contributed Presentations follow, ordered alphabetically by presenters' last names

Editor: **John Haywood**

Plenary Session 1, 1.15-2.05pm, 2 December

Double descent and noise in fitting linear regression models

Alan Welsh
Australian National University
Joint work with Insha Ullah

"Double descent" is used in statistical machine learning to describe the fact that models with more parameters than observations can have better predictive performance (as measured by the test error) than models with fewer parameters than observations. This challenge to the belief that simpler models are generally better, implies we need a rethink of fundamental statistical ideas. We explore the effects of including noise predictors and noise observations when fitting linear regression models. We present empirical and theoretical results that show that double descent occurs in both cases, albeit with contradictory implications: the implication for noise predictors is that complex models are often better than simple ones, while the implication for noise observations is that relatively simple models are often better than complex ones. That is, double descent is not just a high-dimensional big data/machine learning phenomenon but can also occur in small datasets fitted with simple statistical models. We resolve this contradiction by showing that it is not the model complexity but rather the implicit shrinkage by the inclusion of noise in the model that drives the double descent. We also show that including noise observations in the model makes the (usually unbiased) ordinary least squares estimator biased and indicates that the ridge regression estimator may need a negative ridge parameter to avoid over-shrinkage.

Premium Sponsor's Presentation, 3.40-4.10pm, 2 December

The best minds of my generation are teaching machines to be convincingly wrong – how do we make this right?

Joe Robins
SAS New Zealand

Massive investment into artificial Intelligence (AI) technology globally is increasing demand for already scarce data science and statistics practitioner skill sets. This growth of AI also has the potential to boost productivity but brings with it many risks and pitfalls. How do we ensure the workforce is up-skilling to meet demand whilst also balancing the risk vs reward of creeping AI deployments?

In this talk I will share SAS' approach to addressing this, including matching skill sets to demand, tool set design and good practice guidelines.

Plenary Session 2, 11.10am – 12.00pm, 3 December

Data leadership within and beyond your organisation

Kate Kolich
Reserve Bank of New Zealand – Te Pūtea Matua

Data is an integral part of all good decision making. Connecting statistics, data and insight with decision makers is important, but so is having a strong data culture throughout the whole organisation. Being enabled by data and insight is contingent on a data culture that has a defined data strategy aligned with organisation outcomes, ensuring data and insight is fully integrated in all decision making. There are a number of key ingredients for this, including many participants' involvement. In this talk Kate Kolich will share her inclusive approach to data leadership in her role as Assistant Governor/General Manager of Information, Data, and Analytics at Te Pūtea Matua The Reserve Bank of New Zealand (RBNZ). This talk will provide a high level overview of how the RBNZ, as kaitiaki and custodian of New Zealand's financial system data, produces data and insight to enable decision making. Kate will outline how having a data strategy which embraces the needs of data practitioners, consumers and decision makers leads to data innovation and data driven decision making.

ANZJS Read Paper (with discussion), 1.00-2.20pm, 3 December

The incremental progression from fixed to random factors in the analysis of variance: a new synthesis

Marti Anderson
Massey University and PRIMER-e
Joint work with Ray N. Gorley, Antonio Terlizzi

The Zoom link for this live-streamed session is: `https://vuw.zoom.us/j/96030995168`

Ah, the well-known and perplexing phenomenon faced by every practicing researcher who embarks on an analysis of variance (ANOVA) . . . are my factors fixed or random? And then (yikes!) the rather uncomfortable realisation that, yes, it does matter. The choice affects: (i) the expectations of mean squares (EMS); (ii) the estimates of variance components; (iii) the construction of suitable statistics for testing hypotheses; and (iv) the nature and extent of the inferences arising from such tests. Well, good research often starts where there is a fight, and how to calculate the "correct" EMS arising from mixed-model ANOVA designs has been the source of a long-standing tussle. Isn't it amazing that Cornfield & Tukey long ago (1956) suggested that the difference between a fixed and a random factor was not a dichotomy, but rather, a gradation, which depends only on your sampling effort relative to your inference space. When combined with the (little-bit-later) work of Hartley and Rao, we not only get a lovely extension to the most general cases (balanced or unbalanced designs, any types of factors along the gradation, multivariate, etc.), but also achieve a sweet resolution to the dispute. In this talk, I will briefly outline all of this and showcase its virtues in an ecological example: the responses of 151 species of mollusc to a sewage outfall on the Italian coast. I'll unveil the desirable increase in power that taking this synthetic approach can afford.

Plenary Session 3, 9.00-9.50am, 4 December

Insights, killer stats and impact

Ruth Shinoda

Education Review Office — Te Tari Arotake Mātauranga

Ruth Shinoda (Deputy Chief Executive, Education Review Office and Head of the Education Evaluation Centre — Te Ihuwaka) will talk about how to use evidence to have impact on government policy using real world examples. She'll focus on what does and doesn't have an impact, mistakes to avoid (many which she has made!), lessons learnt, and most importantly the difference between evidence, insights and killer stats.

Plenary Session 4, 9.50-10.40am, 4 December

Payment for Success: An introduction to how paying for social outcomes works, and the critical role of statisticians

Kylie Reiri

Partner: Finance, Economics and Analytics &
Lead Partner for the Māori Economy, PwC

What's changing in how we fund social services? How do outcomes-based contracts work – where organisations get paid for the positive changes they create rather than just the work they do? How do we figure out what to pay, and how do we prove these programmes actually make a difference? Why aren't traditional measurement tools like Social Return on Investment enough anymore? And most intriguingly – why do we suddenly need more statisticians than ever in social services? Join us to explore how the future of social impact depends on proving what works.

David Vere-Jones Tribute, 2:30pm, 4 December

A brief personal tribute to David Vere-Jones

David Harte

Statistics Research Associates, Wellington

David Vere-Jones died recently, on 31 October 2024. This is a brief personal tribute to David, who played an enormous role in Statistics, not only in NZ but internationally too. No doubt, much fuller summaries of his academic and personal achievements will appear in due course.

David Vere-Jones was President of the New Zealand Statistical Association from 1981 to 1983 inclusive, and was awarded the prestigious Vaughan Jones Medal for 2014 by the Royal Society of New Zealand. A profile of David was published in the December 2014 Newsletter (issue 122) of the New Zealand Mathematical Society (NZMS), an organisation that David helped to found in 1974:

`https://nzmathsoc.org.nz/downloads/profiles/NZMSprofile122_David_Vere-Jones.pdf?t=1418807255`

The NZMS also published an earlier profile of David in August 1982 (in Newsletter 24), which recognised his election that year as a Fellow of the Royal Society of New Zealand.

Statistical modelling of slow and fast earthquakes in the Hikurangi Subduction Zone, New Zealand

Jessica Allen, University of Otago
Joint work with Ting Wang, Mark Bebbington, Calum Chamberlain,
Charles Williams, Andrea Perez Silva

2 December, Session A1, Presentation 1 (Student)

The Hikurangi Subduction Zone (HSZ) is a large plate boundary system located beneath the North Island of Aotearoa that generates diverse and frequent activity, including mega thrust earthquakes. Improved catalogues of seismicity and slow slip events, developed using matched filter and wavelet analysis respectively, are paired with statistical methods to decipher the interactions between these behaviours and gain insights into the future of the Hikurangi region. Slow slip events are a kind of slow-motion earthquake, and have been associated with both large earthquakes and seismic swarms in the HSZ. Seismic swarms are spatially and temporally clustered sequences without typical mainshock-aftershock decay. Although the boundary between mainshock-aftershock and swarm activity is not clearly defined, swarms rely less on earthquake to earthquake triggering and arise from diverse underlying processes. In our analysis, relevant seismic sequences are first identified non-parametrically and classified into 2D spatial subregions using Gaussian mixture models. Within these subregions, the mutual information between seismicity and slow slip occurrence times provides evidence of precursory, co-occurring, and subsequent slow slip in relation to seismicity. A parametrisation of the mutually exciting Hawkes process is implemented to determine the strength and direction of the linear causal relationships between the two processes. We also use point processes to model event recurrence patterns across seismic sequences to distinguish mainshock-aftershocks from seismic swarms. The epidemic-type aftershock sequence model captures the triggering relationships of mainshock-aftershock activity, while swarm-like sequences are better described using renewal processes.

Teaching GLMs to undergraduates

Andrew Balemi, University of Auckland

2 December, Session B3, Presentation 1

In this talk we will discuss the approach we take at the UoA to teaching regression techniques to our statistics students. We start with the simple linear regression model – with the approach that we commonly encounter straight lines with continuous and/or factor variables. We address what happens when we don't see straight lines. Once these techniques are mastered we move from the linear model (identity link) to others (log-linear and logistic). We emphasise that the default settings in `R` are for commonly encountered analyses and that we can change them as needed, i.e. we on serve the data/research questions – not the model. I will discuss triumphs and failures/frustrations/concerns in this approach.

Introducing the Reserve Bank of New Zealand's new business expectations survey

Alan Bentley, Reserve Bank of New Zealand
Joint work with Matt Haigh

3 December, Session C1, Presentation 5

As inflation targeting pioneers, New Zealand has conducted quarterly surveying of expectations for more than three decades, eliciting the views of households, forecasters, economists, and industry leaders. Te Pūtea Matua Reserve Bank of New Zealand (RBNZ) is now investing in a new broad-industry Business Expectations Survey, Tara-ā-Umanga, to better understand the economic outlook including the inflation expectations of key decision makers.

Salient design features include a stratified random sample, drawn from Stats NZ's Statistical Business Register, of all businesses operating in New Zealand with 6 or more employees. Fifteen design strata will provide domain coverage for 5 industries (Primary, Manufacturing, Construction, Retail, Other) by 3 employment-sizes strata (6-19, 20-99, 100+ employees). A nine-quarter rotating panel design will increase the precision of estimates of temporal changes and allow some longitudinal analysis, whilst having regard for equitable respondent burden and potential for panel conditioning. Web questionnaire content will include three core questions, asked every quarter, alongside occasional and rotating questions.

We present results from a pilot survey designed to test recruitment strategies and other operational procedures, as well as providing initial estimates of response rates, and population parameters (such as mean and median expectations) for each design strata. We also summarise findings from a public consultation on the survey design and learnings from cognitive testing of questions.

Relatedness estimation in pooled samples sequenced using low-depth high-throughput methods

Timothy Bilton, AgResearch
Joint work with Ken Dodds, Andrew Griffiths

2 December, Session A3, Presentation 1

Relatedness information is central to a number of analyses in genetic applications. One such application is genomic selection, where the variance-covariance component of the random effect terms for the linear mixed model fitted to generate breeding values is proportional to the relatedness. In some breeding applications, samples from several (typically related) individuals are often pooled into a single sample when sequencing to reduce cost or because the phenotyping is at the pooled level. Current methods for estimating relatedness for pooled samples found in the literature do not appropriately account for the structure of the pool when estimating relatedness. In addition, researchers are increasing using high-throughput sequencing (HTS) methods for genotyping. HTS methods provide a low cost and efficient approach for genotyping, but the data generated are subject to errors in the form of miscalled bases and heterozygous genotypes being miscalled as homozygous due to low-read depths. Here, we derive an appropriate estimator of relatedness for pooled samples using HTS data. We present the theory behind the estimator and perform a simulation study to explore the properties of the estimator.

That's questionable: Designing and deploying effective models for generating multiple versions of auto-marked questions

Liza Bolton, University of Auckland
Joint work with Anna Fergusson, Lars Thomsen, Charlotte Jones-Todd

2 December, Session B3, Presentation 2

Creating automatically marked question banks with a number of versions is popular for both supporting academic integrity and for providing low-stakes assessment opportunities with instant feedback. Short quizzes can engage students in checking their understanding throughout a course and support their preparation for higher-stakes assessments. While auto-marking can help reduce teaching team workload, the creation and maintenance of high-quality and fair question banks can be very demanding. To support this, a range of computational tools exist for creating auto-marked questions and deploying them to assessment platforms (e.g., the `R` package `exams`, Grün & Zeileis, 2009). However, while there is guidance for the practical implementation of these tools (e.g., Zeileis et al., 2014), there is very little documentation that explains the design process for developing models that can generate tens or even hundreds of versions of questions. This talk has three aims:

1) to explore design principles that support pedagogy-first approaches to creating question-generating models,

2) to share considerations and opportunities with respect to having students analyse data (with iNZight Lite) to answer quiz questions, and

3) to report on how students are actually using quizzes with multiple versions in a large introductory statistics course, including findings based on data about quiz attempts, as well as reflections from the teaching team.

Investigating the impact of gravitational wave glitches on the parameter estimation of extreme mass ratio inspirals

Amin Boumerdassi, University of Auckland & Laboratoire des 2 Infinis (Toulouse)
Joint work with Ollie Burke, Matt Edwards, Avi Vajpeyi, Ruiting Mao

2 December, Session B1, Presentation 1 (Student)

Extreme mass ratio inspirals (EMRI) are a type of astrophysical event in which a compact and massive object such as two black holes of highly non-equal masses merge into each other. These result in the emission of gravitational waves (GW) – the ripples in spacetime famously observed in 2015 by the LIGO collaboration. Many types of GW events can be parameterised by known astrophysical models, which permits estimation with Bayesian inference and MCMC. EMRIs are no exception to this, however parameter estimation of EMRIs is faced with numerous difficulties. These include things such as the 14-dimensional parameterisation of EMRIs, years-long observation times and highly multimodal likelihoods. One possible challenge for EMRI parameter estimation is the presence of GW glitches – a type of frequent, short-duration, high-amplitude noise event present in GW detectors. They are not true GW events, rather coming from local and often unexplained disturbances to the detector. For many types of GW sources, glitches are known to induce biases in the posterior distribution – however to date, no study has been conducted on the impact of glitches on EMRI parameter estimation. This project aims to investigate whether glitches result in biased EMRI posteriors, and to what degree this may be the case.

Comparisons between functional brain networks for schizophrenia and healthy case controls: A Bayesian network approach

Khan Buchwald, Auckland University of Technology
Joint work with Matthieu Vignes, Richard Siegert, Ajit Narayanan, Margaret Sandham

3 December, Session D2, Presentation 1 (STUDENT)

The dysconnection hypothesis of schizophrenia has been advanced to explain its symptomology from a neurophysiological perspective. Network statistical methods have been used to assess the dysconnection hypothesis but not when varying network complexity. This paper leverages network statistical complexity analysis to evaluate the dysconnection hypothesis in schizophrenia. This study obtained fMRI data from the University of California Los Angles Consortium for Neuropsychiatric Phenomics LA5c Study for people diagnosed with schizophrenia (PDS) and healthy case-controls (HC). We used one hundred bootstrap samples of 45 PDS and 45 HC for encoding and retrieval memory trials and fitted a dynamic Bayesian network to each sample. The network properties were assessed at various significance thresholds, and an optimal significance threshold for an edge's inclusion was obtained. After edge pruning, 52.8% of edges were shared between memory encoding in PDS and memory encoding in HC, and 52.7% between retrieval in PDS and retrieval in HC. This was considerably lower than the shared connections within treatment groups. PDS had a similar number of functional connections as HC. The network metrics (Clustering, shortest path length) were higher or lower in PDS and varied according to the threshold selected. Dynamic Bayesian networks can be used to support the dysconnection hypothesis in PDS, and this research suggests that the absence of shared functional connections between PDS and HC is pivotal rather than a loss or a gain in the number of functional connections.

Is AI eating data scientists' lunch?

Lisa Chen, University of Auckland

4 December, Session E2, Presentation 4

As artificial intelligence (AI) advances rapidly, a growing debate questions whether these innovations are diminishing the role and value of data scientists. Today's AI and automation tools perform tasks traditionally handled by data scientists, such as data cleaning, model development, and advanced analytics, often with impressive speed and efficiency. This raises an important question: is AI eating data scientists' lunch?

In this short talk, I'll share insights from my experiences as both a practitioner and educator in this rapidly evolving field. I'll explore how these changes are reshaping data science practice and influencing my teaching approach. Ultimately, rather than becoming obsolete, data scientists—with their unique combination of 'human' insight and 'data' expertise—are ideally positioned to lead in this AI-driven future.

Where missing values really matter for models in public policy – real life examples

Len Cook, Te Ngira
Joint work with Alistair Gray

4 December, Session E3, Presentation 1

Missing values that result from inadequate model specification may be caused by poor validation of model assumptions, or inferring causation from associations that are anecdotal or appear relevant but result from limitations in the then-available quantitative information. Some examples will be given from public policy, including an analysis of the inappropriate methods to produce counts of the number abused in care by a consulting firm. Why must fitness for purpose meet a tougher test when the results are to be placed in the public domain? Some lessons for Social Investment type models will be suggested.

A Bayesian close-kin mark-recapture model using pedigree reconstruction

Sarah Croft, University of Otago
Joint work with Matthew Schofield, Michael Black, Richard Barker

3 December, Session D2, Presentation 2 (Student)

Close-kin mark-recapture (CKMR) models present an attractive, cost-efficient method to estimate animal population abundance. CKMR adapts the mark-recapture model to incorporate genetic samples, either as a replacement or supplement to physical tags, from captured individuals and identify parent-offspring or sibling pairs within the sample. The current CKMR models require sparse data and are limited to large, non-inbred populations. As a result, these models are not appropriate for use in conservation management of many endangered and at-risk species. We are developing CKMR methods for small populations by reconstructing the population's pedigree from the observed individuals using their genotype and supplementary age data. The true underlying pedigree completely describes the population over time and pedigree reconstruction allows for the estimation of both abundance and population structure without making large population assumptions. In this talk I will present an overview of the Bayesian pedigree reconstruction approach for population estimation using dead recovery data and the challenges associated with the full pedigree approach.

Semi-supervised model-based clustering for ordinal response data

Ying Cui, Victoria University of Wellington
Joint work with Louise McMillan, Ivy Liu

3 December, Session C1, Presentation 2 (Student)

This paper introduces a semi-supervised learning technique for model-based clustering. Our research focus is on applying it to matrices of ordered categorical response data, such as those obtained from surveys with Likert scale responses. We use the proportional odds model, which is popular and widely used for analysing such data, as the model structure. Our proposed technique is designed for analysing datasets that contain both labeled and unlabeled observations from multiple clusters. To evaluate the performance of our proposed model, we conducted a simulation study in which we tested the model from six different scenarios, each with varying combinations and proportions of known and unknown cluster memberships. The fitted models accurately estimate the parameters in most of the designed scenarios, indicating that our technique is effective in clustering partially-labeled data with ordered categorical response variables. To illustrate our approach, we use a real-world dataset from the aquaculture area.

What the Zeta!

James Curran, University of Auckland
Joint work with Patrick Buzzini, Tatiana Trejos

4 December, Session E3, Presentation 4

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.

If you were at the IBS-AR conference last year then you might have heard this talk before.

A nearest-neighbour Gaussian process spatial factor model for censored, multi-depth geochemical data

Tilman M. Davies, University of Otago
Joint work with Sudipto Banerjee, Adam P. Martin, Rose E. Turnbull
Poster Presentation

We investigate the relationships between local environmental variables and the geochemical composition of the Earth in a region spanning over 26,000 km^2 in the lower South Island of New Zealand. Part of the Southland–South Otago geochemical baseline survey – a pilot study pre-empting roll-out across the country – the data comprise the measurements of 59 chemical trace elements, each at two depth prescriptions, at several hundred spatial sites. We demonstrate construction of a hierarchical spatial factor model that captures inter-depth dependency; handles imputation of left-censored readings in a statistically principled manner; and exploits sparse approximations to Gaussian processes to deliver inference. The voluminous results provide a novel impression of the underlying processes and are presented graphically via simple web-based applications. These both confirm existing knowledge and provide a basis from which new research hypotheses in geochemistry might be formed.

Vismi: Visualisation tools for multiple imputation

Yongshi Deng, University of Auckland
Joint work with Thomas Lumley
4 December, Session E3, Presentation 2

Multiple imputation has been widely used to handle missing data. There are an increasing number of software packages available for multiple imputation. However, before proceeding with statistical inference, it is crucial for practitioners to assess the quality of multiply-imputed values, particularly when using multiple imputation implementations based on machine learning algorithms. To help with this, we have developed an `R` package `vismi`, which offers comprehensive visual diagnostics for evaluating multiple imputation generated by different packages. In this talk, we will demonstrate various functions of `vismi`, and showcase its practical utility through case studies.

Cluster analysis of unknown samples using Lumi^{TM} drug scan data

Yongshi Deng, ESR
Joint work with Dion Sheppard
Poster Presentation

Lumi^{TM} Drug Scan provides real-time detection of illicit drugs through a portable Near-infrared device. The initial deployed model identified street samples as methamphetamine, MDMA, cocaine, or unknown. A large portion of samples scanned by NZ frontline police officers falls into the unknown category. It is of interest to uncover what other popular substances may constitute these unknown samples. This analysis can provide meaningful descriptive statistics of the street samples seized by the police and offer valuable insights for future model development. In this poster, we will demonstrate how semi-supervised clustering can be used to achieve these goals.

Novel applications of linked administrative data – adding longitudinal capability to the Te Kupenga survey

Tori Diamond, University of Auckland

3 December, Session C1, Presentation 1 (Student)

Can linked administrative data be used to transform New Zealand's only sample survey on Māori wellbeing into a longitudinal study?

This project extends the usefulness of an important survey dataset by linkage to admin data, effectively adding longitudinal capability within a linked administrative data source. This created robust statistical processes to transform an official statistics survey into a nationally representative cohort study. StatsNZ's Integrated Data Infrastructure (IDI) holds administrative and survey datasets containing a range of variables linkable at the individual level. Te Kupenga is a large nationally representative post-censal survey of the Māori population and is the only survey with culturally informed variables but is often under-utilised in research.

The Te Kupenga survey was used as a foundational cohort linking to outcomes and determinants in different datasets from different time periods. Outcomes included Ambulatory Sensitive Hospitalisations (ASH) (post-2013) and COVID-19 vaccinations (post-2020), while determinants included individual, household and geographic variables. Linking a representative survey to admin data created issues of loss to follow-up and missing data, so the original sample is not maintained after linkage. Loss to follow-up and missingness differed depending on variable selection and time periods. So, new universally applicable weights were not possible. However, we created a robust, generally applicable process for re-weighting survey data to account for missingness and loss to follow-up in admin data.

This project demonstrates the approach for turning a sample survey into a longitudinal cohort using admin data and creates methods that can be used for other official statistics surveys.

Ordinal pattern analysis for early bearing fault detection and classification in rotating machinery

Rasika Dilhani, Victoria University of Wellington
Joint work with Alejandro C. Frery

Poster Presentation (Student)

Bearings are critical components in rotating machines, but their demanding operating conditions with high loads and shocks often lead to various failures. These errors can result in significant downtime, costly maintenance, and even complete machine failure. Therefore, early and accurate detection and classification of bearing defects is critical to ensuring operational safety and minimizing maintenance costs. Traditional fault detection methods are mainly based on analyzing physical parameters and trends using vibration, thermal monitoring and current signature analysis techniques. Although these methods have been proven effective, they can be prone to interference and often require significant computational resources. Ordinal pattern analysis has emerged as a promising alternative, providing a robust and computationally efficient approach to analysing time series data. At its core, ordinal pattern analysis involves converting continuous time series data into a sequence of symbols that represent ordering relationships among data points within a specific time window. This approach effectively captures the fundamental dynamics of the signal in a way that is inherently robust to noise and distortion. By analysing these patterns, it becomes possible to identify subtle changes in system behaviour that may indicate the presence of an error. This study investigates the application of ordinal pattern analysis for early detection and classification of bearing defects, focusing on common defect types such as ball, outer ring, and inner ring defects. Using a publicly available dataset from the Case Western Reserve University Bearing Data Center, we demonstrate the effectiveness of ordinal patterns in distinguishing between healthy and failing bearing conditions.

Variance estimators for mixed-effects proportional hazards models fitted to complex samples

Bradley Drayton, University of Auckland

3 December, Session C1, Presentation 3 (Student)

The mixed-effects proportional hazards model for complex samples is designed to analyse correlated time-to-event data collected through complex sampling methods. A significant challenge in this context is variance estimation, which becomes complicated due to model misspecification from sampling weights and cluster correlation from data generation or sampling processes. The current variance estimator, which relies on the information matrix, tends to underestimate variance.

To address this challenge, I developed a robust sandwich estimator and utilised resampling-based variance estimators from the `R survey` package to create five new variance estimators for the fixed effects in these models. These new estimators are included in the `svycoxme` package. I evaluated these new variance estimators against the information-based estimator using two sampling schemes: simple random cluster sampling and stratified, multi-stage sampling. The cluster-level jackknife method performed the best, while the multistage rescaled bootstrap, sandwich estimator, and information-based estimator also showed acceptable performance in many scenarios. Interestingly, the bootstrap method showed unexpected results, with undercoverage for cluster-level effects. We explore the reasons behind this and suggest directions for future research in this area.

Variational autoencoders for stellar core-collapse gravitational waves

Tarin Eccleston, University of Auckland
Joint work with Matt Edwards
2 December, Session B1, Presentation 2 (Student)

We present work towards a rapid stellar core-collapse waveform emulator using a variational autoencoder (VAE) – a follow-up from our previous work on using deep convolutional generative adversarial networks (DCGANs). The main advantage of using VAEs over DCGANs is that they provide a smoother and well-structured latent space representation by assuming a specific prior distribution. VAEs also allow us to perform variational inference and are less prone to mode collapse and unstable training. The pre-trained VAE will be used in match-filtering analysis to detect gravitational wave signals from stellar core-collapse events.

A novel stacked hybrid autoencoder for imputing data gaps in the Laser Interferometer Space Antenna

Matt Edwards, University of Auckland
Joint work with Ruiting Mao, Kate Lee
2 December, Session B1, Presentation 4

The Laser Interferometer Space Antenna (LISA) data stream will contain gaps with missing or unusable data due to antenna repointing, orbital corrections, instrument malfunctions, and unknown random processes. We introduce a new deep learning model to impute data gaps in the LISA data stream. The stacked hybrid autoencoder combines a denoising convolutional autoencoder (DCAE) with a bi-directional gated recurrent unit (BiGRU). The DCAE is used to extract relevant features in the corrupted data, while the BiGRU captures the temporal dynamics of the gravitational-wave signals. We show for a massive black hole binary signal, corrupted by data gaps of various number and duration, that we yield an overlap of greater than 99.9% when the gaps do not occur in the merging phase, and greater than 98% when the gaps do occur in the merging phase. However, if data gaps occur during merger time, we show that we get biased astrophysical parameter estimates, highlighting the need for protected periods.

Predicting the size of any species' genome using the taxonomy tree

Joane Elleouet, Scion
Joint work with Céline Mercier
4 December, Session E1, Presentation 2

Many biological applications require or benefit from knowing the size of the genome of organisms present in an environment. This information provides many insights into the evolution of these organisms in response to ecological processes. Human knowledge of species' genomes has recently increased exponentially and is neatly stored in the well-organised databases of the National Center for Biotechnology Information (NCBI). This enabled the development of the `genomesizeR` package, a new tool to estimate the size of the genome of any fully or partially taxonomically identified organism. Here we describe the statistical models created for this tool. We highlight insightful aspects of statistical model development and challenges associated with highly hierarchical models, touching on the topic of distributional models. We also describe a model validation process allowing to compare strengths and weaknesses of frequentist and Bayesian statistical approaches as well as a non-model-based algorithm.

Introducing rserve-ts: a modernised library for R-to-web communication

Tom Elliott, iNZight Analytics Ltd

4 December, Session E2, Presentation 1

Web front-ends are becoming increasingly popular for building applications due to their ability to deploy seamlessly across both desktop and mobile devices, and a wide range of tools, libraries and frameworks exist that allow developers to create all kinds of applications – RStudio, for example.

`Rserve` is an `R` library providing two-way communication between `R` and other environments – on the web, this means Javascript. However, the `rserve-js` library has become outdated and no longer fits neatly into modern web development ecosystems, making it a difficult pitch to encourage uptake and wider usage.

I have been working on a new ecosystem for integrating `R` into modern web applications that builds on `Rserve` by adding in TypeScript support and other modern patterns. This talk introduces some of my early work updating the `rserve-js` library, demonstrated with a simple React application.

Synthetic bridge over troubled data

Jamas Enright, Stats NZ

4 December, Session F1, Presentation 1

The recent COVID pandemic disrupted a lot of time series. For New Zealand, the impact was limited to either one period or just a few years. With the series returning to pre-COVID levels, we needed a way to side-step the COVID impacted data, to maintain the seasonal patterns from pre-COVID to the post-COVID period. This talk will explain how we create a synthetic bridge that we use in our seasonal decomposition system to replace the disrupted data with data to maintain that seasonal pattern.

Lost (and found) in translation: Examining the diversity and impact of languages selected on student responses to a statistical investigation of automated language translation

Anna Fergusson, University of Auckland
Joint work with Lars Thomsen, Anne Patel

2 December, Session B3, Presentation 4

The data used for teaching statistics is often far removed from students' lives, limiting the ability for students to make personal connections in their responses to learning tasks. As part of a broader initiative to re-design a large introductory statistics course at the University of Auckland, new educational technologies were created to provide greater support for personalised learning. These "data landscapes" utilise APIs to access large data sets and online databases, and require students to make selections to generate the data used in assignment tasks. In one such task, nearly 2000 students were given headlines from a prominent newspaper and asked to select two non-English languages. The data landscape translated the headlines from English into the two selected languages and then back to English, with a similarity score provided for each "round trip translation". Students then used statistical methods to evaluate and compare the use of Google Translate for the two selected languages, before writing a conclusion and reflection on their findings. Analysis of the student responses to the assignment task found that a diverse range of languages were selected, leading to diversity in learning experiences and student responses. However, the impact of using data landscapes was minimal in terms of marking and scores allocated, suggesting that it is feasible to provide personalisation within large scale assessment contexts. The study also suggests that data landscape tasks could enrich the learning experience by providing an opportunity for students to make connections to the data.

Ordinal Patterns, features and their distribution

Alejandro C. Frery, Victoria University of Wellington
Joint work with Andrea Rey, Juliana Gambini, Magdalena Lucini

4 December, Session F3, Presentation 1

In this presentation, we will explore the concept of ordinal patterns and their significance in analysing time series data. Ordinal patterns encode the ranking properties of small vectors, allowing us to transform real-valued time series into sequences of symbols. This transformation simplifies the analysis by focusing on the order of data rather than its actual values. We will discuss the methodology proposed by Bandt and Pompe, which utilizes two key features: permutation entropy and statistical complexity. These features enable us to assess the unpredictability and structural properties of the data, providing valuable insights into the underlying dynamics of complex systems. The application of ordinal patterns spans various fields, including economics, text analysis, neuroscience, and climatology. By examining the marginal and transition properties of these symbols, we can uncover hidden patterns and trends in diverse datasets. Finally, we will highlight recent developments in the statistical properties of ordinal patterns, showcasing their potential for advancing our understanding of complex systems.

The algebra and geometry of Markov bases

Linus Fromm, University of Otago
Joint work with Martin Hazelton
Poster Presentation (Student)

Linear statistical inverse problems can be found in many branches of science. This includes but is not limited to ecology, genetics and network tomography. The aim is to conduct inference on the distribution of some latent variable of interest conditional on corrupted or aggregated observations. This involves the evaluation of a relatively large sum which is infeasible in most circumstances. For that reason, we turn to Markov chain Monte-Carlo (MCMC) sampling. MCMC samplers over discrete data require the use of sets of moves called Markov bases. It is described how to find Markov bases and how the process of finding Markov bases is fundamentally connected to division of multivariate polynomials. We aim to give a geometric intuition for the fundamental theorem of Markov bases which was first proven by Diaconis and Sturmfels in 1998. We will then show that the division algorithm of polynomials is susceptible to changes in monomial orderings. These differences carry over to the bases found using the fundamental theorem of Markov bases. Finally, we will explore how different monomial orderings influence the convergence of samplers.

Dissimilarity measures for time series

Ciprian Doru Giurcăneanu, University of Auckland
Joint work with Miaotian Li
2 December, Session B2, Presentation 4

The problem of measuring the dissimilarity between time series has been discussed in numerous works. It has risen to prominence during recent years when an impressive amount of time series data became available. Various classifications of the existing methods have been proposed, but for the sake of simplicity, we consider the classification that is based on the domain where the dissimilarity is assessed: time domain, frequency domain or cepstrum domain. It is interesting that the cepstral dissimilarity measures are well known in computer science, signal processing and control engineering, but they are less known in statistics. This motivates us to focus on the formulas of the metrics that involve the cepstral coefficients. There are empirical studies which show that, for example, the clustering of time series improves when some of the cepstral coefficients are replaced with zeros in the formulas mentioned above. In this talk, we present principled methods for cepstral nulling and assess their impact on the evaluation of the cepstral metrics. The presentation encompasses novel theoretical results that are illustrated via numerical examples.

The subjective wellbeing of first-in-family university students: A multivariate re-evaluation of common narratives

Adam Glucksman, Victoria University of Wellington
Joint work with Philip Morrison, Louise McMillan

3 December, Session D1, Presentation 2 (Student)

As universities compete for more and more students the proportion of students for whom neither parent has a university degree rises. Many overseas scholars see these ‘first-in-family (FiF)’ students as a risk, pointing to their lower wellbeing and higher attrition rates. This study asks whether New Zealand FiF students also return greater psychological distress than their peers whose parents have university backgrounds.

I model wellbeing outcomes among first-year students at Victoria University of Wellington as collected by the YOU Student Wellbeing Survey in 2019, 2020 and 2021 (n = 4,000). I employ a range of statistical models suited for both continuous and binary response variables, with methods designed to test various interactions and covariates across wellbeing outcomes. Key controls include age, sex, ethnicity, socioeconomic background, and levels of family support. Analytical approaches focus on evaluating the impact of FiF status, both as a main effect and in interaction with covariates, on dimensions of psychological wellbeing.

My results indicate that, after adjusting for covariates, FiF status alone does not have a statistically significant impact on any of the wellbeing measures. By demonstrating that FiF status itself may not be a primary driver of wellbeing differences, this study invites a re-evaluation of common overseas narratives regarding FiF students.

In addition to providing a data-driven perspective on the subjective wellbeing of FiF students, this study underscores the value of interaction-focused analysis in educational research. This allows for a more comprehensive understanding of how demographic and socioeconomic factors jointly influence wellbeing, thus enriching the field’s approach to understanding diverse student populations.

Motivational resources for training in statistics

John Harraway, University of Otago

2 December, Session B3, Presentation 3

Two sets of resources are discussed and links provided for both. The first set comprises 20 recent motivational case study videos about applications of statistics in research from many Otago departments. The videos can be located at:

`https://www.stats.otago.ac.nz/research/Statistics-in-Research/`

Data sets for each study are freely available and investigated using `R` to analyse the data, thus combining statistical ideas in current case studies and first experiences with coding at the same time in our large first year statistics class. The videos are being used in 60 countries. An invited paper at the ICOTS11 Conference (September 2022) in Argentina provides more detail.

The second set of resources comprises three apps for training in Official Statistics and can be found at:

`https://iase-web.org/islp/Resources.php?p=Apps`

These apps are interactive and free to use with concept discussion, references, and marked assessments. Indications are these are also being used in 60 countries. Google docs are being developed for attaching to each video and each app in an attempt to identify exactly who is using these resources in each country, and how they are being used. For example, for training in the workforce, for teaching statistics and programming in the classroom, or for the statistics education of a population. All the material is open source and the resources can be used during Covid (for example), with access both online and in the classroom.

Using spatio-temporal models to investigate stock structure, seasonality and environment influences in the snapper population from the west coast of New Zealand

Oxana Hart, Victoria University of Wellington
Joint work with Arnaud Grüss, Nokuthaba Sibanda, Adam Langley, Matthew Pinkerton

2 December, Session A1, Presentation 2 (Student)

Spatio-temporal modelling is a valuable geostatistical approach that is increasingly being used for fish populations, which accounts for both spatial and spatio-temporal autocorrelation/structure in the data at a very fine scale. As such, spatio-temporal models have the potential to account for a lot of the unmeasured variation in the data. Spatio-temporal models can also include environmental covariates to represent environmental influences on fish density and/or catchability covariates to account for confounding variables affecting fish catchability (detectability).

Our research investigates stock structure, seasonality, and environmental influences in the snapper population from the west coast of New Zealand, using spatio-temporal models fitted to commercial bottom trawl catch-per-unit-effort data collected between 2008 and 2022. Spatio-temporal models provide us with indices of relative abundance and estimates of population range and boundaries for subregions of the west coast of New Zealand, and inform us about spatial patterns of median log-density and interannual variability along the well coast, as well as about "core" and "transition" areas for snapper in the study region. All this information allows us to better understand seasonality, potential seasonal migration and stock structure in the snapper population from the west coast of New Zealand.

Our modelling framework for snapper allows us to enhance understanding of snapper ecology and provides important information to assist snapper population assessments. Our research demonstrates the value of spatio-temporal models in ecological research and illustrates how geostatistical methods can be employed to address complex issues in marine science.

Model-based priors for network tomography

Martin Hazelton, University of Otago

4 December, Session E1, Presentation 3

Network tomography is a challenging type of statistical linear inverse problem. A common example arises in transport engineering, where the goal is to estimate volumes of origin-destination traffic flow based on traffic counts observed at various sites over the network. In that context, the difficulties for statistical inference are exacerbated by the existence of multiple plausible routes connecting most origins and destinations of travel. The observed data provide limited information about the route choice probabilities, and so the availability of an informative prior is critical. In this talk I describe how such a prior can be constructed using classical route choice models founded on game theory. Such models are computationally expensive, and so I also discuss the use of cheap emulators.

Extending spatial capture-recapture with the Hawkes process

Alec van Helsdingen, University of Auckland
Joint work with Charlotte Jones-Todd, Russell Millar

2 December, Session A1, Presentation 3 (Student)

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 independent and consequently cannot incorporate site fidelity (attachment to a particular region) nor the temporal correlation of an animal's location. Recent work has sought to solve these issues by explicitly modelling animal movement. In this talk we propose an alternative solution for camera trapping surveys 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 where and when the animal was most recently detected. Through a mixture of Gaussian distributions, our model expects more detections closer in space to the last detection, and reduces to SCR when an animal is yet to be detected. This formulation, we believe, better reflects animal behaviour because shortly after detection, we expect to next see an individual close to where it was last seen. Thus, our model 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 an overview of Self-Exciting Spatial Capture-Recapture (SESCR) models,

2) demonstrate the additional inference that can be drawn from such models, and

3) apply the framework using a few case studies to compare traditional SCR and SESCR.

Forecasting multiple time series with graph convolutional networks

Guoping Hu, University of Auckland
Joint work with Ciprian Doru Giurcăneanu

2 December, Session B2, Presentation 1 (Student)

Forecasting multiple time series at different levels is often required in many situations, which is commonly known as hierarchical time series forecasting. Supply chain management is a typical application that requires demand forecasting at the store, city, or country level for decision-making. In hierarchical forecasting, top-down, bottom-up, and optimal linear combination methods are the most common methods. While top-down and bottom-up methods use only information from the top and bottom levels, respectively, linear combination methods use individual forecasts from all series and levels and combine them linearly, often outperforming traditional top-down and bottom-up methods. Despite this, these approaches do not make use of the explanatory information that may exist at various levels of the hierarchy directly. In addition to producing accurate forecasts, it is necessary to select a suitable method to generate basic and reconciled forecasts simultaneously. Prediction reconciliation involves adjusting predictions to be consistent across different levels. In this talk, we present a neural network model that utilizes graph convolutional neural networks, recurrent neural networks, and fully connected neural networks to generate accurate and reconciled predictions directly. Specifically, we first use graph convolutional neural networks to extract hierarchical information, then recurrent neural networks to extract the temporal dependencies of all time series in the hierarchy, and finally, train a fully connected neural network to minimize the loss function to generate reconciled forecasts.

A virtual experiment to teach experimental design

Charlotte Jones-Todd, University of Auckland

Poster Presentation

A problem faced by many (bio)statistics students is linking textbook scenarios to real-world examples. This can cause a disconnect between statistical theory and application. For example, students can outline the methodology but struggle to implement it in real-world scenarios, a requirement of their future careers.

One example of this is experimental design where students are often able to discuss the principles of experimental design, but struggle to link and employ them to the design of an experiment and the analysis of real-world data. Having no "hands on" experience of the data collection process distances students from the examples underpinning the statistical concepts.

As much as we might like to provide this "hands on" experience, in reality, it is impossible to do so in a single semester. Therefore, I developed a virtual alternative: an educational game that provides students with the experience of designing an experiment and collecting their own unique dataset in a cost- and hassle-free way.

The game is designed to scaffold the learning process by incorporating interactive checkpoints so that students can simultaneously design their experiment and be immersed In the mechanisms and terms of experimental design. To foster engagement the game is designed with a motivating backstory and a balanced aesthetic. Here you'll see the design and layout of the game and, of course, get an opportunity to play it! Welcome to Farm Rescue: The Tomato Trials!

Accounting for social networks in a self-exciting point process model

Charlotte Jones-Todd, University of Auckland

Joint work with Conor Kresin, University of Otago

4 December, Session E1, Presentation 4

A Hawkes process models self-exciting behaviour in temporal point pattern data. Self-excitement is where the occurrence of an event increases (i.e., excites) the chance of another in quick succession. This phenomenon is apparent in many crime-based event data (e.g., gun and knife violence). In addition to the occurrence of these crime events there are often known links between them. For example, crimes may be linked due to 1) the same perpetrators, or 2) co-perpetrators of another crime being involved. This network of linked crimes evolves and expands over time (e.g., via co-perpetration) and distinct clusters of events form. We develop a Hawkes model that incorporates time-evolving social network dynamics accounting for the dependency between events linked by a given social network.

Interactive visualization of suicide methods in Tokyo, Japan

Takafumi Kubota, Tama University
Joint work with Takahiro Arai

4 December, Session E2, Presentation 2

This study aims to visualize the trends in suicide methods in Tokyo, Japan, using Japan's regional suicide statistics, "Basic Data on Suicide in Local Communities", to provide insights that can inform effective prevention strategies. Suicide is a significant social issue, and analyzing regional data can offer valuable perspectives for targeted interventions. This study focuses on visualizing the trends for different suicide methods using interactive graphs, scatter plots, histograms, parallel coordinate plots, and choropleth maps. These visualizations are generated after data cleaning to depict the occurrence and trends associated with each method.

This application is developed using the `R` packages `Shiny` and `Plotly`. It enables users to explore the data interactively and highlight concern regions between graphs. With `Shiny`, users can select the items of interest, such as region, period, or suicide method, from a menu, while `Plotly` allows for the implementation of interactive graphs that dynamically update based on the selected parameters. This approach facilitates the identification of specific regional trends, such as railway suicides or jumps from high-rise buildings that are more prevalent in Tokyo.

Through the development and analysis of this application, this study aims to enhance the understanding of regional and method-specific suicide trends, providing recommendations for suicide prevention measures. The visualized data is expected to serve as a valuable tool for policymakers and researchers, contributing to the strengthening of suicide prevention efforts.

Speeding up design in Bayesian platform trials

Thomas Lumley, University of Auckland

3 December, Session D2, Presentation 3

Bayesian platform trials are now becoming popular. Although these trials have results and stopping guidelines specified in Bayesian terms, sponsors (in industry or the public sector) often want to know operating characteristics such as probability of finding an effective treatment if there is one, probability of falsely declaring a treatment effective, and probability of continuing for longer than a specified time period with no results. Estimating these characteristics requires simulation. I will present ways to speed up this simulation in `R` and `Stan` using approximations to the posterior and using ratio estimation.

Improving minimum contrast for clustered processes

Bethany Macdonald, University of Otago
Joint work with Tilman Davies, Martin Hazelton
4 December, Session E1, Presentation 1

Spatial point patterns can arise from a vast array of application areas including epidemiology, ecology and geoscience. Of special interest are clustered processes, such as the log-Gaussian Cox process and Neyman-Scott processes. In such models we are interested in estimation of the 'cluster' parameters which describe the behaviour between points.

Minimum contrast is a popular estimation method for cluster processes and related models where exact likelihood based methods are unavailable. This procedure is essentially a method of moments, which involves minimising the differences between a theoretical summary statistic and its non-parametric equivalent. The pair correlation function (PCF) is a popular choice for the summary statistic and is typically estimated using kernel smoothing. However, the kernel estimate is not an unbiased estimator for the theoretical PCF and leads to biased parameter estimates. Additionally, the empirical estimate must be scaled by the squared intensity which is unknown and replaced by an estimator. For clustered processes, the standard estimator introduces further bias. We present a number of improvements to the minimum contrast procedure for clustered processes.

An explainable disease risk prediction model based on deep transfer learning using high-dimensional genomic data

Qingyu Meng, University of Auckland
2 December, Session A2, Presentation 1 (Student)

It is crucial to accurately develop a disease risk prediction model in the pursuit of precision medicine. High-dimensional genomic data sources offer valuable insights into biostatistical fields but present significant analytical challenges due to extensive noise and complex relationships. Deep learning has emerged as a leading approach in various areas, such as computer vision, natural language processing, and speech recognition. The state-of-art framework holds promise for genomic data analysis. However, these models often struggle with the curse of dimensionality and lack of biological interpretability, limiting their effectiveness.

In this study, we introduced a deep neural network (DNN)-based framework for prediction modelling. Firstly, we implement feature selection using a newly proposed groupwise feature importance score. The score can efficiently identify genes with both linear and non-linear genetic variant effects. Then, we developed an explainable transfer-learning DNN method that directly integrates feature selection information and achieves downstream analysis. The technique is a stack-style network, so it is compatible with some other feature selections that can be used. Additionally, our DNN framework is biologically interpretable, focusing on selected predictive genes, computationally efficient, and suitable for genome-wide data. Our method demonstrated superior performance in detecting predictive features and predicting disease risk through extensive simulations and real data applications compared to existing methods.

A tool to evaluate environmental trend estimation methods

Justin Murphy, Stats NZ
Joint work with Tiana Whitehead, Pubudu Senanayake

4 December, Session F1, Presentation 3

Accurately measuring environmental trends is crucial for informing the effectiveness of environmental management policies and identifying where resources need to be allocated. Stats NZ were commissioned by the Ministry for the Environment to assess current and alternative methods for environmental trend estimation. To do this, we created a `Shiny` tool which enables the user to create realistic synthetic data by introducing a known trend to a dataset that retains the other properties of the data, such as noise profiles. This makes it possible to determine how well trend estimation methods can recover trends present in the data.

The tool works by decomposing real environmental time series into trend, seasonal, and noise components, and allows users to replace the estimated trend component with a customizable trend, choosing from multiple patterns. The customizable trend is then recombined with the original seasonal and noise components, producing a synthetic time series. In addition, users can scale the magnitude of the seasonal and noise components and can set the time windows in which trends are to be estimated. The `Shiny` tool facilitates a comparison of the accuracy of trend estimation methods under different conditions (e.g., when the true underlying trend is linear vs non-linear), and the sensitivity of the methods to the magnitude of the seasonal and noise components, and to the time-window chosen for analysis.

Evaluating the predictive performance of regression and machine learning models for rare events: A simulation study on structural recurrence in thyroid cancer

Sajeeka Nanayakkara, University of Otago
Joint work with Jiaxu Zeng, Robin Turner, Matthew Parry and Mark Sywak

2 December, Session B2, Presentation 2 (STUDENT)

In clinical settings, risk prediction models are crucial for aiding decision-making. Various methods, including regression-based and machine learning approaches, are used to develop these models, yet the selection of the most appropriate method remains a challenge. We performed a simulation study to examine the influence of the data-generating process on the relative predictive performance of regression-based and machine learning methods for predicting structural recurrence in thyroid cancer patients, particularly with low event occurrences. The dataset included patients from the endocrine surgical unit of the University of Sydney (2000–2018), with 13 predictors such as demographics, tumour, and lymph-node characteristics. We used eight different methods, including logistic regression with all selected predictors, backward elimination, shrinkage methods (LASSO, ridge, and elastic net), random forests, gradient boosting machines and extreme gradient boosting for data-generating process to simulate outcomes. For each data-generating process, training and test samples were generated through resampling with replacement from the original dataset, with training sample sizes of 500, 2000 and 10,000, and a large test dataset consisting of 100,000 samples. The eight methods were applied for model training on each simulated training sample, and performance was evaluated on the test sample using c-statistics, calibration slope, integrated calibration index, and prediction errors. Shrinkage methods outperformed other methods across all data-generating processes, demonstrating robustness in handling rare events. Conversely, random forests performed poorly, with overfitting issues becoming more pronounced as training sample sizes increased. This study emphasizes the importance of selecting predictive models based on data characteristics and event occurrence.

Methodology for Māori descent in the 2023 Census

Amanda O'Connell, James Maguire, Roimata Timutimu, Stats NZ & TKR
Joint work with Tieta Vesty, Meenu Jose, Pip Bennett

3 December, Session D3, Presentation 1

The Māori descent variables created in the census are some of our highest priority outputs. They help to inform decisions that affect iwi and Māori, to support the aspirations of iwi and Māori, and to calculate the Māori electoral populations. Because of its importance, all usual residents counted in the census need to have a Māori descent value - but not everyone responds. Stats NZ has worked in partnership with Te Kāhui Raraunga Charitable Trust, under the Mana Ōrite Relationship Agreement (MŌRA), to develop a method which accounts for these missing responses. We start with 2023 Census responses and fill in the gaps using historical census data, data collected by other government agencies, parental values and statistical imputation. Using these additional sources has helped us produce quality Māori descent outputs for our users.

The method used for Māori descent was developed alongside the method used for Iwi affiliation in the 2023 Census, which will be discussed in the presentation *Methodology for iwi affiliation in the 2023 Census* by Walter Somerville, James Maguire and Roimata Timutimu.

Investigating the effect of repeated patients in surgical outcomes research

Soyoon Annie Park, University of Auckland

3 December, Session C2, Presentation 1 (STUDENT)

Evaluating post-operative outcomes using retrospectively collected health data often involves patients with multiple eligible operations. Including all operations can violate statistical independence assumptions. A common solution is to include only the first operation. In a retrospective audit of postoperative outcomes, we found significant variation in mortality rates based on how repeat patients were accounted for. Specifically, our first cohort had a 90-day mortality rate of 3.6% when selecting the first operations, compared to 4.2% with random selection. Another cohort had a mortality rate of 2.8%, compared to 3.4% with random selection.

Accurate reporting and analysis of post-operative outcomes is highly important in the field of surgical and peri-operative medicine, with mortality serving as a gold-standard objective measure for assessing performance. However, in longitudinal data sets where individuals may be exposed to an operation of interest multiple times, there is no clear guidance on how to select an index event. Major surgical outcome studies often use the first operation selection approach, possibly to minimise bias. This approach may underestimate mortality risk, as patients undergoing frequent surgeries are less likely to die during the first contact. This effect may vary with the study period, leading to inconsistent bias. To our knowledge, no research has examined the impact of different methods for handling repeat patients on operative mortality. This study investigates these methods using health data from the Ministry of Health and Te Whatu Ora Te Toka Tumai and aims to refine methods to accurately capture the true mortality risk associated with surgical procedures.

Identifying influential factors impacting trajectories of anxiety in adolescents experiencing adverse events

Priya Parmar, University of Auckland
Joint work with Avinesh Pillai, Ben Fletcher, Denise Neumann

3 December, Session D1, Presentation 3

A two-fold presentation exploring the learnings from (1) real-time real-world teaching applications of longitudinal analyses to (2) examine the longitudinal trajectories of anxiety 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 from the Growing Up in New Zealand study.

We reflect upon the outcome of both applied analyst [student] and "client" [clinical researcher] from providing a concurrent collaborative analytic experience 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.

A statistician in the public sector

Gabriele Frigerio Porta, Ministry of Education

3 December, Session C1, Presentation 4

Government agencies produce publications and reports that serve a range of purposes, from supporting the decision-making process to monitoring trends of various aspects of the country. These analyses are based on public data; hence the public sector is home to many statisticians or data professionals. This is an introduction to the multi-faceted work of a statistician in the Ministry of Education. Tertiary System Performance Analysis (TSPA) is a data team embedded in the Tertiary Education policy group at the Ministry of Education. TSPA is responsible for producing a range of publications, resources, and reports on tertiary education that are made available for the public. The team also provides data and insights to support policy work and for the monitoring of key indicators. The team publishes data for public use (on Education Counts) and supplies it to Statistics New Zealand. From data analysis to trend forecasts, from requests and OIAs to costing and research, we will present several aspects of the daily life of a statistician in the public sector.

Navigating the maze of multi-omics methods for classification including microbiome data

Mario Prado, AgResearch

3 December, Session C3, Presentation 1 (STUDENT)

Recent technological advancements have enabled the generation of various types of "omic" datasets, such as genomics, transcriptomics, proteomics, and metabolomics, which facilitates the discovery of new connections and functions in complex biological systems. Multi-omic analysis allows to study samples at a holistic level by integrating different data types. Here, we focus on multi-omics analysis that include microbiomes in the omics cascade, as they represent a fundamental part to the correct functioning of the ecosystem they inhabit.

The number of software methods available for multi-omics analysis has expanded exponentially with the growth of omic datasets, where these tools all have different assumptions, algorithms and are designed for specific applications. Consequently, selection of a suitable tool can be confusing for researchers, while methods for multi-omics analysis with microbiome data are lacking. We provide a review of the classification methods publicly available to integrate two or more omic layers (N-integration) for classification that supports microbiome data as input. As part of this review, we also evaluate a selection of these tools on real data sets – human gut, soil composition, coral holobiont, among others – to understand their performance across various factors (e.g., types of microbiomes, degree of group separation, sample size, number of omic layers). We found that classification tasks can benefit from a multi-omic analysis, their accuracy increased compared to a reductionist perspective. Additionally, we were able to compare the tools' different approaches and how they react to microbiome data as new multi-omic input and its characteristics.

Understanding the lived experiences of impostor phenomenon among leaders in Mathematics and Statistics

Sam Preston, University of Canterbury Business School

3 December, Session D1, Presentation 1 (Student)

Despite clear indicators of success, many leaders experience persistent feelings of inadequacy, question themselves, and fear being exposed as frauds. This is referred to as the Impostor Phenomenon (IP), and it can impact the well-being of leaders and organisations.

In this study, I researched academic leaders in mathematics and statistics and explored their lived experiences of the impostor phenomenon. Existing research has relied mainly on quantitative approaches to understand the prevalence of the impostor phenomenon. Quantitative approaches may overlook subtle contextual factors that shape leaders' experiences, and in some studies, the focus has only been on inherent traits and their role in shaping IP. Thus, this study used an inductive, qualitative research design to explore the lived experiences among leaders in mathematics and statistics in New Zealand to achieve a more nuanced understanding.

To gather data, I conducted 16 semi-structured interviews with professionals who hold or previously held academic leadership positions. Qualitative coding and thematic analysis (research methods) revealed three themes associated with the impostor phenomenon experience. These themes represented the impostor phenomenon as having 1) an essence, 2) some critical contextual considerations, and 3) some alleviating factors.

This study contributes to the literature on the impostor phenomenon by exploring lived experiences, providing a more nuanced understanding of the 'impostor phenomenon' construct. The study provides insight into how mathematics and statistics leaders can mitigate adverse experiences of impostor phenomenon and how creating an overall sense of workplace well-being should be a top priority for organisations.

Statistical evaluation of neural network Hawkes processes

Matt Pyper, University of Otago
Joint work with Conor Kresin

Poster Presentation (Student)

Hawkes processes are widely used to model point process data that exhibit self-excitation or inhibition, with applications spanning seismic activity, high-frequency stock trading, contagious disease spread, and neuronal activity. While Hawkes models provide interpretable parameters, they are often inflexible and computationally expensive to fit. Recently, continuous-time long short-term memory (LSTM) neural networks have been used to estimate the conditional intensity functions of Hawkes processes, enhancing their expressive power. This work presents a toolkit to assess the goodness of fit for neural Hawkes processes and explore the benefits of increased expressivity in a semi-parametric framework. Specifically, a KS-test based on the random time change theorem and residual analysis techniques commonly used for point processes are implemented. Additionally, cost function comparisons are described, including implementations of likelihood, least squares, and the Stoyan-Grabarnik estimator.

Trends in current smoking status at state-age-sex level in Australia: an application of multinomial multilevel time-series modelling

Alice Richardson, Statistical Support Network, Australian National University
Joint work with Sumonkanti Das, Ashis Talukder, Bernard Baffour

4 December, Session F2, Presentation 2

Despite significant declines in smoking rates in Australia in recent decades, smoking remains a leading cause of preventable diseases and death. While the proportion of current smokers has steadily declined and the proportion of never-smokers has increased, no clear trend has emerged for ex-smokers. Importantly, these shifts have not occurred uniformly across all geo-demographic groups. This study aims to estimate trends in current smoking status across small domains defined by seven age groups, two genders, and eight states and territories from 2001–2022.

Direct estimates of smoking status for the small domains were derived from micro-data collected in eight rounds of the Australian National Drug Strategy Household Survey. These estimates were then used to develop multinomial multilevel time-series models within a hierarchical Bayesian framework, employing small-area estimation techniques. The developed models borrow cross-sectional and temporal strengths at various aggregation levels, ensuring numerically consistent trend estimates. Model-based trends for higher aggregation levels are obtained by calculating weighted averages of the detailed level trend predictions. The models are developed using ex-smokers as the reference category, as their proportion remains relatively stable over time. Statistically significant random intercepts and random slopes for linear time trends were identified at the state-age-sex level. Temporal random effects at the state-sex and age-sex levels also substantially contributed to achieving numerically consistent trend estimates.

Findings from this study help identify geo-demographic groups that remain behind in achieving smoking cessation targets. These insights can guide health researchers and policymakers in delivering targeted programs to the most vulnerable groups.

Sampling small older populations: Methods and challenges of a dementia prevalence study

Claudia Rivera-Rodriguez, University of Auckland
Joint work with Ngaire Kerse, Xiaojing Wu

3 December, Session C2, Presentation 4

Dementia is a global health priority. The IDEA programme is a dementia prevalence study aiming to establish the true prevalence of dementia among older adults in Aotearoa New Zealand, with a particular emphasis on diverse ethnic groups. In this talk, I will provide an overview of the methods and challenges of sampling older adults (65+) from small populations. I will present the community sampling strategy in this programme and the challenges that we have encountered along the way.

Methodology for iwi affiliation in the 2023 Census

Walter Somerville, James Maguire, Roimata Timutimu, Stats NZ & TKR
Joint work with Tieta Vesty, Meenu Jose, Pip Bennett

3 December, Session D3, Presentation 2

The iwi affiliation variable is one of the highest-priority variables released in the 2023 Census. It is used to inform decisions that affect iwi and Māori, to provide iwi, iwi-related groups, and Māori with information to support their aspirations, and to meet the Crown's partnership obligations under Te Tiriti o Waitangi. However, non-response in the 2023 Census means that some data is missing. To replace missing values for the iwi affiliation variable in 2023 Census, Stats NZ worked with Te Kāhui Raraunga Charitable Trust, under the Mana Ōrite Relationship Agreement (MŌRA), to investigate and develop a new methodology to incorporate information from alternative data sources. The alternative data sources used include historical census data as well as administrative data from the Ministry of Education and the Ministry of Social Development. The methodology also includes the use of parents', grandparents', or great-grandparents' iwi affiliation where parental relationships are found in Department of Internal Affairs data. The use of alternative data sources increases the percentage of the 2023 Māori descent census usually resident population that have a value for iwi affiliation from 716,208 people (73.2 percent) to 954,042 people (97.5 percent). This presentation will discuss how we've worked together, the method used, and some of the key considerations.

The presentation *Methodology for Māori descent in the 2023 Census* by Amanda O'Connell, James Maguire and Roimata Timutimu presents similar points about the Māori descent variable. The related presentation *Admin data quality for iwi affiliation in the 2023 Census* by Tieta Vesty, Meenu Jose and Roimata Timutimu will have more information about the administrative data sources.

State of the nation: Accessibility of vaping products in New Zealand

Janet Stacey, University of Auckland/ESR
Joint work with Alex Chen

3 December, Session C3, Presentation 2 (Student)

Over the last 4 years, we have investigated the state of the vaping market in New Zealand. This has been a time of evolving regulatory environment and increasing awareness of the associated risks and safety issues, particularly in young people. The Ministry of Health holds a register of Specialist Vaping Retailers. Changes to this register were monitored to understand differences in the number of registered entities as legislation changes occurred. Store information from the register was used to carry out market analysis. Online stores were scraped to obtain information about the products available for sale and trends were determined. Physical store addresses were geocoded, and geospatial analysis was carried out, including developing an understanding of where they are situated in relation to endangered communities. A dashboard was developed to visualise these results. This talk will discuss the state of the vaping market and show the key findings from this work.

Maximum likelihood recursive state estimation

Budhi Arta Surya, Victoria University of Wellington

2 December, Session A3, Presentation 2

In this talk I will present a novel statistical method for estimating a stochastic system from possibly nonlinear, noisy observations. Some distributional identities are developed for derivation of the recursive estimator and for the valuation of estimated standard errors. A series of simulation studies are performed to exemplify the results and to show the accuracy of the estimator.

Pseudo analysis of variance of K-functions

Rolf Turner, University of Auckland

Joint work with Peter Diggle, Lancaster University

4 December, Session F3, Presentation 3

Data consisting of independent point patterns may be collected according to some classification structure or experimental design. In such circumstances it may be of interest to formally test whether the patterns vary in nature according to the classification structure. Such variation could be expressed in term of the K-functions (or other summary functions) of the patterns. Hahn (2012) and Diggle *et al.* (2000) investigated these ideas for one-way classifications. In this talk we extend their ideas to two-way (cross) classifications. A major goal is to test for an effect from one classification factor, allowing for the possible impact of a second factor. This work was fundamentally motivated by the study of point patterns of stomata in plant leaves, Peijian Shi *et al.* (2021).

Democratising interactive statistical graphics

Simon Urbanek, University of Auckland

Joint work with Adam Bartonicek

4 December, Session E2, Presentation 3

Interactive statistical graphics have been long a cornerstone of exploratory data analysis, but typically required specialised software or bespoke tools, and were thus only used by few statisticians. On the other hand the value of adding interactivity for generally available presentation graphics has been recognised, leading to the explosion of interactive web-based publications. In this talk we will present a way to combine both worlds: leverage web-based tools to deliver the capability of using interactive statistical graphics to anyone, enabling a wider range of applications and engaging more people with data.

Reporting randomised factorial trial results: the Topical Analgesia Post-Haemorrhoidectomy trial

Alain C. Vandal, University of Auckland/Te Whatu Ora Counties Manukau

Joint work with James Jin, Weisi Xia, Runzhe Gao, Maree Weston, Lincoln Israel, Andrew Connolly, Primal Singh, Darren Svirskis, Andrew Hill

3 December, Session C2, Presentation 5

A randomised factorial trial is usually designed to detect the smallest clinically meaningful difference for all interventions, absent interaction. When a negative interaction is apparent without a sufficient sample size to test for it, the question arises as to how to best report the results, given the constraints of the statistical analysis plan (SAP). We present the rationale and decisions made in this regard when reporting the results from the Topical Analgesia Post-Haemorrhoidectomy randomised factorial trial, in which a significant difference between a metronidazole+lidocaine and a metronidazole+lidocaine+diltiazem formulation injected unexpected variability in the primary outcome, a pain visual analogue scale measured on the fourth day post-operation. While the SAP dictated the reporting of the main effects only, appropriate translation to clinical practice required separate reporting of the four arms. We describe the trial's design, longitudinal data collection process, blind review, main analysis, and describe how we dealt with this issue.

Admin data quality for iwi affiliation in the 2023 Census

Tieta Vesty, Meenu Jose, Roimata Timutimu, Stats NZ and TKR

Joint work with Walter Somerville, James Maguire, Pip Bennett

3 December, Session D3, Presentation 3

The iwi affiliation variable in the 2023 Census uses alternative data sources to replace missing values. This presentation takes a closer look into various aspects of quality for iwi affiliation data from the administrative (admin) data sources used, from the Ministry of Education (MoE) and the Ministry of Social Development (MSD), in a 2023 Census context. This is the first time that iwi affiliation data from these sources has been used for official statistics. This mahi was undertaken by Stats NZ and Te Kāhui Raraunga Charitable Trust, under the Mana Ōrite Relationship Agreement (MŌRA). The iwi affiliation variable is one of the highest-priority variables released in the 2023 Census. It is used to inform decisions that affect iwi and Māori, to provide iwi, iwi-related groups, and Māori with information to support their aspirations, and to meet the Crown's partnership obligations under Te Tiriti o Waitangi.

The presentation *Methodology for iwi affiliation in the 2023 Census* by Walter Somerville, James Maguire and Roimata Timutimu outlines the entire methodology.

Galaxy cluster gas pressure profile modelling with reversible jump MCMC: A case study using Planck data

Kang Wang, Victoria University of Wellington
Joint work with Yvette C. Perrott, Richard Arnold, David Huijser

2 December, Session B1, Presentation 3 (Student)

This study introduces an innovative approach to modelling galaxy cluster gas profiles by combining Reversible Jump Markov Chain Monte Carlo (RJMCMC) with Nested Sampling. Traditional parametric methods, such as the generalised Navarro-Frenk-White (gNFW) profile, often face challenges like parameter degeneracy. In contrast, our method uses a flexible, semi-parametric nodal model to accurately define the gas pressure profile of galaxy clusters. This node-based model allows for automatic trans-dimensional model selection within a single program execution, eliminating the need to run multiple models and compare Bayes factors. Using data from the Coma, A2255, and A85 clusters observed by the Planck space telescope, our approach significantly improves the ability to describe the pressure-radius relationship compared to conventional parametric models.

Validation of clinically important differences for randomised controlled trials in asthma

Mark Weatherall, University of Otago, Wellington and
Medical Research Institute of New Zealand
Joint work with Richard Beasley, Jon Noble, Allie Eathorne, Pepa Bruce

4 December, Session F2, Presentation 3

Introduction
Planning sample sizes for randomised controlled trials (RCTs) includes nominating a minimum clinically important difference (MCID) for an outcome variable. One way of exploring the MCID of an outcome variable is to compare it with other, established, outcome variables.

Methods
The Asthma Control Questionnaire 5 Item (ACQ-5) is a recognised outcome variable in asthma trials with an MCID of 0.5. This outcome variable was measured in two RCTs in asthma, in association with a biomarker of asthma, the Fractional Expired Nitric Oxide (FeNO). It was also measured in a cohort study of a stepped-care approach to asthma medication, in association with a participant reported outcome measure, the Treatment Satisfaction Questionnaire for Medication (TSQM). Scatter plots and linear regression were used to estimate the strength of association between the ACQ-5 and FeNO and TSQM. Logistic regression with Receiver Operating Characteristic Curve plots and c-statistics assessed the discrimination for the MCID of the ACQ-5 for FeNO and TSQM.

Results
In the RCTs the point estimate for association of FeNO with the ACQ-5 MCID was consistent with past suggestions that a 20% relative change in FeNO may be the MICD. The best c-statistic was 0.62 in a sub-group of participants with poor asthma control and high baseline ACQ-5. In the cohort study the c-statistic was 0.67 for TSQM in relation to ACQ-5.

Discussion
Both the outcome variables were associated with the ACQ-5 but with modest discrimination for its MCID. They may still be useful outcome variables capturing different aspects of asthma.

An adaptive polynomial baseline correction method in voltammetry with application to the prediction of the concentration of cocaine aptamer

Zhijian Wen, ESR

Joint work with Janet Stacey, Yasmin Liu

4 December, Session F3, Presentation 2

A background signal, or baseline, is a typical low frequency that is composited with the target signal that commonly occurs in electrochemical biosensing data. The background signal usually contains various features, such as levels, trends, and shapes. These features are usually uninformative, and if unaccounted for, they may confuse the results of the analysis. Therefore, a correction for baseline is an essential step for analyzing the sensing results. In this project, we will present an adaptive polynomial baseline correction method for the baseline correction of SWV-based electrochemical aptasensors. This method can automatically identify the uninformative regions in the signal and provide a robust mathematical equation to estimate the baseline. We compared our method with other published methods using our aptasensor data. The result shows that our method performs more reliably with acceptable errors. We also used the baseline-corrected aptamer data to develop a statistical model for predicting the cocaine concentrations in saliva. This model shows huge potential to facilitate data automation to detect specific analytes for point-of-care applications.

What makes for a happy tramper?

Ian Westbrooke, NZ Department of Conservation

2 December, Session A3, Presentation 3

Simple classification trees were applied to a large set of online responses from people after taking one of NZ's Great Walks. Two dimensions emerged as explaining much of the variation in overall satisfaction with the DOC's premier multi-night tramping experiences. In this talk Ian Westbrooke, Principal Science Analyst/Statistician at DOC, will reveal what (on average!) makes for a happy tramper.

Improving probabilistic linking in the IDI

Manori Wickramasinghe, Stats NZ

Joint work with Marina Chen

4 December, Session F1, Presentation 2

In probabilistic record linking, accurate name matching is critical particularly in datasets with culturally diverse populations. The IDI used the SOUNDEX phonetic algorithm as a blocking variable, which often struggle to match non-European names, which leads to lower link rates for ethnic groups such as Māori, Pacific, and Chinese. Also, the occurrences of the short names tend to lower the link rate, challenging the accuracy of the record linkage processes. We explore how replacing SOUNDEX with NYSIIS (New York State Identification and Intelligence System) algorithm along with other methods can enhance linkage accuracy for diverse populations. Attendees will gain insights into improving link rates for non-European names and boosting overall record linkage quality.

A logical contradiction when considering an unknown constant to have a probability density function

Robin Willink, University of Otago, Wellington

3 December, Session C3, Presentation 4

The idea of attributing a probability density function (pdf) to an unknown constant is popular among some statisticians. But does this idea have a basis in logic? In this talk, we build on a result published in the measurement literature to show that it leads to a logical contradiction. The premise that an objective set of information about a constant can be accurately represented by a probability distribution seems accepted by some in the measurement community. In that context, this premise has recently been shown to be false: an arbitrary non-linear transformation leads to a logical contradiction when two pdfs that are said to represent disjoint sets of information about the same constant are combined [“On revision of the *Guide to the Expression of Uncertainty in Measurement*: Proofs of fundamental errors in Bayesian approaches”, R. Willink, *Measurement: Sensors* 24 (2022) 100416; “Paradox? What paradox?”, R. Willink, *Accreditation and Quality Assurance*, 29 (2024) 189-192]. This talk will extend this result to apply to the concept of subjective probability also – with far-reaching implications. The conclusion relates to the general idea of attributing a probability distribution to a constant that exists on a continuous scale, i.e., to the attribution of a probability density function. There remains the possibility that a constant on a discrete scale can logically be attributed a probability mass function.

A simple adjustment makes the Wilcoxon rank-sum/Mann-Whitney test uniformly more powerful

Robin Willink, University of Otago, Wellington

4 December, Session E3, Presentation 3

The Wilcoxon rank-sum/Mann-Whitney (WMW) test is the standard non-parametric test for a difference between two populations. In its Wilcoxon form, the test can be seen to be a permutation test involving the sum of ranks in one sample. Ranks are equally spaced, and their summation creates artificial ties in the null distribution, which is obtained under all possible permutations. The presence of these ties is unhelpful because they increase the p-value. If we apply a mild non-linear transformation to the ranks before summing, (for example, if we raise each rank to the power of 1.0001), then we can avoid these ties without affecting the ordering in the null distribution in any other way. The result is a valid test procedure that can produce a smaller p-value but not a larger p-value, meaning that the test is uniformly more powerful than the WMW test. The increase in power depends on the sample sizes. When the samples are small and the standard WMW test has a power of 0.50, the procedure can have a power of, say, 0.55. The increase is negligible when the smaller sample size is 20 or more.

Quantifying the effect of data gaps on structure functions of turbulent time series: are the biases remediable?

Daniel Wrench, Victoria University of Wellington
Joint work with Tulasi Parashar

Poster Presentation (Student)

Structure functions, which represent the moments of the increments of a stochastic process, are essential complementary statistics to power spectra for analysing the self-similar behaviour of a time series. However, many real-world environmental datasets, such as those collected by spacecraft monitoring the solar wind, contain gaps, which inevitably corrupt the statistics. The nature of this corruption for structure functions remains poorly understood – indeed, often overlooked. Here we simulate gaps in a large set of magnetic field intervals from Parker Solar Probe in order to characterise the behaviour of the structure function of a sparse time series of solar wind turbulence. We quantify the resultant error with regards to the overall shape of the structure function, and its slope in the inertial range. Noting the consistent underestimation of the true curve when using linear interpolation, we demonstrate the ability of an empirical correction factor to "de-bias" these estimates. This correction, "learnt" from the data from a single spacecraft, is shown to generalise well to data from a solar wind regime elsewhere in the heliosphere. Given this success, we apply the correction to Voyager intervals from the inner heliosheath and local interstellar medium, obtaining spectral indices similar to those from previous studies. This work provides a tool for analysis of future studies of fragmented solar wind time series, as well as sparsely-sampled astrophysical and geophysical processes more generally.

Evaluating the impact of timely access to concussion services on patient outcomes in Aotearoa

Angeline Xiao, University of Auckland
Joint work with Alain C. Vandal and Braden Te Ao

3 December, Session C2, Presentation 2 (Student)

The Aotearoa Concussion Cost-Effectiveness Services Study examined the cost-efficiency of ACC-funded interdisciplinary concussion services for patients who have persistent symptoms after sustaining a mild traumatic brain injury. It aimed to assess the impacts of timely access (the exposure of interest: early vs late presentation) to such services on symptom reduction, patient self-management, quality of life, functioning, and resource use. This longitudinal study followed participants over a period of 12 months. Mixed-effects regression models were used to determine the effects of presentation time on the outcomes. The assumed causal structure required the use of two-stage regression and inverse probability weighting to yield unbiased estimates of the exposure effects.

Learning cascading failure pattern in massive pipe network data: a plumber's guide

Xun Xiao, University of Otago
Joint work with Zhisheng Ye, Matthew Revie

3 December, Session C3, Presentation 5

In this talk, I will discuss a novel multivariate point process regression model for a large-scale physically distributed network infrastructure with two failure modes, i.e., primary failures caused by the long-term usage and degradation of the asset, and cascading failures triggered by primary failures in a short period. Large-scale field data on pipe failures from a UK-based water utility are exploited to support the rationale of considering the two failure modes. The two failure modes are not self-revealed in the field data. To make the inference of the large-scale problem possible, a time window for cascading failures is introduced, based on which the likelihood of the pipe failure process can be decomposed into two parts, one for the primary failures and the other for the cascading failure processes modulated by the primary failure processes. The window length for cascading failures is treated as a tuning parameter and it is determined through maximizing the likelihood based on all failure data. To illustrate the effectiveness of the proposed model, two case studies are presented based on real data from the UK-based water utility. Interesting features of the cascading failures are identified from massive field pipe failure data. The results provide insights on more advanced modelling and practical decision-making for both researchers and practitioners.

Heaping and seeping, GAITD regression and doubly constrained reduced rank vector generalized linear models, in smoking studies

Thomas Yee, University of Auckland
Joint work with Luca Frigau, Chenchen Ma

4 December, Session F2, Presentation 1

Large-scale health surveys suitable for addiction studies furnish self-reported data that consequently suffer from a form of measurement error called heaping. Also known as digit preference, the aberration is often characterized by spikes at multiples of 10 or 5 upon rounding. To date methods and software for heaped and seeped data have been largely wanting. Identifying three generic problems for simple addiction studies, we solve them by a newly developed technique called Generally Altered, Inflated, Truncated and Deflated regression for counts applied to the most recent NHANES data. In conjunction, we propose the class of Doubly constrained Reduced-rank VGLMs to allow the dimension reduction further simplification. We determine the distribution of smoking initiation age (SIA) and its association with tobacco consumption and smoking duration, e.g., is a lower SIA associated with higher tobacco consumption later in life? Is higher SIA associated with shorter smoking duration among quitters? Together, GAITD regression and DRR-VGLMs hold promise for heaped and seeped data.

Advanced methods for time series data applied to prediction of operating modes for wind turbines

Hannah Yun, University of Auckland
Joint work with Ciprian Doru Giurcăneanu, Gill Dobbie

2 December, Session B2, Presentation 3 (Student)

Wind turbines can be characterised by distinct operating modes that reflect efficiency of the turbine under various conditions. In this talk, we focus on the forecasting problem for univariate discrete-valued time series of operating modes of a wind turbine. We define three prediction strategies to overcome the difficulties associated with missing data. The first strategy is to ignore missing values and to focus solely on available data. The second strategy imputes the missing values by replacing them in the time series with an estimate. The last strategy treats missing values as an additional operating mode. These strategies are evaluated through experiments using five forecasting methods across two real-life datasets. Two of the forecasting methods have been introduced in the statistical literature as extensions of the well-known context algorithm: variable length Markov chains and Bayesian context tree. Additionally, we consider a Bayesian method based on conditional tensor factorisation and two different smoothing techniques from the classical tools for time series forecasting: Exponential smoothing and Whittaker smoother. Each combination of prediction strategy and forecasting method is evaluated in terms of prediction accuracy versus computational complexity. We provide guidance on the methods suitable for forecasting the time series of operating methods in wind turbines. The prediction results demonstrate that high accuracy can be achieved with reduced computational resources.

Branching processes with detection: Probabilistic analysis

Zehua Zang, University of Auckland
Joint work with Jesse Goodman, Simon Harris

3 December, Session C3, Presentation 3 (Student)

The branching process is a stochastic model that describes the evolution of a population, where the offspring of each individual are produced independently of others and the past. This thesis explores a less explored aspect of branching processes by integrating a detection mechanism wherein each individual in a population has a probability of being detected. This augmentation provides insights for applications in fields such as infectious disease research.

We study four models in this thesis: discrete and continuous-time Galton-Watson processes and discrete and continuous-time multi-type branching processes. For these four models, we examine the distribution of the first detection time, establish limit theorems, and analyse the asymptotic behaviour of the detected processes. Our analysis also addresses questions, such as in the case of continuous-time branching processes, we derive an explicit generating function expression for the cluster size at the first detection and the application of a coupling technique in multi-type branching processes. This study extends the theoretical framework of branching processes and emphasises their practical applications in real-world scenarios.

Childhood risk and resilience factors for Pasifika youth respiratory health: Accounting for attrition and missingness

Siwei Zhai, University of Auckland, and Te Whatu Ora Counties Manukau
Joint work with Alain C. Vandal, Catherine A. Byrnes, El-Shadan Tautolo

3 December, Session C2, Presentation 3 (Student)

In New Zealand, 7% of deaths are related to respiratory diseases, with Pacific people at higher risk. Our work investigates the causal effects of early-life risks and resilience factors on early-adulthood lung function amongst Pacific Islands Families Study (PIFS) cohort members (n = 1,398); 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. A previous study had evaluated the effects of early-life nutrition factors on the respiratory health of Pacific youth. The results suggested a positive impact of consuming more fruit and vegetables during childhood on respiratory health later in life. The follow-up study will continue to explore the effects of factors from relevant domains based on the PIFS cohort, where a new integrated model will be applied. A simulation will be conducted to determine this model.

EvoFeat: Genetic programming-based feature engineering approach to tabular data

Hengzhe Zhang, Victoria University of Wellington
Joint work with Qi Chen, Bing Xue, Mengjie Zhang

2 December, Session A2, Presentation 3 (Student)

In recent years, in the context of tabular data classification, the emergence of transformer architecture has led to deep learning methods yielding better results than conventional tree-based models. Most of these findings attribute the success of deep learning to the expressive feature construction capabilities of neural networks. Nonetheless, in real-world practice, manually designed high-order features using traditional machine learning methods are still widely used because features based on neural networks can be prone to overfitting. In this talk, we propose a genetic programming-based feature engineering algorithm to automate the feature construction process through trial and improvement. Importantly, genetic programming provides an opportunity to optimize symbolic models, which gradient-based methods often find hard to optimize. On a large-scale classification benchmark involving 130 datasets, the experimental results demonstrate that the proposed method outperforms existing, fine-tuned state-of-the-art tree-based and deep-learning-based classification algorithms.

Extension of VGAM package

Wenqi Zhao, University of Auckland
Joint work with Thomas Yee

2 December, Session A2, Presentation 2 (Student)

Vector Generalized Linear Models (VGLMs) and Vector Generalized Additive Models (VGAMs) significantly extend the capabilities of GLMs. The `VGAM` package, developed by Yee (2015), implements over 150 family functions, providing a highly flexible framework for modeling diverse datasets. VGLMs are particularly effective for handling multivariate data where multiple responses are collected from the same units of observation. We are now developing a new package called `VGAMplus`, which offers some new distributions, and incorporates LASSO regularization for VGLMs.