SASA
South African Statistical Association

The South African Statistical Association (SASA) Executive Committee are proud to host the 63rd Annual Conference of SASA. SASA 2022 is held in the beautiful town of George, situated in the middle of the scenic Garden Route. With a backdrop of the Outeniqua Mountains, this is the first time in SASA's history that our flagship event is hosted in the region. The SASA conference takes place from 28 November to 2 December 2022 at the King George Protea Hotel.

The SASA conference is the flagship South African statistics event for statisticians, analytics experts and data scientists from South Africa and aboard. This is a preeminent event on the statistics calendar in South Africa and brings statisticians and data experts from around the country and abroad together to share their research, discuss new ideas and to meet and establish collaboration.

For the SASA 2022 conference, SASA is emphasising the home-grown talent, expertise and vast experience held within our local universities, research institutions and industries. SASA 2022 is proud to have our local experts deliver keynote addresses alongside our invited international guests.

We as the SASA EC would like to thank everyone for their support and we wish everyone a wonderful conference.



Organising Committee
Warren Brettenny
Chantelle Clohessy
Inger Fabris-Rotelli
Charl Pretorius
Carel van der Merwe (Treasurer)

Scientific Committee
Allan Clark (Bayesian Statistics Interest Group)
Sugnet Lubbe (Multivariate Data Analysis Group)
Charl Pretorius (Chair)
Inger-Fabris Rotelli (Spatial Statistics Interest Group)
Sheetal Silal (Editor: Conference Proceedings)
Andréhette Verster (Extreme Value Theory Interest Group)

Sponsors

SACNASP
South African Council for Natural Scientific Professions

IBM®

ucdp
UNIVERSITY CAPACITY DEVELOPMENT PROGRAMME

Registration

Registration for the conference with take place in the foyer of the Regency Hall, King George Protea Hotel, at the following times:

Monday (28 November):	08:00 – 10:00 and 14:00 – 17:00
Tuesday (29 November):	08:00 – 10:00 and 14:00 – 17:00
Wednesday (30 November):	07:30 – 13:00

All queries can be directed to the assistants manning the registration desks.

Name tags provided to delegates at registration must be worn at all times to gain access to the venues, tea breaks, lunches and social functions.

Tea and meals

Tea will be served in the Regency Hall foyer and patio area of the King George Protea Hotel during the workshops and conference.

Lunch will be served in the Fairway Terrace Restaurant at the King George Protea Hotel. Please take note that there is a premium on space in the lunch venue and delegates are asked to move through the system in an orderly and prompt manner to simplify the space restriction.

Meetings and social functions

Meeting/function	Time	Venue
SASA Executive Meeting	Tue., 29 Nov., 10:30 – 11:00	Windsor Boardroom
Opening Ceremony	Wed., 30 Nov., 08:30 – 10:30	Foyer, Regency Hall, King George Hotel
Welcoming Function	Wed., 30 Nov., 18:30 – 20:30	Patio at the Fairway Terrace Restaurant
Young Statisticians' Function	Wed., 30 Nov., 20:30 – 22:00	Patio at the Fairway Terrace Restaurant / Rex Tavern at the King George Hotel
SASA AGM	Thu., 1 Dec., 15:30 – 16:15	Regency Hall, King George Hotel
Gala Dinner	Thu., 1 Dec., 19:00 for 19:30	Regency Hall, King George Hotel

Photography

The official photographer is Aviwe Gqwaka.

Internet

Wireless internet will be made available to the delegates at King George Protea Hotel during the conference week.

Presenter guidelines

Speakers

- Double-check the date and time of your presentation.
- Report to the chairperson of the session at least 10 minutes before the start of the session.
- All presentations should be loaded onto the computer in the venue before the start of the session. Each venue will have personnel that will assist you in loading the presentation before the start of the session.
- Keep to the time allocated for your presentation (strictly 15 minutes for your presentation and 5 minutes for questions). The chair of the session will warn you when you have 5 of your allocated 15 minutes remaining, and again when your time is up. Once the chairperson has indicated the end of your presentation, you have to stop immediately.
- You are not allowed to move your time session to any other slot.
- Laser pointers will be available from the session assistants.

Chairpersons

- Keep to scheduled times.
- No changes are to be made to the programme. All presentations must start at the time indicated in the programme.
- Check the attendance of all speakers prior to the start of the session and ensure that all presentations have been loaded on the computer by the assistant.
- Open the session by welcoming the delegates and speakers and be sure to make the following announcements:
  - All cell phones should be switched off.
  - State the theme of the session/stream.
  - For each presentation, state the presenter's name and the title of the presentation.
- Warn speakers 5 minutes before the end of the 15 minutes allocated to the presenters. Allow questions according to time (i.e. the presentation and all questions should not exceed 20 minutes).
- Thank all speakers and delegates at the end of the session
- Report problems and absent speakers to the assistant.

The above instructions are intended as a basic guideline for the sessions. Please use your own initiative in the sessions to keep them running smoothly.

Poster presenters

All posters will be viewed on *Wednesday 30 November from 18:30 to 19:30* on the patio at the Fairway Terrace Restaurant.

Poster presenters must set up their posters during the afternoon tea break on Wednesday. As per the information circulated to delegates prior to the conference, a board which can take a poster of size A1 in portrait format will be available for each delegate. Boards will be demarcated with a delegate's name to assist with the poster competition assessment. Please use the board that has been allocated to you and do not remove names from the boards.

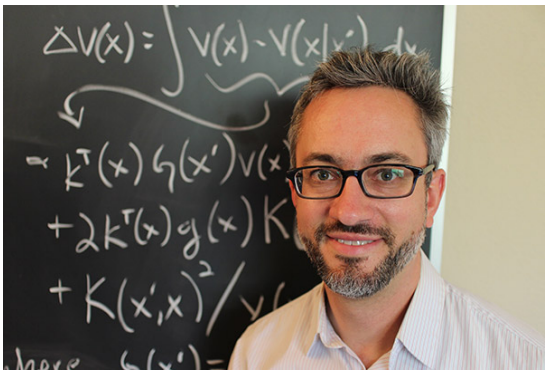
Invited speakers

Ruth Etzioni | *Professor at Fred Hutchinson Cancer Research Center*

Ruth Etzioni received her BS in Mathematics and Computer Science from the University of Cape Town in 1984. A chance encounter with a job advertisement on a department bulletin board led her to the field of statistical modelling and an Honours degree in Statistics and Operations Research at UCT. Following a meeting with Stephen E. Fienberg, the invited speaker at the SASA conference in 1985, she moved to the US where she completed a PhD in Statistics at Carnegie-Mellon University in 1990. She is a full professor at the University of Washington and the Fred Hutchinson Cancer Center where she holds the Rosalie and Harold Rea Brown endowed chair. Her research focuses on the gaps in evidence that inevitably arise when making health care decisions, whether at the individual or policy level. Much of her work has been in the area of cancer diagnostics and early detection, particularly in breast and prostate cancer where she has leveraged disease trends from population registries to learn about unobservable but policy-driving phenomena including disease natural history and overdiagnosis. She works closely with the surveillance program at the National Cancer Institute, serves on several national cancer screening guidelines panels and was a member of a COVID modelling and policy group that informed the Washington State governor's responses to the pandemic. She is a fellow of the American Statistical Association and has been the chair of its Health Policy Statistics Section. She has developed a popular graduate-level course on the modelling of health outcomes using large observational databases on which her textbook, "Statistics for health data science: an organic approach," is based.

Robert Gramacy | *Professor at Virginia Polytechnic and State University*

Prof Robert Gramacy is a Professor of Statistics in the College of Science at Virginia Polytechnic and State University (Virginia Tech/VT) and affiliate faculty in VT's Computational Modeling and Data Analytics program. Previously, Prof Gramacy was an Associate Professor of Econometrics and Statistics at the Booth School of Business, and a fellow of the Computation Institute at The University of Chicago. His research interests include Bayesian modelling methodology, statistical computing, Monte Carlo inference, nonparametric regression, sequential design, and optimisation under uncertainty.



Edzer Pebesma | *Professor at University of Münster, Institute for Geoinformatics*

Edzer Pebesma leads (since 2007) the spatiotemporal modelling lab at the Institute for Geoinformatics of the University of Münster, Germany. His research interests include spatial statistics, analysis of environmental and remote sensing data, open science and reproducibility, interoperability and open source software development. In the past he was editor of Computers & Geosciences and the Journal of Statistical Software, and associate editor of Spatial Statistics. He is an ordinary member of the R foundation, and is a major contributor to the R spatial package ecosystem.

Mark Nasila | *Chief Data and Analytics Officer at FirstRand Risk*

Dr Mark Nasila is the Chief Data and Analytics Officer of FirstRand Risk, and a Singularity University Faculty member. He is also a steering committee member of the National Institute for Theoretical Physics and Computational Sciences (NITheCS). As an experienced AI and data science expert, he ensures the techniques and methodologies he introduces into FNB are at the forefront of where banking is headed, both locally and internationally. He is the developer and the brain behind Manila, an AI system FNB has harnessed to reimagine its risk management and forensic due diligence processes. Dr Nasila has presented at local and international conferences, such as the Business-Tech Digital Banking Conference (SA), Chief Data & Analytics Officer forums in South Africa and the UK, European Simulation & Modelling Conference. He has published and written opinion pieces in Business Day, Mail & Guardian, TechCentral, Daily Maverick, CIO Talk Network, CIO.com, Corinium Intelligence and the European Simulation & Modelling Association. He holds a PhD in Mathematical Statistics from the Nelson Mandela University, and is also an alumni of the SingularityU South Africa Executive programme. He was named one of the Corinium Global Intelligence '2020 Global Top 100 Innovators in data and analytics'.

Sheetal Silal | *Director of Modelling and Simulation Hub, Africa (MASHA)*

Prof Sheetal Silal is the Director of the Modelling and Simulation Hub, Africa (MASHA) and associate Professor in the Department of Statistical Sciences at the University of Cape Town (UCT). She is an Honorary Visiting Research Fellow in Tropical Disease Modelling at the Nuffield Department of Medicine at Oxford University. She received a PhD in Mathematical Modelling of Infectious Diseases in 2014 from UCT. Her primary research area is the development and application of mathematical models to malaria, pertussis, syphilis, COVID-19, and other infectious diseases in South Africa, sub-Saharan Africa and globally, with a focus on using mathematical models to predict the dynamics and control of diseases to evaluate the potential impact of control programmes in reducing morbidity and mortality, and supporting policy development. Prof Sheetal Silal is leading the development of COVID-19 transmission models as part of the South African COVID-19 Modelling Consortium. The Modelling Consortium is a group of researchers from academic, non-profit, and government institutions across South Africa where the mandate of the group is to provide, assess and validate model projections to be used for planning purposes by the Government of South Africa.

Ashwell Jenneker | *Deputy Director General of StatsSA*

Ashwell Jenneker has been the DDG for Statistical Support and Informatics at Statistics South Africa since 2008. Ashwell started his career at Stats SA as a Statistical Training Officer before progressing to Deputy Director of Marketing, Acting Director of Communication for the Census 2001, Executive Manager of Statistical Information Services and Executive Manager of the Data Management and Information Delivery Project (DMID) before his DDG appointment where he is currently responsible for Statistical Operations and Provincial Coordination. Ashwell has a Masters degree in Urban and regional sciences from the Stellenbosch University in South Africa and has a BSc (honours) in mathematics and a higher diploma in education from the University of Western Cape. He is a Fellow of the seventh class of the Africa Leadership Initiative-South Africa.

Renette Blignaut | *2021 SAS® Thought Leader and Professor at University of the Western Cape*

Renette Blignaut manages the UWC Data Science programmes. She as been the HOD or DHOD of the Statistics and Population Studies Department at UWC for 24 of the 30 years in the academic environment. Prior to this she worked in industry for 9 years whilst completing her Honours in Mathematical Statistics (Pretoria), M.Sc. in Mathematical Statistics (UCT) and her PhD (Pretoria) on a part-time basis. Her PhD thesis: "Modelling of claim patterns of members of private medical aid schemes" was completed in 1995. Since then she has been conducting research in various fields namely: data mining, statistical learning, predictive modelling, biostatistics, mobile security, internet access and science education. In 2001, she was promoted to associate professor and in 2015 promoted to full professor. She has published more than 78 accredited articles, 22 technical reports, and has presented research at more than 53 international conferences, 50 national conferences and 56 workshops. She has supervised or co-supervised 29 masters and 6 PhD students.

In 2021, Renette was awarded the South African Statistical Association's Thought Leader Award for significant contributions in academia, industry, government and elsewhere. She is the second female to have been awarded this award.

Renette currently serves as Chairperson of the Data Science Academic Committee of the National Graduate Academy for Mathematical and Statistical Sciences (NGA MaSS). She furthermore serves on the Steering Committee and the Advisory Board of the NGA MaSS. She was appointed on the Governing Board of the Centre for Multi-dimensional Data Visualisation at Stellenbosch University (Apr 2022–Mar 2025).

Workshop presenters

Rajat Mukherjee | *VP Advanced Statistics & Data Sceince at Alira Health*

Rajat has over 25 years of experience in Biostatistics within both pharmaceutical companies, and academic institutions.

Rajat joined Alira Health in April 2022 as VP Advanced Statistics and Data Sciences. In his current role his main responsibilities include definition and implementation of long-term business strategies in collaboration with Alira Health leaders from various practices such as Market Access, RWE, Regulatory and Management consulting and to provide strategic statistical consulting for clinical pipelines and clinical trials. His main areas of expertise include Bayesian and adaptive trial designs, application of machine learning and AI for diagnostics and biomarker discovery and statistical computation to facilitate the incorporation of complex endpoints in clinical trials. Rajat is also keen in developing statistical software to empower statisticians and clinical trialists.

In his previous position Rajat acted as a Senior Research Fellow/Consultant for Cytel, leading innovative clinical trials design activities including trials under the CID (FDA) program. In addition, his responsibilities included consulting in areas of Bayesian adaptive designs, big-data, biomarker discovery and machine learning. He was a part of the executive team for developing consulting business strategies, thought leadership activities (research publication, conferences, etc.), and mentoring fellow consultants.

Rajat holds a PhD degree in Mathematical Statistics from the University of Wisconsin-Madison (USA), an MSc degree in Applied Sciences from Bowling Green State University, Ohio (USA), and a BSc degree in Statistics from the Presidency College, Calcutta (India).

Berthold Lausen | *Professor at University of Essex*

Berthold Lausen has 35 years of experience as medical and trial statistician and data scientist. He contributed to the first conference of the International Federation of Classification Societies (IFCS) in 1987 with a cutting-edge proposal of a three-objects variance estimator and parametric bootstrap to evaluate the confidence and stability of unsupervised learning with DNA-DNA hybridisation data (with P. O. Degens). He made influential contributions to develop machine learning and nonparametric methods for medical statistics and data science as for example maximally selected rank statistics, p-value adjusted classification and regression trees, bagging survival trees (with M. Schumacher and others). He is past president of IFCS (2020-22; former president 2018-19, president elect 2016-17), a former president of Data Science Society (GfKl) (2013-19, vice-president 1995-2001, 2004-13) and the founding vice-president of European Association for Data Science (EuADS) 2015-18. Since 1986 he is member of the International Biometric Society (IBS) and Data Science Society (GfKl), since 1998 a fellow of the Royal Statistical Society (RSS) and since 2015 a founding member of the European Association for Data Science (EuADS).

Since 2010 Lausen is Professor in Biometry and Epidemiology at School of Medicine of the University of Erlangen-Nuremberg and 2012-17 full Professor of Statistics and since 2017 of Data Science at the Department of Mathematical Sciences, University of Essex. 2016-21 Head of Department, while the department grew

substantially (624 students - up 158% from 242 students in 2017, 37 permanent academic staff - up 118% from 17 permanent academics in 2017). He is on sabbatical (2021-23).

He is principal investigator of data science driven Knowledge Transfer Partnerships (KTP) of Innovate UK with Profusion (funded 2014-22), KTP with Mondaq (funded 2016-2022) and a partnership with Rolls-Royce Motor Cars (funded 2021-25). Contributor to the Business and Local Government Data Research Centre, University of Essex, ESRC (funded 2019-22).

He is an associated editor of Behaviormetrika, and a member of the editorial board of Advances Data Analysis and Classification, Archives of Data Science Series A and of Methods of Information in Medicine.

Dianne Cook | *Professor at Monash University*

Dianne Cook is Professor of Business Analytics at Monash University in Melbourne, Australia. Her research is in the area of data visualisation, especially the visualisation of high-dimensional data using tours with low-dimensional projections, and projection pursuit. A current focus is on bridging the gap between exploratory graphics and statistical inference. Technology plays an important role in data visualisation for evaluating effectiveness and for public consumption. Di utilises technology such as virtual environments, Amazon's Mechanical Turk, and eye-tracking in her work, and makes an effort to share her work with open source software. Dianne is a Fellow of the American Statistical Association, elected member of the International Statistical Institute, past-editor of the Journal of Computational and Graphical Statistics, elected member of the R Foundation, and current editor of the R Journal. Education is an important part of her contributions, and mentoring graduate research is a significant activity. Several of her students have won the prestigious American Statistical Association John Chambers Software Award, including Hadley Wickham, Yihui Xie, Carson Sievert, and most recently, Monash student Earo Wang.

David Hofmeyr | *Senior Lecturer at Stellenbosch University*

David is currently employed as a senior lecturer in the department of statistics and actuarial science at Stellenbosch University, a post which he has held since January 2017. David sits on a number of committees, both internal and external to the university, including for the National Graduate Academy for Data Science. Prior to joining Stellenbosch University, David held a one year post doctoral fellowship within the STOR-i Center for Doctoral Training at Lancaster University.

David received his PhD from the STOR-i CDT at Lancaster, where his research was on finding optimal projections for cluster analysis. He has a master of research from Lancaster in statistics and operations research, and a master of science from Edinburgh University in operations research. His bachelor's degree had as majors mathematics and mathematical statistics, and his honours is in pure maths (from WITS and UCT respectively).

David continues to work on the problem of finding optimal projections for clustering, and has also extended his scope into (primarily) the areas of model selection, nonparametric smoothing, and the clustering problem in general. However, he has peripheral interest in a large collection of areas relevant to data science, machine learning and methodological statistics.

Sudipo Banerjee | *President at ISBA*

Sudipto Banerjee is currently Professor and Chair of the UCLA Department of Biostatistics with joint appointments in the UCLA Department of Statistics and the UCLA Institute of the Environment and Sustainability. He also serves as the 2022 President of the International Society for Bayesian Analysis. He has worked on a number of problems on spatial statistics, developing theory and methods related to Bayesian modelling and inference for geographic data with wide-ranging applications in public and environmental health sciences, ecology, forestry, real estate economics and agronomy. He has published 2 textbooks and an edited handbook. His contributions have been recognised through many honours, including the Abdel El-Shaarawi Award from The International Environmetric Society (TIES), the Mortimer Spiegelman Award from the American Public Health Association and the George W. Snedecor Award from the Committee of Presidents of Statistical Societies (COPSS), elected membership of the International Statistical Institute, elected fellowships in the Institute of Mathematical Statistics (IMS), the American Statistical Association (ASA), the International Society for Bayesian Analysis and the American Association for the Advancement of Science (AAAS), a Distinguished Achievement Medal from the ASA's Section on Statistics and the Environment, and the ASA's Outstanding Statistical Application Award.

Workshops

Workshop schedules and venues are given on page 13.

Ruth Etizioni | *Fred Hutchinson Cancer Research Center*

Models for health outcomes using data from population registries and surveys

This workshop will present methods for analysing non-normal outcomes in health data studies with a focus on counts and health care costs. The workshop will cover different regression modelling frameworks tailored to the distributional properties of these outcomes with examples drawn from two major data sources in the US – a national cancer registry and a national health survey that includes annual health expenditure information. Additionally, the G-computation method for marginal effect estimation in non-linear regression models, propensity score analysis with inverse probability weighting for causal effect estimation in observational studies, and methods for accommodating complex survey designs will be covered. The properties, strengths and weaknesses of registries and surveys as sources for health outcomes models will also be discussed. All analyses will be programmed in R and code will be provided to all workshop participants. This workshop will draw on material from the text, "Statistics for Health Data Science: An Organic Approach," co-authored by Dr Etzioni.

Robert Gramacy | *Virginia Polytechnic and State University University*

A practical introduction to Gaussian process regression and Bayesian optimization

Gaussian process regression is ubiquitous in spatial statistics, machine learning, and the surrogate modelling of computer simulation experiments. Fortunately their prowess as accurate predictors, along with an appropriate quantification of uncertainty, does not derive from difficult-to-understand methodology and cumbersome implementation. We will cover the basics, and provide a practical tool-set ready to be put to work in diverse applications. As one example of an one application where Gaussian processes play a fundamental role, we will introduce Bayesian optimization. The presentation will involve accessible slides authored in Rmarkdown, with reproducible examples spanning bespoke implementation to add-on packages.

Edzer Pebesma | *University of Münster, Institute for Geoinformatics*

An introduction to spatial data science with R

This workshop will give an introduction into handling and analysing spatial vector and raster data with R, and exemplify a number of spatial statistical methods including point pattern analysis, geostatistical analysis, and lattice data analysis. The workshop will focus on R packages sf and stars, and a number of analysis packages that are compatible with these. Some prior experience with R is strongly recommended.

Berthold Lausen | *Virginia Polytechnic and State University University*

New developments in data science and data science in Africa

Multivariate Data Analysis Group Workshops

Dianne Cook | *Department of Econometrics and Business Statistics Monash University, Australia*

Visual methods for multivariate data – a journey beyond 3D

This workshop will explain how to use dynamic plots constructed from low-dimensional linear projections, called tours, to examine multivariate data spaces. The tour projections are read similarly to a biplot. There are several tour types, grand, guided, manual, local, and slice, that are useful and you will learn about. These can be helpful when conducting analyses involving non-linear dimension reduction, like t-SNE, and machine learning models, both supervised and unsupervised classification. We will include working with high-dimension, low-sample size data.

Bring your laptop, loaded with R, RStudio and the R package tour, to follow along with me.

David Hofmeyr | *Department of Statistics and Actuarial Science, Stellenbosch University*

Independent component analysis

Components analysis, broadly speaking, refers to the problem of modelling the primary sources of information contained in a set of data. In the linear context, which forms the framework for the present discussion, these sources are represented by linear combinations of the variables on which the observations in the data have been measured. By far the most well known is the Principal Components Analysis (PCA) model, where the objective is to retain as much variation from the data as possible, by minimising the squared residuals between the original observations and their projection onto the principal components subspace.

Although already decades old, and very popular in many fields, the Independent Components Analysis (ICA) model is less well known among the "traditional statistics" community. The problem of ICA is to identify the (statistically) independent sources of information in the data. Perhaps surprisingly, given the seemingly very difficult objective in the abstract, when the data generating distribution can be described through a linear "mixing" of statistically independent source variables, even fairly simple objectives lead to consistent estimation; with the only proviso being that the sources are non-Gaussian.

In this talk I will introduce the ICA problem in greater detail, and discuss some important applications where it has had tremendous impact. I will then introduce some of the more common methods for the estimation of independent components (ICs), as well as some work of my own on enhancing the computational aspects of a more direct approach based on nonparametric pseudo-likelihood maximisation. I will also touch briefly on the problem of estimating ICs in the online context, where estimation needs to be conducted in real time with the receipt of a stream of data seen only one at a time.

Bayes Interest Group Workshop

Rajat Mukherjee | *Alira Health*

Use of informative priors in confirmatory studies, along with a hands-on session in R

The workshop in Bayesian statistics is aimed to provide industry researchers (statisticians as well as domain experts), academicians and students working in medicine and healthcare with an introduction to the topic along with some specific examples and use cases from the pharmaceutical industry. The workshop will focus on translating historical data, for example, from previously conducted randomised clinical trials into informative priors for the parameters of interest which can then be used in the design and analysis of future trials. This approach of Bayesian-Borrowing is gaining interest particularly for investigations in rare diseases and for medical devices. We will discuss a common problem of Prior-Data conflict in this setting and methodologies to control borrowing from historical data in the presence of a conflict. We will also be discussing a recently conducted trial COVID vaccine trial that was conducted in the Bayesian framework. The workshop will conclude with a hands-on session implementing a Bayesian design using the open source R software. Participants are encouraged to install R and the following packages on their laptops prior to attending the workshop: ggplot2, RBesT, parallel, mcmc, mvtnorm, rstan, rstanarm.

Attendance of this workshop will equate to 1 CPD (SACNASP) point.

The workshop will conclude with a address from the ISBA president, Prof Sudipto Banerjee, who will join remotely.

The Bayes Interest Group Workshop is proudly sponsored by Alira Health.

Full Programme

Workshop Day 1: Monday, 28 November

Bayes Interest Group Venue: Regency Hall

09:00 – 10:30	**Rajat Mukherjee** — *Use of informative priors in confirmatory studies* (Session 1)
10:30 – 11:00	Morning tea (Regency Hall Patio and Foyer)
11:00 – 12:30	**Rajat Mukherjee** — *Use of informative priors in confirmatory studies* (Session 2)
12:30 – 13:30	Lunch (Fairway Terrace Restaurant)
13:30 – 15:00	**Rajat Mukherjee** — *Use of informative priors in confirmatory studies* (Session 3)
15:00 – 15:30	Afternoon tea (Regency Hall Patio and Foyer)
15:30 – 17:00	**Sudipto Banerjee** — *ISBA President Address*

Multivariate Data Analysis Group Venue: George Room

09:00 – 10:30	**Dianne Cook** — *Visual methods for multivariate data* (Session 1)
10:30 – 11:00	Morning tea (Regency Hall Patio and Foyer)
11:00 – 12:30	**Dianne Cook** — *Visual methods for multivariate data* (Session 2)
12:30 – 13:30	Lunch (Fairway Terrace Restaurant)
13:30 – 15:00	**David Hofmeyr** — *Independent Component Analysis*
15:00 – 15:30	Afternoon tea (Regency Hall Patio and Foyer)
15:30 – 16:00	MDAG Annual General Meeting

Edzer Pebesma

An introduction to spatial data science using R Venue: Charlotte Room

12:30 – 13:30	Lunch (Fairway Terrace Restaurant)
13:30 – 15:00	Workshop Session 1
15:00 – 15:30	Afternoon tea (Regency Hall Patio and Foyer)
15:30 – 17:00	Workshop Session 2

Workshop Day 2: Tuesday, 29 November

Ruth Etzioni

Models for health outcomes using data from population registries and surveys Venue: Regency Hall

09:00 – 10:30	Workshop Session 1
10:30 – 11:00	Morning tea (Regency Hall Patio and Foyer)
11:00 – 12:30	Workshop Session 2
12:30 – 13:30	Lunch (Fairway Terrace Restaurant)

Robert Gramacy

A practical introduction to Gaussian process regression and Bayesian optimization Venue: Regency Hall

12:30 – 13:30	Lunch (Fairway Terrace Restaurant)
13:30 – 14:30	Workshop Session 1
14:30 – 14:45	Afternoon tea (Regency Hall Patio and Foyer)
14:45 – 16:00	Workshop Session 2

Berthold Lausen

New developments in data science and data science in Africa Venue: Charlotte Room

13:00 – 14:30	Workshop Session 1
14:30 – 14:45	Afternoon tea (Regency Hall Patio and Foyer)
14:45 – 16:00	Workshop Session 2

Wednesday, 30 November

07:30 – 13:00	**Registration**			Venue: Regency Hall
08:30 – 09:40	**Opening Ceremony** 08:30 Welcome: Chantelle Clohessy 08:45 President's address: Warren Brettenny 09:15 Awards: SAS awards for best honours project Postgraduate paper competition winner Sichel Medal Fellowship and Honorary Members Thought Leader (sponsored by SAS)			Venue: Regency Hall
09:40 – 10:40	**Plenary Session** **Ruth Etzioni** — *Statistical models for understanding population cancer trends and informing health policies* (p. 21)			Venue: Regency Hall Chair: Warren Brettenny
10:40 – 11:00	**Morning tea**			Venue: Regency Hall Patio and Foyer
11:00 – 12:00	**Plenary Session** **Edzer Pebesma** — *Spatial statistical questions and big spatial datasets* (p. 21)			Venue: Regency Hall Chair: Inger Fabris-Rotelli
12:10 – 13:10	Stream:	**Data Science**	**General**	**Young Stats (Multivariate Statistics)**
	Venue:	Regency Hall	Charlotte Room	George Room
	Chair:	Sonali Das	Zani Ludick	Tanita Botha
	12:10	Matthew Dicks (p. 21) *A simple learning agent interacting with an agent-based market model*	Ariane Neethling (p. 22) *Theory versus Practice: Reflections on sampling and survey data*	Noëlle van Biljon (p. 23) *Investigating the Heterogeneous Structure of Multivariate Anthropometric Growth Profiles*
	12:30	Sandile Shongwe (p. 21) *Evolutionary support vector regression for monitoring Poisson profiles*	Jeanette Pauw (p. 22) *An introduction to the characterisation of a research design through the MIDA framework with an example from a recent waste collection study*	Zoë-Mae Adams (p. 23) *Exploring sentiment contained in COVID-19 related Tweets with embedded word MCA biplots*
	12:50	Sonali Das (p. 22) *Are winters becoming shorter and warmer? Insights from a functional data analysis investigation*	Zani Ludick (p. 22) *A composer's playground: Statistical linguistic measures & self-similarity matrices in the assessment and generation of musical structure.*	Tanita Botha (p. 24) *Wasserstein distance as discriminator within the Dirichlet family*
13:10 – 14:00	**Lunch**			Venue: Fairway Terrace Restaurant
14:00 – 15:00	**Plenary Session** **Robert Gramacy** — *Deep Gaussian process surrogates for computer experiments* (p. 24)			Venue: Regency Hall Chair: Roelof Coetzer

Wednesday, 30 November

15:00 – 16:00	Stream:	**Educational Statistics**	**Data Science**	**Young Stats (Bayesian Statistics)**
	Venue:	Regency Hall	Charlotte Room	George Room
	Chair:	Inger Fabris-Rotelli	Stefan Britz	Trudie Strauss
	15:00	Inger Fabris-Rotelli (p. 24) *Development of an early career academic supervisor in Statistics in South Africa*	Sagaren Pillay (p. 25) *Principal component analysis (PCA) on annual financial statements of large South African manufacturing enterprises*	Makwelantle Asnath Sehlabana (p. 25) *Modelling the Misuse of Alcohol and Drugs in South Africa Using Bayesian Binary Logistic Regression*
	15:20		Andre Ruben Kleynhans (p. 25) *Robust Self-Paced Learning Algorithm For Finite Mixture Models*	Nobuhle Mchunu (p. 26) *Using joint models to study the association between CD4 count and the risk of death in TB/HIV data*
	15:40		Stefan Britz (p. 25) *Feature Engineering for Tennis Match Outcome Prediction*	Trudie Strauss (p. 26) *Word Frequency Distributions: A Comprehensive Bayesian Approach*
16:00 – 16:20	**Afternoon tea**			Venue: Regency Hall Patio and Foyer
16:20 – 18:00	Stream:	**Educational Statistics**	**Extreme Value Theory**	**Young Stats (General)**
	Venue:	Regency Hall	Charlotte Room	George Room
	Chair:	Paul van Staden	Andréhette Verster	Praise Obanya
	16:20	Aviwe Gqwaka (p. 27) *Efficiency Analysis of South African Schools: A Parametric Approach*	Jan Beirlant (p. 28) *Outlier detection based on extreme value theory and applications*	Lethani Ndwandwe (p. 29) *On testing for the Pareto distribution using U and V statistics*
	16:40	André Zitzke (p. 27) *The value proposition for industry-academic collaboration*	Richard Minkah (p. 28) *Robust Extreme Quantile Estimation for Pareto-Type tails through an Exponential Regression Model*	Ryan Shackleton (p. 30) *COVID-19 and Volatility of International Stock Markets: An FDA Investigation*
	17:00	Ruffin Mpiana Mutambayi (p. 27) *Analysis of Strike action on students' Academic Performance in the Inferential Statistics Module at the University of Fort Hare, South Africa*	Daniel Maposa (p. 28) *Modelling temperature extremes in the Limpopo province: Bivariate time-varying threshold excess approach*	Ané van der Merwe (p. 30) *Negative binomial compounding in a discrete Lindley model with INAR(1) application*
	17:20	Lindo Magagula (p. 27) *Affective Learning: An insight into Mr Lindo's #OperationFinishTheSemesterStrong*	Matthys Lucas Steyn (p. 29) *Open-set Recognition using Excesses of Distance Ratios*	Liliane Tendela (p. 30) *Investigating the effectiveness of an undergraduate mathematics intervention at UWC*
	17:40	Paul Jacobus van Staden (p. 27) *When should Paul visit Paris? A time series case study from an introductory first-year statistics & data science course*	Thakhani Ravele (p. 29) *Estimation of extreme quantiles of GHI: A comparative analysis using an extremal mixture model and a generalised additive extreme value model*	Praise Obanya (p. 30) *Permutation entropy analysis of financial markets*
18:30 – 19:30	**Poster Session** (see p. 19)			Venue: Patio at the Fairway Terrace Restaurant
18:30 – 20:30	**Welcoming Function**			Venue: Patio at the Fairway Terrace Restaurant
20:30	**Young Statisticians' Function**			Venue: Patio at the Fairway Terrace Restaurant / Rex Tavern at the King George Hotel

Thursday, 1 December

08:00 – 10:00	Stream:	**Bayesian Statistics**	**Spatial Statistics**	**Young Stats (Data Science)**
	Venue:	Regency Hall	Charlotte Room	George Room
	Chair:	Allan Clark	Inger-Fabris Rotelli	Annegret Muller
	08:00	Sean van der Merwe (p. 32) *Robust inference in the presence of censoring, skewness, and extreme values*	Ansie Smit (p. 34) *Modelling probabilistic hail hazard in South Africa: following the swath*	Erika Slabber (p. 36) *Factorisation machines as a statistical modelling technique*
	08:20	Annibale Cois (p. 32) *Bayesian meta-regression models for the estimation of population trends in health risk factors*	Michelle de Klerk (p. 34) *Spatial prediction on disjoint spatial lattice data*	Edward Westraadt (p. 36) *Classification of Photovoltaic Module Faults Using a Novel Deep Learning Architecture*
	08:40	Hassan Sadiq (p. 32) *difFUBAR: Scalable Bayesian comparison of selection pressure*	Adeboye Azeez (p. 34) *Bayesian Structured Additive Spatial Model of Intimate Partner Violence among Women in Nigeria*	Nyiko Khoza (p. 36) *Uncertainty problem in sampling*
	09:00	Lulama Kepe (p. 33) *Bayesian Tree Growth modelling. An investigation into individual tree competition.*	Jenny Holloway (p. 35) *Age-stratified COVID-19 epidemiological model*	Liam Carew (p. 36) *Application of CNN-gcForestCS to cassava leaf disease detection*
	09:20	Allan Clark (p. 33) *Bayesian Analysis of Historical Functional Linear Models with application to air pollution forecasting*	Fallo Happy Khanye (p. 35) *An exploratory analysis of location information from the mobileDNA application*	Tshepiso Selaelo Rangongo (p. 37) *Multivariate big data sampling for crop area coverage*
	09:40	Chun-Sung Huang (p. 33) *Optimal window size detection in Value-at-Risk forecasting: A case study on conditional generalised hyperbolic models*	Inger Fabris-Rotelli (p. 35) *Linear hotspot detection for a point pattern in the vicinity of a linear network*	Annegret Muller (p. 37) *Label-dependent splitting for multi-label data*
10:00 – 10:30	**Morning tea**			Venue: Regency Hall Patio and Foyer
10:30 – 12:00	**Keynote Session**			Chair: Warren Brettenny, Venue: Regency Hall
	10:30	**Mark Nasila** — *Emerging technologies, globalization and social challenges are redefining the role of intelligent decisioning* (p. 37)		
	11:15	**Sheetal Silal** — *Towards evidence-based public health management: The role of statistics and modelling in South Africa* (p. 37)		

Thursday, 1 December

12:00 – 13:00	Stream:	**Biostatistics**	**Nonparametric Statistics**	**Young Stats (Spatial Statistics)**
	Venue:	Regency Hall	Charlotte Room	George Room
	Chair:	Sisa Pazi	Jean-Claude Malela-Majika	Renate Thiede
	12:00	Kgethego Sharina Makgolane (p. 38) *Examining factors that contribute to under-five mortality rates in South Africa using count models*	Shawn Liebenberg (p. 39) *An investigation on the use of Bernstein polynomials in entropy estimation*	René Stander (p. 39) *Multiscale decomposition of spatial lattice data to detect hotspots of COVID-19 cases in South Africa*
	12:20	Isaac Singini (p. 38) *Joint modeling for longitudinal and interval censored survival data*	Marien Alet Graham (p. 39) *Statistical process control: A review of current practices and some new recommendations for optimal design schemes*	Kabelo Mahloromela (p. 40) *Covariate construction of nonconvex windows for spatial point patterns*
	12:40	Sisa Pazi (p. 38) *Contributions to acute physiology scoring for South African intensive care units*	Jean-Claude Malela-Majika (p. 39) *Nonparametric precedence chart with repetitive sampling*	Renate Thiede (p. 40) *Measuring Homogeneity of Linear Networks*
13:00 – 14:00	**Lunch**			Venue: Fairway Terrace Restaurant
14:00 – 15:30	**Roundtable:** *Data science, data literacy and the future of statistics* Panelists: Ruth Etzioni, Robert Gramacy, Mark Nasila, Sheetal Silal, Ashwell Jenneker, Renette Blignaut, Pravesh Debba Moderators: Inger Fabris-Rotelli, Warren Brettenny			Venue: Regency Hall
15:30 – 16:15	**Afternoon tea**			Venue: Regency Hall Patio and Foyer
15:30 – 16:15	**SASA Annual General Meeting**			Venue: Regency Hall
16:30 – 17:30	**Biostatistics Interest Group Meeting**			Chair: Isaac Singini, Venue: Charlotte Room
19:00 for 19:30	**Gala Dinner**			Venue: Regency Hall

Friday, 2 December

08:00 – 09:00	Stream:	**Multivariate Statistics**	**Biostatistics**	**Young Stats (Coputational Statistics)**
	Venue:	Regency Hall	Charlotte Room	George Room
	Chair:	Sugnet Lubbe	Nontembeko Dudeni-Tlhone	Arno Otto
	08:00	Thomas Farrar (p. 41) *Two New Auxiliary Models for Estimating Error Variances in Heteroskedastic Linear Regression*	Qondeni Ndlangamandla (p. 42) *Flexible Statistical Modelling of The Determinants of Childhood Anaemia In Tanzania and Angola.*	Innocent Mudhombo (p. 43) *A comparative study of quantile regression and ridge regression based adaptive weights in variable selection and regularized quantile regression*
	08:20	Raeesa Ganey (p. 41) *High-dimensional LDA Biplot through the GSVD*	Najmeh Nakhaei Rad (p. 42) *A Möbius-transformed toroidal distribution for dihedral angles modelling in protein structure*	Ruan Jean du Randt (p. 43) *Estimation of a mixture of semi-parametric partial linear models*
	08:40	Sugnet Lubbe (p. 41) *Biplots for individual differences scaling models*	Nontembeko Dudeni-Tlhone (p. 42) *Safety monitoring of the COVID-19 vaccines in South Africa*	Arno Otto (p. 43) *Skew Laplace candidates emanating from scale mixtures for insightful computational modelling*
09:00 – 09:45	**Keynote Session** **Ashwell Jenneker** — *Census 2022 journey: Updating the nations statistical landscape* (p. 43)			Venue: Regency Hall Chair: Warren Brettenny
09:45 – 10:10	**Morning tea**			Venue: Regency Hall Patio and Foyer
10:10 – 10:30	**Delia North** *Setting up of collaborative systems between academic institutions and industry, to build a strong data skills/talent value chain* (p. 44)			Venue: Regency Hall Chair: Chantelle Clohessy
10:30 – 12:00	**Keynote Session** 10:30 **Pravesh Debba** (SAS Thought Leader 2020) — *Statistical Science: Enriching our lives* 11:15 **Renette Blignaut** (SAS Thought Leader 2021) — *Can statisticians ignore data science, or should it be embraced?* (p. 44)			Venue: Regency Hall Chair: Chantelle Clohessy
12:00 – 13:00	**Lunch**			Venue: Fairway Terrace Restaurant
13:00 – 14:20	Stream:	**Computational Statistics**	**Econometrics and Business Statistics**	**Young Stats (Biostatistics)**
	Venue:	Regency Hall	Charlotte Room	George Room
	Chair:	Leonard Santana	Stefan Janse van Rensburg	Isaac Singini
	13:00	James Allison (p. 45) *On testing for the assumptions of mixture cure models in the presence of covariates*	Tomas Tichy (p. 45) *Analysis of flexibility value related to forest stands*	Awonke Nqayiya (p. 46) *Early prediction of acute kidney injury using machine learning methods*
	13:20	Thobeka Nombebe (p. 45) *Comparing distance-based and traditional parameter estimation techniques for the Lomax distribution.*	Stefan Janse van Rensburg (p. 46) *A score driven volatility model with local leverage*	Zenzile Ntshabele (p. 46) *Comparative performance evaluation of logistic regression and machine learning methods on different data types*
	13:40	Jaco Visagie (p. 45) *On estimating the mode of an angular distribution*		Isaac Singini (p. 47) *Latent Class Joint Model for Longitudinal and Survival Data: an alternative to influence diagnostics for shared parameter joint model*
	14:00	Leonard Santana (p. 45) *Goodness-of-fit tests for Poisson regression models*		
14:30 – 15:00	**Closing Ceremony: Inger Fabris-Rotelli**			Venue: Regency Hall

Posters

Wednesday, 30 November, 18:30 – 19:30
Venue: Patio at the Fairway Terrace Restaurant

Natalie Benschop	*Determining the impact of changing weather conditions on the lung function of children in South Africa: A doctoral study in applied statistics* (p. 49)
Grace Carmichael	*Analysing published data using Meta Analysis and Mendelian Randomisation to determine causal associations.* (p. 49)
Mila Coetzee	*A new similarity metric for linear networks* (p. 49)
Sisipho Hamlomo	*Irregular Local Low Rank Approximation.* (p. 49)
Amy Langston	*Termination versus operation extension for degrading systems* (p. 50)
Jennifer Leigh Liebenberg	*Biplots of Compositional Data from the Tennessee Eastman Process* (p. 50)
Aphiwe Magaya	*Preliminary assessment of resource data for power output predictions of a photovoltaic (PV) system* (p. 50)
Jocelyn Mazarura	*A Gamma-Poisson topic model for short text* (p. 50)
Bonelwa Sidumo	*An approach to multi-class imbalanced problem in ecology using machine learning* (p. 51)
Neill Smit	*Dual-stress accelerated life testing models using the generalised Eyring model* (p. 51)

Abstracts

Wednesday, 30 November

Plenary Venue: Regency Hall
Wednesday, 30 November, 09:45 – 10:30 Chair: Warren Brettenny

09:45

Statistical models for understanding population cancer trends and informing health policies

Ruth Etzioni | *Fred Hutchinson Cancer Research Center*

Trends in the population burden of cancer can be very revealing about the progress of efforts to control the disease. In the US, the National Cancer Institute produces annual estimates of cancer incidence, mortality and survival. I will present a series of models to learn from patterns of these measures over time about the benefits of cancer control interventions such as new screening and treatment approaches. When a new screening test disseminates into population practice as was the case with the PSA test for prostate cancer, this provides an opportunity to also learn about disease natural history. I will share the story of how we used population trends in prostate cancer incidence, mortality, and survival to learn about prostate cancer natural history and inform national policy guidelines for prostate cancer early detection.

Plenary Venue: Regency Hall (Online)
Wednesday, 30 November, 11:15 – 12:00 Chair: Inger-Fabris Rotelli

11:15

Spatial statistical questions and big spatial datasets

Edzer Pebesma | *University of Münster, Institute for Geoinformatics*

A number of very large spatiotemporal datasets have become openly available, in particular from the domain of Earth Observation. These datasets form the basis for creating all kinds of derived "products", often with global coverage and high resolution, and often using machine learning or deep learning algorithms. A number of problems, like assessing the quality of individual predictions or estimating temporal change in areal averages or areal fractions of certain categories, typically remain unsolved. Spatial statistical concepts such as autocorrelation and change of support are usually ignored. In the talk I will discuss to what extent ignoring these concepts is a missed opportunity, and whether this can be mitigated.

Stream: Data Science Venue: Regency Hall
Wednesday, 30 November, 12:10 – 13:10 Chair: Sonali Das

12:10

A simple learning agent interacting with an agent-based market model

Matthew Dicks | *Department of Statistical Sciences, University of Cape Town*
Tim Gebbie | *Department of Statistical Sciences, University of Cape Town*

We consider the learning dynamics of a single reinforcement learning optimal execution trading agent when it interacts with an event driven agent-based financial market model. Trading takes place asynchronously through a matching engine in event time. The optimal execution agent is considered at different levels of initial order-sizes and differently sized state spaces. The resulting impact on the agent-based model and market are considered using a calibration approach that explores changes in the empirical stylised facts and price impact curves. Convergence, volume trajectory and action trace plots are used to visualise the learning dynamics. Here the smaller state space agents had the number of states they visited converge much faster than the larger state space agents, and they were able to start learning to trade intuitively using the spread and volume states. We find that the moments of the model are robust to the impact of the learning agents except for the Hurst exponent, which was lowered by the introduction of strategic order-splitting. The introduction of the learning agent preserves the shape of the price impact curves but can reduce the trade-sign auto-correlations when their trading volumes increase.

12:30

Evolutionary support vector regression for monitoring Poisson profiles

Sandile Shongwe | *Department of Mathematical Statistics and Actuarial Science, University of the Free State*

Many researchers have shown interest in profile monitoring; however, most of the applications in this field of research are developed under the assumption of normal response variable. Little attention has been given to profile monitoring with non-normal response variables, known as general linear models (GLM) which consists of two main categories (i.e., logistic and Poisson profiles). This paper develops a new robust Phase II Poisson profile monitoring tool using support vector regression (SVR) by incorporating some novel input features and evolutionary training algorithm. The new method is quicker in detecting out-of-control (OOC) signals

as compared to the conventional statistical methods. Moreover, the performance of the proposed scheme is further investigated for Poisson profiles with both fixed and variable explanatory variables as well as non-parametric profiles. A diagnostic method with machine learning approach is also used to identify the parameters of change in the profile.

12:50

Are winters becoming shorter and warmer? Insights from a functional data analysis investigation

Sonali Das | *Department of Business Management, University of Pretoria*

A recurrent narrative, in our case, primarily from agricultural scientists and ecologists in the Southern hemisphere, is that 'winters are becoming shorter and warmer', and is noticeably affecting both the physical and biological systems. An ensemble of opportunities within the functional data analysis (FDA) statistical framework is used to explore shifts in annual temperatures by investigating specifically the joint trivariate structure composed of (i) the timing of the onset of winter, (ii) the temperature-trough in the winter season, and (iii) the timing of the onset of spring, in each year.

Stream: General Venue: Charlotte Room

Wednesday, 30 November, 12:10 – 13:10 Chair: Zani Ludick

12:10

Theory versus Practice: Reflections on sampling and survey data

Ariane Neethling | *Mathematical Statistics and Actuarial Science, University of the Free State / GeoTerra Image*
Francois Neethling | *GeoTerra Image*

We all preach the importance of learning theory and applying it in practice. But there is always a tension between theory and practice. In fact, it sometimes appears that the practice of statistics and theory in a scholarly context are poles apart. Thus, the divide between theory and practice remains an enduring challenge – particularly in statistics. So, the need to bridge the gap between theory and practice is instructive.

The challenge for a statistician is to find a connection point in between so that the results are statistically sound, given the need for a solid theoretical grounding on the one hand, and the practical circumstances, limitations and acceptable assumptions on the other. Often creative initiatives are required and this further leads to new research problems or the use of existing techniques in alternative ways. This implies that theory is abstracted practice, and practice is applied theory, and with this in mind in this presentation, various practical challenges experienced in sampling and survey data will be discussed.

12:30

An introduction to the characterisation of a research design through the MIDA framework with an example from a recent waste collection study

Jeanette Pauw | *Department of Statistics, Nelson Mandela University*

When a researcher plans a study, it is best practise that a statistician is involved from the start. The statistician should ideally be involved not only in the data analyses, but also in planning the research design. When the statistician understands the research question, he or she can give best advice on the data collection method and the relevant methods to use when analysing this data to answer the research question. The statistician can then also evaluate the results and give some judgement of how well the original research question was answered.

In this talk I will introduce the Declare Design approach and R package with the same name that was developed in the past five years by Graeme Blair, Jasper Cooper, Alexander Coppock and Macartan Humphreys. At the heart of their approach lies the MIDA framework which characterises a research design by four elements: (M)odel, (I)nquiry, (D)ata strategy and (A)nswer strategy. When a research design has been described within this framework, the quality of the design can be diagnosed through computer simulations. For a specific study, a researcher can then compare and ultimately choose between different designs.

I conclude the talk with an example from a recent project that I was involved in where the MIDA framework was used. The aim of the project was to determine the feasibility of employing waste pickers to remove and sort household waste in an informal settlement that does not have waste removal services.

12:50

A composer's playground: Statistical linguistic measures & self-similarity matrices in the assessment and generation of musical structure.

Zani Ludick | *Department of Mathematical Statistics & Actuarial Science, University of the Free State*
Wilben Pretorius | *Department of Mathematical Statistics & Actuarial Science, University of the Free State*
Trudie Strauss | *Department: Afrikaans and Dutch; German and French, University of the Free State*

When composing music, a careful balance must be struck between coherent temporal structure, musical variety and novelty. Generative machine learning models, and other statistical models, vary in their ability to balance these features. Additionally, assessments of the musical structure of automatically generated music often rely on subjective evaluations, such as time-consuming

listening tests. Although the listener is the ultimate end-user, and subjective evaluations should not be replaced, automatic assessment techniques for identifying elements of structure in compositions can provide an objective alternative which may enhance the training of machine learning and other models. These measures may also provide musically interesting and useful material for the human composer to manipulate.

Self-similarity matrices (SSMs) can be used to capture and represent structural patterns in music. Identified blocks, paths and corners in SSMs correspond respectively with areas of homogeneity, repetition and novelty. Statistical linguistic measures, such as measures of repetitiveness and complexity also lend themselves to the description of musical structures.

We investigate the links between statistical linguistic measures, such as the moving-average type-token ratio (MATTR) and self-similarity matrices in the assessment of musical structure and demonstrate how they can assist in automatic music generation.

Stream: Young Stats (Multivariate Statistics)

Venue: George Room

Wednesday, 30 November, 12:10 – 13:10

Chair: Tanita Botha

12:10

Investigating the Heterogeneous Structure of Multivariate Anthropometric Growth Profiles

Noëlle van Biljon | *Department of Statistical Sciences, University of Cape Town*

Francesca Little | *Department of Statistical Sciences, University of Cape Town*

Heather Zar | *Department of Paediatrics and Child Health, University of Cape Town*

Marilyn Lake | *Department of Paediatrics and Child Health, University of Cape Town*

Conventional methods for modeling longitudinal growth data focus on the analysis of mean longitudinal trends or the identification of abnormal growth based on standardized z-scores. Latent Class Mixed Modelling (LCMM) takes into account the underlying heterogeneity in growth profiles and allows for the identification of groups of subjects that follow similar longitudinal trends. LCMM is based on a combination of linear mixed-effect-, structural equation- and multinomial logistic modelling. We used LCMM to identify underlying latent profiles of growth for a multivariate response of height and weight measurements taken from birth until the age of five years for a sample of 1143 children from the Drakenstein Child Health Study (DCHS). Instead of focusing on an optimal number of latent classes, we examined the added structural information provided by additional classes. When considering the presence of two (k=2), three (k=3) or up to six latent classes (k=6) specific features within the profiles were preserved while new features relating to level and shape of the profiles were also identified. While smaller values of k allowed us to identify large groups of individuals that follow broadly similar growth trajectories, by increasing k we were able to identify slightly more nuanced features that could be used to distinguish subgroups within the larger classes. With the identification of these classes, a better understanding of distinct childhood growth trajectories and their predictors may be gained, informing interventions to promote optimal childhood growth.

12:30

Exploring sentiment contained in COVID-19 related Tweets with embedded word MCA biplots

Zoë-Mae Adams | *Centre for Multi-Dimensional Data Visualisation (MuViSU) at the Department of Statistics and Actuarial Science, Stellenbosch University*

Johané Nienkemper-Swanepoel | *Centre for Multi-Dimensional Data Visualisation (MuViSU) at the Department of Statistics and Actuarial Science, Stellenbosch University*

Niël le Roux | *Centre for Multi-Dimensional Data Visualisation (MuViSU) at the Department of Statistics and Actuarial Science, Stellenbosch University*

Sugnet Lubbe | *Centre for Multi-Dimensional Data Visualisation (MuViSU) at the Department of Statistics and Actuarial Science, Stellenbosch University*

Social media platforms are continually gaining popularity which results in vast amounts of shared data in the form of images, videos and text. Twitter is a micro-blogging platform which allows the sharing of short messages that are labelled according to a specific key word (i.e. tag), representing a relevant topic or theme. These messages reflect personal opinions with subjective content which could provide insight to grasp the underlying attitude towards specific topics. During the global COVID-19 pandemic users could easily share messages by using social media platforms containing information on for example regulations on lockdown or vaccination. Twitter's application programming interface (API) allows the procurement of posts made on the platform for a specific Twitter tag, timeframe and location within a specified radius. In this study the unstructured pieces of text, Tweets, are processed and the sentiment of the remaining words are classified using two lexicons. Multi-dimensional visualisation enables the exploration of the associations between the Twitter users based on the resultant sentiment scores of their posts. A multiple correspondence analysis (MCA) biplot is embedded with the extracted words to enable the simultaneous interpretation of the underlying sentiment of the processed Tweets. This paper presents two case studies of COVID-19 related Tweets. The first case study considers posts made by South African users in three cities (Cape Town, Johannesburg and Durban), with the second case study evaluating the sentiment towards COVID-19 on a global scale by considering three predominantly English-speaking countries (South Africa, Australia and United Kingdom). In this presentation these ideas will be explored and illustrated with an interactive display in a Shiny application.

12:50

Wasserstein distance as discriminator within the Dirichlet family

Tanita Botha | *Department of Statistics, University of Pretoria, Pretoria, South Africa*
Johan Ferreira | *Department of Statistics, University of Pretoria, Pretoria, South Africa and Centre of Excellence in Mathematical and Statistical Sciences, University of the Witwatersrand, Johannesburg, South Africa.*
Andriëtte Bekker | *Department of Statistics, University of Pretoria, Pretoria, South Africa and Centre of Excellence in Mathematical and Statistical Sciences, University of Witwatersrand, Johannesburg, South Africa.*

The Dirichlet distribution is a cornerstone consideration when working with data on the unitary simplex. Several generalizations of the Dirichlet distribution has been developed with more flexible structures which can be applied to data with forms that may illustrate departures from the usual Dirichlet, such as multimodality and models which may exhibit positive correlation. To gain a deeper insight into the impact of applying generalisations of the Dirichlet to data that exhibit structural departures, the Wasserstein distance is considered and investigated between different members of the Dirichlet family. Since this distance gives a natural measure of the distance and difference between two (multivariate) distributions this paper explores the differences between several (competitive) Dirichlet constructions to highlight and examine the effect that these structural changes may theoretically result in.

Plenary

Venue: Regency Hall
Wednesday, 30 November, 14:00 – 14:45
Chair: Roelof Coetzer

14:00

Deep gaussian process surrogates for computer experiments

Robert Gramacy | *Virginia Polytechnic and State University*

Deep Gaussian processes (DGPs) upgrade ordinary GPs through functional composition, in which intermediate GP layers warp the original inputs, providing flexibility to model non-stationary dynamics. Recent applications in machine learning favor approximate, optimization-based inference for fast predictions, but applications to computer surrogate modeling – with an eye towards downstream tasks like calibration, Bayesian optimization, and input sensitivity analysis – demand broader uncertainty quantification (UQ). We prioritize UQ through full posterior integration in a Bayesian scheme, hinging on elliptical slice sampling the latent layers. We demonstrate how our DGP's non-stationary flexibility, combined with appropriate UQ, allows for active learning: a virtuous cycle of data acquisition and model updating that departs from traditional space-filling design and yields more accurate surrogates for fixed simulation effort. But not all simulation campaigns can be developed sequentially, and many existing computer experiments are simply too big for full DGP posterior integration because of cubic scaling bottlenecks. For this case we introduce the Vecchia approximation, popular for ordinary GPs in spatial data settings. We show that Vecchia-induced sparsity of Cholesky factors allows for linear computational scaling without compromising DGP accuracy or UQ. We vet both active learning and Vecchia-approximated DGPs on numerous illustrative examples and a real simulation involving drag on satellites in low-Earth orbit. We showcase implementation in the deepgp package for R on CRAN.

Stream: Educational Statistics

Venue: Regency Hall
Wednesday, 30 November, 15:00 – 16:00
Chair: Inger Fabris-Rotelli

15:00

Development of an early career academic supervisor in Statistics in South Africa

Inger Fabris-Rotelli | *Department of Statistics, University of Pretoria*
Michael von Maltitz | *Department of Statistics and Actuarial Science, University of the Free State*
Ansie Smit | *Department of Geology, University of Pretoria*
Daniel Maposa | *Department of Statistics & Operations Research, University of Limpopo*
Sonali Das | *Department of Business Management, University of Pretoria*
Danielle Roberts | *Department of Statistics, University of Pretoria*
Gao Maribe | *Department of Statistics, University of Pretoria*

There is an increasing pull of a Mathematical Sciences graduate to enter industry rather than pursue further postgraduate studies. Within South Africa, there is an urgent need to address this even more as the field of Statistics is particularly affected by the 4IR pull, resulting in a crisis in academic capacity building. Two primary factors can be isolated as those preventing the correction of this: First, academic salaries in Statistical sciences are not comparable to what industry would pay at the same level of qualification (especially in light of the growth of 'Data Science'). Second, the South African National Research Foundation (NRF) student funding is not attractive to a student whom industry is already offering more to, as to pay fees and living expenses on the NRF bursaries are unrealistic for a full-time student. In response to these crises is Statistics, the NRF, since 2016, has provided funding to support postgraduate students who may be trained to enter academia after their PhD. The grant provided larger bursaries than the standard

NRF bursaries, as well as funding to bring in expertise to train young staff. In spite of this initiative, the lack of supervisory skills and capacity, especially at the Doctoral level, is evident across South African Statistical sciences departments. In 2020, a group of 8 novice, and near-novice, doctoral supervisors in academic Statistical science in South Africa initiated discussions that delved into the current state of academic Statistics in the country, specifically with regards to the nurturing of an early career academic supervisor in Statistics. These discussions have resulted in a clear need for actionable actions by and for early career academics in Statistics that involve a guiding rubric for the doctoral thesis, coupled with a reference guideline for early career supervisors in South Africa.

Stream: Data Science

Venue: Charlotte Room

Wednesday, 30 November, 15:00 – 16:00

Chair: Stefan Britz

15:00

Principal component analysis (PCA) on annual financial statements of large South African manufacturing enterprises

Sagaren Pillay | *Statistics South Africa*

The South African manufacturing industry is rapidly growing and provides numerous opportunities to investors both locally and abroad. This present study analyses the financial ratios of large manufacturing enterprises by using the technique of principal component analysis. The research presents an analysis of the reduction of the many financial ratios to fewer unbiased ratios that will assist stakeholders in assessing the performance of the manufacturing industry. The resulting ratios can be used to analyse the financial performance of manufacturing enterprises with a minimal loss of information.

15:20

Robust Self-Paced Learning Algorithm For Finite Mixture Models

Andre Ruben Kleynhans | *Department of Statistics, University of Pretoria*
Sollie Millard | *Department of Statistics, University of Pretoria*
Frans Kanfer | *Department of Statistics, University of Pretoria*

A robust self-paced learning algorithm for the finite mixture model is proposed. Self-paced learning (SPL) is a training strategy that mitigates the impact of non-typical observations. SPL introduces observations in a meaningful order by considering the likelihood of each observation. The proposed algorithm considers a finite mixture model that includes a distributional structure for non-typical observations. The properties of this algorithm are presented through a simulation study along with an application on real data. A comparison is made with the properties of popular approaches. The algorithm shows a reduction in parameter estimation bias which indicates an improvement in the estimation accuracy of the parameters.

15:40

Feature Engineering for Tennis Match Outcome Prediction

Stefan Britz | *Department of Statistical Sciences, University of Cape Town*

As with many sports in the professional era, tennis provides a surfeit of data that can be used to both analyse past performances and predict future match outcomes. However, in its raw form information from past matches provides limited benefit to statistical learning models aimed at predicting the winner of a match. To fully harness the power of predictive algorithms, predictor variables (features) must be created and curated (engineered) from the available data using domain knowledge – in this case the understanding of tennis matches and tournaments – in such a way that predictive patterns are possibly generated.

The focus of this talk is on engineering features from ATP tour data from 2010 onwards in order to predict the outcomes of matches in a statistical learning framework. Predictive models popular in the modern literature are applied: elastic-net regularized logistic regression, random forests, xgBoost, feed-forward neural networks, and support vector machines. Classification accuracy is used to assess model performance, whilst variable importance measures appropriate to each model are used to determine the influence of the engineered features on prediction.

Stream: Young Stats (Bayesian Statistics)

Venue: George Room

Wednesday, 30 November, 15:00 – 16:00

Chair: Trudie Strauss

15:00

Modelling the Misuse of Alcohol and Drugs in South Africa Using Bayesian Binary Logistic Regression

Makwelantle Asnath Sehlabana | *Department of Statistics and Operations Research, University of Limpopo*
Daniel Maposa | *Department of Statistics and Operations Research, University of Limpopo*
Alexander Boateng | *Department of Statistics and Actuarial Science, Kwame Nkrumah University of Science and Technology*

The misuse of alcohol and drugs is a continuous life threat globally, including in South Africa. For that reason, researchers continue to investigate the risk factors associated with alcohol and drugs misuse. Most studies in literature employed the classical logistic regression model to investigate these risk factors. However, some of the issues pertaining to the classical methods

are accounted for in the Bayesian framework. Likewise, the Bayesian logistic regression model can also account for problematic issues to the classical logistic regression model. Several studies used the Bayesian logistic regression model to investigate the risk factors associated with alcohol and drugs misuse. Usually, most Bayesian studies utilize default prior probability distributions such as Jeffereys' prior and Zellner's informative g-prior distributions. This study aims to evaluate the effectiveness of a modified Zellner's informative g-prior distribution and subdue separation in modelling the misuse of alcohol and drugs. The model developed through the use of a modified Zellner's g-prior distribution is compared to the models developed through the use of a hyper g-prior distribution and mixtures of g and n prior distribution. Comparisons are based on precision and average prediction error. Although the models yielded similar results, the modified version of Zellner's informative g-prior distribution resulted in narrow credible intervals, and a small everage prediction error. Separation is also accounted for in the model. In this study, the modified version of Zellner's informative g-prior distribution is evidently effective. All models are developed using the Bayesian adaptive sampling (BAS) R package. Further research may include evaluating some of the recommended prior distributions for the Generalised Linear Models (GLM) and comparison of Bayesian binary logistic regression developed in this study with logistic regression in Machine Learning algorithms.

15:20

Using joint models to study the association between CD4 count and the risk of death in TB/HIV data

Nobuhle Mchunu | *Biostatistics Unit, South African Medical Research Council (SAMRC)*
Henry Mwambi | *University of KwaZulu-Natal, School of Mathematics, Statistics and Computer Science*
Dimitris Rizopoulos | *Department of Biostatistics, Erasmus University Medical Center*
Nonhlanhla Yende-Zuma | *Centre for the AIDS Programme of Research in South Africa (CAPRISA), University of KwaZulu-Natal*
Tarylee Reddy | *Biostatistics Unit, South African Medical Research Council (SAMRC)*

The association structure linking the longitudinal and survival sub-models is of fundamental importance in the joint modeling framework and the choice of this structure should be made based on the clinical background of the study. However, this information may not always be accessible and rationale for selecting this association structure has received relatively little attention in the literature. To this end, we aim to explore four alternative functional forms of the association structure between the CD_4 count and the risk of death and provide rationale for selecting the optimal association structure for our data. We also aim to compare the results obtained from the joint model to those obtained from the time-varying Cox model. We used data from the Centre for the AIDS Programme of Research in South Africa (CAPRISA) AIDS Treatment programme, the Starting Antiretroviral Therapy at Three Points in Tuberculosis (SAPiT) study, an open-label, three armed randomised, controlled trial between June 2005 and July 2010 (N=642). We utilized the Deviance Information Criterion (DIC) to select the final model with the best structure. Combined integrated therapy arms had a reduction of 55% in mortality (HR:0.45, 95% CI:0.28–0.72) compared to the sequential therapy arm. The joint model with a cumulative effects functional form was chosen as the best association structure. In particular, our joint model found that the area under the longitudinal profile of CD_4 count was strongly associated with a 21% reduction in mortality (HR:0.79, 95% CI:0.72–0.86). Whereas results from the time-varying Cox model showed a 19% reduction in mortality (HR:0.81, 95% CI:0.77–0.84). We have shown that the "current value" association structure is not always the best structure that expresses the correct relationship between the outcomes in all settings, which is why it is crucial to explore alternative clinically meaningful association structures that links the longitudinal and survival processes.

15:40

Word Frequency Distributions: A Comprehensive Bayesian Approach

Trudie Strauss | *Department of Mathematical Statistics and Actuarial Sciences,University of the Free State*

Many parametric models and statistical laws have been suggested to model word frequency distributions, the most famous being Zipf's law (the inverse proportionality between the rank of a word and its frequency). Zipf's law has been generalised and extended to model the relationship between the frequency spectrum and its groups, and also to include more parameters. As such, there are currently several proposed models of Zipf-like distributions that seem to describe word frequency distributions relatively well. These models, with their respective advantages and disadvantages in particular settings, have been shown to hold in different contexts: particular ranges of sample size, certain genres of text, etc.

In this study, we implement a Bayesian approach by fitting a class of general models based on Zipfian distributions that encompasses these current propositions for word frequency distributions. We further identify several linguistically relevant features that may be calculated from word frequency distributions of languages and express natural language as a multidimensional array of these measures. Ultimately, through a Bayesian analysis, we compare the theoretical model with empirical data based on 200 languages from 22 language families, to determine the space that these measures occupy within the posterior distributions. Comparing the patterns and distributions of empirical natural languages to what may be expected from these theoretical models allows us to determine whether some of these linguistically relevant measures also behave in a universal manner across languages and we investigate the linguistic implication.

Stream: Educational Statistics

Wednesday, 30 November, 16:20 – 18:00

Venue: Regency Hall
Chair: Paul van Staden

16:20

Efficiency Analysis of South African Schools: A Parametric Approach

Aviwe Gqwaka | *Department of Statistics, Nelson Mandela University*
Warren Brettenny | *Department of Statistics, Nelson Mandela University*
Gary Sharp | *Department of Statistics, Nelson Mandela University*

South African learners have ranked low in global assessments of reading skills and mathematics literacy. To remedy this, government has sought to adequately equip schools in their education provision services. Thus, to get an indication of the state of the South African education sector, understanding the level of performance of schools is apt. To do this, an efficiency analysis is conducted. Here the ability of a school to minimise its use of available resources while maximising learner performance is quantified and assessed. This is done using the parametric approach, stochastic frontier analysis (SFA), where observed performances are compared to a theoretical best practice or frontier. Deviations from this frontier are attributed to effects not in control of the school (random shock) and those that are (inefficiency). Use of this approach allows for the identification of best and wort performing schools. These findings could then assist policy-makers to perhaps review their resource allocations, where they can better attend to the needs of those schools deemed to be inefficient.

16:40

The value proposition for industry-academic collaboration

André Zitzke | *SAS Institute South Africa*
Murray de Villiers | *SAS Institute B.V.*

The onset of industrial revolutions has, in the past, brought about rapid change to both active- and inactive participants. The current data-driven Fourth Industrial Revolution poses significant challenges and opportunities to industry, government and learning providers. This paper explores the various strategies, design- and execution considerations for Industry-Academic collaboration. In particular it proposes specific courses of action for some of the participants in the Fourth Industrial Revolution learning value chain.

17:00

Analysis of Strike action on students' Academic Performance in the Inferential Statistics Module at the University of Fort Hare, South Africa

Ruffin Mpiana Mutambayi | *Department of Statistics, University of Fort Hare*
Adeboye Azeez | *Department of Statistics, University of Forte Hare*
Happiness Tshepo | *Department of Statistics, University of Fort Hare*
Akinwumi Odeyemi | *Department of Statistics, University of Fort Hare*

This study aims to analyse the impact of strike actions on students who enrolled for the Statistics module as one of the prerequisite modules. The study was done at the University of Fort Hare, and 142 students participated in the study.

The collection of the data was done through a questionnaire and descriptive Statistics, inferential Statistics combined with quantile regression analysis were used to analyse the data.

The results reveal that the marks of students were normally distributed (p-value: 0.057), and there were no outliers (p-value: 0.515). It was also found that the academic performance of students at quantiles 0.25 (p-value: 0.0005), 0.5 (p-value: 0.0001), and 0.75 (p-value: 0.0652) for 'nationality' were having an impact on the performance of students in statistics. Moreover, quantiles 0.50 (p-value: 0.0304) and 0.75 (p-value: 0.0107) for 'appointment of industrial arbitration panels to review at the interval a measure to eradicate strike actions' were also statistically significant.

17:20

Affective Learning: An insight into Mr Lindo's #OperationFinishTheSemesterStrong

Lindo Magagula | *Department of Statistics, University of Pretoria*

Learning in the affective domain is often neglected in most educational programs. The disuse is because affective learning needs to be a better-understood concept amongst educators. It goes beyond the scope of practice for education practitioners who would instead invest in cognitive learning. In this presentation, I highlight the preliminary finding of a study conducted on STK120 students. The Department of Statistics presents this course to first-year students majoring in various degrees. In this study, over and above the teaching of content, a small fraction of teaching time (at the beginning of class) was taken and dedicated to students' emotional well-being, after that was evaluating the impact on students' willingness to learn. This was done in an interactive way by making use of the turning point clicker app and a word cloud visualization of responses. It is imperative to remain vigilant on the extent to which the proposed measures of affective learning should be implemented.

17:40

When should Paul visit Paris? A time series case study from an introductory first-year statistics & data science course

Paul Jacobus van Staden | *Department of Statistics, University of Pretoria*

At many universities time series analysis is only taught at final-year undergraduate or at postgraduate level. But the proliferation

of time series data in, for example, financial markets, social media analytics and, more recently, epidemiology (due to the COVID-19 pandemic), necessitates that students already be introduced to time series analysis at a first-year level.

This talk presents a case study in which students from the presenter's first-year introductory statistics and data science course have to analyze a time series dataset from Paris, France. Basic tools including time plots and time series decomposition are sufficient for these students to learn about data dependency in time series as well as time series patterns and components such as trend, seasonality, cyclical fluctuations and noise. However, the nature of the chosen dataset lends itself to further scrutiny in that students discover, with the assistance of Dr Google and Professor Wikipedia, how spurious conclusions can be made in the absence of statistical intuition.

So when should Paul visit Paris? Maybe in October...

Stream: Extreme Value Theory Venue: Charlotte Room

Wednesday, 30 November, 16:20 – 18:00 Chair: Andréhette Verster

---

16:20

Outlier detection based on extreme value theory and applications

Jan Beirlant | *Department of Mathematical Statistics and Actuarial Science, University of the Free State*
Shrijita Bhattacharya | *Department of Statistics and Probability, Michigan State University*
François Kamper | *Department of Statistics and Actuarial Science, Stellenbosch University; Swiss Data Science Center, EPFL & ETH*

Whether an extreme observation is an outlier or not depends strongly on the corresponding tail behaviour of the underlying distribution. We develop an automatic, data-driven method rooted in the mathematical theory of extremes to identify observations that deviate from the intermediate and central characteristics. The proposed algorithm is an extension of the method proposed in Bhattacharya et al. (2019) for the specific case of heavy tailed Pareto-type distributions to all max-domains of attraction. We propose some applications such as a tail-adjusted boxplot which yields a more accurate representation of possible outliers, and the identification of outliers in a multivariate context through an analysis of associated random variables such as local outlier factors. Examples and simulation results illustrate the finite sample behaviour of the algorithm and its applications.

16:40

Robust Extreme Quantile Estimation for Pareto-Type tails through an Exponential Regression Model

Richard Minkah | *Department of Statistics and Actuarial Science, University of Ghana*
Tertius de Wet | *Department of Statistics and Actuarial Science, Stellenbosch University*
Abhik Ghosh | *Indian Statistical Institute*

The estimation of extreme quantiles is one of the main objectives of statistics of extremes (which deals with the estimation of rare events). In this paper, a robust estimator of extreme quantile of a heavy-tailed distribution is considered. The estimator is obtained through the minimum density power divergence criterion on an exponential regression model. The proposed estimator was compared with two estimators of extreme quantiles in the literature in a simulation study. The results show that the proposed estimator is stable to the choice of the number of top order statistics and show lesser bias and mean square error compared to the existing extreme quantile estimators. Practical application of the proposed estimator is illustrated with data from the pedochemical and insurance industries.

17:00

Modelling temperature extremes in the Limpopo province: Bivariate time-varying threshold excess approach

Daniel Maposa | *Department of Statistics and Operations Research, University of Limpopo*
Anna M. Seimela | *Department of Financial Management, University of Limpopo*

A common problem that arises in extreme value theory (EVT) when dealing with several variables is to find an appropriate method to assess their joint or conditional multivariate extremal dependence behaviour. The method for choosing an appropriate threshold in peaks-over threshold approach is also another problem of endless debate. In this era of climate change and global warming, extreme temperatures accompanied by heat waves and cold waves pose serious economic and health challenges particularly in small economies or developing countries like South Africa. The present study attempts to address these problems, in particular, to deal and capture dependencies in extreme values of two variables, by applying bivariate conditional extremes modelling with a time-varying threshold to Limpopo province's monthly maximum temperature series. The present study is carried out in the Limpopo province of South Africa for the period 1994-2009. With the aim to model extremal dependence of maximum temperature at four meteorological stations, two modelling approaches are applied: bivariate conditional extremes model and time-varying threshold. The latter approach was used to capture the climate change effects in the data. The main contribution of this paper is in combining these two approaches in bivariate extremal dependence modelling of maximum temperature extremes in the Limpopo province. The findings of the study revealed both significant positive and negative extremal dependence in some pairs of meteorological stations. Among the major findings were the significant strong positive extremal dependence of Thabazimbi on high temperature values at Mara and the strong negative extremal dependence of Polokwane on high temperature values at Messina. The findings of this study

play an important role in revealing information useful to meteorologists, climatologists, agriculturalists, planners in the energy sector among others.

17:20

Open-set Recognition using Excesses of Distance Ratios

Matthys Lucas Steyn | *Department of Statistics and Actuarial Science, Stellenbosch University; Department of Data Analysis and Mathematical Modelling, Ghent University*

This talk discusses methods that combine statistical learning and extreme value theory to perform open-set recognition. Open-set recognition is an extension of supervised classification where not all classes are known during training. During prediction of the test data, open-set models must correctly classify the observations from known classes and detect observations from classes that were unknown during training. The term 'open set' refers to the fact that only a subset of the possible classes is available during training, and that the number of possible classes is unknown. A method is presented to perform open-set recognition by using the extreme values of a dissimilarity score. It is shown that the ratio of distances locally around the target point can be used to express how dissimilar a target point is from the known classes. The class of generalized Pareto distributions with bounded support is then used to model the peaks of the distance ratio above a high threshold. The resulting distribution provides a probabilistic framework to perform open-set recognition. It is demonstrated how the probabilistic model can be combined with convolutional neural networks to perform open-set recognition on image and audio data. The talk is concluded with demonstrations of the proposed method on two image recognition datasets and one audio dataset.

17:40

Estimation of extreme quantiles of GHI: A comparative analysis using an extremal mixture model and a generalised additive extreme value model

Thakhani Ravele | *PhD candidate, Department of Mathematical and Computational Sciences, University of Venda*

Caston Sigauke | *Senior Lecturer, Department of Mathematical and Computational Sciences, University of Venda*

Lordwell Jhamba | *Senior Lecturer, Department of Physics, University of Venda*

Solar power poses challenges to the management of grid energy due to its intermittency. To have an optimal integration of solar power on the electricity grid it is important to have accurate forecasts. This study discusses the comparative analysis of semi-parametric extremal mixture (SPEM), generalised additive extreme value (GAEV) or quantile regression via asymmetric Laplace distribution (QR-ALD), additive quantile regression (AQR-1), additive quantile regression with temperature variable (AQR-2) and penalised cubic regression smoothing spline (benchmark) models for probabilistic forecasting of hourly global horizontal irradiance (GHI) at extremely high quantiles (τ = 0.95, 0.97, 0.99, 0.999 and 0.9999). The data used are from the University of Venda radiometric in South Africa and are from the period 1 January 2020 to 31 December 2020. Empirical results from the study showed that the AQR-2 is the best fitting model and gives the most accurate prediction of quantiles at τ = 0.95, 0.97, 0.99 and 0.999, while at 0.9999-quantile the GAEV model has the most accurate predictions. Based on these results it is recommended that the AQR-2 and GAEV models be used for predicting extremely high quantiles of hourly GHI in South-Africa. The predictions from this study are valuable to power utility decision-makers and system operators when making high risk decisions and regulatory frameworks that require high-security levels. This is the first application to conduct a comparative analysis of the proposed models using South African solar irradiance data, to the best of our knowledge.

Stream: Young Stats (General)

Venue: George Room

Wednesday, 30 November, 16:20 – 18:00

Chair: Praise Obanya

16:20

On testing for the Pareto distribution using U and V statistics

Lethani Ndwandwe | *Department of Statistics, North West University*

James Allison | *Department of Statistics, North West University*

Marius Smuts | *Department of Statistics, North West University*

Jaco Visagie | *Department of Statistics, North West University*

We propose new classes of tests for the Pareto type I distribution. These tests utilise the empirical characteristic function and is based on a lesser-known characterisation of the Pareto distribution. The finite sample performances of the newly proposed tests are evaluated and compared to some of the existing tests, where it is found that the new tests are competitive in terms of empirical powers.

16:40

COVID-19 and Volatility of International Stock Markets: An FDA Investigation

Ryan Shackleton | *Department of Computer Science, University of Pretoria*
Sonali Das | *Department of Business Management, University of Pretoria*
Rangan Gupta | *Department of Economics, University of Pretoria*

The COVID-19 pandemic in 2020 resulted in the biggest decline in the financial markets since the global financial crisis of 2007-2009, and not surprisingly, brought the discussion on volatility of international financial markets back to the fore. The purpose of this study is to investigate the daily realised volatility, as a metric of risk, derived from intraday data of international stock markets, with a particular focus on the COVID-19 period. The stock markets investigated include Brazil, China, Europe, India, the United Kingdom, and the United States, representing a mix of both developing and developed countries. The empirical investigation was carried out using the Functional Data Analysis (FDA) framework, and details on the smoothing approach and rate-of-change methods used will be presented. Results from this exercise revealed the following important findings for investors and policymakers: First, COVID-19 had significant effect on five of the six markets; second, five of the six markets had similar realised volatility profiles indicating that both developed and emerging markets were affected on a similar scale; third, the volatility of international financial markets started returning to pre-COVID-19 levels in early-May 2020.

17:00

Negative binomial compounding in a discrete Lindley model with INAR(1) application

Ané van der Merwe | *Department of Statistics, University of Pretoria*
Johan Ferreira | *Department of Statistics, University of Pretoria*

Analysing time series data remains a relevant and evolving matter of interest, where oftentimes the assumption of normality is made for the error terms. In the case when data are of a discrete nature, the Poisson model may be assumed for the structure of the error terms. In order to address the equidispersion restriction of the Poisson distribution, various alternative considerations have been investigated in an integer environment. This paper, inspired by the integer autoregressive process of order 1, considers a Poisson-negative binomial noncentral Lindley model for the error terms; which emanates from a novel noncentral Lindley type construction. The systematic construction of this model is discussed and juxtaposed against alternate candidates, and thus with a threefold contribution: development of a continuous noncentral type Lindley model, a discrete counterpart, and its competitive nature within a time series framework. The development and obtained insight from this work is meaningful within the broad spectrum of noncentrality, and shows promise as valuable contenders in the distribution theory space.

17:20

Investigating the effectiveness of an undergraduate mathematics intervention at UWC

Liliane Tendela | *Department of Statistics and Population Studies, University of the Western Cape*
Retha Luus | *Department of Statistics and Population Studies, University of the Western Cape*
Renette Blignaut | *Department of Statistics and Population Studies, University of the Western Cape*

Prior to 2018, the Faculty Natural Sciences at the University of the Western Cape had experienced an extended time to degree amongst their students enrolled for mainstream mathematics. To address the low throughput, the Mathematics Department introduced an intervention programme aimed at second year mathematics students. The intervention, later named the 'turnaround project', commenced in 2018 with a weeklong bootcamp for first semester second-year mathematics. The methods used in the turnaround project included interactive engagement in lectures, additional seminars, workshops, tutor consultations, and additional training sessions. The main objective of this presentation is to report on the effectiveness of the intervention programme using statistical learning methodologies. The data provided includes background information, matric results and the university mathematics marks. This information was used to compare the performance of students prior to the intervention and after the intervention. Initial results indicate that the intervention was successful in improving throughput. A decision tree was used to identify factors contributing to the improvement due to the intervention.

17:40

Permutation entropy analysis of financial markets

Praise Obanya | *Unit for Data Science and Computing, North-West University*
Carel Olivier | *Pure and Applied Analytics, Department of Mathematics and Applied Mathematics, North-West University*
Tanja Verster | *Centre for Business Mathematics and Informatics, North-West University*

Permutation entropy is a time series analysis tool that captures permutation patterns/ordinal relationships between the individual values of a given time series. Being a pattern recognition tool which provides deeper insights into the underlying driving mechanism of a time series, it can also be used to determine the degree of complexity of an economic system, thus aiding in the identification of chaotic and stochastic behaviours. Permutation entropy has been successfully applied in various fields such as biomedicine and physical sciences.

Due to the complexity of financial markets such as the capital, commodity, derivatives, foreign exchange, futures markets, amongst others, it is difficult to determine the level of predictability of each sector of the market and hence its efficiency level. This problem can be addressed through the application of permutation entropy.

In this research, we present an overview of permutation entropy and its successful application in a variety of research fields. We also investigate the application of permutation entropy in determining the predictability and efficiency of the commodity markets. The results obtained and the interpretation are discussed in details.

Thursday, 1 December

Stream: Bayesian Statistics
Thursday, 1 December, 08:00 – 10:00

Venue: Regency Hall
Chair: Allan Clark

08:00

Robust inference in the presence of censoring, skewness, and extreme values

Sean van der Merwe | *Mathematical Statistics and Actuarial Science, University of the Free State*

This presentation discusses how modern statistical modelling software has enabled the fitting of models that are both flexible enough to accommodate data features and still simple enough to answer research questions via intuitive inferences. Inference regarding location is extremely popular in statistics, but often faces difficulties such as adjusting for skewness and extreme observations. It is often desired to do inference for the typical case instead of the raw mean. A t density is naturally robust to occasional extreme observations, while skew variants are particularly robust to a heavy tail on one side. Further, these distributions are not limited in domain. Fitting of these distributions was historically challenging but is currently seeing a surge in use. This work expands the theory of a particular skew-t variant to accommodate censoring and derives a new prior distribution with excellent properties. The implementation is explained and illustrated for a variety of real problems.

08:20

Bayesian meta-regression models for the estimation of population trends in health risk factors

Annibale Cois | *Division of Health Systems and Public Health, Stellenbosch University*

The accurate quantification of trends in the distribution of risk factors is critical in public health. Reliable estimates are key for planning prevention activities and treatment services, especially in low-income countries where the optimal allocation of limited resources is a priority. However, empirical data – usually self-report from population surveys – are often sparse (available for selected subpopulations and time points), heterogeneous (collected with inconsistent methods across data sources), and of varying characteristics in terms of precision and risk of bias.

Bayesian meta-regression is an alternative to frequentist approaches to make sense of sub-optimal data by integrating in a principled way additional sources of information and broad epidemiological and biological evidence. We present an application of Bayesian meta-regression to estimate age- and sex-specific trends in alcohol consumption – a major risk factor for cardiovascular and other diseases – in the South African adult population. The model accounts for the censored nature of the consumption data and the ubiquitous under-reporting of alcohol use in surveys. It allows for time and age non-linearity and for complex constraints in the parameter space, derived from biological knowledge and administrative records on alcohol sales. Mild assumptions of smoothness in age and time trends and relationship with auxiliary variables allow the model to make estimates where data are sparse or unreliable. The Bayesian estimator – implemented using Stan Modelling Language and its default NUTS sampling algorithm – accounts for uncertainty beyond sampling error, and the availability of draws from the posterior distribution makes it straightforward to recover estimates of various linear and non-linear functions of the model parameters. We show how this approach compares favourably to classical rescaling methods used to recover estimates of population alcohol consumption from downward-biased survey data.

08:40

difFUBAR: Scalable Bayesian comparison of selection pressure

Hassan Sadiq | *Department of Statistics and Actuarial Science, Stellenbosch University*

While many phylogenetic methods exist to characterise evolutionary pressure at individual codon sites, relatively few allow the direct comparison between different a priori selected sets of branches. Indeed, this was only recently addressed by an approach, developed in the frequentist framework, that proposes a site-wise likelihood ratio test to test such hypotheses.

Previously, we have demonstrated that approximate grid-based Bayesian approaches to characterising site-wise variation in selection parameters can outperform individual site-wise likelihood ratio tests. Such grid-based approaches can exhibit poor computational scaling when the number of site-wise parameters expands, but here we show that a simple sub-tree likelihood caching strategy can ameliorate this.

We propose difFUBAR, implemented in MolecularEvolution.jl — a new framework for phylogenetic models of molecular evolution developed in the Julia language for scientific computing. difFUBAR allows the demarcation of two branch sets of interest and, optionally, a background set, and estimates joint site-specific posterior distributions over α, ω_1, ω_2 and ω_{BG} using a Gibbs sampler. Evidence for hypotheses of interest can then be quantified directly from the posterior distribution, and we standardly report $P(\omega_1 > \omega_2|Data)$, $P(\omega_2 > \omega_1|Data)$, $P(\omega_1 > 1|Data)$, $P(\omega_2 > 1|Data)$.

We characterise the statistical performance of this approach on previous simulations, comparing it to the site-wise likelihood ratio test approach, and we demonstrate how our subtree-likelihood caching approach improves the speed of the approach, outperforming site-wise likelihood ratio testing. We also showcase difFUBAR on datasets of mammalian immunoglobulin sequences.

09:00

Bayesian Tree Growth modelling. An investigation into individual tree competition.

Lulama Kepe | *Department of Statistics, Nelson Mandela University*
Keith Little | *Dept. of Forestry, NMU*
Johan Hugo | *Dept. of Statistics, NMU*

At some stage after canopy closure, individual trees in a plantation begin to compete for the same resources. To investigate this competition, a Bayesian mixed effects model, similar in characteristics to a SIRE model used for estimating breeding values, as well as variance components, in mixed linear model settings, is proposed. In a similar way to the inclusion of inbreeding coefficients, it is therefore proposed that published competition indices used in tree growth modelling, be included in this Bayesian mixed model. As different competition indices are introduced into the model, posterior probabilities will be observed and compared to what is visually observed on the plot, i.e. if the tree with the highest posterior probability of being the strongest grower, is in fact the largest tree on the plot as well.

09:20

Bayesian Analysis of Historical Functional Linear Models with application to air pollution forecasting

Allan Clark | *University of Cape Town*
Yovna Junglee | *University of Toronto*
Birgit Erni | *University of Cape Town*

Historical functional linear models are used to analyse the relationship between a functional response and functional predictors. Here we develop a functional data analysis model that handles multiple functional covariates with measurement error and sparseness that can be used to predict functional response surfaces.

The method uses the connection between non-parametric smoothing and Bayesian methods to reduce sensitivity to the number of basis functions used to model the functional regression coefficients of the model. We investigate two methods of estimation. First, propose to smooth the predictors independently from the regression model in a two-stage analysis, and secondly, jointly with a regression model. The efficiency of the MCMC algorithms is increased by implementing a Cholesky decomposition to sample from high-dimensional Gaussian distributions and taking advantage of the orthogonal properties of the functional principal components used to model the functional covariates.

A simulation study suggests substantial improvements in both the recovery of the functional regression surface and the true underlying functional response with higher coverage probabilities, when compared to a classical model under which measurement error is unaccounted for. We also found that a two-stage analysis outperforms the joint model under certain conditions.

A major challenge with the collection of environmental data is that they are prone to measurement error. Hence, our methodology provides a reliable functional data analytic framework for modelling such data. As an application of our method, we forecast the level of daily atmospheric pollutants at certain locations in the City of Cape Town. The forecasts provided by the Bayesian two-stage model are highly competitive where compared to the functional autoregressive models which are traditionally used for functional time series.

09:40

Optimal window size detection in Value-at-Risk forecasting: A case study on conditional generalised hyperbolic models

Chun-Sung Huang | *University of Cape Town, Cape Town, South Africa*
Chun-Kai Huang | *Curtin University, Bentley, Western Australia*
Jahvaid Hammujuddy | *University of KwaZulu-Natal, Durban, South Africa*
Knowledge Chinhamu | *University of KwaZulu-Natal, Durban, South Africa*

The conventional parametric approach for financial risk measure estimation involves determining an appropriate quantitative model, as well as a suitable historical sample period in which the model can be trained. While a lion's share of the existing literature entertains the identification of the most appropriate model for different types of financial assets, or across conflicting market conditions, little is known about the optimal choice of a historical sample period size (or window size) to train the model and estimate model parameters. In this paper, we propose a method to identify an optimal window size for model training when estimating risk measures, such as the widely-utilised Value-at-Risk (VaR) or Expected Shortfall (ES), under the generalised hyperbolic subclasses. We show that the accuracy of VaR estimates may increase significantly through our proposed method of optimal window size detection. In particular, our results demonstrate that, by relaxing the usual restriction of a fixed window size over time, superior VaR forecasts may be produced as a result of improved model parameter estimates.

Stream: Spatial Statistics
Thursday, 1 December, 08:00 – 10:00

Venue: Charlotte Room
Chair: Inger-Fabris Rotelli

08:00

Modelling probabilistic hail hazard in South Africa: following the swath

Ansie Smit | *UP Natural Hazard Centre, Department of Geology, University of Pretoria*
Palesa Makena | *Department of Statistics, University of Pretoria*
Cameron Drotsky | *Department of Statistics, University of Pretoria*
Kabelo Mahloromela | *Department of Statistics, University of Pretoria*
Liesl Dyson | *Geography, Geoinformatics & Meteorology, University of Pretoria*
Inger Fabris-Rotelli | *Department of Statistics, University of Pretoria*
Christine Kraamwinkel | *Department of Statistics, University of Pretoria*
Andrzej Kijko | *UP Natural Hazard Centre, Department of Geology, University of Pretoria*

Hail is an extreme meteorological event causing global, yearly cumulative insured losses up to US$1 billion. During the austral summer rainfall season, convective thunderstorms are almost a daily occurrence in South Africa and can be accompanied by large hail events. Yet, observational hail data remains sparse, difficult to collect, and are affected by low spatial and temporal resolution, incompleteness and heterogeneous observational systems. This interdisciplinary research project focuses on investigating various aspects of the South African hail climatology and probabilistic hail hazard analyses. Thus far, the project yielded a hail climatology using pseudo-soundings from ERA-Interim reanalysis data and the HAILCAST (EIH) model, predicting the average hail stone diameter expected per day. The estimated hail day frequency compares well to historical climatologies, but the horizontal resolution of ERA-Interim are problematic in areas of steep topography. The project now focus on developing hazard estimates using EIH predictions and include area-characteristic maximum event size, probabilities of exceedance and return periods. The methodology accounts for aleatory and epistemic uncertainty associated with the predicted event sizes and the applied distributions. Maximum likelihood and Bayesian estimates are calculated by employing likelihood functions from convolution and mixture distributions. Another research stream investigates if the spatial dependency in hail event translates from EIH to the probabilistic hail hazard parameters. The spatial autocorrelation between the frequency of hail events from EIH and the spatial autocorrelation between the mean activity rate when uncertainty was taken into consideration are investigated using Spatial Autoregressive models (SAR). The interdisciplinary methodologies applied and preliminary results achieved from the project will be discussed.

08:20

Spatial prediction on disjoint spatial lattice data

Michelle de Klerk | *Lightstone and Department of Statistics, University of Pretoria*
Inger Fabris-Rotelli | *Department of Statistics, University of Pretoria*

Modeling on spatially disjoint lattice data presents challenges in the determination of appropriate spatial dependency. When considering the feeder areas of points-of-interest based on drive-time to residential areas, spatially overlapping areas will typically be observed in metro areas. Spatially disjoint areas will be identified for households which fall outside of metro drive-time catchment areas. We present an approach for spatial regression making use of covariates, to model on such a spatial data set. The methodology is applied to healthcare points-of-interest considering distance and location as well as sociodemographic covariates (population density, income etc.) and environmental covariates (rainfall, temperature, and proportion of healthcare services in an area). Current applications being investigated include retail sales of over-the-counter medication at pharmaceutical stores and identifying service areas of public hospitals and laboratories.

08:40

Bayesian Structured Additive Spatial Model of Intimate Partner Violence among Women in Nigeria

Adeboye Azeez | *University of Fort Hare*
Ruffin Mpiana Mutambayi | *University of Fort Hare*
Akinwumi Odeyemi | *University of Fort Hare*
James Ndege | *University of Fort Hare*

Intimate partner violence (IPV) is a major public health issue that affects millions of women around the world. It has enormous global health burden and negative socioeconomic impact on affected individuals of different social classes, religious, and cultural groups. The purpose of this study was to identify disparities in physical or sexual IPV against women using Bayesian structured addictive regression (STAR) models to determine the factors associated with IPV in Nigeria. We used 2018 Nigerian Demographic and Health Survey (DHS) data, which is a national cross-sectional survey database. The sample size used in this study was 6387 women aged 14-59 years. We applied Bayesian Geo-Additive models in this analysis. Of the 6387 women included in the study, about 24% were aged 25-29 years, 69.5% were living in urban area, 96.6% were married and have working partners, 76.2% had partners with no alcohol history and about 39% had secondary education. The risk of women experiencing IPV was 25.2% higher (POR = 1.252, CrI = 1.218–1.290) among women whose partner drinks alcohol, 11.3% higher (POR = 1.113, CrI = 1.023–1.217) among women who had no education, and 11.5% higher (POR = 1.115, CrI = 1.062–1.172) among women with only primary education than among those women who had up to higher education. There was 8.6% risk higher among working partners than among those partners were not working. The prevalence of IPV among older women in this study demonstrates that it is a serious issue with significant variation across and

within states showing IPV vulnerability. This cut across socioeconomic status and there is a need for multi-sectoral approaches to preventing and responding to IPV.

09:00

Age-stratified COVID-19 epidemiological model

Jenny Holloway | *CSIR*
Inger Fabris-Rotelli | *Department of Statistics, University of Pretoria*
Renate Thiede | *Department of Statistics, University of Pretoria*
Claudia Dresselhaus | *Department of Statistics, University of Pretoria*
Nada Abdelatif | *SAMRC*
Charl Janse van Rensburg | *SAMRC*
Pravesh Debba | *CSIR*
Nontembeko Dudeni-Tlhone | *CSIR*
Raeesa Manjoo-Docrat | *Department of Statistics and Actuarial Science, University of Witwatersrand*
Elona Mbayise | *Department of Statistics and Actuarial Science, University of Witwatersrand*
Warren Brettenny | *Department of Statistics, Nelson Mandela University*

Deterministic compartmental models have been used extensively to model the dynamics of COVID-19. The assumption with these models is that there is homogenous mixing in individuals in that they all have the same chance of moving to the other compartments. Stratified compartmental models consider epidemiological differences between sub-populations by assuming different rates of transitions. COVID-19 has affected age groups differently in terms of the severity of infection, hospitalization rates and death, especially in the first and second waves in South Africa. We present an age-stratified susceptible-exposed-infected-recovered (SEIR) model with adjusted contact matrices, as well as spatial variability between districts of South Africa. Contact matrices between the different age groups were included in the model and used to produce adjusted basic reproduction numbers (R0) per age group. Spatial weights, at the district level, were also incorporated using Facebook mobility data. The results will show age-dependent effects as well district variability for the first and second waves of COVID-19 in South Africa.

09:20

An exploratory analysis of location information from the mobileDNA application

Fallo Happy Khanye | *Statistics and Population Studies Department, University of the Western Cape*
Julia Keddie | *Statistics and Population Studies Department, University of the Western Cape*
Renette Blignaut | *Statistics and Population Studies Department, University of the Western Cape*

According to studies, global positioning systems are more trustworthy than other techniques for gathering location data and have opened up new research opportunities, particularly in Big Data. The purpose of this study is to investigate mobileDNA users' behaviour in terms of where and when they utilized their smartphone on a daily basis, including the applications they access. We designed an interactive dashboard to study mobileDNA user mobility and mobile application activity in order to gain insights that will help understand location data in a social science setting. We discovered that the application usage sequence of users on a daily path varies from day to day, and that most users using the mobileDNA application travel to more than one city or town in a day . Our data has a significant limitation in the collection of GPS records. There are gaps in the recorded location sequence when a user is inactive on their mobile device at the time the data is recorded, which is every 15 minutes. On this basis, it is recommended to collect the GPS location data irrespective of the user being active or not at the time of data collection. This will allow for more flexibility and accuracy in the analysis process. Further research is needed to cluster mobileDNA users based on the amount of time spent on their smartphones and then discover whether there are any commonalities in the application use sequences of users in that cluster.

09:40

Linear hotspot detection for a point pattern in the vicinity of a linear network

Inger Fabris-Rotelli | *Department of Statistics, University of Pretoria*
Jaocb Modiba | *Department of Statistics, University of Pretoria*
Alfred Stein | *University of Twente, Netherlands*
Gregory Breetzke | *Department of Geography, Geoinformatics and Meteorology, University of Pretoria*

The analysis of point patterns on linear networks is receiving current attention in spatial statistics. This refers to the analysis of points in a spatial domain that coincide with a linear network like a road network. The linear network is modelled as a set of lines that are connected at their ends or are intersecting, that is, modelled as mathematical graphs. Limited research so far has been conducted on spatial points that fall on the Euclidean space containing the linear network. This study addresses new steps by exploring points in the vicinity of the network that do not necessarily fall on the linear network. We present a novel method that is motivated by crime locations amongst a road network. The aim is to detect spatial hotspots around a linear network, where crime locations are considered as a point pattern lying in the vicinity of the linear road network. A new connectivity measure is also introduced to define the line segment neighbours of a line segment. The methodology is applied to crime data in Khayelitsha, South

Africa. We detect a pattern of crime locations within the network that can be well interpreted. We conclude that our method is well applicable and could potentially help governmental organisations to allocate measures to reduce criminality.

Stream: Young Stats (Data Science) Venue: George Room
Thursday, 1 December, 08:00 – 10:00 Chair: Annegret Muller

08:00

Factorisation machines as a statistical modelling technique

Erika Slabber | *Evolution Finance (Senior Data Analyst) and PhD student at the North-West University*
Tanja Verster | *Business Mathematics and Informatics, North-West University*
PJ de Jongh | *Business Mathematics and Informatics, North-West University*

Factorisation machines originated from the field of machine learning literature and have gained popularity because of the high accuracy obtained in several prediction problems, in particular in the area of recommender systems. This article will provide a motivation for the use of factorisation machines, discuss the fundamentals of factorisation machines, and provide examples of some applications and the possible gains by using factorisation machines as part of the statistician's model-building toolkit. Data sets and existing software packages will be used to illustrate how factorisation machines may be fitted and in what context it is worth being used.

08:20

Classification of Photovoltaic Module Faults Using a Novel Deep Learning Architecture

Edward Westraadt | *Department of Statistics, Nelson Mandela University*
Warren Brettenny | *Department of Statistics, Nelson Mandela University*
Chantelle Clohessy | *Department of Statistics, Nelson Mandela University*
Ernest van Dyk | *Department of Physics, Nelson Mandela University*

Faults arising in photovoltaic (PV) systems can result in major energy loss, system shutdown and safety breaches. It is thus crucial to detect such faults to improve the efficiency, reliability and safety of such PV systems. This study seeks to develop a new deep learning architecture for the classification of faults in large PV installations. This study extends on past research, as well as the published work of Dunderdale et al. (2020) at Nelson Mandela University, in an effort to find the most efficient technique for classifying PV faults using thermal imagery. This study will attempt to design, construct and test a new convolutional neural network (CNN) architecture, with the specific purpose of classifying PV faults into a total of fourteen IEC-aligned fault categories. These categories are designed based on the size, shape and intensity of the faults within the thermal images. The results obtained are compared to various popular CNN architectures, namely: InceptionV3, ResNet50, and Xception.

The research of Bommes et al (2021) provides a semi-automated process which performs the task of detection and classification of PV faults in large-scale systems. The use of a novel CNN architecture within such a system, with the specific purpose of PV fault classification, may be extremely beneficial to both researchers and PV plant operators or maintainers. Made up of aspects from both the physical science and statistics fields, this study could yield useful tools for larger scale PV installations.

08:40

Uncertainty problem in sampling

Nyiko Khoza | *Department of Statistics, University of South Africa*

Whenever there is large data (i.e., big data) the only feasible way to get an inference about the data is to sample the population. The subsection representing a population that contains all substantial elements of the population is called a sample. Since it is not ideal to evaluate a large population, we sample the population and make an inference about the population from the sample. However, we show that not all samples are representative of the population because of the uncertainty that exists. We evaluate the uncertainty about the sample in estimating the population using variable importance of the HPSPLIT procedure. Instead of randomly selecting a sample from a population without knowing its uncertainty when estimating the population, we select the most representative sample about the population. This allows the study to make accurate inferences about the population.

09:00

Application of CNN-gcForestCS to cassava leaf disease detection

Liam Carew | *Department of Statistical Sciences, University of Cape Town*
Stefan Britz | *Department of Statistical Sciences, University of Cape Town*

Cassava is one of the most consumed carbohydrates in the world, providing a reliable source of income and nutrition to inhabitants of Latin America, Africa and Asia. However, its production is greatly affected by pathogenic infection with cassava mosaic disease (CMD) posing the greatest threat to cassava farmers in Africa and Asia. Given that developing nations are estimated to be hit hardest by climate change and projected to have the largest population increases in coming decades, optimisation of cassava yield in these areas is imperative to ensure food security. Traditionally, crop health is determined by manual inspection which can be

laborious, error-prone and require technical expertise. This produces a costly barrier of entry for smallholding farmers who make up majority of global cassava production. Development of automated disease detection systems using convolutional neural networks (CNNs) deployable on mobile phones have shown to be a cost-efficient and effective method for cassava monitoring, mainly owing to their advanced feature extraction capabilities. However, CNNs require complex hyperparameter tuning and can be computationally intensive to train. gcForestCS (multi-grained cascade forest with confidence screening) presents a novel statistical learning method that can be trained using CPU, and requires less complex hyperparameter tuning than deep learning while producing competitive performance for lower-dimensionality datasets. Taking advantage of the feature extraction capabilities of CNNs and the competitive performance of gcForestCS for lower-dimensionality datasets, CNN-gcForestCS was investigated as an alternative to deep learning for cassava leaf disease detection. Performance of this method was compared to gcForestCS and deep learning with and without training set class balancing. Results thus far have shown that the best deep learning model (85.2%) outperforms the best CNN-gcForestCS model (84.83%).

09:20

Multivariate big data sampling for crop area coverage

Tshepiso Selaelo Rangongo | *Department of Statistics, University of Pretoria*
Inger Fabris-Rotelli | *Department of Statistics, University of Pretoria*
Renate Thiede | *Department of Statistics, University of Pretoria*

Big data can result in more than sufficient information if used efficiently and effectively. Big data poses challenges in storage, management, processing, analysis and visualisation. Techniques for handling big data, specifically geospatial data, have advanced over the years. However, most require high computational power and time. The use of metadata is a solution. Metadata provides a descriptive, administrative, structural and statistical summary of data. This paper constructs metadata of a remote sensing image dataset for crop classification, and proposes a novel multivariate stratified sampling algorithm which selects the most informative images to minimise the number of images used for taining. The proposed sampling algorithm performs effectively on a big spatial image dataset of crop types. The results are assessed by measuring the number of images sampled and as well as matching the proportionality of the population crop percentages.

09:40

Label-dependent splitting for multi-label data

Annegret Muller | *Department of Statistics and Actuarial Science Stellenbosch University*

Multi-label classification extends single-label classification to settings where each data case may be associated with a set of labels. Effective exploitation of correlation amongst labels can be a vital attribute of an accurate multi-label learning method. In this work a new tree-based ensemble method for multi-label classification, Label-Dependent Splitting (LDsplit), is proposed. The algorithm fits an ensemble of trees based on differently ordered label subsets. For each tree, different labels are used at different levels of the tree, as determined by the label order applicable to that tree. The tree-levels are made up of nodes that are split using any binary classifier. As data cases filter down a tree, they are split in a label-dependent manner, thereby implicitly incorporating label correlations into the model in a simple manner. Both a random and conditional label ordering strategy are implemented. Empirical evidence shows that despite the simple framework, both strategies offer very competitive performance compared to other multi-label learning methods.

Keynote

Venue: Regency Hall
Thursday, 1 December, 10:30 – 12:00
Chair: Warren Brettenny

10:30

Emerging technologies, globalization and social challenges are redefining the role of intelligent decisioning

Mark Nasila | *Chief Data and Analytics Officer at FirstRand Risk*

The era of emerging technologies is changing the way we live, work and connect with and relate to one another. At the same time, businesses face changing expectations from their customers, and governments have to contend with new demands from their citizenry. We are at an inflection point where emerging technologies are on the cusp of becoming mainstream, and the borders between physical, digital and biological technologies are becoming harder to delineate. Neurotechnology, artificial intelligence and robotics have begun to bleed into one another. These emerging technologies, globalization and social challenges have reimaged the role of intelligent decisioning. This presentation will cover: this new reality that is driving demand for alternative intelligent decisioning models; data science applications leveraging alternative intelligent decisioning; how organizations can inform their strategies with modern intelligent decisioning; designing for Human – Machine (augmented intelligence) partnerships Operating models that will define the future; and skill to drive strategies enabled by future intelligent decisioning models.

11:15

Towards evidence-based public health management: The role of statistics and modelling in South Africa

Sheetal Silal | *Director of Modelling and Simulation Hub, Africa (MASHA)*

Be it COVID-19, TB, HIV or routine early childhood immunisation, the public health system is continually needing to adapt and

adjust health provision policy to maximise the delivery of health services to the population. Increases in health surveillance, clinical trials and digitised systems have resulted in numerous datasets becoming available to support health planning. Statistics and mathematical modelling are in a unique position to leverage these datasets to provide a scientific evidence base to decision-makers. This talk will present past and current applications of statistical and math modelling in various aspects of public health in South Africa, and reflect on the role we can all play in shaping the future of health provision.

Stream: Biostatistics

Venue: Regency Hall
Thursday, 1 December, 12:00 – 13:00
Chair: Sisa Pazi

12:00

Examining factors that contribute to under-five mortality rates in South Africa using count models

Kgethego Sharina Makgolane | *Department of Statistics and Operations Research, University of Limpopo*

Under-five mortality remains a major health challenge in most sub-Saharan African countries including South Africa, despite the significant progress made in child survival and the government's efforts to reduce under-five mortality rate. South Africa has failed to achieve the 4th Millennium Development Goal which aimed at reducing under-five mortality rate by two-thirds between the year 1990 and 2015. Due to this failure, the 3rd Sustainable Development Goal which aims to have no more than 25 deaths per 1 000 live births by the year 2030 was implemented. The aim of this study was to identify factors that contribute to under-five mortality rate using count models. To identify these factors, the study utilised a secondary data set obtained from the South African Demographic and Health survey for 2016. Generalized linear models namely, logistic regression, Poisson regression and negative binomial regression models were employed for the analysis of under-five mortality rate. The results revealed that baby postnatal check-up within the first two months, the child's health checked prior discharge, childbirth size, toilet facility at home, maternal education, province, type of residence and water source were significantly associated with the risk of experiencing under-five mortality. In conclusion, the study suggests that the Department of Health and various concerned agencies take these factors into account when planning to reduce under-five mortality rate and achieve the 3rd Sustainable Development Goal by 2030.

12:20

Joint modeling for longitudinal and interval censored survival data

Isaac Singini | *Department of Statistics, Faculty of Agriculture and Natural Sciences, University of Pretoria, South Africa*
Ding-Geng Chen | *College of Health Solutions, Arizona State University, Phoenix, AZ 85004, USA*

Joint models for longitudinal and survival data are a class of models that jointly analyse an outcome repeatedly observed over time such as a bio-marker and associated event times. These models are useful in two practical applications; firstly focusing on survival outcome whilst accounting for time-varying covariates measured with error and secondly focusing on the longitudinal outcome while controlling for informative censoring. For the survival sub-model this is done by recording the moments of the event of interest and calculation the time span between the event and some initial onset time. The joint modelling framework has mainly focused on right censored data in the survival outcome for the last decade. This has been for two-Stage joint model, shared parameter joint models and latent class joint models. There have been many theoretical developments in the last five decades that have focused on censoring mechanisms in oder to correctly model time to event data e.g. left or right censoring, however interval censoring has seldom been implemented in the joint modeling framework. This has been due to the fact that many are unaware of the impact of inappropriately dealing with interval censoring within the joint modeling framework. The other complexity has been that the necessary software is at that handles interval censored data in the joint modeling framework is not readily available. In this chapter we fill the gap between theory and practice by illustrating our theoretical technique using the interval censored data in the joint model using a cardiology multi-centre clinical trial. We implement our approach with examples using R statistical software.

12:40

Contributions to acute physiology scoring for South African intensive care units

Sisa Pazi | *Department of Statistics, Nelson Mandela University*
Gary Sharp | *Department of Statistics, Nelson Mandela University*
Elizabeth van der Merwe | *Adult Critical Care Unit, Livingstone Hospital*

This study describes research which was conducted in fulfilment of a Doctor of Philosophy degree. The research comprised of four studies, the first of which sought to investigate the epidemiology of acute kidney injury (AKI) at a tertiary hospital in the Eastern Cape. The Simplified Acute Physiology Score III (SAPS III), a severity-of-illness score, was found to be one of the statistically significant risk factors for AKI. As the SAPS III was developed without data from Africa, this opened the opportunity to scrutinise the model, which then led to the second research study, the purpose of which was to assess the SAPS III model in the South African context. The results of the second study provided the motivation to develop a model more suited for the South African context, which led to the third study, the aim of which was to develop a model similar to the SAPS III model but using South African data. The results of that study indicated that the proposed adaptive model was superior to the SAPS III model. Furthermore, a comparative analysis conducted as part of the fourth study indicated that the proposed model was superior to some machine learning models. To broaden the usage of the proposed adaptive model, future research includes collecting data from multiple hospitals in South Africa. The collected data will then be used to externally validate the proposed adaptive model.

Stream: Nonparametric Statistics

Venue: Charlotte Room
Thursday, 1 December, 12:00 – 13:00
Chair: Jean-Claude Malela-Majika

12:00

An investigation on the use of Bernstein polynomials in entropy estimation

Shawn Liebenberg | *Department of Statistics, North-West University*

Entropy estimation has become an important component in many fields of research. Among the many developed procedures for estimating entropy, spacing and kernel density based procedures have become the most prominent. Kernel density estimation is plagued by boundary bias that potentially carry over to the corresponding entropy estimators. This study introduces two new Bernstein based entropy estimators and aims to investigate Bernstein polynomial density estimation in entropy estimation as a remedy to the boundary bias problem. It was found that the Bernstein based entropy estimators performed very well against the spacing and kernel density estimators used in this study.

12:20

Statistical process control: A review of current practices and some new recommendations for optimal design schemes

Marien Alet Graham | *Department of Science, Mathematics and Technology Education, University of Pretoria*
Jean-Claude Malela-Majika | *Department of Statistics, University of Pretoria*

Statistical process control (SPC) refers to a collection of statistical techniques that are widely employed to monitor and enhance the quality of processes. SPC primarily entails adopting monitoring schemes which detect changes in a process that could impact the output quality. These monitoring schemes were initially employed primarily on manufacturing processes; their applications have since expanded to include the food industry, education, engineering, environmental research, biology, genetics, epidemiology, medical, finance, law enforcement, and sports, amongst others. The problem is not the exponential growth of these schemes but rather that recent research still uses outdated performance measures, such as the mean of the run-length distribution (ARL). In addition, many recent publications do not verify the necessary assumptions when applying well-known parametric monitoring schemes (which have underlying assumptions, such as normality, that are rarely met in practice), nor do they comply with the required Phase-I sample sizes for these schemes to operate effectively. In general, when standards are unknown (Case U), the unconditional ARL is considered during Phase-II monitoring. The impact of bias in the Phase-I sample may result in remarkably high rates of early false alarms. We explore the idea of restricting the probability of unconditional early false. This new method is referred to as the "lower percentile-based design". Consideration is given to the design and implementation of several nonparametric techniques employing a prefixed value of some lower percentile point of the in-control run-length distribution. The optimal scheme has the lowest value for a specific higher percentile point of the out-of-control run-length distribution. We illustrate the new design and implementation methodologies with actual data, provide a summary and conclusion, and make recommendations for future research.

12:40

Nonparametric precedence chart with repetitive sampling

Jean-Claude Malela-Majika | *Department of Statistics, University of Pretoria*

In most real-world applications, such as production and manufacturing processes, the underlying process distribution does not always follow a normal distribution. In such cases, statistical process control literature recommends the use of nonparametric (or distribution-free) control charts. This paper introduces a new distribution-free precedence chart using repetitive sampling. The performance of the proposed chart is investigated in terms of the average run-length (ARL) profile. The expressions of the in-control process probability and ARL of the proposed chart are introduced using integral formulas. The out-of-control performance of the new chart is compared to that of the existing precedence charts with and without runs-rules. A numerical example is provided using real-life data to demonstrate the application and implementation of the new chart.

Stream: Young Stats (Spatial Statistics)

Venue: George Room
Thursday, 1 December, 12:00 – 13:00
Chair: Renate Thiede

12:00

Multiscale decomposition of spatial lattice data to detect hotspots of COVID-19 cases in South Africa

René Stander | *Department of Statistics, University of Pretoria*
Inger Fabris-Rotelli | *Department of Statistics, University of Pretoria*
Ding-Geng Chen | *Department of Statistics, University of Pretoria*

During a pandemic such as COVID-19, it is important to know where positive cases are clustered for local governments to implement measures to control the spread of the disease. The detection of such hotspot areas is an important part of spatial analysis. In

current literature, several different methods have been used such as measures for local spatial association and spatial scan statistics. In this work, we propose a new approach, making use of the Discrete Pulse Transform (DPT) on spatial lattice data along with the multiscale Ht-index as a measure of saliency on the extracted pulses to detect significant hotspots.

12:20

Covariate construction of nonconvex windows for spatial point patterns

Kabelo Mahloromela | *Department of Statistics, University of Pretoria*
Inger Fabris-Rotelli | *Department of Statistics, University of Pretoria*
Christine Kraamwinkel | *Department of Statistics, University of Pretoria*

Window selection for spatial point pattern data is complex. Often, the point pattern window is given a priori. Otherwise, the region is chosen using some objective means reflecting that the window is representative of a larger region. Common approaches used are the smallest rectangular bounding window and convex windows. The chosen window should however cover the true domain of the process. Choosing too large a window results in estimation and inference in regions where the possibility of observations has not been confirmed. We propose a new algorithm for selecting a point pattern domain based on spatial covariate information without the restriction of convexity, allowing for a better fit to the true domain. The proposed algorithm is applied in the setting of rural villages in Tanzania. As a spatial covariate, remotely sensed elevation data is used. The algorithm is able to detect and filter out high relief areas and steep slopes; observed characteristics that make the occurrence of a household in these regions improbable. A modified kernel smoothed intensity estimate using the Euclidean shortest path distance is proposed to estimate the intensity on the resultant nonconvex window, producing more representative intensity estimation.

12:40

Measuring Homogeneity of Linear Networks

Renate Thiede | *Department of Statistics, University of Pretoria*
Inger Fabris-Rotelli | *Department of Statistics, University of Pretoria*
Pravesh Debba | *Council for Scientific and Industrial Research*
Christopher W Cleghorn | *School of Computer Science and Applied Mathematics, University of the Witwatersrand*

Spatial linear networks exhibit a variety of patterns. Like spatial point patterns, they can be homogeneous or heterogeneous. While there exist a wide variety of tests for the homogeneity of point patterns, no statistical tests currently exist to quantify the homogeneity of spatial linear networks. This research provides a statistical methodology to test for homogeneity in spatial linear networks. A simulated spatial linear network is approximated by a point pattern, which is obtained by overlaying a grid on the extent of the linear network and representing the midpoint of each line in each grid cell by a point. Existing tests for homogeneity of point patterns are then applied. This research investigates the optimal grid size using measures of internal uniformity and robustness to error. Furthermore, the power of the tests are investigated.

Friday, 2 December

Stream: Multivariate Statistics
Friday, 2 December, 08:00 – 09:00

Venue: Regency Hall
Chair: Sugnet Lubbe

08:00

Two New Auxiliary Models for Estimating Error Variances in Heteroskedastic Linear Regression

Thomas Farrar | *Department of Mathematics & Physics, Cape Peninsula University of Technology; Department of Statistics & Population Studies, University of the Western Cape*
Renette Blignaut | *Department of Statistics & Population Studies, University of the Western Cape*
Retha Luus | *Department of Statistics & Population Studies, University of the Western Cape*
Sarel Steel | *Department of Statistics & Population Studies, University of the Western Cape; Department of Statistics & Actuarial Science, Stellenbosch University*

Two new models are proposed for estimating error variances in heteroskedastic linear regression models. These are, respectively, the Auxiliary Linear Variance Model and the Auxiliary Nonlinear Variance Model, which use the squared Ordinary Least Squares residuals as their response and are built around a correct specification of the conditional mean response. The dimensionality of the parameter vector is reduced by assuming a functional relationship between the error variances and the predictor variables. Several different sub-models emerge depending on how one deals with the heteroskedastic function.

Practical problems in applying the models are discussed, such as parameter estimation, tuning of hyperparameters, and feature selection. Methods of parameter estimation include inequality-constrained least squares and quadratic programming for the linear model and maximum quasi-likelihood estimation for the nonlinear model. Methods of hyperparameter tuning include K-fold cross-validation and quasi-generalised cross-validation. Methods of feature selection include feature-wise heteroskedasticity testing, best subset selection, and LASSO.

The new error variance estimation methods are assessed under a variety of experimental conditions in terms of four distinct mean squared error metrics, and are found to outperform existing methods under some conditions. The nonlinear model is particularly effective if the form of the heteroskedastic function is known; the linear model is more reliable otherwise. The new variance models are found to be competitive methods for handling heteroskedasticity in linear regression.

08:20

High-dimensional LDA Biplot through the GSVD

Raeesa Ganey | *School of Statistics and Actuarial Science, University of Witwatersrand*
Sugnet Lubbe | *Statistics, Stellenbosch University*

Discriminant analysis is a multivariate technique concerned with separating distinct sets of observations. However, a common limitation of trace optimisation in discriminant analysis is that the within cluster scatter matrix must be nonsingular, which restricts the use of data sets when the number of variables is larger than the number of observations, p > n. In this presentation, we show that by applying the generalised singular value decomposition (GSVD), we can achieve the same goal of discriminant analysis regardless on the number of p. This originates from the work done by Howland, Jeon and Park (2003). Furthermore, we describe an attempt to construct a meaningful biplot from the GSVD approach.

Reference: P Howland, M Jeon and H Park, “Structure preserving dimension reduction for clustered text data based on the generalised value decomposition”, Society for Industrial and Applied Mathematics, 2003.

08:40

Biplots for individual differences scaling models

Sugnet Lubbe | *MuViSU, Department of Statistics and Actuarial Science, Stellenbosch University*
Niël le Roux | *MuViSU, Department of Statistics and Actuarial Science, Stellenbosch University*

Indscal models deal typically with two mode, three way data. The typical format is a set of K $n \times n$ distance matrices, for instance K judges each rating differences between n items. Parallel to classical scaling, also known as Principal Coordinate Analysis, a set of positive semi-definite symmetric matrices are formed by double centring the squared distance matrices. In general, for any set of K positive semi-definite symmetric matrices, the Indscal model finds the best, in the least squares sense, representation of the n objects in r, usually 2, dimensions and an associated set of r weights for each of the K judges. For $r = 2$, two plots can be made: the subject space, based on the K sets of weights and a compromise group stimulus space, representing the n objects / items. Assuming that the dissimilarities between the objects were generated by observations on p variables, we want to simultaneously represent the n objects and p variables in a biplot. In this paper we will discuss how to represent the variables with the objects in the group stimulus space. Representing the variables as biplot axes, allows for the prediction of the p variable values for any point in the r-dimensional biplot space. We will also discuss how to do the converse: finding the r-dimensional coordinates for p (new) observations on the variables.

Stream: Biostatistics
Friday, 2 December, 08:00 – 09:00

Venue: Charlotte Room
Chair: Nontembeko Dudeni-Tlhone

08:00

Flexible Statistical Modelling of The Determinants of Childhood Anaemia In Tanzania and Angola.

Qondeni Ndlangamandla | *Department of Biostatistics*
Henry Mwambi | *University of KwaZulu-Natal*
Shaun Ramroop | *University of KwaZulu-Natal*
Nonhlanhla Yende | *CAPRISA*

Anaemia is one of the major causes of morbidity and mortality in children aged five or less in Africa, affecting 25% of the world's population. In developing countries, it accounts for more than 89% of the disease burden. Although it affects all population groups, but children under five years of age and women of reproductive age are more valnurable. This study aims to determine the factors associated with childhood anaemia in Tanzania and Angola. A survey logistic regression (SLR) and GAMM (generalized additive mixed models) were fitted to identify factors associated with childhood anaemia, both of which consider the survey weights, stratification, and clustering within primary sampling unit. The GAMM model offered more flexibility than the standard survey logistic model and provided a better fit to the data. Hence, the results presented are based only on GAMM. According to GAMM factors that are highly associated with childhood anaemia in both countries are: Child age, parents or guardians' age, parents or guardians' education level, stunting, standard of living and the gender of the child. In Tanzania, Male children were found to be more likely to be anemic compared to females (p-value =0.042, OR = 1.115, 95%CI (0.809;0.991)). Children with mothers' who had secondary education had reduced chances of having anaemia compared to children who's their mothers had no education ((p-value < 0.001, OR = 0.647, 95% CI (0493;0.8450). Children who were not suffering from stunting were less likely to be anemic in contrast to those who were suffering from severe stunting (P-value <0.001, OR=0.628, 95% CI (0.534;0.738)). The models also reveals that poor kids were more likely to suffer from anaemia compared to kids from middle class (P-value = 0.012, OR = 0.920, 95% CI (0.795, 1.064). The child and parent age were fitted as non-parametric terms in the model, both have p-values < 0.001 and have the effective degrees of freedom (EDF) of 5.11 and 1.00 respectively.

08:20

A Möbius-transformed toroidal distribution for dihedral angles modelling in protein structure

Najmeh Nakhaei Rad | *Department of Statistics, University of Pretoria*
Thasmika Mohan | *Department of Statistics, University of Pretoria*
Ding-Geng Chen | *Department of Statistics, University of Pretoria*

In this paper we propose a novel distribution and its skew version on the torus by applying a Möbius transformation to an existing distribution. These distributions can then be used as a proposal distribution for Markov-Chain Monte-Carlo sampling to predict the 3-D structure of protein molecules. We discuss the related properties of the novel models and substantiate our contributions using two real datasets and a simulation study in the performance assessment of the estimating approach.

08:40

Safety monitoring of the COVID-19 vaccines in South Africa

Nontembeko Dudeni-Tlhone | *Next Gen Enterprises and Institutions, Council for Scientific and Industrial Research (CSIR)*
Jenny Holloway | *Next Gen Enterprises and Institutions, Council for Scientific and industrial Research (CSIR)*

As the COVID-19 pandemic began, nations worldwide focused on non-pharmaceutical tactics to slow the spread of SARS-CoV-2 virus, whilst scientists began to find ways to better understand, treat and control the spread of COVID-19 through pharmaceutical means.

When the COVID-19 vaccines became available to the market, their safety and effectiveness were questioned. Also, vaccination resistance among some segments of the global population began to emerge on social media, as a result, post-market surveillance gained attention. Meanwhile, health regulatory bodies such as the South African Health Products Regulatory Authority (SAHPRA) were progressing with post-market surveillance of COVID-19 vaccines licenced for use in South Africa.

Potential adverse drug effects are investigated at various stages of clinical trials. Trials of this kind, however, usually recruit a small number of participants, eliminating potential future drug candidates. Also, these trials may be conducted over short periods of time, which may not be sufficient to identify effects with prolonged latency. As a result, post-market surveillance is carried-out by health regulatory bodies to ensure public safety.

SAHPRA conducts medical product safety testing on a regular basis, but during the COVID-19 pandemic, they required additional capacity to provide timely evaluation of the safety aspects of these vaccines. SAHPRA then collaborated with the CSIR for analytics support and safety monitoring of the COVID-9 vaccines including Janssen COVID-19 Vaccine and Comirnaty (Pfizer-BioNTech). This involved analysing the characteristics of reports of adverse events following immunisation (AEFIs, implementing procedures for safety signal detection and responding to other additional analytic demands. This paper highlights some of the pharmacovigilance actions carried out to aid SAHPRA in responding to the safety concerns generated by the COVID-19 vaccinations and provides a snapshot of the outcomes.

Stream: Young Stats (Coputational Statistics)

Venue: George Room
Friday, 2 December, 08:00 – 09:00
Chair: Arno Otto

08:00

A comparative study of quantile regression and ridge regression based adaptive weights in variable selection and regularized quantile regression

Innocent Mudhombo | *Department of Statistics, University of South Africa*
Edmore Ranganai | *Department of Statistics, University of South Africa*

We compare the performance of adaptive weights in variable selection and regularized quantile regression (QR) in the presence of collinearity and collinearity influential points. The adaptive weights based on the ridge regression parameter are compared to ridge quantile regression (QRR) based parameters. The QRR based adaptive weights penalize variable coefficients differently at each regression quantile (RQ) level in contrast to the RR based weights which are global weights. These adaptive weights in regularization penalties for variable selection and regularized QR procedures are namely, QR-ALASSO and QR-AE-NET. The performance of the adaptive weights is measured in terms of how the respective QR-ALASSO and QR-AE-NET procedures perform in variable selection and prediction. We carry out a simulation study to compare the performance of these adaptive weights in the presence of mixed, moderate and high collinearities, as well as collinearity influential points. Results show that in the majority of cases, the QRR based adaptive weights outperform the RR based adaptive weights in prediction performance and correctly fitting models, though in fewer cases the latter is superior.

08:20

Estimation of a mixture of semi-parametric partial linear models

Ruan Jean du Randt | *Department of Statistics, University of Pretoria*
Sollie Millard | *Department of Statistics, University of Pretoria*
Frans Kanfer | *Department of Statistics, University of Pretoria*

This presentation considers a semi-parametric finite mixture of partial linear models with Gaussian errors. The semi-parametric structure allows for flexible modelling of the expected value of the response variable. These models are used in cases where the regression structure includes both parametric and non-parametric covariate structures. We demonstrate the properties of the profile likelihood expectation maximization algorithm (PL-EM) using a simulation study. The estimation algorithm is also demonstrated on real data.

08:40

Skew Laplace candidates emanating from scale mixtures for insightful computational modelling

Arno Otto | *Department of Statistics, University of Pretoria*
Andriëtte Bekker | *Department of Statistics, University of Pretoria*
Johan Ferreira | *Department of Statistics, University of Pretoria*

The search for appropriate and flexible models for describing complex data sets, often with departure from normality, remains a main interest in various computational research fields. In this paper, the focus is on developing flexible skew Laplace scale mixture distributions to model these data sets. Each member of the collection of distributions is obtained by dividing the scale parameter of a conditional skew Laplace distribution by a purposefully chosen mixing random variable. Highly-peaked, heavy tailed skew models with relevance and impact in different fields are achieved. Finite mixtures consisting of the members of the skew Laplace scale mixture models are investigated, further extending the flexibility of the distributions by being able to potentially account for multimodality. The maximum likelihood estimates of the parameters for all the members of the developed models are obtained via an EM algorithm. The models are fit to bodily injury claims of Massachusetts to show the applicability and compared to other existing flexible distributions. Various goodness of fit measures are used to validate the performance of the proposed models as valid alternatives.

Keynote

Venue: Regency Hall
Friday, 2 December, 09:10 – 10:00
Chair: Warren Brettenny

09:10

Census 2022 journey: Updating the nations statistical landscape

Aswell Jenneker | *Deputy Director General of StatsSA*

Census 2022 revolutionised the collection and processing of statistics. This came off the back of the COVID pandemic which delayed the collection by few months as well as numerous challenges to bring an updated population count to the fore.

The presentation will explore the new digital method of collection, as well as give a broad overview of the current state of the society as reflected in our data, which will be further illuminated with the planned release of Census 2022 data in 2023.

Academia and Industry

Venue: Regency Hall
Friday, 2 December, 10:10 – 10:30
Chair: Chantelle Clohessy

10:10

Setting up of collaborative systems between academic institutions and industry, to build a strong data skills/talent value chain

Delia North | *Statistics sector, School of Mathematics, Statistics and Computer Science, UKZN*
André Zitzke | *SAS*
Temesgen Zewotir | *Statistics sector, School of Mathematics, Statistics and Computer Science, UKZN*

It is well-evidenced that there is an acute shortage data science/data analytics skilled resources all over the world, and in South Africa in particular. There is further evidence of major unemployment amongst the youth, even for those who have graduated with a degree in higher education from a degree generating institution in the country. This disconnect, between the skills being developed and the skills required by industry, leads to substantial unrealised potential amongst the youth of the country.

This talk will focus on how a university and SAS have partnered to define a set of Industry-integrated Skills Development programs, aimed at increasing the flow of "job ready" data analysts into the workplace.

Keynote

Venue: Regency Hall
Friday, 2 December, 10:30 – 12:00
Chair: Chantelle Clohessy

10:30

Statistical Science: Enriching our lives

Pravesh Debba | *2020 SAS® Thought Leader, CSIR*

We see statistics being applied on a daily basis, through for example, weather reports, financial markets and pharmaceuticals. Each of these applications has benefits to the general public. Fields like big data analytics and data science have seen an exponential growth in the job markets due to data being created in all forms. The last 2 years has seen 90% of the world's data being collected.

However, we are sometimes slow and cautious to react to these opportunities and to demonstrate the ability of statistical science to provide real solutions to national and world-wide problems. Yet we have seen the impact of COVID-19 and the impact of loadshedding on our daily lives, even to the extent to which we have adapted to operate. Communication through media and social platforms also form a vital role in disseminating information that is credible by the journalists and reporters. They therefore rely on the use of scientific information for their storytelling.

In this talk, a series of case studies will be presented to demonstrate the way in which statistical science can be used to assist decision makers by providing them with supporting evidence in undertaking key decisions and to assist the public with a better understanding and awareness of relevant issues. This helps both the decision maker and public to better plan for the future and what lies ahead.

Some of the work that would be presented is by the SEPIMOD (Spatial Epidemiological Modelling) group that was formed during COVID-19 outbreak.

11:15

Can statisticians ignore data science, or should it be embraced?

Renette Blignaut | *2021 SAS® Thought Leader, University of the Western Cape*

In the 1960s Peter Naur used the word datalogy (datalogi), the science of data processes, instead of the term computer science. In 1961, John Tukey, described a field "data analysis" which might be the closer to the field of "data science" as it is known today. The term "data science" appears in the preface of Naur's 1974 book "Concise Survey of Computer Methods". In 1985, Chien-Fu Jeff Wu used the term "data science" as an alternative name for statistics. In 1997, Wu gave a lecture entitled "Statistics = Data Science?". Wu advocated that statistics be renamed to data science and statisticians be called data scientists. Twenty-five years later - are we still grabbling with this?

What are you - a statistician or a data scientist? What are the differences and what are the similarities? This presentation will explore the history and evolution of the term and discipline "data science".

Stream: Computational Statistics

Venue: Regency Hall
Friday, 2 December, 13:00 – 14:20
Chair: Leonard Santana

13:00

On testing for the assumptions of mixture cure models in the presence of covariates

James Allison | *Department of Statistics, North-West University*
Jaco Visagie | *Department of Statistics, North-West University*
Ingrid van Keilegom | *KU Leuven*

Mixture cure models have become popular models for lifetimes in various fields including medicine and finance. Although tests for the assumptions underlying these models exist in the absence of covariates, no test can be found in the literature which can be used in the presence of covariates. We propose a test that can be employed in the mentioned setting. The test utilises transformed data involving the Kaplan-Meier estimate of the distribution function of the lifetimes. We present a Monte Carlo study in order to demonstrate the finite sample performance of the proposed tests.

13:20

Comparing distance-based and traditional parameter estimation techniques for the Lomax distribution.

Thobeka Nombebe | *Department of Statistics, North West University*
Leonard Santana | *Department of Statistics, North West University*
James Allison | *Department of Statistics, North West University*
Jaco Visagie | *Department of Statistics, North West University*

We investigate the performance of a variety of estimation techniques for the scale and shape parameter for the Lomax distribution. These methods include the L-moment estimator, the probability weighted moments estimator, the maximum likelihood estimator, maximum likelihood estimator adjusted for bias, method of moments estimator and three different minimum distance estimators. The comparisons will be done by considering the variance and the bias of these estimators. Based on an extensive Monte Carlo study we found that the so-called minimum distance estimators are the best performers for small sample sizes, however for large sample sizes the maximum likelihood estimators outperform these minimum distance estimators. We conclude with a practical example applied in the context of duration models.

13:40

On estimating the mode of an angular distribution

Jaco Visagie | *Statistics Department, North-West University*
Fred Lombard | *Department of Statistics, University of Johannesburg*
Charl Pretorius | *Business Mathematics and Informatics, North-West University*

We propose estimators for the mode of an angular distribution, each adapted from a corresponding class of estimators defined on the real line. In addition to point estimation, the construction of confidence intervals using the bootstrap is considered. The asymptotic properties of the proposed estimators are outlined and a Monte Carlo study is included in order to compare the finite sample performance of the proposed estimators.

14:00

Goodness-of-fit tests for Poisson regression models

Leonard Santana | *Subject Group Statistics, North-West University, Potchefstroom*
Simos Meintanis | *Department of Economics, National and Kapodistrian University of Athens, Athens, Greece*
Joseph Ngatchou-Wandji | *EHESP Sorbonne Paris Cité & Institut E'lie Cartan de Lorraine, Nancy, France*
Marius Smuts | *Subject Group Statistics, North-West University, Potchefstroom*

We propose goodness-of-fit tests for models of count responses with covariates. We primarily focus on the null hypothesis that the observed data are from a Poisson regression model, however the proposed method is general enough to allow for the responses to follow any discrete distribution, conditional on covariates. The test criteria are formulated by using the probability generating function. In this talk, Monte Carlo results are presented to motivate the use of this test, and some asymptotic theory is also mentioned. An application on a real-world data set is also reported.

Stream: Econometrics and Business Statistics

Venue: Charlotte Room
Friday, 2 December, 13:00 – 14:20
Chair: Stefan Janse van Rensburg

13:00

Analysis of flexibility value related to forest stands

Tomas Tichy | *Department of Finance, Technical University of Ostrava*

A classic stand-level optimization problem of forest stands from the point of view of (mathematics and) economics dates back to Samuelson (1976), who maximized the net present value of an infinite series of timber regeneration, growth, and harvest cycles, though rather ignoring managerial exibility. However, current structure of various subsidizes and exibilities in the forest usage,

together with bark beetle infestation shocks to the market brings new challenges for the pricing and provide support for real options methodology. From the statistical point of view, a key question is which process is followed by relevant quantities and how to estimate its parameters with limited times series. In this contribution we analyze several such possibilities under simple models. Obviously, the presence of the American constraint makes the flexibility valuation problem more. Illustrative examples are provided.

13:20

A score driven volatility model with local leverage

Stefan Janse van Rensburg | *Department of Statistics, Nelson Mandela University*

Asymmetric conditional volatility models abound. Few models can, however, accommodate proper leverage effects whereby the condition volatility decreases in response to positive shocks. As an alternative, local leverage represents a partial form of leverage. We propose a quasi score-driven volatility model that accommodates local leverage. The model retains the robustness of existing score-driven volatility models. We discuss preliminary issues surrounding estimation and inference. A simple empirical application demonstrates the viability of the model.

Stream: Young Stats (Biostatistics)

Venue: George Room

Friday, 2 December, 13:00 – 14:20

Chair: Isaac Singini

13:00

Early prediction of acute kidney injury using machine learning methods

Awonke Nqayiya | *Department of Statistics, Nelson Mandela University*
Sisa Pazi | *Department of Statistics, Nelson Mandela University*
Gary Sharp | *Department of Statistics, Nelson Mandela University*

Acute kidney injury is a severe renal disease. It occurs when there is a rapid decrease in kidney's ability to purify blood and is common in critically ill patients. It is associated with adverse complications such as permanent failure in kidney's function, requirement of dialysis for blood purification, and death. Early prediction of AKI can alert medical caregivers about patients at risk of AKI so that they may timely implement suitable care, and potentially prevent AKI. The purpose of this study was two-fold. First, this study sought to identify risk factors associated with acute kidney injury. The second objective was to develop a machine learning model for prediction of acute kidney injury. The data set used consisted of 849 observations, 19 covariates and one response variable. The 849 patients were adult patients who were admitted to an intensive care unit during a period of one year from the 3rd of January 2017. The response variable was the acute kidney injury status, indicating whether a patient had acute kidney injury or not. Logistic regression was used to identify statistically significant risk factors associated with acute kidney injury. The statistically significant risk factors were then used to develop five machine learning models for prediction of acute kidney injury. Performance metrics such as area under receiver operating characteristic curve, accuracy, sensitivity, specificity, and precision were used to compare the models. The results showed that all the models exhibited good performances, with the random forest model being the superior model.

13:20

Comparative performance evaluation of logistic regression and machine learning methods on different data types

Zenzile Ntshabele | *School of Statistics and Actuarial Science, Wits*
David Rose | *School of Statistics and Actuarial Science, Wits*
Charles Chimedza | *School of Statistics and Actuarial Science, Wits*

Different techniques have evolved in response to solving classification problems. These techniques include traditional statistical and machine learning methods. Their performance differs depending on the characteristics of the data used, making it difficult to determine which method will perform best under certain conditions, and this is a challenge that this research attempts to address. This study used various dimension sizes of balanced and unbalanced simulated data to compare the performance of logistic regression (LR) using both logit and link functions to that of machine learning techniques such as artificial neural networks (ANN), support vector machines (SVM), and weighted k-nearest neighbor (WkNN) using accuracy, Kappa statistic, Area under the ROC curve (AUC), sensitivity, and specificity. The study examined the impact of binning and skewing predictors, and feature selection on model performance. Binning predictors into at most two classes improved the LR's performance. The performance of the WkNN also improved with binned predictors when the probability of success was at least 0.5. Binning predictors did not affect the SVM's performance while the ANN was better off with numeric predictors. The WkNN performed better with skewed predictors than the competing models given that the probability of success was at most 0.5. Nonetheless, when the probability increased above 0.5, the models performed similarly. In the case of imbalance, the LR and SVM could still produce similar results with fewer predictors, but the performance of the LR fitted on balanced data improved after predictor selection. Predictor selection improved the WkNN's performance as the success class increased. The research also discovered that machine learning models performed better with high-dimensional unbalanced data than the LR; however, the LR outperformed machine learning models once the data was balanced. The LR with logit and probit links functions performed identically.

13:40

Latent Class Joint Model for Longitudinal and Survival Data: an alternative to influence diagnostics for shared parameter joint model

Isaac Singini | *Divison of Biostatistics & Epidemiology, Stellenbosch University, Tygerberg, Cape Town, South Africa.*
Ding-Geng Chen | *College of Health Solutions, Arizona State University, Phoenix, AZ 85004, USA*
Freedom Gumedze | *Statistical Sciences Department, University of Cape Town, Cape Town, South Africa*

Joint models for longitudinal and survival data are a class of models that jointly analyse an outcome of interest repeatedly observed over time along with the associated event times. There are two main classes of these models, namely; shared parameter and latent class joint models. These models are useful in two practical applications; firstly focusing on survival outcome whilst accounting for time varying covariates measured with error and secondly focusing on the longitudinal outcome while controlling for informative censoring.

Interest on the estimation of these joint models has grown in the past two and half decades with minimal effort directed towards developing influence diagnostics.

In this study we compared Cook's statistics for detecting influential subjects to classes identified by the latent class joint model which in effect would classify influential subjects through population heterogeneity. We used simulation studies to evaluate our approach and we only present one scenario in this paper. We then illustrated this approach using data from a multi-centre clinical trial on TB pericarditis. Simulation studies and the motivating dataset confirmed our hypothesis that latent class joint models can be used as an alternative to diagnostics to identify influential subjects in the shared parameter joint models for longitudinal and survival data. This is done by classifying heterogeneous classes using a latent variable. This is especially useful if there's little focus on the association between the bio-marker and risk for and event in which case the longitudinal and event processes are explained by some latent population heterogeneity. Modeling population heterogeneity has practical applications in personalised medicine by predicting the risk for an event adjusted for the longitudinal process.

Poster Abstracts

Determining the impact of changing weather conditions on the lung function of children in South Africa: A doctoral study in applied statistics

Natalie Benschop | *Discipline of Statistics, University of Kwazulu Natal*

Children are more vulnerable to climatic changes than adults, and because of their ongoing organ development, at greater risk for acute and chronic health outcomes as they age. Short-term fluctuations in temperature and humidity, and longer-term increasing heat trends, are likely to compromise respiratory health among children. However, the effect of changing weather patterns on child health outcomes generally remains an understudied topic throughout Africa. This is due (in part) to a lack of sufficient data, as well as interactions that are assumed to exist between environmental exposure variables more broadly. We present an overview of the various data collected, with the aim of investigating the impact of short-term changes in weather conditions, on differing respiratory health outcomes of South African children. We highlight some of the immediate challenges that need be addressed in the modelling of this complex data set, through the development and application of enhanced statistical techniques.

Analysing published data using Meta Analysis and Mendelian Randomisation to determine causal associations.

Grace Carmichael | *Department of Statistics, University of Cape Town*
Jared Tavares | *Department of Statistics, University of Cape Town*
Noëlle van Biljon | *Department of Statistics, University of Cape Town*
Francesca Little | *Department of Statistics, University of Cape Town*

Determining causality in the relationship between two variables is necessary to implement correct treatment and prevention, especially in the health sciences. Conventionally, assessing causal relationships requires a randomised controlled trial (RCT). However, there are situations in which these trials are unethical or impractical to implement, which is problematic for determining causal associations. Mendelian randomisation is a method for determining causality between two variables without needing RCTs. It involves using genetic variants (Single Nucleotide Polymorphisms, or SNPs) as proxies for the exposure of interest and assessing the relationship between these SNPs and the outcome of interest. This use of genetic proxies mitigates the effect of confounding variables on the relationship, allowing the determination of causality.

Here we demonstrate the use of summarised data to determine relationships between obesity as the exposure and asthma as the outcome. The methods implemented were a systematic review followed by a meta-analysis, which uses a three-level mixed effect model to estimate the relationship between the exposure and outcome. The meta-analysis is followed by a Mendelian randomisation where an inverse variance weighted model and a Mendelian randomisation-egger (MR-Egger) regression model are fitted to summarised genetic data obtained from Genome-Wide Association Studies (GWAS). The results from the Mendelian randomisation were similar to the results from the meta-analysis.

A new similarity metric for linear networks

Mila Coetzee | *Department of Statistics, University of Pretoria*
Inger Fabris-Rotelli | *Department of Statistics, University of Pretoria*
Renate Thiede | *Department of Statistics, University of Pretoria*
René Stander | *Department of Statistics, University of Pretoria*

Social mobility networks are becoming increasingly important as both technology and politics continue to promote global connectivity. Understanding how social mobility networks and existing infrastructure interact is of particular interest, informing efficient urban planning and economic resource allocation. Modelling social networks generally presents both computational limitations, by being distribution-based, and application difficulties by neglecting correlation with surrounding infrastructure. This paper therefore proposes a novel method of comparing social mobility data and road networks that both takes spatial context into account and is distribution free. A generic spatial similarity test for spatial data is extended to investigate similarity between two linear networks. The first linear network represents the social mobility data. Nodes are origin and destination locations while edges represent the shortest, and therefore the most likely, path between these locations. The second linear network is a road network located in the same area where the social mobility data is collected. Comparisons across different time periods are done to investigate any temporal aspect of the spatial similarity. The analysis thus provides insight into both the interaction between social mobility and informal roads.

Irregular Local Low Rank Approximation.

Sisipho Hamlomo | *Department of Statistics, Rhodes University*
Marcellin Atemkeng | *Department of Mathematics, Rhodes University*
Jeremy Baxter | *Department of Statistics, Rhodes university*

In the big data regime, many fields currently apply image processing to project a high rank dataset to a low-rank dataset. This is a process of feature extraction that preserves the most relevant information in a given data set. The data is compressed to save memory which reduces the need for high computational resources in postprocessing. A well-known low-rank approximation that

belongs to the linear class is the singular value decomposition (SVD). The SVD technique decomposes the data into separate sets of relevant features, noisy and redundant components. In this work, we aim to project the data matrix into an irregular sampling space of rank r_b data matrices. The extra subscript b indicates that the rank of the data matrix varies across irregular blocks of pixels, which makes the compression and accuracy dependent on the irregular block of pixels in the data matrix. We anticipate that at the same accuracy threshold, our proposed method will be competitive with regular SVD while its compression factor is higher.

Termination versus operation extension for degrading systems

Amy Langston | *Department of Statistics, Rhodes University*
Maxim Finkelstein | *Department of Mathematical Statistics, University of the Free State and Department of Management Science, University of Strathclyde*
Ji Hwan Cha | *Department of Statistics, Ewha Womans University,*

Optimal cost-wise termination of operation for degrading engineering systems is considered that maximizes the expected profit for the infinite horizon. Termination reduces probability of the future failure that usually results in substantial loss. The cases of non-repairable and minimally repairable systems are discussed. The black-box scenario without additional information on stochastic process of degradation and that with additional information are considered. The novel sequential optimal inspection policy is described. It is proved that the termination time can be postponed in an optimal way if the degradation observed at inspection is smaller than the defined level. The developed approach allows for optimal usage of system's "resource." The detailed numerical examples illustrate the findings.

Biplots of Compositional Data from the Tennessee Eastman Process

Jennifer Leigh Liebenberg | *School of Mathematical and Statistical Sciences, North-West University*
Roelof Coetzer | *School of Mathematical and Statistical Sciences, North-West University*
Marike Cockeran | *School of Mathematical and Statistical Sciences, North-West University*
Thobeka Nombebe | *School of Mathematical and Statistical Sciences, North-West University*

Compositional data refer to proportions, or fractions, of a whole. It is important to consider the analysis of compositional data, as it can be found in many fields, from physics to finance, chemistry to colour theory. Some examples of compositional data in these fields are particle size distributions, alloy composition and chemical milling bath composition, colour compositions of paintings, chemical compositions of basalt specimens, as well as household expenditure. Compositional data are subject to certain properties that complicate the analysis thereof. That is, a compositional data set contains a large number of variables which are subject to a unit sum constraint. The sum constraint imposes constraints upon the variance-covariance matrix of the variables and therefore invalidates most standard statistical approaches. Log-ratio transformations of the data simplify the process of data analysis as it remedies the dependency created by the sum constraint. Additionally, principal components analysis (PCA) is often used to reduce the number of variables in the dataset and further simplifies the analysis of compositional data. Performing a PCA gives one access to an arsenal of tools that are invaluable in the analysis of multivariate, and consequently compositional data. One such tool is the PCA biplot which displays the data in reduced (typically two) dimensions. Geometric properties of the biplots can be used to estimate statistical properties of the log-transformed data such as distances between observations, standard deviations of the variables, correlations between the log ratios, and relationships between variables. These biplot properties form the basis of this study. The aim of the study is to give an overview of compositional data and compositional biplots, and to demonstrate biplot properties using simulated Tennessee Eastman process (TEP) data.

Preliminary assessment of resource data for power output predictions of a photovoltaic (PV) system

Aphiwe Magaya | *Department of Statistics, Nelson Mandela University*
Chantelle May Clohessy | *Department of Statistics, Nelson Mandela University*
Warren Brettenny | *Department of Statistics, Nelson Mandela University*
Ernest van Dyk | *Department of Physics, Nelson Mandela University*

This study forms a part of an investigation of machine learning models for the prediction of power output of a photovoltaic (PV) system. The hourly power output (kWh) dataset was collected from a 1MW system installed on the Nelson Mandela University South Campus in Gqeberha for the period of January 2018 to December 2022. Weather data including the following variables, global horizontal radiation (W/m2), diffuse radiation (W/m2), direct radiation (W/m2), temperature (C), wind speed (m/s), wind direction (°), precipitation (mm), air pressure (hPa), and humidity (%) were also collected for the same period. The study will make use of these variables and machine learning models for the prediction of the power output. A comparison of the machine learning regression algorithms namely Artificial Neural Networks, Support Vector Machine, Random Forest, Decision Tree, and k-Nearest-Neighbours, will be done using a cross-validation approach to identify the best-performing model. The preliminary findings of this study will be presented.

A Gamma-Poisson topic model for short text

Jocelyn Mazarura | *Department of Statistics, University of Pretoria*
Alta de Waal | *Department of Statistics, University of Pretoria*
P de Villiers | *Electrical, Electronic and Computer Engineering, University of Pretoria*

Topic modelling is a subfield of natural language processing whose objective is to discover latent topics in large unlabelled corpora. Over the years, short texts, such as tweets and reviews, have become increasingly relevant due to the growing popularity of social media and online shopping. Traditional topic models assume that a document is generated from multiple topics. Whilst this assumption may be acceptable for long texts, such as e-books and news articles, many studies have shown that the one-topic-per-document assumption imposed by mixture models, such as the Dirichlet-multinomial mixture (DMM) model, fits short texts better. Most topic models are constructed under the assumption that documents follow a multinomial distribution. The Poisson distribution is an alternative distribution to describe the probability of count data. It has been successfully applied in text classification, but its application to topic modelling is not well documented, specifically in the context of a generative probabilistic model. The main contributions of this work are a new Gamma-Poisson mixture (GPM) model and a collapsed Gibbs sampler, which enables the model to automatically learn the number of topics contained in the corpus. The results show that the GPM performs better than the DMM at selecting the number of topics in labelled corpora. Furthermore, the GPM produces better topic coherence scores, thus making it a viable option for the challenging task of topic modelling of short text.

An approach to multi-class imbalanced problem in ecology using machine learning

Bonelwa Sidumo | *Department of Statistics, North-West University*

Ecologists collect their data manually by visiting multiple sampling sites. Since there can be multiple species in the multiple sampling sites, manually classifying them can be a daunting task. Much work in literature has focused mostly on statistical methods for classification of single species and very few studies on classification of multiple species. In addition to looking at multiple species, we noted that classification of multiple species result in multi-class imbalanced problem. This study proposes to use machine learning approach to classify multiple species in population ecology. In particular, bagging (random forests (RF) and bagging classification trees (bagCART)) and boosting (boosting classification trees (bootCART), gradient boosting machines (GBM) and adaptive boosting classification trees (AdaBoost)) classifiers were evaluated for their performances on imbalanced multiple fish species dataset. The recall and F1-score performance metrics were used to select the best classifier for the dataset. The bagging classifiers (RF and bagCART) achieved high performances on the imbalanced dataset while the boosting classifiers (bootCART, GBM and AdaBoost) achieved lower performances on the imbalanced dataset. We found that some machine learning classifiers were sensitive to imbalanced dataset hence they require data resampling to improve their performances. After resampling, the bagging classifiers (RF and bagCART) had high performances compared to boosting classifiers (bootCART, GBM and AdaBoost). The strong performances shown by bagging classifiers (RF and bagCART) suggest that they can be used for classifying multiple species in ecological studies.

Dual-stress accelerated life testing models using the generalised Eyring model

Neill Smit | *Centre for Business Mathematics and Informatics, North-West University*
Lizanne Raubenheimer | *Department of Statistics, Rhodes University*

Traditional life testing and obtaining reliability estimates for durable products are usually not feasible due to time constraints. A possible solution is the use of accelerated life tests. When performing accelerated life tests, products are tested under a stress environment that is more severe than their normal operating environment to induce early failures. This accelerated failure data can then be used to extrapolate the life characteristics of the products under their normal operating conditions. A functional relationship between the parameters of the life distribution and the stress variables, known as a time transformation function, is assumed. In this paper, Bayesian accelerated life testing models that incorporate two stress variables are compared. The generalised Eyring model is used as the time transformation function, which incorporates one thermal stress variable and one non-thermal stress variable. The Weibull, Birnbaum-Saunders, and log-normal distributions are used as the life distributions. The models are applied to a real dataset, where different prior settings are considered as part of a sensitivity analysis. Due to the intractable nature of the posterior distributions for these models, Markov chain Monte Carlo methods are employed to generate posterior samples for inference. The deviance information criterion is used to compare the fit of the models and the predictive reliability is calculated.

Index of contributors

Page numbers in bold refer to abstracts where the listed author is the presenting author.