EARLY CAREER & STUDENT STATISTICIANS CONFERENCE PROCEEDINGS 2021

All Presentation and Poster Sessions
at the
7th Statistical Society of Australia Early Career & Student Statisticians Conference

Sessions are listed in order of their appearance during the conference

DISCLAIMER:

These abstracts were provided by all presenters who were allocated a presentation or poster session at the conference during 26th July to 1st August 2021. Although every effort has been made to ensure accurate reproduction of these abstracts, the conference organizers cannot be held accountable for inaccuracies that may have occurred in their reproduction.

CONTENT

WELCOME LETTER FROM THE ORGANISING COMMITTEE

We welcome you to the 2021 Early Career & Student Statisticians Conference (ECSSC), hosted by the Statistical Society of Australia. The biennial conference, formerly known as the Young Statisticians Conference, is aimed at developing, maintaining and improving contact and support amongst ECSSs applying statistics in various scientific disciplines, including agriculture, medicine, economics, bioinformatics, artificial intelligence and machine learning, and environmental sciences.

We have wonderful keynote speakers and career panellists and are offering ECSSs the opportunity to connect with industry leaders. This is your opportunity to gain insights from their expertise and experience. You will also have several opportunities to connect with your colleagues via our social events, and many chances to win cash prizes, including: the *Louise Ryan Award for Best Presentation*, the *Alison Harcourt Award for Best Poster*, and finally the *Sue Finch Award for Best Data Visualisation*.

For the first time ever, the ECSSC includes a High School Career Day. The day is dedicated to our future budding statisticians who will be presented with essential career information, such as career pathways in statistics, understanding different career stages (i.e., from university student to industry), and getting preliminary insights into some statistical skills, including visualisation techniques.

We are happy with the diversity of abstracts submitted to ECSSC. Check them out in the subsequent pages! While you are sitting through the wonderful keynote presentations and the below presentations, share your experiences via social media using the hashtag #ECSSC2021. Follow us on the following social media channels:

- Twitter (https://twitter.com/ssa_ecssn)
- Facebook (https://www.facebook.com/SSA.ECSSN/)
- Instagram (https://www.instagram.com/ssa_ecssn/)

We are all looking forward to meeting you at the ECSSC2021. See you soon!

Janan Arslan
ECSSC2021 Chair

ORGANISING COMMITTEE

Chair	Janan Arslan
Program Chair	Linh Nghiem
Committee Secretary	Ben Harrap
Co-abstract Submissions Manager	Shih Ching Fu
Co-abstract Submissions Manager	Jordan Hedi
Website Manager	John Yeung
Assistant Website Manager	Cameron Patrick
Content Marketing Manager	Splithoof Rivera
Social Media Manager	Mahantesh Biradar
Social Program Manager	Melissa Middleton
Treasurer	Sherri McRae
Assistant Treasurer	Fui Swen Kuh
Sponsorship Manager	Luca Maestrini
Co-sponsorship Manager	Catriona Croton
ABS Representatives	Tiernan Byrne, Soraya McPhail & Phil Newbold
Executive Officer	Marie-Louise Rankin
Event Coordinator	Jodi Phillips

OUR SPONSORS

PLATINUM SPONSORS

The Australian Bureau of Statistics (ABS) is the independent statutory agency of the Australian Government responsible for statistical collection and analysis, and for giving evidence-based advice to federal, state, and territory governments. The ABS collects and analyses statistics on economic, population, environmental and social issues, publishing many on their website. The ABS also operates the national Census of Population and Housing that occurs every five years. https://www.abs.gov.au/

GOLD SPONSORS

The Biostatistics Collaboration of Australia (BCA) is a collaboration between 5 Australian Universities. We deliver an online program of Postgraduate Biostatistics Courses. Our focused curriculum was developed with a commitment to provide Australia with well-trained professional biostatisticians and upgrade the skills of clinical researchers. The courses provide a sound mathematically-based grounding in statistical methods with a strong emphasis on applications in all areas of health and medical research. https://www.bca.edu.au/ecssc2021/

AMSI is the collaborative enterprise of Australia's mathematical sciences. The national and international initiatives of AMSI fall largely into three classes – research and higher education, school education and engagement with the industrial and commercial world. AMSI has built a record of achievement in these areas and is recognised by government and industry as a leading provider of services, activities and strategic initiatives. The common aim that AMSI shares with partners is the radical improvement of levels of mathematical capacity and facility in the Australian community. https://amsi.org.au/

Survey Design and Analysis Services Pty Ltd (SDAS) is the authorised distributor of Stata, Stat/Transfer, QDA Miner, WordStat and Arbutus Software in Australia, Indonesia and New Zealand. Software, books, journals and support to businesses, government organisations, non-profits, universities and personal users are offered by Survey Design with the most competitive pricing available. Survey Design sells the tools to help auditors, analysts, academics, researchers and professionals get more out of their data. "We provide the tools, so you can find the answers". https://www.surveydesign.com.au/

The Australian Research Council (ARC) Centre of Excellence for Mathematical and Statistical Frontiers (ACEMS) brings together for the first time a critical mass of Australia's best researchers in applied mathematics, statistics, mathematical physics, and machine learning. With partner researchers, ACEMS engages in research programs that combine innovative methods for the analysis of data with theoretical, methodological, and computational foundations, provided by advanced mathematical and statistical modelling. ACEMS focuses on the impact of new insights for end-users working in the Collaborative Domains of Healthy People, Sustainable Environments, and Prosperous Societies. https://acems.org.au/home

OUR PARTNERS

The objective of the Statistical Society of Australia (SSA) is to further the study, application and professional practice of statistics in Australia and Internationally. With close to 1000 members, the SSA offers opportunity to augment your skills, network with your colleagues, and contribute to the statistical community. The ECSSC2021 is one of the many conferences that SSA members are involved in. If you would like to be a part of this society, please visit https://statsoc.org.au/.

PROCEEDINGS CITATION

Please use the following citation when referencing works from this conference:

Author, A. (2021, July). Presentation Title. In *Early Career & Student Statisticians Conference Proceedings 2021*. 7th Conference of the Early Career & Student Statisticians Network (Australia), (pp. xx-xx). Statistical Society of Australia. URL: https://statsoc.org.au/ECSSC

CONFERENCE PROGRAM

Day	Time in AEST	Activities	Speakers
24 July	11 AM - 3PM	Short course 1	Anqi Fu (Stanford University)
25 July	11 AM - 3PM	Short course 2	Chul Moon (Southern Methodist University)
26 July	11 AM - 11:15 AM	Opening speech	
	11:15 AM - 12:50 PM	Abstract/poster presentation 1	Michael James Leach (Monash University), Elena Tartaglia (CSIRO), Atousa Grahramani (Victoria University), Vanessa Pac Soo (University of Melbourne), Raphael Udeh (University of Newcastle), Ravindi Nanayakkara (La Trobe University), Taya Annabelle Collyer (Monash University)
	1:00 PM - 1:45 PM	SurveyDesign information session	David White (SurveyDesign)
	2:00 PM - 3:00 PM	Keynote Speaker 1	Peter Taylor (University of Melbourne)
	3:00 PM - 4:00 PM	Welcome Event	
27 July	11 AM - 12:30PM	Abstract/poster presentation 2	Chathuri Lakshika Samarasekara (RMIT), Shih Ching Fu (Curtin University), Josh Jacobson (University of Wollongong), Jason Whyte (University of Melbourne), Udani Wijewardhana (Swinburne), John Samuel Warmenhoven (University of New South Wales)
	12:30 PM - 1:30 PM	BCA information session	Andrew Forbes (BCA)
	2:00 PM - 3:00 PM	Keynote speaker 2	Paola Oliva-Altamirano (Innovation Lab)
	3:00 PM - 3:30 PM	Virtual Pub	
28 July	11 AM - 12:30PM	Abstract/poster presentation 3	Xu Ning (ANU), Muzhi Zhao (ANU), Phillip Oluwatobi Awodutire (University of Port Harcourt), Zhi Yang Tho (ANU), Catriona Croton (University of Southern Queensland), Cameron Patrick (University of Melbourne)
	12:30PM - 1:30 PM	Roundtable discussion	
	1:30 PM - 3:00 PM	Career panel	Alessandra Monneris (Bureau of Meteorology), Calvin Hung (QuantumBlack), Ha Dinh (Shopify), Phil Newbold (ABS)
	3:00 PM - 3:30 PM	Virtual Pub	
	11 AM - 12 PM	Keynote speaker 3	Minh-Ngoc Tran (University of Sydney)

29 July	12 PM - 1:30 PM	ABS Information Session	Phil Newbold (ABS)
	1:30 PM - 3:05 PM	Abstract/poster presentation 4	David Allen (National Acoustic Laboratories), Parinaz Mehipour (University of Melbourne), Gizem Ashraf (University of Melbourne), Puxue Qiao (St. Vincent's Institute of Medical Research), Luca Maestrini (UTS), David Warne (Queensland University of Technology), Ben Harrap (University of Melbourne)
	3:00 PM - 3:30 PM	Virtual Pub	
	6:00 PM - 8:00 PM	Games Night	
30 July	11 AM - 12:30PM	Abstract/poster presentation 5	Chun Fung Kwok (St. Vincent's Institute of Medical Research), Fan Cheng (Monash University), Oisin Fitzgerald (UNSW), Mohomed Amsar Mohomed Abraj (Queensland University of Technology), Jeffrey Pullin (University of Melbourne), Sharm Thuraisingam (University of Melbourne)
	12:30 PM - 1:30 PM	ACEMS information session	Peter Taylor (ACEMS)
	2 PM - 3 PM	Keynote speaker 2	Kendra Vant (Xero)
	3:00 PM - 3:30 PM	Virtual Pub	
31 July	11 AM - 2:30 PM	High School Engagement Day: Introduction to programming and statistical analysis workshop	Emi Tanaka (Monash University), Kevin Wang (CSL), Daniel Fryer (La Trobe University), Patrick Robotham (Nimble Australia)
	2:45 PM - 3:45 PM	Statistical Careers & Education Panel	Elena Tartaglia (CSIRO Data61), Phil Newbold (ABS)
1 Aug	11 AM - 12PM	Keynote speaker 5	Helena Jia (Educational Testing Services, USA)
	12 PM - 1PM	Closing ceremony	

KEYNOTE PRESENTATIONS

PETER TAYLOR
Redmond Barry Professor
University of Melbourne

Modelling the Bitcoin blockchain: what can probability and statistics teach us?

In 2009 the pseudonymous Satoshi Nakamoto published a short paper on the Internet, together with accompanying software, that proposed an `electronic equivalent of cash' called Bitcoin. At its most basic level, Bitcoin is a payment system where transactions are verified and stored in a distributed data structure called the blockchain. The Bitcoin system allows electronic transfer of funds without the presence of a trusted third party. It achieves this by making it `very hard work' to create the payment record, so that it is not computationally-feasible for a malicious player to repudiate a transaction and create a forward history with the transaction deleted.

The Nakamoto paper contained a simple model used to show that the above-mentioned malicious player would be very unlikely to succeed. Unfortunately, this calculation contained an error, which I shall quickly discuss and show how to correct. As its name suggests, the blockchain is comprised of discrete blocks. Blocks are added to the blockchain by `miners' working across a distributed peer-to-peer network to solve a computationally d cult problem. With reference to historical data, I shall describe some models for the block mining process. I shall finish with some brief comments about how stochastic modelling can be used to address the current concerns that the transaction processing rate of the Bitcoin system is not high enough.

PAOLA OLIVA-ALTAMIRANO
Principal Data Scientist. Innovation Lab,
Our Community/SmartyGrants

Metrics in the wild! How to deal with biases when building auto-classification systems

When designing auto-classification systems we often trust standard metrics in-built in Machine Learning models. How much do you trust them? Or in other words, to which extent do these scores reflect the success of your algorithm? Knowing your data, the decisions that your model will influence, and the role that scoring plays in enhancing or mitigating biases should be essential for Statistics practitioners and product builders. In this talk I will share our lessons learned when building classifiers in the social sector, the biases we have encountered in multilabel text classifications, and our constant battle to design ethical, human centre products.

MINH-NGOC TRAN

Associate Professor

University of Sydney

Bayesian computation: why/when Variational Bayes, not MCMC or SMC?

Bayesian inference has been increasingly used in statistics and related areas as a principled and convenient tool for reasoning with uncertainty. Bayesian computation is often a challenging task and modern applications of Bayesian inference, such as Bayesian deep learning, have been called for more scalable Bayesian computation techniques. In this talk, we will give a quick introduction to Variational Bayes for scalable Bayesian inference. We then provide a general discussion on its pros and cons, recent advances and some potential research directions.

KENDRA VANT

Executive General Manager - Data

Xero

Commercial machine learning at scale - the joys and the pitfalls

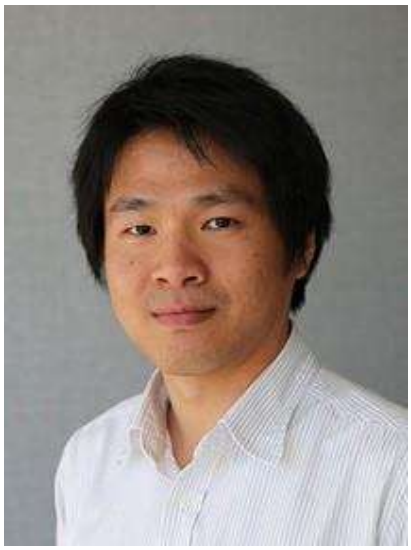
The art and science of applying machine learning techniques inside a for profit company is a world away from pursuing algorithm improvement and fundamental in a research setting. I will talk about the end to end process of building smart products within a SaaS company today.

HELENA JIA
Executive Director,
National and International Assessments,
Educational Testing Services, USA

Statistics, Psychometrics and Data Analytics in Educational Survey Assessments

Educational survey assessments are used in measuring and monitoring learning and educational progress for youth as a group, rather than as individuals. This talk offers an overview of how practitioners apply methodologies in statistics, psychometrics and data analytics in the design of educational survey assessments, as well as in the estimation of group scores nationally and internationally. I will describe several statistical approaches that are involved in the group score estimation--item response theory models, latent regression models and missing data imputation/plausible values. Recent and future research and methodology development as survey assessment data collection transitioning from paper-based to digitally-based platform will be discussed as well.

CAREER PANEL



CALVIN HUNG
Principal Data Scientist
QuantumBlack, a McKinsey Company

Dr Calvin Hung is a Principal Data Scientist at QuantumBlack. QuantumBlack is an advanced analytics firm and part of McKinsey & Company, operating at the intersection of strategy, technology and design to improve performance outcomes for organisations. Calvin holds a BSc in Physics and Mathematics, a BEng in Aerospace and a PhD in Robotics, all from the University of Sydney. His research focused on applying Deep Learning techniques to help robots navigate real world environments. After 3 years of postdoc Calvin joined the finance industry, developing systems to improve financial well-being of customers. Calvin joined QuantumBlack in 2017 and has since worked on a variety of transformation programs across retail, telco, mining and pharmaceutical industries.

HA DINH
Data Scientist
Shopify

Ha Dinh is working as a Data Scientist for the People Analytics team at Shopify. Before becoming a Data Scientist, Ha was studying Vietnamese literature during high school, and then Marketing during her undergraduate years. She decided to pursue a Master of Data Science at the University of British Columbia to better support her passion to help organizations make better decisions through insights from behaviour. Outside of work, Ha spends her time with the Women in Data Science Vancouver to empower women and people who are interested in the field.

PHIL NEWBOLD
Data Scientist
Methodology Division, Australian Bureau of Statistics

Phil Newbold studied Mathematics and Statistics at La Trobe University, then took part in the AMSI summer research scholarship, working on an R package for statistical ecology. He then spent two years working as a consultant in digital analytics. Early this year Phil took the opportunity of entering the ABS Graduate program and is currently working in the Machine Intelligence and Novel Data Sources section working predominantly on natural language processing. In his private time, Phil enjoys running, travel, hiking and golf.

ALESSANDRA MONERRIS BELDA
Manager of Satellite Operations, Australian Bureau of Meteorology

Dr Alessandra Monerris Belda holds a Masters' and a PhD in Telecommunication Engineering with major in Earth Observation (EO). Before completing her PhD, Alessandra took on the position of founding Executive Director at SMOS-BEC, a satellite research and data distribution centre in Spain. Later on, Alessandra moved to Australia and joined various EO research projects on the validation of satellite soil moisture products, cosmic-ray probes, and the utilisation of Global Navigation Satellite Systems Reflectometry for the retrieval of geophysical parameters. She was a consultant for the European Commission and is currently working at the Bureau of Meteorology as the manager of satellite operations

CONTRIBUTED PRESENTATIONS

Monday, 26 July 2021

Leach, M J., & Roughhead, E E. & Pratt, P L.

Title: *Data visualisation can reveal exposure misclassification in case-crossover studies of medication safety* greatly adding to the

Background

In epidemiology, the case-crossover design is a self-controlled observational study design for quantifying the effects of short-term exposures on acute outcomes. It compares each person's exposure status immediately prior to an outcome (case window) with their own exposure status at an earlier point or points in time (control window[s]). The case-crossover design is used in medication safety studies yet is prone to exposure misclassification. This study aimed to describe a novel data visualisation method for identifying exposure misclassification in case-crossover studies of medication safety.

Methods

A case-crossover study was conducted among Australian Government Department of Veterans' Affairs Beneficiaries aged >65 years who were hospitalised for hip fracture between 2009 and 2012. The exposure of interest was use of tricyclic antidepressants (TCAs), as indicated by dispensing records. The case window, control window one and control window two were set at 1-50 days, 101-150 days and 151-200 days before hip fracture, respectively. In a data visualisation, each individual's TCA dispensings were sorted and plotted over the 200 days before hip fracture. Exposure misclassification was determined via visual inspection and consensus decision-making among three researchers.

Results

There were 8,828 older patients, 348 of whom contributed data to the analyses with either control window. Based on visual inspection of the data visualisation and consensus decision-making, 14% and 45% of subjects were potentially misclassified with control window one and control window two, respectively. While there was no association between TCAs and hip fracture using control window one (odds ratio [OR]=1.18, 95% confidence interval [CI]=0.91–1.52), risk was significantly increased (OR = 1.43, 95% CI=1.11–1.83) when control window two was used.

Discussion

TCA exposure misclassification was more likely to be present with the earlier control window: control window two. Data visualisation is a useful means of identifying exposure misclassification in case-crossover studies of medication safety.

Keywords: Data visualisation, case-crossover study, medication safety

Tartaglia, E.

Title: *Understanding the role of causal inference from observational datasets in developing government policy*

One of the aims of public policy is to encourage behaviours towards a desired outcome. To develop effective and evidence-based policy, policymakers need to understand the likely impact of a policy. In other words, they need to understand the causal effect of an intervention. The 'gold standard' for showing causal effects is the randomised control trial (RCT). However, there are many situations where RCTs are impossible or unethical. Instead, governments rely on administrative data, which is a type of observational data that typically contains various biases. Controlling for biases is critical when estimating causal effects. Although it is impossible to guarantee perfect bias removal, using causal diagrams and adjustment techniques can help us identify data limitations and better estimate causal effects. I recently worked with the Department of Education, Skills and Employment (DESE) to incorporate causal inference techniques into the analytics underpinning their policy advice and formulation. Example questions of interest to DESE are 'what is the effect of childcare attendance on student readiness to enter primary school?' and 'what is the effect of high school completion on income later in life?'. In this talk, I will highlight the utility of causal inference in policy formulation and address some implementation challenges.

Keywords: Causal inference, Education, DAGs, Administrative data

---

Broadbridge, P. & **Nanayakkara, R.** & Olenko, A.

Title: *Probing Multifractionality of Cosmic Microwave Background Radiation*

The Cosmic Microwave Background (CMB) is the radiation residue from the Big Bang. It is the remnant radiation from the universe after 380,000 years from its birth. The first measurements of the CMB were made by Arno Penzias and Robert Wilson in 1960. The space missions that have studied the CMB so far are Cosmic Background Explorer (COBE), Wilkinson Microwave Anisotropy Probe (WMAP) and Planck. The Planck mission was launched in 2009 to study the CMB thoroughly. The main motivation of this study is to investigate the non-Gaussianity of the CMB data from the Planck mission.

In literature, isotropic spherical Gaussian fields are considered as the main stochastic model underlying the CMB data. We consider a multifractional approach to study spherical random fields with cosmological applications. The Hölder exponent is used to measure the roughness in a rigorous mathematical way. In this study, pointwise Hölder exponent values are computed for the one- and two-dimensional regions of the CMB data using the HEALPix ring and nested ordering visualization structures. The established methodology is also used to distinguish probable CMB anomalies in the cleaned CMB maps.

The results suggest that there exists some multifractionality in the CMB data. The implemented computing techniques and the obtained results are useful for stochastic modeling and analysis of other geoscience, environmental and spherical data.

The talk is based on joint results with Professors Philip Broadbridge and Andriy Olenko (La Trobe University, Australia).

Keywords: CMB, Multifractionality, Random Fields, Spherical Statistics

---

Collyer, T A.

Title: *What is statistical expertise?*

Statistical expertise is vital for the appropriate conduct of health-related research, but there is no simple way to define statistical expertise, and no straightforward way to identify scientists who are 'expert'. In this talk I present data from interviews with 45 academics in 8 countries, and draw upon a range of concepts from Science and Technology Studies, the Sociology of Professions and Sociology of Scientific Knowledge to sketch a picture of statistical expertise as it appears to function within population health research. The key findings are: 1) Statistical literacy carries connotations of legitimacy, productivity and employability. 2) Individuals' statistical confidence is not necessarily connected to the way a statistician might judge their ability. 3) The identity of statistical expert takes diverse forms in diverse settings, because 'statistical expertise' is relative. 4) Despite the positive connotations identified in 1), researchers with and without statistical expertise report pressure to delegate analysis to others as they advance through their careers. My findings call into question the notion of statistical expertise as being something researchers simply possess or lack and suggests that efforts to change statistical practice will be most effective when tailored to particular scientific contexts.

Keywords: Sociology, Expertise, Science Studies, Statistics

Samarasekara, C L. & Wang, Y. & Stone, L.

Title: *On the Estimation of the Resource Selection Probability Functions from Presence Background Data*

Background

Estimating the absolute probability of presence of a species from presence-background data has been a controversial topic in species distribution modelling. There are many arguments regarding the conditions that need to be satisfied in order to achieve this task. In this paper, we try to address the issue from a new perspective by proposing an approach which combines both statistics and machine learning techniques.

Methods

We develop a new method based on the Constrained Lele and Keim (CLK) procedure and presence background learning (PBL) algorithm, the latter proposed in the context of learning classifiers (for positive and unlabelled data). Extensive simulation studies have been conducted to assess the performance of the proposed method, in comparison with the popular Lele and Keim (LK) method. Explicitly three categories of functions have been considered that satisfy either the resource selection probability function (RSPF) condition or the local certainty condition and both conditions.

Results

The simulation studies show that when a "local knowledge" condition is satisfied, the proposed refined CLK method is able to accurately estimate the actual probability of presence, outperforming the popular LK method. The simulations also show that when the RSPF condition is satisfied, the LK method is fragile and often fails to give reliable estimates even when its underlying RSPF condition is met.

Discussion

We reinstate that it requires the knowledge the species' population prevalence in order to estimate the absolute probability of presence accurately from the presence background data. The local knowledge condition proposed in this paper extends the prototypical presence location condition (i.e. local certainty defined in the machine learning context), and serves as the more generalised condition for accurately estimating the absolute probability of presence in species distribution modelling.

Keywords: Machine Learning, Species Distribution Models, RSPF

Fu, S C.

Title: *Comparison of Spatial Models for Analysing On-farm Strip Experiments based on Geographically-Weighted Regression and Kriging*

The main premise of precision agriculture is that the optimal application of additives such as irrigation or pesticides is expected to vary within a field. Or conversely, an assumption that the response of crops is completely uniform in a field is unrealistic and may lead to wasteful farming practices. Targeted delivery of additives has therefore both economic and environmental payoffs. Consequently, it is important to accurately characterise the spatial variations of a response of interest such as the expected increase in crop yield per unit of applied fertiliser. However, to be of practical use to farmers, any detected variations must be manifest at scales meaningful for intervention.

This study undertook several numerical experiments, using both real and synthetic data, aimed at measuring the relative effectiveness of four geostatistical algorithms for predicting spatially varying yield and estimating treatment effects. The algorithms under investigation were simple kriging (SK), regression kriging (RK), geographically weighted regression (GWR), and geographically weighted regression plus kriging (GWR+K). In particular, these computational experiments were conducted in the context of large scale on-farm strip experiments in agriculture.

Results indicated that whilst RK and GWR are good at providing yield predictions and treatment effect sizes respectively, neither fully captures the non-stationarity and autocorrelation that is common in on-farm strip experiment data. The GWR+K approach however was able to return both reasonable yield predictions of moderate precision and also meaningful treatment effect sizes for targeted locations in the crop field. Given the high relative popularity of RK in the literature, this suggests that GWR+K may be underutilised in the analysis of large scale crop experiments. This study also investigated the effect of different kernel bandwidths on the result of GWR, finding that larger bandwidth reduced variance at the cost of increased bias.

Keywords: geographically weighted regression, kriging, spatial statistics

---

Wijewardhana, U. & Meyer, D. & Jayawardana, M. & Apputhurai, P.

Title: *Statistical modelling with citizen science data for local shorebirds on the Mornington Peninsula*

Background:
Climate change has been identified as a possible threat to the biodiversity of shorebirds on the Mornington Peninsula with specific species vulnerabilities likely to be important in both protected and unprotected areas.

Methods:
Due to the lack of comprehensive survey data, this study uses citizen science bird counts, extracted from the Atlas of Living Australia, to determine for which species climate change is likely to be more of a threat and to determine which species benefit more from protected areas. In all cases, analyses are conducted using temporal models fitted using the Integrated Laplace Approximation (INLA) method.

Results:
The trends for six local shorebird species were compared to the Australian Pied Oystercatcher, with significantly steeper upward trends identified for the Black-fronted Dotterel, Red-capped Dotterel, and Red-kneed Dotterel. Also, while protected areas were associated with greater numbers of the locally threatened Hooded Plover and Sooty Oystercatcher than unprotected areas, there was a significant upward trend for Sooty Oystercatchers which was particularly obvious in protected areas. However, warmer years saw a significant reduction in bird counts for the Hooded Plover on the Mornington Peninsula.

Implications:
This work suggests that, with some limitations, statistical models can be used with citizen science data for monitoring the persistence of local shorebirds and for investigating factors that are impacting these data. The results for the locally threatened Hooded Plover and Sooty Oystercatcher in protected areas are particularly encouraging, although the higher temperatures expected in the future may adversely affect the Hooded Plover.

Keywords: climate change, species vulnerabilities, protected areas, citizen science data, temporal models, INLA

Warmenhoven, J S.
Title: *A non-conventional entry into the world of statistics...*

This presentation centres on my own non-conventional pathway into statistics. And more specifically, how I've attempted to improve application of methods in my own scientific discipline (primarily health and sport) and the difficulties I've experienced doing this, my experiences working with statisticians and how I currently operate as a part of a multi-disciplinary data science team in my current role as a data science fellow.

This presentation will also look at things that have helped me on my journey into statistics and will explore the concepts of 1) established institutional relationships (with some examples of my own experiences collaborating with international universities who have built relationships across faculties), 2) the concept of "discipline wandering" being a healthy part of developing as an early career academic (and giving some examples where this has helped me both face-to-face and digital contexts) and 3) not being afraid to ask what appear to be silly questions and write long emails (I promise to provide some compelling evidence for long emails being a necessary skill as an ECR).

I would also like to provide some thoughts moving forwards around how statistics can be better utilised in the age of data science. With this specifically focusing on 1) further growing connections across disciplines (and very specifically at the postgraduate student level, relative to some of my own experiences), 2) collaborating between fields before we even have any questions (and the importance of statistician involvement in the development of multi-site data repositories) and 3) allowing more avenues to bring people like myself into your disciplines to assist in translation of better statistical practice (and exploring ways of formalising this as a pathway).

Keywords: Career, Inter-disciplinary, Collaboration, Practice

Ning, X.

Title: *Penalised Quasi-Likelihood Asymptotics in Generalised Linear Mixed Models*

We establish asymptotic results for the penalized quasi-likelihood (PQL) estimator of the parameters and random effects in a generalized linear mixed model (GLMM) for clustered data when both the number of independent clusters and the cluster sizes (the number of observations in each cluster) go to infinity. Under typical regularity conditions, the estimator is shown to be – conditional on the random effects -asymptotically normal with a very simple covariance matrix. The result holds for the most general form of a generalized linear mixed model, as long as all covariates under consideration are included as both fixed and random effects. These results allow simple asymptotic confidence and prediction intervals to be formed for the PQL estimator, the latter of which has not been derived for unstructured GLMMs for any predictor currently in the literature.

Keywords: PQL asymptotic; GLMM; conditional inference

Escobar-Bach, M. & Maller, R A. & Van Keilegom, I. & **Zhao, M**.

Title: *Estimation of the Cure Rate for Distributions in the Gumbel Maximum Domain of Attraction Under Insufficient Follow-up*

Estimators of the cured proportion from survival data which may include observations on cured subjects can only be expected to perform well when the follow-up period is sufficient. When follow-up is not sufficient, and the survival distribution of those susceptible to the event belongs to the Frechet maximum domain of attraction, a nonparametric estimator for the cure proportion proposed by Escobar-Bach and Van Keilegom (2019) incorporates an adjustment that reduces the bias in the usual estimator. Besides the Frechet, an important class of limiting distributions for maxima is the Gumbel class. In the present paper, we show that a very wide class of commonly used survival distributions - the generalized Gamma distributions - are in the Gumbel domain of attraction. Extrapolation techniques from extreme value theory are then used to derive, for distributions in this class, a nonparametric estimator of the cure proportion that is consistent and asymptotically normally distributed under reasonable assumptions and performs well in simulation studies with data where follow-up is insufficient. We illustrate its use with an application to survival data where patients with differing stages of breast cancer have varying degrees of follow-up.

Keywords: survival, cure, extreme value, nonparametric

Croton, C.

Title: *Communicating with Clinical Collaborators: a Vet's POV*

Connecting new information to existing knowledge structures can improve understanding, and so communicating the statistical consulting process to clinical collaborators and clients by explicitly linking it to the process of treating patients can assist. From patient history taking to understanding your research question, clinical examination to checking your dataset, and when to call in the specialist (statistician), this talk focuses on communicating with clinical researchers. Warning: it is from the POV of a biostatistician who used to be a veterinarian and still finds statisticians hard to understand at times!

Keywords: communication, clinician, biostatistics, consulting

---

Patrick, C.

Title: *Tales from the trenches of statistical consulting: five tips for early career statistical consultants*

Statistical consulting can be one of the most diverse career paths for an applied statistician. The skills required and the challenges faced are not just statistical. What can you expect from a day in your life as a consultant? This talk will present some examples from real life statistical consulting projects to get an idea of what to expect, and provide some tips for those starting a career in statistical consulting.

Keywords: statistical consulting, applied statistics, tips

Thursday, 29 July 2021

Allen, D.

Title: *Lessons learned from the development and implementation of a Bayesian MIMIC model to predict hearing aid benefit*

Familiarity with a range of statistical methods and approaches are an important part of development for all statisticians and researchers, so that the best approaches can be applied to the research problem at hand. This can be particularly challenging where methods are novel to other researchers or to the institution in which the statistician is based. This presentation will explore the experiences of the presenter working as statistician and project lead on a Government-funded research project in an established hearing research institution.

The study that will be discussed is ongoing and aims to identify predictors of successful hearing rehabilitation among clients fitted with hearing aids under a national Australian hearing services program. Data were collected using surveys delivered before hearing aid fitting and at 2 and 8 weeks after hearing aid fitting.

The large number of both input predictors and outcome measures led to difficulties in interpretability using naïve analysis methods. As a result, a Multiple Indicator, Multiple Indicator Cause (MIMIC) model was developed. The model was fit using Hamiltonian Monte Carlo and the No-U-Turn-Sampler, and iteratively refined to address emerging performance, ethical, and research issues.

Lessons learned from the development and implementation of this approach will be presented, along with suggestions for other early career researchers seeking to explore and implement methods that may be unfamiliar to them or to their colleagues.

Keywords: biostatistics, bayesian, career

Mehipour, P. & Watson, J & Zaloumis, S. & Dini, S. & Commons, R. & Simpson, J.

Title: *Bayesian Within-host Modelling of Red Blood Cell Dynamics and Primaquine-induced Haemolysis in G6PD Deficiency*

Background:
Glucose-6-phosphate dehydrogenase (G6PD) deficiency is a prevalent genetic condition in malaria-endemic countries. Primaquine is the only widely available drug that targets Plasmodium vivax malaria parasites in the liver, however, this drug can cause dose-dependent haemolysis in G6PD deficient individuals. In this research, we develop a within-host model of red blood cell (RBC) dynamics to explore the safety of primaquine regimens for vivax malaria patients with G6PD deficiency.

Methods:
The within-host RBC model captured deviations from the normal process of RBC production and destruction by modifying parameters, RBC lifespan, release time of reticulocytes into the circulation, and production rate of RBC. The compartmental mechanistic RBC model was fitted using a two-stage Bayesian hierarchical framework to a regimen-adaptive trial of ascending primaquine doses in G6PD deficient individuals. In the first stage, the RBC model was fitted to each individual's measurements separately. The posterior distributions from this stage were used as proposal distributions for determining the population-level parameters in stage 2.

Results:
The dataset includes longitudinal haemoglobin and reticulocyte measurements from 24 Thai healthy male volunteers with seven different G6PD deficiency variants who were given ascending doses of primaquine over approximately 15-20 days. Posterior predictive checks from stage 1 show that the RBC model reproduced the change in haemoglobin and reticulocyte counts for all individuals. Preliminary results from stage 2, suggest the posterior median for the reduction in RBC lifespan due to primaquine is 30 days (95% CrI: 28-31) and that primaquine increases the production rate of RBCs in the bone marrow 5.8-fold (3.7-7.1) compared to the steady state.

Discussion:
The within-host RBC model captured the effect of primaquine on the haemoglobin and reticulocyte profiles of the G6PD deficient individuals. Population average parameters from stage 2 will be used to simulate hypothetical profiles and investigate alternative primaquine dosing schemes.

Keywords: Vivax malaria, Bayesian analysis, Within-host

Warne, D J. & Baker, R E. & Simpson, M J.
Title: *Using model approximations to accelerate Bayesian computation*

Almost all fields of science rely upon statistical inference to estimate unknown parameters in theoretical and computational models. While the performance of modern computer hardware continues to grow, the computational requirements for the simulation of models are growing even faster. This is largely due to the increase in model complexity, often including stochastic dynamics, that is necessary to describe and characterize phenomena observed using modern, high resolution, experimental techniques. Such models are rarely analytically tractable, meaning that extremely large numbers of stochastic simulations are required for parameter inference. We present new computational Bayesian techniques that accelerate inference for expensive stochastic models by using computationally inexpensive approximations to inform feasible regions in parameter space, and through learning transforms that adjust the biased approximate inferences to closer represent the correct inferences under the expensive stochastic model. We demonstrate a speed improvement of an order of magnitude without any loss in accuracy. This represents a substantial

improvement over current state-of-the-art methods for Bayesian computations when appropriate model approximations are available.

Keywords: Bayesian, Inference, SMC, ABC, preconditioning, moment-matching

---

Harrap, B.

Title: *Things I Didn't Learn at University*

I learned a lot about statistics at university. You'd expect that when doing a master's in biostatistics. But I learned a lot about how statistics works and very little about what it's actually like 'doing' statistics. In this talk I will cover some of the most valuable lessons I learned in the first few years after graduating and working as an applied statistician, including:

- The perils of data you didn't collect and other data tales
- Why collaboration makes you a better statistician
- Making mistakes is important (and inevitable)
- Communication, communication, communication
- For the love of everything that is holy, please comment your code
- Which statistical software is the best one

This talk will be a light-hearted take on my real-world experiences, which hopefully helps someone in their transition to the workplace

Keywords: applied statistics, career, practice

Friday, 30 July 2021

Kwok, C F.
Title: *Enhancing Markov-chain spatial simulation with web mapping technology*

This work investigates the use of web mapping technology to enhance spatial simulation and analysis. We consider the Markov chain on graphs from network analysis and combine it with a simulator operating on an interactive map. The map provides a good approximation of the physical constraints in the real world and gives additional spatial information. They are incorporated into the model to generate predictions with improved spatial resolution and help researchers answer spatial questions at a refined scale. We will present some use cases with the enhanced model in the context of policy evaluation and experimental design. We will also show how our approach gives a natural and more flexible way to generate spatial-related queries and study complex impulse responses.

Keywords: interactive-visualisation, spatial-simulation, map, Markov-chain

---

Cheng, F. & Hyndman, R. J. & Panagiotelis, A
Title: *Manifold Learning with Approximate Nearest Neighbors*

Manifold learning algorithms are valuable tools for the analysis of high-dimensional data, many of which include a step where nearest neighbors of all observations are found. This can present a computational bottleneck when the number of observations is large or when the observations lie in more general metric spaces, such as statistical manifolds, which require all pairwise distances between observations to be computed. We resolve this problem by using a broad range of approximate nearest neighbor algorithms within manifold learning algorithms and evaluating their impact on embedding accuracy. We use approximate nearest neighbors for statistical manifolds by exploiting the connection between Hellinger/Total variation distance for discrete distributions and the L2/L1 norm. Via a thorough empirical investigation based on the benchmark MNIST dataset, it is shown that approximate nearest neighbors lead to substantial improvements in computational time with little to no loss in the accuracy of the embedding produced by a manifold learning algorithm. This result is robust to the use of different manifold learning algorithms, to the use of different approximate nearest neighbor algorithms, and to the use of different measures of embedding accuracy. The proposed method is applied to learning statistical manifolds data on distributions of electricity usage. This application demonstrates how the proposed methods can be used to visualize and identify anomalies and uncover underlying structure within high-dimensional data in a way that is scalable to large datasets.

Keywords: statistical manifold, dimension reduction, anomaly detection

---

Pullin, J. & McCarthy, D.

Title: *Comparing statistical methods for identifying marker genes in single-cell RNA sequencing data*

In the last 5 years, scientists have become able to routinely measure expression of tens of thousands of genes in individual cells, the fundamental unit of life. This new data will generate many significant biological insights, but like most modern biological data requires substantial statistical analysis.

One critical analysis task is to identify the known cell types present in a given dataset. This task is often performed using so-called marker genes. First, cells are partitioned into groups using a clustering algorithm. Then, statistical methods are used to define a small subset of the genes measured in the experiment most strongly associated with each group of cells: the marker genes. The data for these genes is visualized and interpreted by a biologist to identify the cell type of each group of cells. Today, there are over 40 different methods - from the t-test to methods published this year - for identifying marker genes. Despite this abundance of methods, however, there has been no published comparison of the competing methods making it difficult for researchers to choose the right method.

In my talk I will describe the different statistical methods currently used to identify marker genes, highlighting differences in their methodology and assumptions. I will then present the results of a comprehensive simulation study designed to assess the performance of the methods. We use the popular single-cell RNA sequencing data simulation tool Splatter, to simulate 45 different situations with parameters estimated from 3 different single-cell datasets. We assess the ability of the methods to accurately identify genes which show the most cluster-specific expression profiles. Results from the simulation study demonstrate substantial differences in the performance of competing methods and highlight that the performance of methods is often situation-specific. I will conclude with best practices for the use of marker gene identification methods.

Keywords: bioinformatics, simulation study, benchmarking, single cell

---

Thuraisingam, S.

Title: *Surviving a PhD with a toddler during a pandemic*

Focus of presentation:

Undertaking a PhD is a long and challenging journey. Having a baby is heart-warming, but exhausting. Doing all this during a pandemic is seriously difficult! Despite the blood, sweat and tears (many tears), there have been some character-building opportunities and important lessons learnt that have helped me keep on track during this challenging time. In this presentation I share my tips for surviving a PhD with an energetic toddler during a pandemic.

Findings:
The lessons I have learnt over this last year can be summarised as ten top tips: (1) plan, plan, plan, (2) sleep when the opportunity comes (if it ever does), (3) have small and achievable goals, (4) utilise time spent on menial chores for thinking/solving, (5) make time to chill out and exercise, (6) try to leave your PhD work in a place that is easy to pick up from when you return, (7) accept that your PhD work time may ebb and flow and is now dependent on someone else (your little cherub), (8) be present when spending time with your little one (9) stay connected to peers, supervisors, friends and family, and (10) be kind to yourself, this is hard!

Conclusions/Implications:
Doing a PhD is difficult. Life can throw curve balls too- including pandemics! Remember that nothing is permanent and soon this time will pass. Hopefully, the tips presented here may help you in your PhD journey and inspire you to identify and test your own strategies for surviving. Hang in there, we will get there!

Keywords: PhD, pandemic, raising a family

CONTRIBUTED POSTER PRESENTATIONS

Monday, 26 July 2021

Ghahramani, A.

Title: *Use of social media analytics for raising awareness of cardiovascular diseases risk factors in the female population of Australia*

Despite the popularity of social media in diverse populations and the importance of social media data, few studies analysed social media to increase awareness in health promotion campaigns and no study used social media analytics to raise awareness of cardiovascular diseases (CVD) risk factors in general and specifically for women. However, some studies analysed the developed social media platforms by the researchers and the effectiveness of social media intervention in existing health promotion campaigns has not been evaluated yet. There are challenges in how to utilise Artificial intelligence to extract and analyse the information from complex data sources such as social media. This study is designed to utilise R programming to bridge the gap between research and industry by using Machine Learning techniques as fast & cost-effective methods for social media analytics to analyse real-world big datasets in real-time. The outcome of this research will help enhance social media strategies through detecting the Key Performance Indicators (KPI) to measure the effectiveness of social media campaigns for raising awareness of CVD risk factors.

Keywords: social media analytics, health promotion, awareness

---

Pac Soo, V. & Bratt, S. & De Silva, A. & Peyton, P.

Title: *Between-centre differences in overall patient outcomes and in trial treatment effects in multicentre perioperative trials*

Background

In multicentre trials, patient outcomes have been shown to differ substantially due to differences across centres or due to characteristics of patients treated. It has been suggested that this variability may reduce the chance of finding a significant treatment effect in multicentre trials.

Aims

Our aim was to examine whether there are differences in the outcome or treatment effect on the outcome between centres and if heterogeneity affects the overall treatment effect. We used data from a large international multicentre randomised controlled trial (RCT) in anaesthesia using the primary trial endpoint and hospital length of stay (LOS).

Methods
We used mixed-effects logistic (primary endpoint) and Weibull (hospital LOS) regression to estimate the between-centre and treatment effect heterogeneity. We used meta-analysis to assess the heterogeneity of treatment effect between centres (I2). We used a non-inferential visual aid to allow comparison of the observed and expected differences between centres for exploring treatment effect heterogeneity. The overall treatment effect was estimated using fixed and random-effects logistic and Weibull models and meta-analytic approaches.

Results
There were between-centre differences in both outcomes. There was no important heterogeneity in the primary endpoint and substantial heterogeneity in the treatment effect for the hospital LOS between centres. The overall treatment effects were similar when centre-effects were included or ignored.

Conclusions
In this multicentre study, we did not find support for the hypothesis that between-centre differences in outcome or treatment effect affect the overall treatment effect in an RCT.

Keywords: RCT, multicentre trials, random-effects, fixed-effects

---

Udeh, R. & Advani, S. & De Guadiana Romualdo, L. G. & Dolje-gore, X.
Title: *Calprotectin Metadata in COVID-19*

Background:
COVID-19 has been shown to present with varied clinical course, hence the need for more specific diagnostic tools that could identify severe cases and predict outcomes during COVID-19 infection. Recent evidence has shown an expanded potential role for calprotectin, both as a diagnostic tool and as a stratifying tool in COVID-19 patients in terms of severity. Therefore, this systematic review and meta-analysis aims to evaluate the levels of calprotectin in severe and non-severe COVID-19 and also identify the implication of raised calprotectin levels.

Methods:
Databases searched include MEDLINE, EMBASE, The Cochrane Library, Web of science and MedRxiv. Stata was employed in meta-analysis to compare the serum/fecal levels of calprotectin between severe and non-severe COVID-19 infections.

Results:
A pooled analysis of data in the eight quantitative studies from 613 patients who were RT-PCR positive for COVID-19 (average age = 55 years; 52% males) showed an overall estimate as 1.34 (95%CI: 0.77, 1.91). Stata was further employed to carry out an in-depth investigation of the in-between study heterogeneity.

Conclusion:
Calprotectin levels have been demonstrated to be significantly elevated in COVID-19 patients who develop the severe form of the disease, and it also has prognostic importance. Also very relevant therapeutically, as a potential druggable target for tasquinimod in COVID-19.

Keywords: COVID-19, Calprotectin, Biomarker, Systematic review, Meta-analysis

Jacobson, J. & Cressie, N. & Zammit-Mangion, A.
Title: *Multivariate Spatial-Dependence Modelling of Satellite Data*

Spatial statistical prediction methods like kriging, leverage spatial covariation in the underlying process to produce de-noised and gap-filled predictions along with their statistical uncertainty. Multivariate spatial prediction methods leverage covariation both between and within the underlying multivariate spatial processes, which they do by fitting covariogram/variogram models and cross-covariogram/cross-variogram models to noisy and spatially incomplete multivariate data. Bivariate satellite data from NASA's Orbiting Carbon Observatory-2 (OCO-2) mission are featured, namely total column carbon dioxide (XCO2) data and solar-induced fluorescence (SIF) data. SIF is an indicator of photosynthetic activity and exhibits a lagged dependence with XCO2. Bivariate covariograms and cross-covariograms are fitted to the OCO-2 data over North America, and the talk finishes with a discussion of their scalability to global data.

Keywords: spatial, multivariate, variogram, cross-covariance, remote-sensing

---

Whyte, J M.
Title: *Keyword trends for chemicals may lead regulatory response. Could this hint at tomorrow's (unknown) poisons?*

Modern economies depend on - and produce - thousands of chemicals. History leads us to expect that some of these chemicals will have some (currently unknown) adverse effect on human or environmental health (more concisely, AE). Regulators aim to prevent AEs by conducting risk assessments, typically informed by experimental determination of chemical properties (e.g. persistence, toxicity). However, this approach is impractical when limited resources constrain a regulator's ability to gather data. Consequently, regulatory action may occur only after sufficient evidence of an AE has accumulated. One notable recent example of such "reactive" management is given by PFOS, once a widely used (fire-fighting foams to household goods) member of the PFAS (per- and poly- fluoroalkyl substances). The time between PFOS introduction and regulation has led to global concerns around the chemical's concentration and distribution in soil and water.

We would expect to apply timelier regulation in such cases if we could design a means of anticipating risk which does not depend on scarce data. We investigate such a "proactive" approach by interrogating features of scientific publications. We hypothesise that features of the time course of published chemical-AE associations may encourage a regulator to determine that chemical's properties. Through such focused efforts, regulators may better anticipate the emergence of potentially troublesome chemicals.

Our approach begins by querying Web of Science over a fixed publication year range to find papers featuring PFOS in their title, abstract, or keywords. Following "cleaning" of the

accompanying keyword lists, we record associations between PFOS and other terms, obtaining tallies for each year. Control charts allow us to judge those associations showing sustained growth over time. We investigate what information on PFOS is revealed by this approach, and consider how this could have assisted a regulator's response to an emerging problem.

Keywords: emerging contaminants, text mining, control charts

Awodutire, P O. & Nduka, E. & Ogoke, U.

Title: *A new two-parameter distribution: Properties, Regression Modelling and Applications to COVID 19 data*

In this study, a new two parameter continuous distribution is derived by compounding the Topp-Leone and one parameter distribution is proposed. Some statistical properties of this new distribution are discussed which includes moments, moment generating functions, renyi entropy and order statistics. The maximum likelihood estimation method was employed to estimate the unknown model parameters both under complete and censored situations in the present of covariates. In addition, simulation studies were conducted to assess the performance of the model parameters. Finally, the usefulness of this distribution is illustrated by application to COVID-19 data and discussed.

Keywords: Topp-Leone, COVID-19, covariates, maximum likelihood estimation

---

Tho, Z Y. & Zou, T. & Hui, F.

Title: *Joint Mean and Correlation Regression Modelling for Multivariate Data*

In the analysis of multivariate or multi-response data, researchers are often not only interested in studying how the mean (say) of each response evolves as a function of covariates, but also and simultaneously how the correlations between responses are related to one or more similarity/distance measures. For instance, when analysing community composition data, ecologists are interested in understanding how each species' distribution is related to environmental covariates, as well as how correlations between species (potentially reflected of biotic interactions) are related to phylogenetic and trait similarity. To address such research questions, we propose a novel joint mean and correlation regression model, which is applicable to a wide variety of correlated discrete and (semi)continuous responses. The model involves regressing the mean of each response against a set of covariates and the correlations between responses against a set of similarity measures observed either directly from the data or induced from available predictors information associated with each response. We develop a constrained optimisation procedure which iterates between solving estimating equations for the mean regression coefficients and correlation regression parameters. Under a general setting where the number of responses can tend to infinity with the number of clusters, we demonstrate that the proposed joint estimator is consistent and asymptotically normal, with differing rates of convergence. Simulations demonstrate the strong finite sample performance of the proposed estimator in terms of point estimation and inference. We apply the proposed joint mean and correlation regression model to a dataset of over dispersed counts of 38 Carabidae ground beetle species sampled throughout Scotland, with results showing in particular that beetle total length and breeding season have statistically important effects in driving the correlations between beetle species.

Keywords: Correlation regression, Joint mean-covariance model

Ashraf, G. & Arslan, J. & Crock, C. & Chakrabarti, R.

Title: *Sport-related Eye Trauma Study (SETS): Five-year audit of sports-related eye injuries at a tertiary eye hospital in Australia: 2015-2020*

Aim:
To examine outcomes of sports-related ocular injuries in an Australian tertiary eye hospital setting.

Methods:
Retrospective audit from the Royal Victorian Eye and Ear Hospital from 2015-2020. Patient demographics, diagnosis, and injury causation were recorded from baseline and follow-up. Outcomes included visual acuity (VA), intra-ocular pressure (IOP), ocular injury and management.

Results:
1793 individuals (mean age 28.7 years; 80.4% males and 19.6% females) presented with sports-related ocular trauma. Most injuries (99.2%) were unilateral. The top three injury-causing sports were soccer (n=327, 18.2%), Australian Rules Football (n=306, 17.1%) and basketball (n=215, 12.0%). The top three injury mechanisms were projectile (n=976, 54.4%), incidental body contact (n=506, 28.2%), and sporting equipment (n=104, 5.8%). The most frequent diagnosis was traumatic hyphaema (n=725).

Best documented baseline VA was ≥ 6/12 in 84.8%, 6/30-6/12 in 7.5% and < 6/30 in 7.7% of cases. Follow-up VA was ≥ 6/12 in 95.0%, 6/30-6/12 in 2.3% and < 6/30 in 2.7% of cases.

Multivariate logistic regression showed that the greatest risk of globe rupture was associated with martial arts (OR 16.2); orbital blow-out fracture with skiing (OR 14.4); hyphaema with squash (OR 4.2); and retinal tears with foam dart projectiles (OR 5.6) - $p < 0.05$ for all.

Topical steroids were the most common non-surgical treatment (n=693, 38.7%). CT orbits and facial bones were the most common investigation (n=184, 10.3%). The mean baseline IOP in the injured eye was 16.1mmHg; n=103 (5.7%) cases required topical anti-ocular hypertensive medication. 27 (1.5%) individuals were admitted to hospital and n=26 (1.5%) required surgery. Football contributed the most surgical cases (n=5, 19.2%).

Conclusion:
The top three ocular injury causing sports were soccer, Australian Rules Football, and basketball. The most frequent injury was traumatic hyphaema. Projectiles posed the greatest risk for injury.

Keywords: trauma, ophthalmology, emergency, sports, ocular, injury

Qiao, P.

Title: *A Bayesian model for inferring intratumour heterogeneity in copy number variation*

Cancer has long been understood as a somatic evolutionary process driven by genetic and epigenetic alterations. The accumulation of somatic mutations over time results in a clonal structure in cancer cell populations. Characterization of intratumoral heterogeneity is critical to understanding the natural histories of cancer cell populations and to guiding patient treatment. Although genetic variation is a well-studied source of intratumoral heterogeneity, the functional differences between clones captured by gene expression profiles remain unclear.

To investigate the genetic and transcriptional heterogeneity between clones in cell populations, we develop a Bayesian clustering model to cluster individual cells from single-cell RNA-sequencing data into sub-clones based on their copy number variations (CNVs). By integrating allelic imbalance and gene expression information, the model jointly identifies subclonal copy number profiles and assigns single-cells transcriptome profiles to their clone-of-origin. Due to the observation that nearby genes are likely to have the same copy number, which leads to spatial dependencies, sub-clonal copy number profiles are modelled as latent Markov chains. Thus, the key outputs utilized by the proposed model are cell assignment (represented as a matrix of indicators) and clonal copy number profiles (represented as Markov chains). We implement inference via Gibbs sampling.
Even though the current model focuses mainly on CNVs, it has a very flexible design in the sense that it can easily be extended to incorporate other genetic variation information, for example, somatic single-nucleotide alterations. We demonstrate through numerical experiments that our model is able to accurately identify the presence of copy number change and recover the underlying subclonal structure with improved clustering performance.

Keywords: single-cell, copy number, Hidden Markov Model

Dang, K D. & **Maestrini, L.**

Title: *Variational approximations for structural equation models*

Structural equation models are commonly used to capture the structural relationship between sets of observed and unobservable variables. In Bayesian settings, fitting and inference for these models are typically performed via Markov chain Monte Carlo methods that may be computationally intensive, especially for models with a large number of manifest variables or complex structures. Variational approximations can be a fast alternative; however, they have not been adequately explored for this class of models. We develop a mean field variational Bayes approach for fitting basic structural equation models. We show that this variational approximation method can provide reliable inference while being significantly faster than Markov chain Monte Carlo. Classical mean field variational Bayes may underestimate the true posterior variance, therefore we propose and study bootstrap to overcome this issue. We discuss different inference

strategies based on bootstrap and demonstrate how these can considerably improve the accuracy of the variational approximation through real and simulated examples.

Keywords: approximate inference; latent variables; nonparametric bootstrap

Fitzgerald, O. & Perez-Concha, O. & Rudd, L. & Metke-Jimenez, A. & Gallego-Luxan, B. & Jorm, L.

Title: *Deep learning for glycaemic control in the ICU*

Background:
A major goal of intensive care medicine is maintenance or achievement of physiological control. Decision support systems that provide personalised predictions of patient outcomes under various treatment scenarios are an area of active research. An ability to forecast future patient state accurately and reliably is required for the deployment of such systems. The optimal forecasting model would be 1) scalable to large datasets 2) robust to noise/sparsity in data availability 3) probabilistic, enabling anomaly detection and characterising risk of critical events. Deep learning, and in particular, neural ordinary differential equations (ODEs) offer a flexible framework for the development of physiological time series forecasting models that accounts for these requirements.

Methods:
Using the open-source electronic medical record (EMR) MIMIC-III critical care database we compare several deep learning approaches to the development of personalised forecasting models for blood glucose. The analysis dataset contains 12,047 patients, 571,063 blood glucose measurements, along with patient demographic, physiology, lab results, and treatment information. Due to the sporadic nature of the measurement process, we used a neural ODE approach to produce probabilistic forecasts for future blood glucose measurement. We evaluate the performance of the models by examining the calibration properties of the probabilistic predictions and point prediction accuracy.

Results:
Experiments show that the proposed neural ODE approach achieves a level of point predictive accuracy comparable with previous research, with prediction intervals achieving near nominal coverage levels. As with previous research performance is poorest in the tails of the blood glucose distribution.

Discussion:
Deep learning offers a flexible approach to the development of probabilistic forecasting systems for physiological time series. Ongoing research includes interrogation of model assumptions and assessment of generalisability across different datasets and populations.

Keywords: deep learning; medicine; longitudinal data analysis

Mohomed Amsar, M A. & Mery, H T. & Wang, Y G.

Title: *Modelling of Anisotropic Spatial Random Fields Using Mixture Copulas*

In spatial studies, one of the key assumptions is that the spatial process is isotropic. Although this assumption considerably reduces the complexity in the modelling and interpolation process, ignoring the directional dependence in spatial studies can lead to an inappropriate understanding of the physical phenomenon under study as well as making incorrect inferences. A novel spatial copula modelling approach is developed to model univariate anisotropic spatial random field using mixture copulas, which captures both spatial and joint dependence of multiple directions over two-dimensional locations. The joint dependence between multiple directions is captured using mixture copulas, and the spatial copula for the univariate anisotropic spatial random field is constructed as the convex combination of mixture copulas. The proposed model applied to the heavy metal cadmium concentration data that showed a reasonable directional dependence. Also, the performance of the novel model is compared with an existing spatial copula method which assume an isotropic spatial random field. The results show that the proposed method that incorporates mixture copulas with directional dependence outperformed in the reproduction of variable across locations.

Keywords: Anisotropic dependence; mixture copula; spatial copula

INDEX OF PRESENTERS