Contents

Welcome 3

General information 4

Presenter guidelines 6

Keynote speakers 7

Workshops 14

SCHEDULE

Programme at a glance 21

Parallel sessions at a glance 22

Workshop days: Monday and Tuesday 23

Wednesday, 25 November 2026 25

Thursday, 26 November 2026 28

Friday, 27 November 2026 31

Poster session 33

ABSTRACTS

Wednesday, 25 November 2026 35

Thursday, 26 November 2026 62

Friday, 27 November 2026 90

Posters 101

Notes 119

Welcome

The School of Statistics and Actuarial Science at the University of the Witwatersrand is delighted to welcome you to the 67th Annual Conference of the South African Statistical Association between 23 and 27 November 2026. The conference brings together statisticians from across South Africa and beyond for an in-person gathering focused on statistical theory, multivariate analysis, Bayesian statistics, biostatistics, statistical machine learning and related areas.

Preceding the main conference, a series of workshops offers hands-on engagement and skills development, while a line-up of national and international speakers provides global perspectives on contemporary statistical research and practice.

We hope you enjoy a meaningful opportunity to network, exchange ideas and advance the statistical community, spanning postgraduate students, early career researchers, established academics and industry practitioners.

Scientific Programme Committee

Prof. Johan Ferreira (chair)
Dr. Justine Nasejje
Dr. Raeesa Docrat
Dr. David Rose
Dr. Salha Mamane
Dr. Jacob Majakwara
Dr. Honest Chipoyera

Local Organising Committee

Ms. Anna Kaduma Gumbie (chair)
Dr. Raeesa Ganey
Ms. Yoko Chhana
Ms. Nothabo Ndebele
Ms. Annapurna Hazra
Dr. Raeesa Docrat
Mr. Miguel Torres

Supported by

UNIVERSITY OF THE WITWATERSRAND, JOHANNESBURG

SACNASP
South African Council for Natural Scientific Professions
Integrity in science

NITheCS
National Institute for Theoretical and Computational Sciences

MaSS
DSTI-NRF CENTRE OF EXCELLENCE IN MATHEMATICAL & STATISTICAL SCIENCES

Macmillan Learning

General information

Venue

The conference is held at the Birchwood Hotel & OR Tambo Conference Centre, 44 Viewpoint Road, Boksburg, 1459; one of Gauteng's premier conference facilities and conveniently located near OR Tambo International Airport.

Registration desk

The registration desk is located in the **Le Grande Centre** on Monday and Tuesday, and the **Atrium Centre** from Wednesday to Friday, and is open at the following times:

Monday (23 November):	08:00 – 12:00; 13:45 - 15:00
Tuesday (24 November):	08:00 – 12:00; 13:45 - 15:00
Wednesday (25 November):	08:00 – 12:00; 13:45 - 15:00
Thursday (26 November):	08:00 – 12:00; 13:45 - 15:00
Friday (27 November):	08:00 – 09:30

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

Tea and meals

Tea and lunch are served in Phomolong/Le Grande Centre, respectively.

Meetings and social functions

Meeting / function	Time	Venue
SASA Executive Meeting	Tue., 24 Nov., 17:00 – 18:30	Boardroom, Atrium Centre
Opening Ceremony	Wed., 25 Nov., 08:00 – 09:30	Baobab, Atrium Centre
HoD Meeting	Wed., 25 Nov., 17:10 – 18:00	Weaver
Poster session	Wed., 25 Nov., 17:30 – 18:45	Baobab, Atrium Centre
Meet and Greet / Cocktail Function	Wed., 25 Nov., 19:00 – 21:00	Phomolong
Young Statistician Event	Wed., 25 Nov., 21:00 – 23:30	Phomolong
SIG Meetings	Thu., 26 Nov., 12:40 – 13:30	Consult with SIG Chair
SASA AGM	Thu., 26 Nov., 17:00 – 18:00	Baobab, Atrium Centre
Gala Dinner	Thu., 26 Nov., 18:30 – 22:00	Serengeti Lapa
Closing Ceremony	Fri., 27 Nov., 12:40 – 13:30	Baobab, Atrium Centre

Contact

For general enquiries during the conference, please contact the organising committee at sasa2026@sastat.org.

Full program online

The full program for the conference can be found online here, or by scanning this code:

Conference proceedings

SASA produces peer-reviewed proceedings from the annual conference. Presenters who wish to submit a contribution for review and publication are invited to send their paper to the Proceedings Editor. Papers should contain original research on theoretical or applied statistics, and only papers presented as oral presentations at SASA 2026 by one of the contributing authors may be submitted. The submission deadline is **18 January 2027**; queries can be directed to the Proceedings Editor at proceedings.editor@sastat.org.

Venue outline



Presenter guidelines

Slot lengths	Slot	Presentation	Questions
Oral presentation	20 minutes	15 minutes	5 minutes
Speed session	10 minutes	7 minutes	3 minutes

Oral presentations

- Each oral slot is 20 minutes: a **15-minute presentation** followed by 5 minutes of questions.
- Please report to your venue at least 10 minutes before the start of the session and ensure that your presentation is loaded on the computer.
- You are not allowed to move your presentation to any other slot.
- Where possible, please bring a laser pointer with you if you need it.

Presentations marked with an asterisk (*) in the programme are participating in the student competition.

Oral presentations in the speed session

Each speed slot is 10 minutes: a **7-minute presentation** followed by 3 minutes of questions. Slides for the whole speed session are loaded on one laptop in advance. Please also follow the other guidelines for oral presentations given above.

Chairpersons

- Please report to the venue **at least 10 minutes before the start** of the session.
- **Ensure that all presentations have been loaded on the computer** before the session starts.
- No changes to the programme are allowed and all presentations must start *exactly* at the time indicated in the programme. Please do not rearrange presentations in the session if a presenter does not show up.
- At the start of the session, please state the theme of the session. For each presentation, state the presenter's name and the title of the presentation.
- For full length oral presentations, each slot is 20 minutes: a 15-minute presentation followed by 5 minutes of questions.
- For speed oral presentations, each slot is 10 minutes: a 7-minute presentation followed by 3 minutes of questions.
- Time warning signs are provided for each venue. Please use these to warn presenters of their remaining time.
- At the end of the session, please thank all the presenters.

Poster presenters

All posters will be viewed on *Wednesday, 25 November 2026 from 17:30 to 18:45* in Baobab, Atrium Centre. Poster numbers are listed in the programme.

- All posters must be A0 in size and in portrait format.
- There will be no printing facilities at the conference venue.
- Please ensure that your poster is set up at least 15 minutes before the start of the poster session. Poster display boards and tacks/drawing pins will be provided.
- Each presenter will be allocated to a specific board. You must be available at your poster for discussion and questions for the full duration of the poster session.
- Please remove your poster at the end of the poster session.

Keynote speakers

SASA 2026 welcomes a line-up of national and international keynote speakers, bringing global and local perspectives on contemporary statistical research and practice. Keynote sessions take place in the Baobab, Atrium Centre.

Keynote talk	Speaker	When	Chair
Decoding the SDGs: A Bayesian perspective	Prof. Kerrie Mengersen	25 November, 09:30	Janet van Niekerk
Calibration Is the Product: What Statistics Does for AI Weather Forecasting in Africa	Prof. Rendani Mbuvha	25 November, 10:50	Raeesa Docrat
Functional Decomposition: Statistical Foundations for Explainable Machine Learning	Prof. Marvin N. Wright	26 November, 08:30	Justine Nasejje
Grading what we can no longer see: Rethinking assessment after GenAI adoption	Dr. Gao Maribe	26 November, 14:00	Michael von Maltitz
Principal Component Analysis Biplots in Symbolic Data Analysis	Prof. Sugnet Lubbe	27 November, 08:00	Raeesa Ganey
Monitoring Traffic Crashes with Bayesian Spatio-Temporal Models: Evidence from Cape Town, South Africa	Prof. Şebnem Er	27 November, 11:40	Inger Fabris-Rotelli

International speakers



Prof. Kerrie Mengersen

Queensland University of Technology, Australia

Distinguished Professor Kerrie Mengersen has spent more than 30 years advancing Bayesian statistics and its application within data science, helping to bridge statistical theory with machine learning and artificial intelligence. A leader in Australia's data science ecosystem, she established one of the country's earliest university-based data science centres and founded the Australian Data Science Network, which connects more than 40 research organisations.

Her research contributions include spatio-temporal modelling, Markov chain Monte Carlo (MCMC) methods, expert prior elicitation, and the integration of sensor and citizen science data, as well as privacy-preserving approaches such as federated learning. These developments have underpinned impactful real-world applications, including the award-winning Australian Cancer Atlas and Virtual Reef Diver, supporting advances in health policy and environmental conservation.

With over 400 publications, a range of national and international honours, and leadership across global scientific communities, Professor Mengersen continues to shape the field of data science. She is also a dedicated mentor, having supervised more than 90 postgraduate researchers.

Decoding the SDGs: A Bayesian perspective

25 November, 09:30, Baobab, Atrium Centre

The United Nations (UN) Sustainable Development Goals are a set of 17 interconnected global goals adopted by the United Nations in 2015 to create a more peaceful and prosperous world by 2030. They are a "universal call to action" to end poverty, protect the planet, and ensure that all people have peace and prosperity. A critical activity common to each of the SDGs is trustworthy acquisition, analysis and interpretation of data – for rigorous monitoring, evidence-based decision-making, transparent communication and demonstrable impact. Statisticians, therefore, play a key role in achieving the SDGs. In this presentation, I will describe some of the small efforts of our team to employ Bayesian modelling and analysis to advance the SDGs. Five case studies will be presented in the context of SDG 3 (Good health and wellbeing), SDG 10 (Reduced inequalities), SDG 6 (Clean water and sanitation), SDG 8 (Decent work and economic growth), SDG 16 (Peace, justice and strong institutions), SDG 13 (Climate action) and SDG 14 (Life below water). The case studies will spotlight our work on developing the Australian Cancer Atlas, high-dimensional anomaly detection, AI to combat modern slavery, predicting terrorism, and protecting the Great Barrier Reef. For each study, I will also touch on challenges in modelling and computation, and our efforts to address to address these issues. This work has been undertaken by a large number of researchers in our Centre for Data Science at QUT. I will acknowledge key contributors in the presentation.



Prof. Marvin N. Wright

Leibniz Institute for Prevention Research and Epidemiology (BIPS), Germany

Marvin N. Wright builds machine learning that doesn't just predict, but explains. His research sits where modern machine learning meets classical statistics, and centers on some of the recurring challenges in health data science: how to make complex models interpretable, how to generate realistic synthetic data, and how to draw valid statistical conclusions from algorithms designed only to predict.

He is the creator of ranger, a widely used random forest implementation, and his work on explainable AI, survival analysis, and generative modeling has appeared in venues including ICML, AISTATS, Statistics in Medicine, and Bioinformatics. A common thread runs through it: pushing machine learning in the health sciences beyond prediction, toward transparency, trust, and valid inference.

Marvin is Professor of Machine Learning in Statistics at the University of Bremen and heads the Department of Statistical Methods in Epidemiology at the Leibniz Institute for Prevention Research and Epidemiology – BIPS, Germany, where he leads an Emmy Noether Research Group funded by the German Research Foundation (DFG).

Functional Decomposition: Statistical Foundations for Explainable Machine Learning

26 November, 08:30, Baobab, Atrium Centre

Machine learning models are increasingly used in epidemiology and medicine, where the goal is rarely prediction alone: we want to know which factors drive an outcome, how, and for whom. In practice, researchers pick from a toolbox of methods such as feature importance, partial dependence and SHAP. These are rarely defined as estimators of well-specified population quantities, their relationships to each other are poorly understood, and they can disagree substantially on the same model and data.

This talk presents functional decomposition as a common language for these methods. Writing a regression or classification function as a sum of main effects and interaction effects of increasing order, and fixing an identification constraint, we obtain partial dependence functions and SHAP values from one and the same object, which unifies global and local explanations. This clarifies which target quantity each method actually estimates, why Shapley values can mislead by conflating main and interaction effects, and how interactions can be quantified explicitly rather than absorbed into single-feature attributions. I also present tree-based estimators for arbitrary components of the decomposition that remain consistent under feature dependence, in contrast to path-dependent TreeSHAP, and reduce complexity from quadratic to linear in the sample size for moderately deep trees.

I illustrate the approach with examples from large population-based cohort studies, and then turn to functional prediction targets. For time-to-event models, the non-additivity of hazard and survival functions invalidates standard additive explanations; separating time-dependent from time-independent effects reveals how much a factor contributes and also when. More generally, we generalize functional decomposition to Hilbert-valued prediction functions, with kernel-based output representations that yield time-specific, time-resolved and time-aggregated explanations.

Local speakers



Prof. Sugnet Lubbe

Stellenbosch University, South Africa

Sugnet Lubbe obtained her PhD in mathematical statistics at Stellenbosch University in 2001. After spending 13 years in industry, she moved to academia in 2009 as Associate Professor in the Department of Statistical Sciences at UCT. In 2017 she moved to the Department of Statistics and Actuarial Science, Stellenbosch University as Professor of Statistics.

Sugnet's research is focused on multivariate data analysis and more specifically on visualisation and biplots. She is the Director of the Centre for Multi-Dimensional Data Visualisation (MuViSU), with 28 research members from four South African universities, eleven international universities, and four members from industry.

Sugnet is an elected member of the International Statistics Institute, has served on the executive committee of the International Association of Statistical Computing, and is chair of the Multivariate Data Analysis Special Interest Group (MDAG). She received a B2 NRF rating in 2023.

Principal Component Analysis Biplots in Symbolic Data Analysis

27 November, 08:00, Baobab, Atrium Centre

Principal Component Analysis (PCA) biplots provide a powerful and elegant framework for low-dimensional visualisation of multivariate data matrices. The observations (rows) are represented as points, typically in a two-dimensional space. The variables (columns) can be represented as calibrated axes.

In this paper, we extend the classical PCA biplot paradigm to the more general framework of Symbolic Data Analysis (SDA). Unlike traditional data analysis, where each observation is described by a single numerical value per variable, SDA accommodates more complex data structures such as intervals and histograms. Similar to PCA, which allows only numerical variables, our focus is on constructing a PCA biplot where variables may be numeric, interval-scaled, or histogram-scaled.

Units projected onto the plane spanned by the first two principal components reflect variation not only in location but also in spread and shape of the distributions within each observation. In parallel with traditional PCA biplots, we construct biplot axes corresponding to the original variables. These axes are calibrated in the original units of measurement, ensuring interpretability and continuity with classical PCA biplots. The resulting display offers a dual perspective: one can simultaneously examine the relative positions and distributional characteristics of units, as well as the contributions and interpretations of variables.



Prof. Rendani Mbuvha

University of the Witwatersrand, South Africa

Rendani Mbuvha is an Associate Professor of Actuarial Science at the University of the Witwatersrand, following previous appointments as an Associate Professor at the University of Manchester and a Google DeepMind Academic Fellow in Machine Learning at Queen Mary University of London. He holds a PhD in Electrical Engineering from the University of Johannesburg, where he was awarded the Google Africa PhD Fellowship in 2019.

His research spans the intersection of machine learning, climate risk, weather forecasting, and actuarial science, specifically targeting decision-making under uncertainty in data-scarce environments. He is the Principal Investigator for FineCast, a Bezos Earth Fund AI for Climate and Nature Grand Challenge project at Wits developing next-generation AI forecasting systems for African National Meteorological Services.

Rendani led the development of the Actuarial Society of South Africa's (ASSA) Climate Index and co-founded Afri-Climate AI, a pan-African research community behind the Google.org supported Forecast4Africa initiative. He is a Fellow of both ASSA and the Institute and Faculty of Actuaries, holding the Chartered Enterprise Risk Actuary (CERA) designation.

Calibration Is the Product: What Statistics Does for AI Weather Forecasting in Africa

25 November, 10:50, Baobab, Atrium Centre

AI weather models such as GraphCast, GenCast and the Functional Generative Network increasingly rival physics-based numerical weather prediction at a fraction of the computational cost. This offers African forecasting centres a route around long-standing infrastructure constraints. But a raw model output is not an operational forecast. This talk argues that the step between the two is fundamentally statistical.

Drawing on the Forecast4Africa and FineCast programmes, I show how Isotonic Distributional Regression (IDR) transforms deterministic predictions and ensemble summaries into predictive distributions. IDR is a nonparametric method built around stochastic ordering constraints. Applying a common calibration framework supports comparisons between AI and physical models, using Continuous Rank Probability Score decompositions, Diebold–Mariano tests and verification against three observational products. Our evaluation finds a median improvement of approximately 5% over ECMWF's IFS across observational products and seasons, with important regional differences. After calibration, deterministic GraphCast can approach the skill of a generative ensemble, although ensemble advantages emerge at longer lead times.

I then show how this statistical framework translates into practice. Work with the Ethiopian Meteorological Institute illustrates how gauge-level calibration produces daily heavy-rainfall probabilities. In South Africa, collaborations with the Agricultural Research Council connect forecast development to agricultural decisions and the evaluation of forecast value.

I close with open problems for statisticians: multivariate calibration, rare extremes, changing forecasting systems and continuous verification. The central argument is that calibration, uncertainty quantification and evaluation are core components of an operational weather forecast product.



Prof. Şebnem Er

University of Cape Town, South Africa

Şebnem Er is an Associate Professor in the Department of Statistical Sciences at the University of Cape Town, South Africa. Her research lies at the intersection of applied multivariate statistics, Bayesian modelling, spatial and spatio-temporal analysis, and data science, with particular interests in transport safety and public sector decision support. Her work often combines methodological development with real-world applications.

Her recent research has focused on Bayesian spatio-temporal models for traffic crash analysis in Cape Town, including work on collision counts at road intersections and citywide monitoring of crash severity trends. More broadly, she is interested in developing statistical methods that support evidence-based policy and planning in complex urban systems. In addition to transport analytics, her academic interests include stratified sampling, spatial statistics, and multivariate analysis applied in many different areas.

At the University of Cape Town, she teaches and convenes undergraduate and postgraduate statistics courses, coordinates the Master's in Data Science degree, and supervises Master's and PhD students. Through both her teaching and research, she aims to strengthen the role of statistics in addressing socially important challenges in South Africa and beyond.

Monitoring Traffic Crashes with Bayesian Spatio-Temporal Models: Evidence from Cape Town, South Africa

27 November, 11:40, Baobab, Atrium Centre

Urban road crashes remain a major public safety challenge, particularly in rapidly growing cities where crash patterns vary across space, time, and severity. Recent literature has increasingly moved beyond simple count models towards spatio-temporal and multivariate Bayesian frameworks. These models can capture dependence across locations, and time, and between severity outcomes. For example, prior work has used multivariate Bayesian models to jointly analyse crash severity categories, while newer studies have emphasized spatio-temporal variation, severity specific patterns, and proactive hotspot monitoring.

Given this background, this talk presents a spatio-temporal Bayesian multinomial mixture model for monitoring monthly crash severity outcomes such as fatal, serious, slight, and no injury crashes across police stations in the City of Cape Town, South Africa, from January 2015 to October 2023. The model jointly estimates spatially varying crash type probabilities, captures temporal evolution through structured and unstructured random effects, enables probabilistic detection of atypical police stations, and identifies police stations exhibiting increasing, decreasing, or stable trends for certain crash types. Using a nine year police station level crash dataset, the results reveal emerging temporal patterns, and areas with increasing fatal and serious crashes, thereby providing an evidence base for data driven road safety planning and intervention.

The talk also situates this work within my broader research on Bayesian modelling of road safety in Cape Town. In related work, published in Applied Spatial Analysis and Policy in 2025, we modelled crash counts at intersections in a particular ward in Cape Town using a Bayesian spatio-temporal zero-inflated Negative Binomial model. That study showed that node degree, traffic signals, major road context, location on Main Road, and minibus taxi (main public transport vehicle for the majority) route presence were associated with increased collision counts. Taken together, the two studies illustrate a complementary modelling strategy for road safety surveillance where the earlier paper focuses on crash counts at intersections in a local area, the present work addresses overall monitoring of crash severity profiles and temporal trends. This progression reflects an important development in the literature, from modelling where crashes occur to also understanding how severity distributions evolve over time and where unusual patterns may signal emerging risk. In a South African urban context, where transport safety resources are limited and interventions must be targeted, such Bayesian models may offer a framework for evidence based monitoring, prioritisation, and policy support.



Dr. Gao Maribe

Momentum Group and Delta Four Two, South Africa

Dr Gao Maribe is a Data Scientist and an enterprise AI leader. He currently serves as Momentum Group's Executive AI Lead and leads Delta Four Two, an independent AI advisory practice focused on responsible, production-scale AI adoption.

Previously, he served as Group Head of Data Management before transitioning into an Executive role for Data and AI at Old Mutual. He also worked as an Advanced Analytics Associate at McKinsey & Company. At the University of Pretoria, he has taught and supervised in statistics, data science and computing. Gao holds a PhD in Mathematical Statistics from the University of the Free State.

His research spans extreme-value theory, large-scale data analytics and applied machine learning. He has contributed to the World Economic Forum's work on the evaluation and governance of AI agents, and has spoken at DataFest, AI in Action, The Open Letter's forum on AI-enabled business models, and actuarial-industry platforms on enterprise implementation, insurance innovation and Africa's AI opportunity.

Grading what we can no longer see: Rethinking assessment after GenAI adoption

26 November, 14:00, Baobab, Atrium Centre

GenAI can now produce essays, analyses, software and mathematical explanations that resemble the work of capable students and practitioners. This has created an immediate academic-integrity problem, but it also reveals a deeper and more interesting problem: the quality of an observed output is no longer, on its own, a reliable measure of the human capability that educational assessment is intended to infer. This talk examines how AI changes the relationship between performance, learning and evidence of competence. It argues that education should not attempt to preserve a world in which answers were scarce. Instead, institutions must distinguish foundational human capability from AI-assisted performance, while protecting learners from cognitive debt: the hidden gap that can arise when polished output improves faster than durable understanding, independent reasoning and the ability to transfer knowledge. The talk proposes a practical assessment architecture that combines unaided performance, AI-assisted production, oral defence, transfer to an unfamiliar problem and critical evaluation of machine-generated results. These complementary evidence channels allow educators to assess what a learner knows, what the learner can accomplish with contemporary tools, and whether the learner can recognise when a plausible answer should not be believed. The conclusion is a particular challenge for scientific domains such as statistics: if AI has altered the data-generating process behind student work, then the construct validity and identifiability of our existing educational inferences might need to be reconsidered.

Workshops

A programme of pre-conference workshops runs across two days, covering interpretable machine learning, AI-native research workflows, climate risk, responsible AI in supervision, teaching portfolios, multivariate data sourcing, biplots, infectious disease modelling, and machine learning in SAS Viya. Workshops take place in **Acacia 1**, **Acacia 3**, and **Weaver** at the **Le Grande Centre**. Please attend the workshop you selected when you registered.

Workshops at a glance

Workshop	Presenter	Time	Venue
Day 1: Monday, 23 November 2026			
Interpretable Machine Learning	Prof. Marvin N. Wright	Morning	Acacia 1
Who Is Doing the Thinking? Responsible AI Use in Teaching, Research and Postgraduate Supervision	StatsNetSA team	Morning	Weaver
AI for Climate Risk	Prof. Rendani Mbuvha	Afternoon	Acacia 1
An AI-Native Research and Development Lifecycle	Dr. Gao Maribe	Afternoon	Acacia 3
Completing Your Teaching Portfolio	Prof. Michael von Maltitz	Afternoon	Weaver
Day 2: Tuesday, 24 November 2026			
Introduction to Bayesian Spatial and Spatio-Temporal Modelling in R	Prof. Kerrie Mengersen	Full day	Acacia 1
Sourcing Multivariate Data (MDAG)	Prof. Şebnem Er, Dr. Arno Otto and Jason Pillay, Dr. Shawn Liebenberg	Morning	Acacia 3
Spatio-Stochastic Modelling of Infectious Diseases (SEPIMOD)	Dr. Raeesa Manjoo-Docrat and facilitators	Morning	Weaver
Biplots: Where We Come From and Where Are We Heading	Prof. Sugnet Lubbe	Afternoon	Acacia 3
Machine Learning in SAS Viya for Learners	Prof. Tanja Verster and Prof. Renette Blignaut	Afternoon	Weaver

Day 1: Monday, 23 November 2026

Interpretable Machine Learning

Prof. Marvin N. Wright *(Leibniz Institute for Prevention Research and Epidemiology (BIPS))*

Acacia 1, morning

Machine learning models often achieve strong predictive performance, but this comes at the cost of transparency: complex models behave as black boxes, making it difficult to understand why they arrive at a given prediction or which features drive their behaviour. Interpretable machine learning (IML) addresses this gap with a toolbox of methods that make black-box models understandable, and this workshop provides a practical, application-oriented introduction to that toolbox.

The workshop covers a broad range of interpretability methods, including:

- Feature effects, such as partial dependence plots and accumulated local effects
- Feature importance measures, including permutation importance
- Local explanation methods for individual predictions
- Shapley values and their use in attributing predictions to input features

Sessions combine short theoretical introductions with hands-on practical exercises in R, allowing participants to apply each method directly and build intuition for when and how to use it.

Prerequisites: A background in statistics is assumed; some basic machine learning experience is helpful. Participants should bring a laptop with R installed (or access to RStudio Server). Basic familiarity with fitting statistical or machine learning models in R is helpful but not required.

Who Is Doing the Thinking? Responsible AI Use in Teaching, Research and Postgraduate Supervision

StatsNetSA team *(StatsNetSA Workshop)*

Weaver, morning

The rapid emergence of Generative AI is transforming teaching, learning, research, and postgraduate supervision within higher education. While AI tools offer opportunities to enhance productivity and learning, they simultaneously challenge traditional approaches to assessment, academic integrity, research training, and the development of critical thinking. These developments have created growing uncertainty among academics and supervisors regarding how AI should be responsibly integrated into higher education practice without undermining intellectual engagement and scholarly ownership.

This half-day workshop, hosted by the StatSNetSA core group, provides a practical and reflective platform for academics, supervisors, and postgraduate students to engage with the evolving role of AI in academia. It introduces an adaptation of an existing guiding rubric for doctoral research, expanded to address the opportunities and challenges of GenAI — moving beyond simplistic detection-or-prohibition approaches towards AI literacy, transparency, verification, ethical engagement, and independent scholarly thinking.

AI for Climate Risk: Statistical Foundations for Forecasting in Data-Scarce Environments

Prof. Rendani Mbuvha *(University of the Witwatersrand)*

Acacia 1, afternoon

Climate change is generating new demands on statistical modelling, particularly in regions like Africa, where historical climate records are sparse, irregular, or entirely absent at the spatial and temporal resolution needed for robust risk assessment. This workshop introduces participants to the statistical and machine learning methods being developed to address this challenge, with a focus on probabilistic forecasting, uncertainty quantification, and climate risk index construction.

Drawing on real-world applications, the session will cover:

- Statistical foundations of probabilistic weather and climate forecasting;
- Methods for uncertainty quantification and decision-making with limited or noisy data;
- Techniques for constructing climate risk indices and early-warning indicators;
- Case studies applying these methods across African National Meteorological Services; and
- Opportunities and open challenges at the intersection of statistics, machine learning, and climate science.

An AI-Native Research and Development Lifecycle for R and Python Statistics Practitioners

Dr. Gao Maribe *(Momentum Group and Delta Four Two)*

Acacia 3, afternoon

Artificial intelligence (AI) is beginning to change all knowledge domains, including statistical research from a sequence of manually intensive activities into an integrated, iterative process spanning problem formulation, mathematical reasoning, simulation, software development, validation and scientific communication. Faster production of code and mathematical text, however, does not necessarily produce more reliable science. Without an explicit research lifecycle, AI can accelerate incorrect assumptions, generate plausible but invalid arguments, conceal implementation errors and create results that are difficult to reproduce. This workshop introduces an AI-native research lifecycle designed specifically for statisticians. Participants will learn how to translate a research question into explicit assumptions, estimands, conjectures and falsifiable claims; use AI coding agents to construct modular simulation studies in R/python; and build tests that distinguish mathematical, statistical, numerical and software failure modes. The workflow emphasises invariants, limiting cases, analytically tractable benchmarks, sensitivity analysis, adversarial counterexamples and independent critique. A substantial mathematical-statistical case study will show how a difficult research problem can be revisited using contemporary agentic methods. The end-to-end process connects code, results and exposition to produce a documented R package (or Python), reproducible experiments, automated figures and tables, and a LaTeX manuscript with a clear claim-to-evidence trail. The aim is not to delegate statistical reasoning to AI, but to create an auditable system in which AI expands the speed and breadth of investigation while the statistician remains responsible for validity, interpretation and proof.

Completing Your Teaching Portfolio

Prof. Michael von Maltitz *(University of the Free State)*

Weaver, afternoon

While the teaching portfolio might not be compulsory at every institution, the momentum is heading that way. This workshop helps academic statisticians improve their teaching and prepare for subsequent promotion applications by assisting participants in completing their teaching portfolios.

The workshop is tailor-made for academic statisticians, presented by an academic statistician — and not an education specialist, with terminology and ideas beyond understanding. The basic structure of a teaching portfolio is provided as part of the session:

- Structure of a teaching portfolio
- The purpose of a teaching portfolio
- How to write up a Teaching & Learning philosophy
- The relationships with research and engaged scholarship (or academic citizenship)

Day 2: Tuesday, 24 November 2026

Introduction to Bayesian Spatial and Spatio-Temporal Modelling in R

Prof. Kerrie Mengersen *(Queensland University of Technology)*

Acacia 1, full day

Spatial data are now pervasive in many fields of science and social science, and are used extensively by businesses, government agencies and community groups. This workshop is intended for beginners who would like to get started in Bayesian modelling of spatial and spatio-temporal data. The workshop will be a blend of theory, practical sessions and case study discussions. Specific topics that will be covered in the workshop include the following:

- Overview of spatial analysis and why it is important
- Basics of spatial and spatio-temporal data
- Overview of Bayesian modelling and computation
- Overview of Bayesian spatial statistical models
- Practical session: Bayesian spatial modelling
- Overview of Bayesian spatio-temporal modelling
- Practical session: Bayesian spatio-temporal modelling
- Visualisation
- Case studies in Bayesian spatial and spatio-temporal analysis

A first-year university-level knowledge of statistics will be assumed, but no knowledge of Bayesian analysis or spatial modelling is required. Practical sessions will focus on a range of examples and will be run as group-based discussions on potential spatial and spatio-temporal models, computational methods, and the interpretation of output. R code for the examples will be provided to participants who wish to run them in their own time after the workshop.

Sourcing Multivariate Data

Prof. Şebnem Er, Dr. Arno Otto and Jason Pillay, Dr. Shawn Liebenberg *(MDAG Workshop)*

Acacia 3, morning

This workshop covers three case studies of multivariate data. Each speaker gives guidance on procuring a particular kind of public multivariate data, followed by applications illustrating the characteristics of that data, equipping participants to source a wide variety of multivariate data for teaching and research.

Topic 1: Sourcing and combining publicly available data in R for evidence-based analysis (Şebnem Er). A practical introduction to sourcing, combining and analysing open statistical and spatial data, using traffic safety as a motivating case study, with attention to the modifiable areal unit problem and to exposure and measurement bias.

Topic 2: Sourcing complex environmental data (Arno Otto and Jason Pillay). Public multivariate environmental repositories, preprocessing strategies that turn irregularities and extreme observations into informative features, and modelling approaches that accommodate heterogeneity, skewness, missing values and outliers.

Topic 3: Sourcing multivariate climate data (Shawn Liebenberg). Working with GSOD and ERA5 reanalysis data, including reproducible downloads, common challenges with large climate datasets, visualisation, and a modelling example.

Spatio-Stochastic Modelling of Infectious Diseases Using Compartmental Models

Dr. Raeesa Manjoo-Docrat and facilitators *(SEPIMOD Workshop)*

Weaver, morning

This half-day workshop, hosted by SEPIMOD at the SASA Conference, focuses on the statistical formulation and analysis of compartmental models for infectious diseases. It introduces deterministic and stochastic compartmental models (e.g. SIR/SEIR) from a statistical perspective, with emphasis on model specification, identifiability, and inference. Participants will explore how stochasticity is incorporated through discrete and continuous-time processes, and how spatial structure can be introduced via metapopulation and hierarchical modelling frameworks.

A central COVID-19 case study demonstrates likelihood- and simulation-based inference, model fitting to surveillance data, short-term forecasting, and uncertainty quantification and sensitivity analysis.

- Statistical formulation of compartmental models
- Stochastic processes for infectious disease dynamics
- Spatial modelling frameworks and dependence structures
- Parameter estimation and model fitting (likelihood-based and simulation-based approaches)
- Case study: COVID-19 modelling and inference
- Prediction, forecasting, and uncertainty quantification
- Sensitivity analysis and model robustness

Facilitators: Jenny Holloway, Nontembeko Dudeni-Tlhone, Nada Abdelatif, Inger Fabris-Rotelli and Miguel Torres.

Biplots: Where We Come From and Where Are We Heading

Prof. Sugnet Lubbe *(Stellenbosch University)*

Acacia 3, afternoon

The Principal Component Analysis biplot was introduced by Gabriel in 1971 as a tool for visualisation of multi-dimensional data in two dimensions. While the underlying linear algebra was elegant, interpreting biplots purely through inner products initially limited their practical use. Gower later revolutionised the field by reframing the biplot as a multivariate scatter plot with calibrated variable axes, transforming it into a powerful visualisation tool. This workshop explores that evolution and its future in two parts.

Part I focuses on the fundamental concepts underlying biplots and the linear algebra associated with the plots — the objectives of a biplot, how matrix decompositions provide a framework for representing observations and variables in a single display, the major classes of biplots, and a range of applications across different disciplines. Part II shifts attention to recent advances in biplot methodology and where the field is heading, presenting modern extensions that adapt the biplot framework to increasingly complex data structures. This section steers away from the underlying mathematics to emphasise practical implementation.

Machine Learning: Building Predictive Models Using Model Studio in SAS Viya for Learners

Prof. Tanja Verster and Prof. Renette Blignaut *(Data Science Interest Group)*

Weaver, afternoon

This workshop provides a hands on, practical introduction to predictive model development, with a strong focus on real world credit risk applications. Using predominantly retail credit examples, participants will gain experience in data preparation, variable selection, and the development and validation of predictive models. The course emphasises the building and comparison of credit models using logistic regression alongside a range of machine learning techniques, including random forests, gradient boosting, and neural networks. All examples are demonstrated using the cloud native SAS Viya for Learners (SAS VFL) platform, with particular emphasis on SAS Viya for Learners Model Studio and its intuitive, visual modelling pipelines. Participants will learn how to upload data, build, train, compare, and evaluate models step by step, turning raw data into actionable insights using an industry leading analytics platform. The workshop also illustrates automated pipelines in SAS Viya for Learners Model Studio, showcasing its AutoML capabilities that automate data preparation, feature engineering, model training, tuning, and comparison within a defined time limit—allowing learners to rapidly jump-start end-to-end data science projects.

Programme at a glance

Time	Mon 23 Nov	Tue 24 Nov
08:00–09:30	Registration (see General information)	Registration (see General information)
09:00–10:30	Workshops	Workshops
10:30–11:00	*Tea*	*Tea*
11:00–12:30	Workshops	Workshops
12:30–13:30	*Lunch*	*Lunch*
13:30–15:00	Workshops	Workshops
15:00–15:30	*Tea*	*Tea*
15:30–17:00	Workshops	Workshops
17:00–18:00		SASA EC Meeting

Time	Wed 25 Nov	Thu 26 Nov	Fri 27 Nov
08:00–09:30	Opening ceremony	Keynote 3 (08:30–09:30)	Keynote 5 (08:00–09:00)
09:30–10:30	Keynote 1	Parallel sessions 4 (09:40–10:40)	Parallel sessions 7 (09:10–10:10)
10:30–10:50	*Tea*	*Tea* (10:40–11:00)	*Tea* (10:10–10:30)
10:50–13:00	Keynote 2 (10:50–11:50) Parallel sessions 1 (12:00–13:00)	Parallel sessions 5 (11:00–12:40)	Parallel sessions 8 (10:30–11:30) Keynote 6 (11:40–12:40)
13:00–14:00	*Lunch*	*Lunch & SIG meetings* (12:40–14:00)	Closing ceremony (12:40–13:30)
14:00–15:20	Parallel sessions 2	Keynote 4 (14:00–15:00)	*Lunch* (13:30–14:30)
15:20–15:40	*Tea*	Parallel sessions 6 (15:00–16:40)	
15:40–17:00	Parallel sessions 3	*Tea* (16:40–17:00)	
17:00–18:00	HoD meeting (17:10–18:00)	SASA AGM	
17:30–18:45	Posters		
18:30 onwards	Meet and Greet (19:00–21:00) and Young Statistician Event (21:00–23:30)	Gala dinner (18:30–22:00)	

Workshops · Keynotes · Parallel sessions and posters · Tea · Lunch · Registration, ceremonies, meetings and social events

Parallel sessions at a glance

Time	Stream	Venue	Chair	Programme	Abstracts
Parallel sessions 1 — Wed 25 Nov					
12:00 – 13:00	**Biostatistics 1**	Weaver	Lwanda Kondlo	p. 25	p. 35
12:00 – 13:00	**Bayesian statistics 1**	Yellowwood	Jess Rees	p. 25	p. 38
12:00 – 13:00	**Spatial statistics 1**	Cedarwood	Jamie-Lee Nel	p. 25	p. 40
12:00 – 13:00	**Biostatistics 2**	Cypress	Justine Nasejje	p. 25	p. 41
Parallel sessions 2 — Wed 25 Nov					
14:00 – 15:20	**Financial statistics 1**	Weaver	Jacob Majakwara	p. 26	p. 43
14:00 – 15:20	**Bayesian statistics 2**	Yellowwood	Gandhi Jafta	p. 26	p. 46
14:00 – 15:20	**Classification and machine learning**	Cedarwood	Humphrey Brydon	p. 26	p. 48
14:00 – 15:20	**Recent advancements in modelling for complex data**	Cypress	Najmeh Nakhaeirad	p. 26	p. 50
Parallel sessions 3 — Wed 25 Nov					
15:40 – 17:00	**Compositional data**	Weaver	Stephan van der Westhuizen	p. 27	p. 52
15:40 – 17:00	**Biostatistics 3**	Yellowwood	Najmeh Nakhaeirad	p. 27	p. 54
15:40 – 17:00	**Multivariate statistics**	Cedarwood	Ruan Buys	p. 27	p. 56
15:40 – 16:40	**Speed session 1**	Cypress	Chantelle Clohessy	p. 27	p. 58
Parallel sessions 4 — Thu 26 Nov					
09:40 – 10:40	**Official and business statistics**	Weaver	Oabona Machete	p. 28	p. 62
09:40 – 10:40	**Financial statistics 2**	Yellowwood	Energy Sono	p. 28	p. 64
09:40 – 10:40	**Environmental statistics**	Cedarwood	Tyriq Emmanuel Pillay	p. 28	p. 66
09:40 – 10:40	**Sport statistics**	Cypress	Paul van Staden	p. 28	p. 68
Parallel sessions 5 — Thu 26 Nov					
11:00 – 12:40	**Spatial statistics 2**	Weaver	Raeesa Manjoo-Docrat	p. 29	p. 69
11:00 – 12:40	**Emerging topics in biostatistics 4**	Yellowwood	Din-Geng Chen	p. 29	p. 73
11:00 – 12:40	**Statistics in education**	Cedarwood	Rene Stander	p. 29	p. 75
11:00 – 12:20	**Speed session 2**	Cypress	Inger Fabris-Rotelli	p. 29	p. 78
Parallel sessions 6 — Thu 26 Nov					
15:00 – 16:20	**Sampling and survey methods and methodology**	Weaver	Thomas Farrar	p. 30	p. 83
15:00 – 16:20	**Theoretical statistics**	Yellowwood	Johan Ferreira	p. 30	p. 85
15:00 – 16:20	**Biostatistics 5**	Cedarwood	Miguel Torres	p. 30	p. 87
Parallel sessions 7 — Fri 27 Nov					
09:10 – 10:10	**Process monitoring**	Weaver	Roelof Coetzer	p. 31	p. 90
09:10 – 10:10	**Forecasting and timeseries**	Yellowwood	Tendai Makoni	p. 31	p. 92
09:10 – 10:10	**Stochastic processes**	Cedarwood	Etienne Pienaar	p. 31	p. 94
Parallel sessions 8 — Fri 27 Nov					
10:30 – 11:30	**GARCH modelling**	Weaver	Lerato Mahlangu	p. 32	p. 95
10:30 – 11:30	**Applied statistics**	Yellowwood	Stefan Britz	p. 32	p. 97
10:30 – 11:30	**Spatial statistics 3**	Cedarwood	Kabelo Mahloromela	p. 32	p. 99

Monday, 23 November 2026 WORKSHOP DAY 1

Time	Acacia 1	Acacia 3	Weaver
08:00 – 09:00	*Registration*		
09:00 – 10:30	**Interpretable Machine Learning** Prof. Marvin N. Wright Session 1		**Who Is Doing the Thinking? Responsible AI Use in Teaching, Research and Postgraduate Supervision** StatsNetSA team *StatsNetSA* Session 1
10:30 – 11:00	*Morning tea*		
11:00 – 12:30	Session 2		Session 2
12:30 – 13:30	*Lunch*		
13:30 – 15:00	**AI for Climate Risk: Statistical Foundations for Forecasting in Data-Scarce Environments** Prof. Rendani Mbuvha Session 1	**An AI-Native Research and Development Lifecycle for R and Python Statistics Practitioners** Dr. Gao Maribe Session 1	**Completing Your Teaching Portfolio** Prof. Michael von Maltitz Session 1
15:00 – 15:30	*Afternoon tea*		
15:30 – 17:00	Session 2	Session 2	Session 2

Tuesday, 24 November 2026 WORKSHOP DAY 2

Time	Acacia 1	Acacia 3	Weaver
08:00 – 09:00	*Registration*		
09:00 – 10:30	**Introduction to Bayesian Spatial and Spatio-Temporal Modelling in R** Prof. Kerrie Mengersen • **full day** Session 1	**Sourcing Multivariate Data** Prof. Şebnem Er, Dr. Arno Otto and Jason Pillay, Dr. Shawn Liebenberg *MDAG* Session 1	**Spatio-Stochastic Modelling of Infectious Diseases Using Compartmental Models** Dr. Raeesa Manjoo-Docrat and facilitators *SEPIMOD* Session 1
10:30 – 11:00	*Morning tea*		
11:00 – 12:30	Session 2	Session 2	Session 2
12:30 – 13:30	*Lunch*		
13:30 – 15:00	Session 3	**Biplots: Where We Come From and Where Are We Heading** Prof. Sugnet Lubbe Session 1	**Machine Learning: Building Predictive Models Using Model Studio in SAS Viya for Learners** Prof. Tanja Verster and Prof. Renette Blignaut *Data Science Interest Group* Session 1
15:00 – 15:30	*Afternoon tea*		
15:30 – 17:00	Session 4	Session 2	Session 2
17:00 – 18:30	*SASA Executive Meeting*		

Wednesday, 25 November 2026 CONFERENCE DAY 1

08:00 – 09:30	**Opening Ceremony** 08:00 Welcome 08:10 President's address 08:25 Awards: best honours project, postgraduate paper competition, Fellowships and Honorary Members, Thought Leader, Sichel Medal, IMSA address	*Baobab, Atrium Centre*
09:30 – 10:30	**Plenary** **Decoding the SDGs: A Bayesian perspective** Prof. Kerrie Mengersen Chair: Janet van Niekerk	*Baobab, Atrium Centre*
10:30 – 10:50	*Morning tea*	*Atrium Centre*
10:50 – 11:50	**Plenary** **Calibration Is the Product: What Statistics Does for AI Weather Forecasting in Africa** Prof. Rendani Mbuvha Chair: Raeesa Docrat	*Baobab, Atrium Centre*

Stream:	**Biostatistics 1**	**Bayesian statistics 1**	**Spatial statistics 1**	**Biostatistics 2**
Venue:	**Weaver**	**Yellowwood**	**Cedarwood**	**Cypress**
Chair:	Lwanda Kondlo	Jess Rees	Jamie-Lee Nel	Justine Nasejje
12:00	**A structural equation modelling framework for integrating spatially misaligned data in antimicrobial resistance research** p. 35 Tshedza Makasela	**Population-level marginal summaries from smooth two-slope mixed models with random changepoints** p. 38 Sean van der Merwe	**Generating spatial linear networks from homogeneous point patterns** p. 40 Renate Thiede	**Conformal prediction for protein structure models** p. 41 Seitebaleng Makgai
12:20	**Modelling multispell birth interval dynamics in Southern Africa** p. 36 Thambeleni Portia Nevhungoni*	**Deep learning–based subject selection for incorporating historical data in clinical trials** p. 39 Akalu Banbeta Tereda	**The complexity of spatially modelling antimicrobial resistance: Integrating clinical, environmental, and household survey data** p. 40 Qondeni Ndlangamandla*	**Beyond the midpoint: Fuzzy imputation for survival estimation with interval-censored data** p. 42 Kumbirai Manungo
12:40	**Comparative evaluation of joint modelling approaches for longitudinal CD4 count and time-to-death among severely ill HIV-positive participants: A randomized controlled clinical trial study** p. 37 Lwando Kondlo	**Bayesian truth inference over hierarchical label spaces: Modelling diagnostic consensus in ICD-11** p. 39 Jess Rees	**Domain selection for spatial modelling in public health** p. 41 Jamie-Lee Nel*	**Scalable and interpretable cause-specific competing risks modeling with neural additive models** p. 42 Justine Nasejje
13:00 – 14:00	*Lunch*			*Phomolong*

PROGRAMME

Stream:	Financial statistics 1	Bayesian statistics 2	Classification and machine learning	Recent advancements in modelling for complex data
Venue:	**Weaver**	**Yellowwood**	**Cedarwood**	**Cypress**
Chair:	Jacob Majakwara	Gandhi Jafta	Humphrey Brydon	Najmeh Nakhaeirad
14:00	**Exploring extreme risk in selected cryptocurrencies using APARCH-EVT and APARCH-extended Burr XII models** p. 43 Thabani Ndlovu	**Approximate Bayesian computation for parameter estimation in generalised Eyring accelerated models** p. 46 Neill Smit	**A self-supervised transformer and perceptron framework for incomplete multivariate time series in precision medicine** p. 48 Jeanette Maria Louw	**A circular–linear regression model for interval-valued circular data** p. 50 Najmeh Nakhaeirad
14:20	**Bayesian discrete-time survival analysis of loan default: Class-imbalance solutions in South African banks** p. 44 John Masekoameng	**Bayesian analysis for k-out-of-n failure time distributions in dependent environments** p. 46 Lizanne Raubenheimer	**Optimising the classification boundary with imbalanced data** p. 49 Humphrey Brydon	**New flexible methods for circular distributional regression** p. 50 Luke Smythe
14:40	**Assessment of credit distributional model risk under parametric and non-parametric bootstrap methods** p. 45 Modisane Seitshiro	**B-FAIR: Bayesian spatial fairness** p. 47 Gandhi Jafta	**Feature space engineering for machine learning on bioinformatics data: A CNS drug penetrance case study** p. 48 Tshemollo Rapolai*	**Semiparametric mixture of nonparametric regressions** p. 51 Kagisho Nkadimeng
15:00	**Unifying cure rate models in credit risk: The Box–Cox transformation approach** p. 45 Jacob Majakwara	**Bayesian negative-binomial neural networks for climate-informed malaria prediction: A comparative evaluation of statistical and machine learning models** p. 47 Makwelantle Asnath Sehlabana	**Contaminated count regression models** p. 49 Arno Otto	**INLA goes full circle** p. 51 Janet Van Niekerk
15:20 – 15:40	*Afternoon tea*			*Weaver and Atrium Centre*

Wednesday, 25 November 2026 CONFERENCE DAY 1, LATE AFTERNOON

Stream:	Compositional data	Biostatistics 3	Multivariate statistics	Speed session 1
Venue:	**Weaver**	**Yellowwood**	**Cedarwood**	**Cypress**
Chair:	Stephan van der Westhuizen	Najmeh Nakhaeirad	Ruan Buys	Chantelle Clohessy
15:40	**Transformation-free random forest modelling of compositional data** p. 52 Stephan van der Westhuizen	**Joint latent class modelling of time-to-event and discrete longitudinal data** p. 54 Alizandre Klynsmith*	**Dynamic visualisation: A bridge between multivariate data tours and GPAbin biplots** p. 56 Johané Nienkemper-Swanepoel	15:40 **Anomaly identification and portfolio construction in equity markets: An unsupervised machine learning approach** p. 58 Adrian Crisp 15:50 **Traffic data availability accessibility** p. 58 Diamando Katsaitis*
16:00	**A likelihood-based approach to compositional random forests** p. 53 Khaled Masoumifard	**Deep learning enhanced non-linear repeated measurement modelling in latent class joint models of time-to-event and longitudinal data with application to psychometric scores in Alzheimer's patients** p. 55 Tristan Harris	**Tracking changing rainfall regimes through animated biplots** p. 57 Raeesa Ganey	16:00 **Spatial dependency modelling of disjoint spatial areas: A SAPRIN urban node analysis** p. 59 Ephent Selahle 16:10 **Traffic simulation on spatial linear networks** p. 59 Dylan Dabbadie*
16:20	**Fault detection and diagnostic analysis in multivariate compositional data** p. 53 Mduduzi Maphosa	**Tail-aware latent subgrouping for survival: Identifying patient subgroups by extreme-event risk rather than by median prognosis** p. 55 Brenda Mac'Oduol	**Interpreting evidential deep learning through gradient-guided biplots** p. 57 Ruan Buys	16:20 **Spatial modelling of wastewater data: The case of Ekurhuleni** p. 60 Abongile Nombona 16:30 **Comparing three CNN architectures: Baseline compound scaling and residual learning for photovoltaic fault classification** p. 60 Nico Bezuidenhout
16:40	**A Contaminated Model for Overdispersed Multinomial Microbiome Count Data** p. 54 Ockert van Heerden	**Model-Free Screening for Treatment-Interactive Variables with Multivariate Responses in High-Dimensional Data** p. 56 Shiferaw Bizuayehu		16:40 **A contaminated Kumaraswamy regression model for responses on the unit interval** p. 61 Ronewa Mathuba

17:10 – 18:00	**HoD Meeting**	*Boardroom, Atrium Centre*
17:30 – 18:45	**Poster presentations** (27 posters, see p. 33)	*Baobab, Atrium Centre*
19:00 – 21:00	**Meet and Greet / Cocktail Function**	*Phomolong*
21:00 – 23:30	**Young Statistician Event**	*Phomolong*

PROGRAMME

Thursday, 26 November 2026 CONFERENCE DAY 2

08:30 – 09:30	**Plenary** **Functional Decomposition: Statistical Foundations for Explainable Machine Learning** Prof. Marvin N. Wright Chair: Justine Nasejje			*Baobab, Atrium Centre*
Stream:	**Official and business statistics**	**Financial statistics 2**	**Environmental statistics**	**Sport statistics**
Venue:	**Weaver**	**Yellowwood**	**Cedarwood**	**Cypress**
Chair:	Oabona Machete	Energy Sono	Tyriq Emmanuel Pillay	Paul van Staden
09:40	**Optimizing household energy production, distribution, and consumption for sustainability: A data-driven approach to renewable energy integration (net metering), grid optimization, and energy demand forecasting using time-series analysis and machine-learning models** p. 62 Hausitoe Nare	**Redundancy and synergy in an emerging equity market: A mutual information and partial information decomposition approach to the JSE40** p. 64 Palesa Kale	**Modelling seasonal compound extreme risk using vine copula towards enhancement of renewable energy production** p. 66 Maashele Kholofelo Metwane	**A mathematical model to select medley relay teams in competitive swimming** p. 68 Paul van Staden
10:00	**Women's labour-force participation in Botswana: A multivariable analysis of the 2022 Population and Housing Census** p. 63 Oabona Machete	**A regime-dependent structural-deep framework for probabilistic financial forecasting: Formulation and analysis** p. 64 Energy Sonono	**A loggamma generalised linear model for NO2 emissions data from South Africa's Eskom's coal-fired power stations when the data are non-normal and the variance is non-constant** p. 67 Mpendulo Mamba	**An integer linear programming approach to team selection for the mixed medley relay** p. 68 David Forbes
10:20	**Assessing the effect of sample size on supervised and voting ensemble learning models for binary classification of hiring decisions** p. 63 Tshegofatso Botlhoko	**Evaluating VaR for stock indices using a contaminated framework** p. 65 Lehlonoholo Mphanya*	**Performance of static and dynamic occupancy models** p. 67 Tyriq Emmanuel Pillay*	**Sports analytics: Defining a performance measure for a cricket batsman** p. 68 Justin Friend
10:40 – 11:00	*Morning tea*			*Atrium Centre*

PROGRAMME

Thursday, 26 November 2026 CONFERENCE DAY 2 (CONTINUED)

Stream:	Spatial statistics 2	Emerging topics in biostatistics 4	Statistics in education	Speed session 2
Venue:	**Weaver**	**Yellowwood**	**Cedarwood**	**Cypress**
Chair:	Raeesa Manjoo-Docrat	Din-Geng Chen	Rene Stander	Inger Fabris-Rotelli
11:00	**Assessing the impact of TB vaccination strategies in South Africa: A stochastic spatial compartmental modelling approach** p. 69 Raeesa Manjoo-Docrat	**ML-validated tensor RPCA for high-dimensional medical image data analysis** p. 73 Ding-Geng Chen	**Early prediction of student success for a South African higher education institution** p. 75 Tayla Wannenberg*	**11:00 Spatial linear network regression of crime data** p. 78 Karabo Matjila **11:10 Spatial variable selection in mixed graphical models: Antimicrobial resistance in KwaZulu-Natal** p. 79 Marquis Pretorius
11:20	**The impact of spatial connectivity structures on model fit, inference, and data integration in spatial epidemiology and wastewater-based epidemiology: A systematic review** p. 70 Claris Siyamayambo	**High-dimensional likelihood ratio tests for equality of means, outliers and profile analysis applied to the analysis of TB data** p. 73 Carlos Coelho	**Efficacy of a student-led teaching intervention in mathematical statistics, framed by Fink's Taxonomy of Significant Learning** p. 76 Morné Sjölander	**11:20 Modelling spatially misaligned wastewater data in epidemiology** p. 79 Esther Mazarura **11:30 Investigating risk prediction models for intermittent disease progression using machine learning** p. 80 Thlabane Mogashoa
11:40	**Spatial Bayesian clustering of voting patterns** p. 71 Nontembeko Dudeni-Tlhone	**Bayesian assurance for diagnostic accuracy studies: An application to the SAHeart dataset** p. 74 Lizelle Venter	**From AI use to intellectual independence: Rethinking PhD supervision in the generative AI era** p. 76 Danielle Roberts	**11:40 Spatio-temporal modelling of extreme precipitation using INLA** p. 80 Akani Msimeki* **11:50 Non-linear association joint models for clinical studies** p. 81 Sable Walters-Somers*
12:00	**Extending cross-type K-functions for polygonal and linear features: Applications to natural hydrogen in Mpumalanga, South Africa** p. 72 Calvin Botha*	**A Bayesian joint framework for matrix-valued mediation: Application to breast cancer** p. 74 Kasahun Takele Geneti	**Empowering learners to thrive in a data-driven future** p. 77 Delia North	**12:00 Exploring an enriched innovation intensity INAR(1) model for count data** p. 81 Muano Patrick Nemasea* **12:10 Virological failure and subsequent care outcomes among people living with HIV on a dolutegravir-based regimen in South Africa: A retrospective cohort study** p. 82 Mbali Precious Nkosi*
12:20	**The fuzzy lattice catchment areas for diarrhoea mapping** p. 72 Amy Pieters	**Testing before the plateau in mixture cure models** p. 75 James Allison	**Telling a better story about statistics: Tackling anxiety amongst business students** p. 77 Rene Stander	
12:40 – 14:00	*Lunch & SIG meetings*			*Phomolong*

PROGRAMME

Thursday, 26 November 2026 CONFERENCE DAY 2, AFTERNOON

14:00 – 15:00	**Plenary** **Grading what we can no longer see: Rethinking assessment after GenAI adoption** Dr. Gao Maribe Chair: Michael von Maltitz		*Baobab, Atrium Centre*
Stream:	**Sampling and survey methods and methodology**	**Theoretical statistics**	**Biostatistics 5**
Venue:	**Weaver**	**Yellowwood**	**Cedarwood**
Chair:	Thomas Farrar	Johan Ferreira	Miguel Torres
15:10	**Multiple inverse sampling without replacement for post-stratification** p. 83 Thomas Farrar	**A flexible cosine-perturbed wrapped Lindley model** p. 85 Delene van Wyk-de Ridder	**Modelling seasonality of stroke types and associated risk factors in South Africa using a Poisson quantile regression approach: Hospital-based study 2014-2018** p. 87 Lyness Matizirofa
15:30	**Multi-objective stratified sampling: A GA for optimal allocation using the survey design effect** p. 83 Georgi Borros	**Bayesian parametric modelling of credit risk using the three-parameter log-logistic distribution** p. 85 Nthabiseng Khabele	**Competing risks vs. censoring: Methodological insights from RR-TB trials** p. 87 Lateef Amusa
15:50	**Transforming spatial data onto lattices to assess pathogen surveillance systems** p. 84 Ciara Klinkert	**Parameter estimation techniques for the hybrid Weibull–exponentiated Rayleigh distribution** p. 86 Tolulope Adeniji	**Statistical modelling of antimicrobial resistance using wastewater based epidemiology** p. 88 Miguel Torres*
16:10	**Developing data-driven mental health risk indices using machine learning: A comparison with traditional index construction methods** p. 84 Sbongiseni Mthethwa	**Moment-weighting to improve the power of empirical likelihood ratio-based tests for normality under symmetric short-tailed alternatives** p. 86 Lwando Dlembula	**An automated analysis of a novel flow cytometry assay for TB vaccines** p. 89 Sian Wood*
16:30 – 17:00	*Afternoon tea*		*Atrium Centre*
17:00 – 18:00	**SASA AGM**		*Baobab, Atrium Centre*
18:30 – 22:00	**Gala Dinner**		*Serengeti Lapa*

PROGRAMME

Friday, 27 November 2026 CONFERENCE DAY 3

08:00 – 09:00	**Plenary** **Principal Component Analysis Biplots in Symbolic Data Analysis** Prof. Sugnet Lubbe Chair: Raeesa Ganey		*Baobab, Atrium Centre*
Stream:	**Process monitoring**	**Forecasting and timeseries**	**Stochastic processes**
Venue:	**Weaver**	**Yellowwood**	**Cedarwood**
Chair:	Roelof Coetzer	Tendai Makoni	Etienne Pienaar
09:10	**A distribution-free method for anomaly detection in industrial processes** p. 90 Roelof Coetzer	**Regime-aware volatility forecasting of South African gold sales: A comparative analysis of Markov-switching and LSTM models** p. 92 Tendai Makoni	**Self-normalised sequential monitoring and break estimation for panel data** p. 94 Charl Pretorius
09:30	**Refining wastewater surveillance through regression kriging: A geostatistical approach to improve spatial prediction and public health applications** p. 91 Odey Mofokeng	**Analysing the efficiency of deseasonalisation methods in removing seasonal bias in outliers** p. 93 Rofhiwa Netshiomvani*	**Assessing the impact of screening sensitivity and follow-up duration on parameter identifiability in a five-state hidden continuous-time Markov model** p. 94 Tshepo Maja
09:50	**Novel multivariate multiple regression synthetic EWMA control chart for monitoring general linear profiles** p. 91 Maonatlala Thanwane	**The path to the 2030 SDG target: Statistical forecasting of maternal mortality in South Africa using Box-Jenkins time series analysis** p. 93 Gomolemo Rakale*	**Evaluating first passage for Brownian motion with jumps** p. 95 Etienne Pienaar
10:10 – 10:30	*Morning tea*		*Atrium Centre*

PROGRAMME

Friday, 27 November 2026 CONFERENCE DAY 3 (CONTINUED)

Stream:	GARCH modelling	Applied statistics	Spatial statistics 3
Venue:	**Weaver**	**Yellowwood**	**Cedarwood**
Chair:	Lerato Mahlangu	Stefan Britz	Kabelo Mahloromela
10:30	**Stochastic aspects of market risk under different Basel Accord regimes** p. 95 Lerato Mahlangu	**Compression-based distance measures in text classification** p. 97 Keara-Linn Stapelberg	**Exploratory investigation of Discrete Pulse Transform features for roof-covering-material classification** p. 99 Mark de Lancey*
10:50	**Dynamic prices and volatility linkages across sectoral stock markets in South Africa: A cointegration and GARCH- based approach** p. 96 Pepenene Mofolo	**Structural connectivity in the African air transport network** p. 97 Stefan Britz	**Defining homogeneity for spatial linear networks** p. 99 Arthur Antonio
11:10	**Volatility and dependence in COVID 19 mortality across South African epicenters: A bivariate copula based NB INGARCH approach** p. 96 Thembhani Hlayisani Chavalala	**Intervention modelling of time series tourist arrivals data from the non-SADC African and unspecified group from July 2010 to February 2024: In the context of the COVID-19 pandemic** p. 98 Musara Chipumuro	**Quantifying local anisotropy in point pattern analysis** p. 100 Kabelo Mahloromela
11:40 – 12:40	**Plenary** **Monitoring Traffic Crashes with Bayesian Spatio-Temporal Models: Evidence from Cape Town, South Africa** Prof. Şebnem Er Chair: Inger Fabris-Rotelli		*Baobab, Atrium Centre*
12:40 – 13:30	**Closing Ceremony**		*Baobab, Atrium Centre*
13:30 – 14:30	*Lunch*		*Phomolong*

PROGRAMME

Poster session

Wednesday, 25 November 2026, 17:30 – 18:45 | Atrium Centre. Presenters stand at their boards throughout the session. Each entry below links to its abstract.

Board	Poster	Presenter	Page
P01	The effect of including direct-travel neighbourhood matrices in hierarchical Bayesian models for disease mapping	Johannes Moller*	101
P02	Modelling extreme maximum temperatures using block maxima and peaks-over-threshold approaches	Anna Seimela	101
P03	Comparative modelling of tail risk in global equity markets: Evidence from heavy-tailed distributions and back-testing	Sthembiso Dlamini*	102
P04	Investigating the strengths and weaknesses of the SARIMAX intervention model in time series intervention analysis	Thabiso Masena	103
P05	Through-the-cycle to point-in-time probability of default conversion under IFRS 9: Integrating macroeconomic variables into credit risk modelling	Hope Alum	104
P06	A cointegration-GARCH–EVT–copula framework for modelling tail risk in equity markets	Moses Mokhatholane	105
P07	Analysis of statistical, deep learning and hybrid models for foreign exchange rates	Nokuthula Mokoena	106
P08	AI techniques for risk analysis in banking sector: A review and perspective with emphasis on credit risk modelling	Themba Khohlakala	107
P09	Empowering potential of the smartphone in South Africa	Fallo Happy Khanye	108
P10	Human activity recognition: Personalisation through user calibration	Thutego Potsane*	108
P11	Comparison of convolutional neural network architectures for photovoltaic fault classification using infrared imagery	Justin Marshall*	109
P12	Bayesian modelling of extreme rainfall in flood-prone coastal and escarpment regions of KwaZulu-Natal	Felicia Sethosa	109
P13	Clustering daily solar irradiance patterns	Wehan Smith	110
P14	Forecasting of daily electricity demand in South Africa using machine learning models	Hodisane Clement Mamaila	110
P15	Modelling relative humidity with upper boundary observations: A comparison of alternative beta distributions	Mbikazi Ketelo	111
P16	Classification of statistical methodologies in South African medical literature: A review of SAMJ and SAJCC	Kutloano Mavovana*	111
P17	Robust calibration-based imputation scheme and estimators for finite population mean	Maria Lekganyane	112
P18	Comparison of Cox proportional hazards and Fine–Gray competing risk models for loss to follow-up among tuberculosis patients in Kenya	Nelson Ouma*	112
P19	CRISP-NAM for survival analysis under competing risks using risk set sampling	Michael Falk*	113
P20	From risk factors to risk trajectories: Integrating survival analysis, machine learning and explainable AI for first exposure to intimate partner violence in South Africa	Tshiamo Medupe*	113
P21	Modelling and forecasting of daily gold prices using an ARIMA–GARCH hybrid model	Happy Maluleke	114
P22	Dynamic prediction of disease progression using longitudinal biomarkers: Joint modeling and landmarking	Melkamu Molla Ferede	114

Board	Poster	Presenter	Page
P23	**Factors associated with consistent condom use among sexually active HIV-negative PrEP users in South Africa: Results from the 2022 South African HIV household survey**	Goitseone Maseko	115
P24	**Assessment and comparison of error metrics used for photovoltaic energy forecasting**	Tatenda Nyoni*	115
P25	**Prevalence and risk factors associated with HIV infection among pregnant antenatal care attendees in Limpopo Province, South Africa**	Donald Tshabalala	116
P26	**Calibrating Basel III capital adequacy for South Africa: A vector error correction and copula-based analysis of the trade-off between bank safety and lending capacity**	Keneuoe Rachel Rasupu	116
P27	**Comparative analysis of the performance of the standard SIR model across COVID-19 waves in South Africa**	Tolulope Adeniji	117
P28	**Evaluating growth trajectories in HIV-exposed uninfected children using a calibrated simulated HIV-unexposed comparator population**	Mthobisi Zondi	117
P29	**Assessing copula symmetry and tail dependence in multi-source temperature data from South Africa**	Vusi Masingi*	118

ABSTRACTS

Wednesday, 25 November 2026

Abstracts are listed in programme order: by day, then by session, then by time. Posters follow at the end. The presenting author is underlined. Each abstract links back to its place in the programme.

Biostatistics 1

Wednesday, 25 November 2026, 12:00 – 13:00 · Weaver · Chair: Lwanda Kondlo

A structural equation modelling framework for integrating spatially misaligned data in antimicrobial resistance research

Tshedza Makasela, Kabelo Mahloromela, Inger Fabris-Rotelli, Rene Stander

University of Pretoria

Wednesday 25 November, 12:00 – 12:20 · Weaver · Biostatistics 1 ↑ programme p. 25

Antimicrobial resistance (AMR) is a growing global public health concern that threatens the effective prevention and treatment of infectious diseases. Monitoring AMR requires the integration of multiple data like clinical records, population surveys and wastewater surveillance to have a better understanding of resistance trends across populations over time. However, these datasets are collected at different spatial scales and locations, resulting in misalignment that makes it difficult for statistical analysis.

This study will propose a framework for integrating spatially misaligned data. The aligned data will then be analysed using a Structural Equation Model (SEM) to model the relationships between observed variables and latent variables associated with AMR processes. As part of future work, this framework could then be extended to a Spatio-Temporal Structural Equation Model to account for both spatial and temporal dependence.

Keywords: Antimicrobial Resistance (AMR); Spatially Misaligned Data; Structural Equation Modelling; Spatial Data Integration; Spatial Dependence

Modelling multispell birth interval dynamics in Southern Africa

Thambeleni Portia Nevhungoni, Thslidzi Mulaudzi, Alphonce Bere

South African Medical Research Council; University Of Venda

Wednesday 25 November, 12:20 – 12:40 · Weaver · Biostatistics 1 ↑ programme p. 25

Birth intervals, defined as the time between successive live births, are closely linked to maternal and child health outcomes, particularly in low- and middle-income settings. Existing studies often rely on cross-sectional approaches or conventional time-to-event models that treat successive birth intervals as independent, fail to account for dependence among repeated intervals contributed by the same woman, or are formulated in continuous time despite birth intervals being measured on a discrete monthly scale. In addition, conventional approaches may not adequately capture variation in birth interval dynamics across geographic settings and time. A discrete-time multiple-spell frailty framework provides a natural approach for modelling these recurrent events while accounting for within-woman dependence and allowing flexible specification of the hazard across successive birth spells. This study aims to employ discrete-time multiple-spell survival modelling frameworks for recurrent birth events in Southern Africa. A subject-specific frailty approach will be employed to account for within-woman dependence, alongside a complementary marginal generalized estimating equation (GEE) approach to obtain population-averaged covariate effects. Flexible hazard specifications and covariate effects will be allowed across successive birth spells, with spatial dependence and cohort-based temporal effects also considered using nationally representative data from South Africa, including the 2016 Demographic and Health Survey (DHS) and wave 1 to 5 of the National Income Dynamics Study (NIDS), together with data from selected Southern African countries. By addressing timing, recurrence, dependence, and contextual variation through complementary modelling approaches, the study seeks to provide a coherent and flexible approach to understanding birth interval dynamics in Southern Africa.

Keywords: Birth intervals; Recurrent events; Discrete-time survival analysis; Multiple-spell survival models; Frailty models; Generalized estimating equations; Spatial statistics; Southern Africa

Comparative evaluation of joint modelling approaches for longitudinal CD4 count and time-to-death among severely ill HIV-positive participants: A randomized controlled clinical trial study

Lwando Kondlo

University of Fort Hare

Wednesday 25 November, 12:40 – 13:00 · Weaver · Biostatistics 1 ↑ programme p. 25

Background: Longitudinal CD4 count is an important indicator of HIV disease progression, while death is a key clinical time-to-event outcome. In severely ill HIV-positive participants, changes in CD4 count over time are likely to be associated with the risk of death. Analysing CD4 measurements and time-to-death separately may therefore not fully capture the relationship between the two processes. Joint modelling provides a framework for analysing longitudinal and time-to-event outcomes simultaneously. However, different joint modelling approaches may vary in their ability to describe patient trajectories and predict mortality.

Methods: This study uses data from 1,771 treatment-naïve, severely ill HIV-positive participants enrolled in the PHIDISA II randomized controlled trial. Participants were recruited from six research sites, comprising three urban and three rural sites. CD4 counts were collected at approximately six-month intervals, and participants were followed for mortality. Three joint modelling approaches were evaluated: a shared-parameter joint model, a joint latent class model, and a Bayesian joint model. The models were used to characterise individual CD4 trajectories and their relationship with time-to-death. Model performance was assessed using model fit, discrimination, calibration, predictive accuracy, and robustness. Dynamic prediction was also used to estimate mortality risk based on participants' accumulated CD4 measurements.

Results: Participants had a mean age of 35.4 years, 68% were male, and the median baseline CD4 count was 105 cells/mm^3 (IQR: 41–157). Previous analyses of PHIDISA II demonstrated differences in mortality and AIDS-event outcomes across study sites. The present analysis evaluates the association between longitudinal CD4 trajectories and time-to-death and compares the predictive performance of the three joint modelling approaches.

Conclusion: This study provides a comparative assessment of alternative joint modelling approaches for longitudinal CD4 count and time-to-death among severely ill HIV-positive participants. The findings will highlight the strengths and limitations of each approach and inform the selection of appropriate methods for dynamic mortality prediction and the analysis of longitudinal and time-to-event outcomes in resource-limited clinical research settings.

Keywords: Joint modelling; CD4 count; HIV; mortality; longitudinal data; time-to-death

Bayesian statistics 1

Wednesday, 25 November 2026, 12:00 – 13:00 · Yellowwood · Chair: Jess Rees

Population-level marginal summaries from smooth two-slope mixed models with random changepoints

Divan A. Burger, Sean van der Merwe, Emmanuel Lesaffre

Cytel Inc., Waltham, MA, USA; Department of Mathematical Statistics and Actuarial Science, University of the Free State, Bloemfontein, Free State, South Africa; I-BioStat, KU Leuven, Leuven, Belgium

Wednesday 25 November, 12:00 – 12:20 · Yellowwood · Bayesian statistics 1 ↑programme p. 25

Two-slope mixed-effects models with subject-specific changepoints are widely used to describe biphasic longitudinal trajectories. Treatment-arm summaries are often reported using a single "typical subject" curve obtained by plugging mean parameter values into the subject-specific model. This can be misleading when the changepoint timing varies across individuals, because averaging subject-specific curves is not the same as evaluating a single curve at average parameter values. A separate but related issue is that individual trajectories may not change slope through a literal kink; they may bend gradually, and the apparent sharpness of this bend may differ across subjects. We propose population-average estimands for two-slope mixed models with random changepoints and smooth subject-level transitions. These estimands target the treatment-arm mean trajectory and treatment-arm instantaneous slope, obtained by averaging over fitted heterogeneity in changepoint timing and transition sharpness. We use a baseline-anchored normal-convolution soft hinge, which preserves the broken stick interpretation, includes the abrupt changepoint model as a limiting case, and allows analytic evaluation of the changepoint-marginalized kernels under a truncated normal changepoint distribution for fixed transition sharpness, followed by deterministic one-dimensional quadrature over the fitted transition sharpness distribution. In an application to HIV-1 RNA data from AIDS Clinical Trials Group Protocol 315, with censoring and heavy-tailed residuals, the marginal summaries reveal a more gradual population-level transition than conventional plug-in summaries; the marginal slope remains more negative for longer, and the marginal mean continues to decline after the plug-in curve has begun to flatten. These results illustrate the main reporting implication: in random changepoint models, representative subject curves should not be treated as population-average trajectories or slopes.

Keywords: Random changepoint; Two-slope mixed model; Population-average estimand; Smooth hinge; HIV-1 RNA

Deep learning–based subject selection for incorporating historical data in clinical trials

Akalu Banbeta Tereda, Legesse Kassa Debusho

Department of Statistics, School of Science, College of Science, Engineering and Technology, University of South Africa

Wednesday 25 November, 12:20 – 12:40 · Yellowwood · Bayesian statistics 1 ↑programme p. 25

Background: Inclusion of historical data may reduce the necessary sample size of a current trial and/or improve precision and statistical power. Bayesian dynamic borrowing methods, such as the meta-analytic and power prior approaches, are used to borrow information from historical data. However, these benefits require that the historical and current data be similar. Conventional criteria for historical trial selection, such as Pocock's criteria, assess comparability between trials. This study examines historical subject selection for information borrowing using an autoencoder, a deep learning-based approach.

Methods: A simulation study was performed using a Gaussian outcome with covariates. Motivated by previous works showing the risks of ignoring covariates, the borrowing methods were evaluated without covariates, both before and after subject selection. The approach was further illustrated based on an analysis of Clinical Opiate Withdrawal Scale (COWS) scores after three days of detoxification in patients with opioid dependence.

Results: When covariate distributions differed across studies, naive pooling and the dynamic borrowing methods showed inflated type I error rates. After subject selection, the dynamic borrowing methods yielded type I error rates closer to 5% level, whereas naive pooling continued to produce inflated type I error rates. After subject selection, the modified power prior that assumes similarity of the power parameters generally provided the highest statistical power.

Conclusion: Neural network–based subject selection improved comparability between historical and current subjects and enhanced the performance of dynamic borrowing methods. The results suggest that subject selection can complement the robustness provided by dynamic borrowing approaches.

Keywords: Bayesian inference; borrowing methods; deep learning; subject selection

Bayesian truth inference over hierarchical label spaces: Modelling diagnostic consensus in ICD-11

Jess Rees, Amy Rouillard, Bruce Bassett

School of Computer Science and Applied Mathematics, University of the Witwatersrand; Wits MIND Institute, University of the Witwatersrand

Wednesday 25 November, 12:40 – 13:00 · Yellowwood · Bayesian statistics 1 ↑programme p. 25

How should one infer ground truth labels when only multiple noisy labels of unknown reliability are available? This situation is common in clinical diagnosis, where human clinicians and AI models can differ significantly in opinion. This setting presents a substantial statistical challenge: a massive, hierarchical label space with ~470 diagnoses observed in an exemplar South African public hospital ward, and up to 28,000 diagnoses coded in the ICD-11 taxonomy.

To address this challenge, we develop a Bayesian approach that leverages the directed acyclic graph (DAG) structure of the ICD-11 label space to perform truth inference. Using simulations, we show that our model successfully recovers the true diagnosis and accurately ranks diagnoser reliability in problems with up to 2,000 labels. We also evaluate the model's robustness, finding that while diagnosis prediction and reliability ranking remain stable under various violations of model assumptions, standard parameter uncertainties are more sensitive and require careful interpretation. Finally, we show that our model improves truth recovery compared to standard baselines, whose diagnoser parameters are not estimable at this scale. Thus, our approach offers a reliable tool for modelling diagnostic consensus and evaluating AI-assisted diagnosis.

Keywords: Truth inference; Bayesian inference; hierarchical label spaces; clinical diagnosis; annotator reliability

Spatial statistics 1

Wednesday, 25 November 2026, 12:00 – 13:00 · Cedarwood · Chair: Jamie-Lee Nel

Generating spatial linear networks from homogeneous point patterns

Renate Thiede, Arthur Antonio, Inger Fabris-Rotelli

University of Pretoria

Wednesday 25 November, 12:00 – 12:20 · Cedarwood · Spatial statistics 1 ↑ programme p. 25

The concept of homogeneity is well-defined for point patterns, namely constant intensity across an observation area, and simulating homogeneous point patterns can be considered a solved problem. However, homogeneity has not yet been defined for more complex structures like line patterns or spatial networks. Recent research has begun to extend the definition of homogeneity to line patterns. Spatial networks, which can be defined as a set of points (nodes) and lines (edges), present the additional challenge that their nodes and edges must be connected. This research seeks to simulate explicitly homogeneous spatial networks. Specifically, we investigate how spatial networks can be generated from homogeneous point patterns, such that the spatial networks inherit the property of homogeneity from the underlying point patterns. We evaluate and compare the following simulation approaches: (1) Point-based simulation generates a homogeneous Poisson point pattern and then stochastically connects the points, creating an edge between each connected pair of points. (2) Tessellation-based simulation applies tessellation methods to a homogeneous Poisson point pattern and converts the tessellation to a spatial network. By exploring these simulation methods, this research aims to establish a framework for simulating spatial networks that preserve the statistical properties of homogeneity from the underlying point pattern, ultimately providing a benchmark for analysing real-world networks.

Keywords: spatial statistics; linear networks; simulation; point patterns

The complexity of spatially modelling antimicrobial resistance: Integrating clinical, environmental, and household survey data

Qondeni Ndlangamandla, Kabelo Mahloromela, Inger Fabris-Rotelli

University of Pretoria; South African Medical Research Council

Wednesday 25 November, 12:20 – 12:40 · Cedarwood · Spatial statistics 1 ↑ programme p. 25

Antimicrobial resistance (AMR) is a growing public health challenge that is influenced by complex interactions between human, environmental, and socioeconomic factors. Understanding the spatial distribution of AMR requires the integration of diverse data sources that capture different aspects of transmission and exposure. This study explores a spatial modelling framework that combines clinical AMR data, surface water data, and household survey data to investigate geographic patterns of resistance and their potential determinants. Clinical data provide information on the occurrence of resistant infections, while surface water data capture environmental reservoirs of resistant organisms. Household survey data contribute contextual information on water access, sanitation, hygiene practices, socioeconomic conditions, and healthcare utilization, all of which may influence AMR risk. Spatial analysis of AMR is inherently challenging due to variations in data resolution, sampling designs, spatial coverage, and the complex relationships between human populations and environmental systems. Linking environmental measurements with household- and individual-level outcomes requires methods that can account for spatial dependence, heterogeneity, and uncertainty. Geographic Information Systems (GIS) and spatial statistical models provide opportunities to integrate these diverse datasets, identify AMR hotspots, characterize spatial clustering, and assess potential environmental and household-level drivers of resistance. This work highlights both the opportunities and methodological complexities of spatially modelling AMR using multiple data sources and demonstrates the value of a One Health approach for strengthening AMR surveillance and informing targeted interventions.

Keywords: Antimicrobial resistance; One Health; spatial modelling; GIS; spatial epidemiology

Domain selection for spatial modelling in public health

Jamie-Lee Nel, René Stander, Inger Fabris-Rotelli, Kabelo Mahloromela

University or Pretoria

Wednesday 25 November, 12:40 – 13:00 · Cedarwood · Spatial statistics 1 ↑ programme p. 25

Spatial aggregation plays an important role in the analysis of public health data, where observations are aggregated into regular or irregular grid cells to identify disease hotspots or anomalies, detect spatial patterns, and support prediction and surveillance. The choice of grid size influences the resulting analysis, yet there is no widely accepted method for selecting an optimal grid size for point pattern data. If the grid is too coarse, important spatial variation may be smoothed over and lost. In contrast, a grid that is too fine can produce many cells with few or no observations, resulting in a sparse representation of the underlying spatial structure. This research aims to develop a method to select an optimal grid size for analysing spatial point pattern data. Several existing two-dimensional histogram bin-width selection methods exist. However, these approaches were developed for independent observations and do not explicitly account for the spatial structure or spatial dependence that may be present in spatial data. The current research is motivated by Entropy-based Local Indicators of Spatial Association (ELSA), which quantify local spatial association by incorporating information from neighbouring spatial units. The proposed methodology will build on this framework to define an optimisation algorithm to identify an optimal grid size for point pattern data. Selecting an appropriate grid size can improve the representation of spatial information and support more reliable analysis and data-driven decision-making.

Keywords: ELSA; Grid size; Point patterns

Biostatistics 2

Wednesday, 25 November 2026, 12:00 – 13:00 · Cypress · Chair: Justine Nasejje

Conformal prediction for protein structure models

Seitebaleng Makgai, Mohammad Arashi

University of Pretoria; Ferdowsi University of Mashhad

Wednesday 25 November, 12:00 – 12:20 · Cypress · Biostatistics 2 ↑ programme p. 25

Generative models for protein structure increasingly produce distributions over possible conformations rather than a single predicted structure. While such conformational ensembles provide a richer representation of structural variability, the observed spread among sampled conformations may conflate genuine biological flexibility with uncertainty arising from the predictive model. Moreover, these models typically do not provide calibrated statements about the reliability of their predictions. Conformal prediction offers a principled framework for uncertainty quantification, but its standard formulation presents several challenges in the protein-structure setting. In particular, it provides marginal rather than family-conditional coverage; calibration proteins are not necessarily exchangeable with a target protein because of evolutionary relatedness; and conformational ensembles are probability measures on a high-dimensional, non-Euclidean space, making the construction and interpretation of prediction sets non-trivial. In this study, we address these challenges by developing a conformal prediction framework for generative protein-structure models that accounts for evolutionary relatedness and the geometry of conformational space. The proposed approach aims to provide calibrated uncertainty statements for predicted conformational ensembles while accommodating the dependence induced by phylogenetic relationships.

Keywords: Conformal prediction; Protein structure prediction; Uncertainty quantification; Phylogenetic dependence; Conformational ensembles

Beyond the midpoint: Fuzzy imputation for survival estimation with interval-censored data

Kumbirai Kimpton Manungo, Chioneso Show Marange, Odilo Ndiweni

Department of Mathematical Statistics and Actuarial Science, University of the Free State; Department of Computational Sciences, University of Fort Hare, Alice 5700, South Africa

Wednesday 25 November, 12:20 – 12:40 · Cypress · Biostatistics 2 ↑ programme p. 25

The analysis of failure time data often involves cases where the exact time of event occurrence is unobserved but known to lie within an interval, giving rise to interval censored data. Classical approaches such as the Kaplan–Meier (KM) and Turnbull estimators are widely used but fail to accurately capture within-interval uncertainty, resulting in stepwise and sometimes inaccurate survival curves. This study explores fuzzy logic as an alternative method for estimating survival functions under such censoring. A Monte Carlo (MC) simulation was conducted to assess five fuzzy midpoint-based imputation methods, namely Fuzzy Centroid, Fuzzy α -Cut, Fuzzy Arithmetic, Fuzzy Fully Random Variable, and Fuzzy Weighted Average, in comparison with the traditional Midpoint imputation method. Statistical performance metrics were used to evaluate the methods' performance across different sample sizes and prevalence rates. For empirical validation, the ACTG181 and Breast Cancer clinical datasets were analysed using bootstrap resampling. Results show that the Fuzzy α -Cut, Fuzzy Arithmetic, and Fuzzy Fully Random Variable imputation methods outperform the classical Midpoint approach, yielding smoother and more reliable survival curves, particularly at small sample sizes. The Fuzzy Centroid and Fuzzy Weighted Average methods performed comparably to the classical midpoint.

Keywords: Interval Censoring; Fuzzy Modelling; Survival Function; Kaplan–Meier Estimator; Monte Carlo Simulation

Scalable and interpretable cause-specific competing risks modeling with neural additive models

Justine B. Nasejje, Steve Mphaya, Julian Brooks, Marvin N. Wright

University of the Witwatersrand, Johannesburg, South Africa; University School for Advanced Studies IUSS Pavia, Pavia, Italy; University of Leeds, Leeds, United Kingdom; Leibniz Institute for Prevention Research and Epidemiology – BIPS; University of Bremen, Bremen, Germany

Wednesday 25 November, 12:40 – 13:00 · Cypress · Biostatistics 2 ↑ programme p. 25

Competing risks survival analysis is widely used to analyse time-to-event outcomes in the presence of multiple mutually exclusive event types. Classical approaches, including cause-specific Cox and Fine–Gray models, provide interpretable inference but often rely on restrictive modelling assumptions. Deep neural networks can capture complex nonlinear relationships and improve predictive performance; however, they frequently sacrifice interpretability and may require computationally intensive optimization over complete risk sets. In this study, we extend the Competing Risks Interpretable Survival Prediction with Neural Additive Models (CRISP-NAM) framework within a cause-specific hazard formulation by introducing a sampling-aware risk-set optimization strategy inspired by Cox-Time. For each cause-specific event, the proposed approach replaces the complete risk set with a small sample of eligible controls to approximate the corresponding full risk cause-specific likelihood. The neural additive structure is retained, allowing nonlinear covariate relationships to be estimated and visualized separately for each competing event. Because the number of sampled controls per event can remain fixed as the cohort size increases, the proposed method reduces computational dependence on increasingly large risk sets and provides a scalable training strategy. We evaluate the framework using simulation studies and benchmark competing risks datasets, considering sampled risk-set sizes ranging from 1 to 10 controls per event. The results show that sampling-aware CRISP-NAM achieves predictive discrimination, cumulative incidence estimation, and recovery of cause-specific feature effects comparable to those obtained using complete risk sets. Overall, the proposed framework provides a practical and scalable approach to fitting interpretable cause-specific competing risks models while maintaining predictive accuracy and model transparency.

Keywords: Competing risks; cause-specific hazards; survival analysis; neural additive models; risk-set sampling; interpretable machine learning

Financial statistics 1

Wednesday, 25 November 2026, 14:00 – 15:20 · Weaver · Chair: Jacob Majakwara

Exploring extreme risk in selected cryptocurrencies using APARCH-EVT and APARCH-extended Burr XII models

Thabani Ndlovu, Energy Sonono

North-West University

Wednesday 25 November, 14:00 – 14:20 · Weaver · Financial statistics 1 ↑ programme p. 26

Accurate estimation of extreme market risk remains challenging for cryptocurrencies because of their pronounced volatility and heavy-tailed return distributions. This study compares the performance of two hybrid conditional risk models, the Generalised Autoregressive Conditional Heteroskedasticity–Extreme Value Theory (GARCH–EVT) model and the GARCH–Marshall–Olkin Extended Burr XII (GARCH–MOEBXII) model, for modelling tail risk in cryptocurrency returns. Although the GARCH–EVT framework is widely regarded as the benchmark for extreme risk modelling, its performance depends critically on selecting an appropriate threshold for the Generalised Pareto Distribution (GPD), for which no universally accepted criterion exists. As a threshold-free alternative, we investigate the Marshall–Olkin Extended Burr XII (MOEBXII) distribution, whose additional Marshall–Olkin parameter enhances flexibility in modelling heavy-tailed behaviour. Daily returns of Bitcoin, Ethereum, BNB, Tether (USDT), and USD Coin (USDC) are analysed. An Asymmetric Power ARCH (APARCH) model is first fitted to capture conditional volatility, after which the standardised residuals are modelled using either the GPD or the MOEBXII distribution. Value-at-Risk (VaR) and Expected Shortfall (ES) forecasts are evaluated using the Kupiec and Christoffersen VaR backtests together with the Acerbi–Székely ES backtest. The results show that, despite the unresolved threshold selection problem, the GARCH–EVT model consistently outperforms the GARCH–MOEBXII model by producing fewer VaR exceedances and superior backtesting performance at the 95%, 99%, and 99.5% confidence levels. These findings confirm that GARCH–EVT remains a robust and reliable framework for quantifying extreme cryptocurrency risk.

Keywords: Extreme-Risks; Cryptocurrency; APARCH; Heavy-Tailed; Risk-Measures; Backtest; Exceedances

Bayesian discrete-time survival analysis of loan default: Class-imbalance solutions in South African banks

John Lehlaka Masekoameng, Sizwe Vincent Mbona, Anisha Ananth, Retius Chifurira

Department of Statistics, Faculty of Applied Sciences, Durban University of Technology, South Africa; School of Agriculture and Sciences, University of KwaZulu-Natal, South Africa

Wednesday 25 November, 14:20 – 14:40 · Weaver · Financial statistics 1 ↑ programme p. 26

Loan default remains a significant challenge for financial institutions, elevating credit risk and undermining the sustainability of lending operations. Accurate prediction of when borrowers are likely to default enables improved credit risk management and decision-making. This study modelled the time to loan default among South African borrowers using discrete-time survival analysis and assessed whether class balancing improved predictive performance. Anonymised data for 5,000 borrowers from three commercial banks, observed from 2018 to 2022, were converted to 54,489 borrower-month records. Four models were fitted: frequentist and Bayesian models using the original class-imbalanced data and corresponding models using data balanced through random undersampling. Performance was evaluated using discrimination, Brier scores, calibration error, information criteria and, for Bayesian models, convergence diagnostics. Overall, 81.1% of borrowers defaulted during follow-up. Class balancing reduced the Brier score from 0.0680 to 0.0327 and improved calibration error from 0.0458 to 0.0017. The Bayesian class-balanced model achieved excellent convergence (maximum R-hat = 1.000), the lowest calibration error and close agreement with the frequentist balanced model. Higher credit scores reduced default risk, whereas higher inflation increased risk. Longer loan terms and higher income categories were also associated with higher odds of default. Discrete-time survival analysis using class-balanced data therefore provides a robust, interpretable and well-calibrated framework for estimating default timing. The approach can strengthen credit risk monitoring, early warning systems, and lending decisions in the South African banking sector.

Keywords: Bayesian discrete-time survival analysis; class imbalance; credit risk; loan default; South Africa

Assessment of credit distributional model risk under parametric and non-parametric bootstrap methods

Modisane Seitshiro, Hopolang Mashele

North-West University; University of South Africa

Wednesday 25 November, 14:40 – 15:00 · Weaver · Financial statistics 1 ↑ programme p. 26

Despite Value-at-Risk (VaR) having well-documented weaknesses, it remains the principal risk measure used by practitioners and regulatory bodies to calculate risk capital for banks. This study forecasts credit Value-at-Risk (CVaR) and assesses distributional model risk (DMR) within a credit portfolio framework. Distribution model risk arises whenever a practitioner calibrates or estimates a model that differs from the true, unobservable data-generating process. We assess the DMR for a credit risk model by estimating bootstrap upper confidence levels and comparing them with the point-estimate CVaR. Random variables drawn from well-known parametric distributions and non-parametric resamples are adopted as proxies for the systematic common factor in the Vasicek Asymptotic Single Risk Factor (ASRF) model. Bootstrap Monte Carlo simulations with B = 1000 replicates are performed to quantify the accuracy of the CVaR estimator. Four bootstrap confidence-interval techniques are considered: the basic percentile (BP), standard hybrid percentile (SHP), bias-corrected percentile (BCP), and a newly proposed modified hybrid percentile (MHP). The MHP embeds a Cornish–Fisher skewness correction and a finite-sample margin-of-error adjustment, making it sensitive to the degree of asymmetry in the bootstrap replicate distribution. Empirical analysis is conducted on a synthetic credit portfolio calibrated to South African banking sector parameters drawn from South African Reserve Bank (SARB) published data and the National Credit Regulator (NCR) quarterly reports. The results demonstrate that the proposed MHP method outperforms all alternatives. Thus, it overestimates CVaR under heavy-tailed and asymmetric systematic factors, providing a conservative capital cushion for stressed markets, while underestimating it under Gaussian factors, appropriately relaxing the capital requirement during calm market conditions. These findings carry direct policy implications for capital adequacy assessment under the regulatory treatment of model risk.

Keywords: Credit Value-at-Risk; Distributional Model Risk; Bootstrap Confidence Intervals; Vasicek Single Factor Model; Modified Hybrid Percentile

Unifying cure rate models in credit risk: The Box–Cox transformation approach

Jacob Majakwara

University of the Witwatersrand, Johannesburg

Wednesday 25 November, 15:00 – 15:20 · Weaver · Financial statistics 1 ↑ programme p. 26

Cure rate models have become increasingly important in credit risk modelling because they distinguish borrowers who will eventually default from those who remain non-defaulting throughout the loan lifetime. This work provides the first comprehensive Monte Carlo evaluation and empirical application of the Box–Cox transformation cure rate model in credit risk. The simulation results demonstrate favourable finite-sample properties of the maximum likelihood estimator across a range of transformation parameter values. The empirical analysis illustrates the flexibility of the Box–Cox transformation model in accommodating both mixture and promotion time cure model formulations within a unified framework.

Keywords: Credit risk; Cure rate models; Box-Cox transformation cure8 rate model; Mixture cure rate; Promotion time cure rate

Bayesian statistics 2

Wednesday, 25 November 2026, 14:00 – 15:20 · Yellowwood · Chair: Gandhi Jafta

Approximate Bayesian computation for parameter estimation in generalised Eyring accelerated models

Neill Smit, Lizanne Raubenheimer

North-West University; Rhodes University

Wednesday 25 November, 14:00 – 14:20 · Yellowwood · Bayesian statistics 2 ↑ programme p. 26

Accelerated life testing is widely used to estimate the life characteristics of high-reliability products when conventional reliability assessment is infeasible. In accelerated life tests, products are subjected to elevated stress levels to induce early failures. These accelerated failure times can be extrapolated, allowing inference on lifetime behaviour under normal operating conditions. In this paper, parametric accelerated life testing models are considered, with the generalised Eyring relationship employed as the time transformation function.

Within a Bayesian framework, the resulting likelihood functions are often analytically intractable, particularly under dual-stresses and flexible distributional assumptions. Consequently, approximation techniques are required for posterior inference. We investigate approximate Bayesian computation for parameter estimation in this setting. A comprehensive simulation study is conducted to assess the statistical performance of the proposed approach. Results are evaluated against established inferential procedures.

Keywords: Accelerated life testing; Approximate Bayesian computation; Generalised Eyring relationship; Parameter estimation; Reliability

Bayesian analysis for k-out-of-n failure time distributions in dependent environments

Lizanne Raubenheimer, Thomas Mazzuchi, Shahram Sarkani

Rhodes University; The George Washington University

Wednesday 25 November, 14:20 – 14:40 · Yellowwood · Bayesian statistics 2 ↑ programme p. 26

The use of k-out-of-n systems has become increasingly popular especially for fault tolerant systems. The failure time distribution for a k-out-of-n system can be constructed as the distribution for the $(n+1-k)$th order statistic. This is easily constructed using an independence structure; however, it is well known that many of these systems are subject to common failure environments. As such we specify a Marshal Olkin Multivariate Exponential distribution to model common failures and develop expressions for the time to failure using a shock model approach. Expressions are developed for the symmetric model. For the Bayesian analysis, independent gamma priors are used for the shock rates, and an ordered Dirichlet prior for the normalized shock contributions.

Keywords: Dependent environments; Dirichlet prior; Marshal Olkin exponential distribution

B-FAIR: Bayesian spatial fairness

Gandhi Jafta, Inger Fabris-Rotelli, Emma Ruttkamp-Bloem, Janet van Niekerk, Iketle Maharela

University of Pretoria

Wednesday 25 November, 14:40 – 15:00 · Yellowwood · Bayesian statistics 2 ↑programme p. 26

Spatial dependence is often ignored in algorithmic fairness, even though location may be tied to both protected attributes and merit-relevant features. This creates a basic problem for standard fairness metrics, because peer comparisons can be distorted by spatial structure rather than by the substantive relationship we want to study. To address this, we introduce Bayesian Individual Spatial Fairness (B-FAIR), a framework that defines fairness at the individual level while explicitly modelling protected context as a continuous stochastic spatial surface. In B-FAIR, the protected demographic surface is treated as a latent spatial process, and the resulting uncertainty is carried through to the final fairness estimate. The B-FAIR functional couples peer similarity, demographic disparity, and outcome differences into a single posterior quantity, so that fairness is not reduced to a point estimate but expressed as a distribution with uncertainty. Since neither the protected-context model nor the peer graph depends on administrative zones, the framework is less sensitive to areal aggregation and avoids the usual zone-based critique.

Keywords: Algorithmic Fairness; uncertainty Quantification; Bayesian Modelling

Bayesian negative-binomial neural networks for climate-informed malaria prediction: A comparative evaluation of statistical and machine learning models

Makwelantle Asnath Sehlabana

University of Limpopo

Wednesday 25 November, 15:00 – 15:20 · Yellowwood · Bayesian statistics 2 ↑programme p. 26

Accurate prediction of malaria incidence is important for timely disease surveillance and targeted public-health interventions, particularly where transmission is influenced by climatic and environmental conditions. This study compared the predictive performance of four modelling approaches: a frequentist Negative Binomial (NB) model, a Bayesian Negative Binomial (Bayesian-NB) model, an optimised Multi-Layer Perceptron (MLP), and a Bayesian Negative-Binomial Neural Network (BNN). Model performance was evaluated using root mean square error (RMSE), mean absolute error (MAE), symmetric mean absolute percentage error (sMAPE), seasonal mean absolute scaled error (MASE), and the percentage of predictions within $\pm 20\%$ of observed cases. The BNN demonstrated the strongest overall predictive performance, achieving the lowest RMSE, MAE, and seasonal MASE, followed closely by the optimised MLP. The conventional NB model showed intermediate performance, while the Bayesian-NB generally recorded the highest prediction errors for RMSE, MAE, and MASE. In contrast, the Bayesian-NB achieved the lowest sMAPE, closely followed by the NB model. While the NB and Bayesian-NB models demonstrated higher $\pm 20\%$ prediction accuracy than the neural-network approaches, overall ranking based on the combined performance measures identified the BNN as the best-performing model, followed by the NB, MLP, and Bayesian-NB. These findings demonstrate the potential of Bayesian Negative-Binomial neural networks for modelling overdispersed malaria counts and suggest that integrating Bayesian inference with neural-network architectures can provide a valuable approach to climate-informed malaria prediction.

Keywords: Bayesian neural network; Negative Binomial regression; malaria prediction; climate-informed modelling; machine learning

Classification and machine learning

Wednesday, 25 November 2026, 14:00 – 15:20 · Cedarwood · Chair: Humphrey Brydon

A self-supervised transformer and perceptron framework for incomplete multivariate time series in precision medicine

Jeanette Maria Louw, Brenda Otieno Mac'Oduol

University of Pretoria

Wednesday 25 November, 14:00 – 14:20 · Cedarwood · Classification and machine learning ↑ programme p. 26

Multivariate time series data (MTSD) are pervasively incomplete in clinical settings, posing a fundamental barrier to reliable decision-making. Deep learning approaches have addressed missingness through imputation and interpolation, but they are optimised for either speed or accuracy. Excellence in a single dimension is insufficient for practical deployment in healthcare environments because it demands both reliability and responsiveness. In response, this paper introduces SelfPerFormer. This architecture employs a novel asymmetric dual-encoder transformer to capture complex dynamics from incomplete observations, augmented with dataset specific modules designed for signals with dense temporal resolution, followed by a lightweight multi-layer perceptron decoder that reconstructs MTSD profiles. Self-supervised representation learning enables the model to extract meaningful structure despite pervasive missingness, eliminating dependence on complete training data. The architectural design avoids sequential bottlenecks, making the model highly parallelisable. Comprehensive evaluation demonstrates that SelfPer-Former achieves up to four times faster training, eleven times faster inference and five to thirty five percent better accuracy than its strongest competitor. The framework seamlessly extends to downstream tasks, achieving strong performance using learnt representations or in combination with predicted MTSD as input. The SelfPerFormer demonstrates that a principled encoder design eliminates the need for complex decoders, enabling efficient, accurate and extensible time series analytics. This research endeavours to strengthen the analytical foundations essential for precision medical interventions, effectively bridging the gap between complex medical data and actionable clinical insights, thereby advancing towards truly personalised medicine.

Keywords: Multivariate time series; SelfPerFormer; transformer; missing data; precision medicine

Feature space engineering for machine learning on bioinformatics data: A CNS drug penetrance case study

Tshemollo Rapolai, Seite Makgai, Mohammad Arashi

University of Pretoria

Wednesday 25 November, 14:40 – 15:00 · Cedarwood · Classification and machine learning ↑ programme p. 26

Most drug candidates fail simply because they cannot cross into the central nervous system (CNS), making CNS penetrance one of the central bottlenecks in neuro-pharmacology. Machine learning offers a way to predict this property computationally, but the datasets available for training are small and highly constrained, which makes building reliable models genuinely difficult. This work looks at how chemical representations (SMILES strings) can be transformed into structured molecular feature spaces suited to this kind of prediction. Sample sizes are so limited that the emphasis throughout is on careful feature extraction and deliberate model tuning to keep variance under control rather than chasing complexity. The result is an approach to mapping molecular features onto CNS penetrance outcomes that stays reliable even when data is scarce.

Keywords: SMILES; Machine Learning; Feature Extraction

Optimising the classification boundary with imbalanced data

Humphrey Brydon, Retha Luus

University of the Western Cape

Wednesday 25 November, 15:00 – 15:20 · Cedarwood · Classification and machine learning ↑ programme p. 26

Class imbalance poses a persistent challenge in supervised classification, especially concerning rare events. The default 0.5 classification boundary, prevalent in most modelling applications, is predicated upon an assumption of balanced distributions and perfectly symmetrical misclassification costs. Consequently, this boundary is biased towards an over-represented majority class. Since the targeted minority class is usually the class of interest under imbalanced scenarios, particularly in fields such as fraud detection, rare medical disease diagnosis, and critical failure prediction, this shortcoming is problematic.

Boundary optimisation for imbalanced targets is an established concept, and a widely adopted approach relies on maximising Youden's Index, derived from the ROC curve, to optimise the trade-off between sensitivity and specificity. However, Monte Carlo simulations demonstrated that Youden's Index frequently produces unstable results when the target events are extremely rare. This instability limits its practical reliability. Alternatively, PR curves and the associated F1-score are frequently recommended for imbalanced targets because they focus primarily upon minority class success. The current study extends previous research by determining the optimal classification boundary utilising the PR curve. By leveraging the PR curve, which emphasises the minority class, the study aims to establish a comparative thresholding mechanism.

The primary objective is to determine the most stable metrics across diverse imbalance scenarios, ranging from balanced data (1:1) to extreme imbalance (99:1). Predictive performance based on the threshold selected according to the P4 metric, will be benchmarked against predictive performance based on classification thresholds determined by the default boundary (0.5), Youden's Index, and the standard F1-score.

To examine the interaction between data-level and model-level interventions, our methodology evaluates six distinct modelling approaches. These approaches systematically combine data-level sampling, the aforementioned thresholding techniques, and a post-hoc prediction bias correction adjustment.

Keywords: class imbalance; supervised classification; boundary optimisation

Contaminated count regression models

Arno Otto, Andriëtte Bekker, Antonio Punzo, Cristina Tortora

University of Pretoria; University of Catania; San Jose State University

Wednesday 25 November, 15:00 – 15:20 · Cedarwood · Classification and machine learning ↑ programme p. 26

Count data frequently depart from equidispersion and are often further contaminated by atypical observations that do not come from the dominant data-generating process. Classical two-parameter count models offer flexible dispersion but leave no parameters free to accommodate the effect of atypical observations, while existing heavy-tailed alternatives gain robustness at the cost of dispersion flexibility. We propose a family of contaminated count distributions that addresses this limitation by representing the data as a two-component mixture in which both components retain flexible dispersion and share a common mean, but differ only in dispersion. The construction adds two contamination parameters, one governing the proportion of atypical points and another controlling the dispersion increment of the contaminant component relative to the reference component. The nature and extent of the contamination are estimated rather than assumed. The availability of covariates that can improve inference about the mean of the contaminated count distribution motivates the further proposal of the contaminated count regression model. The contaminated count distribution is shown to be identifiable, and parameters are estimated by maximum likelihood. The mixture structure assigns each observation an a posteriori probability of arising from the contaminant component, so that atypical observations are detected without a subjectively specified threshold.

Keywords: contamination; outliers; counts; dispersion

Recent advancements in modelling for complex data

Wednesday, 25 November 2026, 14:00 – 15:20 · Cypress · Chair: Najmeh Nakhaeirad

A circular–linear regression model for interval-valued circular data

Najmeh Nakhaeirad, Thasmika Mohan, Sphiwe Skhosana, S Yaser Samadi

University of Pretoria; Southern Illinois University, Carbondale, IL, USA

Wednesday 25 November, 14:00 – 14:20 · Cypress · Recent advancements in modelling for complex data ↑ programme p. 26

Many real-world phenomena are naturally observed in interval form, although conventional statistical methods are primarily designed for single-valued data. A common practice is to replace each interval with its midpoint or mean, potentially discarding valuable information about within-interval variation. This limitation is especially important for circular variables, for which interval-valued observations are represented as arcs and must be analysed using methods that preserve their periodic structure. We propose a circular–linear regression framework that simultaneously accounts for circularity and the information contained within interval-valued observations. The model accommodates a circular interval-valued response and both circular and linear covariates, which may also be interval-valued. A maximum-likelihood estimation procedure is developed to retain the internal distributional characteristics of the observed intervals instead of relying on aggregated summaries. The method is assessed through simulation studies and applied to South African wind-farm data to model variability in wind direction. In both the simulated and empirical analyses, the proposed approach outperforms the classical centre–range method, demonstrating its potential as an effective framework for analysing circular interval-valued data.

Keywords: Circular interval-valued data; circular regression; Maximum likelihood estimation; Internal distribution

---

New flexible methods for circular distributional regression

Luke Smythe, Najmeh Nakhaeirad, Sphiwe Skhosana, Thomas Kneib

University of Pretoria; Georg-August-Universität Göttingen

Wednesday 25 November, 14:20 – 14:40 · Cypress · Recent advancements in modelling for complex data ↑ programme p. 26

Circular response data arise in many scientific and applied settings where the outcome is directional or periodic, such as wind direction, orientation, seasonal timing and movement data. Existing circular regression methods typically focus primarily on modelling the conditional mean direction, while allowing other distributional characteristics to remain fixed. This research develops a new class of circular distributional regression models in which both the conditional mean direction and concentration of a von Mises response are allowed to vary flexibly with covariates. The methods are inspired by linear distributional models formulated under a GAMLSS framework. Four novel flexible estimation approaches are developed and investigated. The first uses multivariate local-linear smoothing to estimate unrestricted predictor surfaces for both distributional parameters. The second introduces a structured additive local-linear formulation that decomposes these predictors into lower-dimensional nonlinear effects and selected interactions. The third develops a structured P-spline formulation in which smooth distributional effects are represented through penalised spline bases. The fourth introduces a MARS-basis von Mises distributional regression approach, where adaptive hinge bases are identified from circular pilot regressions and subsequently re-estimated under the exact von Mises likelihood. A fully parametric von Mises regression model is included as a benchmark. The proposed models are evaluated through an extensive Monte Carlo simulation study covering nonlinear, interaction-based, circular-covariate and higher-dimensional settings. Performance is assessed using circular prediction error for the mean direction, root mean squared error for the concentration parameter, out-of-sample log-likelihood and computational efficiency. The research is novel in developing and systematically comparing new flexible distributional regression estimators specifically for circular responses. By modelling both location and concentration as covariate-dependent quantities, the proposed framework extends circular regression beyond conventional mean-direction modelling and provides a foundation for more general circular distributional modelling.

Keywords: Circular statistics; distributional regression; von Mises distribution; GAMLSS; local-linear regression; P-splines; MARS; nonparametric regression

Semiparametric mixture of nonparametric regressions

Kagisho Nkadimeng, Sphiwe Skhosana, Najmeh Nakhaeirad

University of Pretoria

Wednesday 25 November, 14:40 – 15:00 · Cypress · Recent advancements in modelling for complex data ↑ programme p. 26

Finite mixture regression models provide a flexible framework for analysing heterogeneous populations in which unobserved subgroups may exhibit distinct relationships between covariates and a response variable. However, conventional mixtures of parametric regressions may be restrictive when component-specific relationships are nonlinear or their functional forms are unknown. In this presentation, we investigate the semiparametric mixture of nonparametric regressions (SMNR), in which the mixing proportions and component variances are parametric, while the component-specific regression functions are nonparametric. Estimation is performed using a global expectation-maximisation (EM)-type algorithm that combines posterior component probabilities with kernel smoothing. Typically, the Nadaraya-Watson (NW) kernel estimator is used for estimating the nonparametric component regression functions. However, the NW estimator has some shortcomings such as boundary bias and a tendency to oversmooth local features, since it fits local constants rather than local trends. To overcome some of these shortcomings, we make use of the local-linear (LL) kernel estimator, which reduces boundary bias and provides better bias-variance properties by fitting local linear regressions. We conduct a simulation study to compare the finite-sample performance of the two estimators under various scenarios. To demonstrate the practical usefulness of the LL-based EM-type algorithm, we apply it to real data to estimate the SMNR model.

Keywords: EM algorithm; Kernel regression; Mixture of regression models; Semiparametric mixture models

INLA goes full circle

Janet van Niekerk, Xiang Ye, Haavard Rue

University of Pretoria

Wednesday 25 November, 15:00 – 15:20 · Cypress · Recent advancements in modelling for complex data ↑ programme p. 26

Regression models for circular variables constitute a class of statistical models that has received comparatively little development and attention within the general statistical framework. This is largely because constructing a linear predictor through linear combinations of covariates and random effects tends to break the circular nature of the response variable, making standard regression approaches difficult to apply. In this paper, we introduce a new approach to address this issue, resulting in well-defined regression models for circular responses when the data are concentrated. The proposed framework naturally extends to joint regression settings involving multiple circular and non-circular responses, and accommodates a combination of linear covariates, circular covariates, and various random effects. This approach provides an efficient implementation for jointly modeling the complex temporal dynamics of wind speed and direction in the presence of additional covariates. This approach is disseminated through an R package called INLACircular . We demonstrate the performance of the approach using several simulated and real data examples.

Keywords: Circular; regression; Bayesian; INLA

Compositional data

Wednesday, 25 November 2026, 15:40 – 17:00 · Weaver · Chair: Stephan van der Westhuizen

Transformation-free random forest modelling of compositional data

Stephan van der Westhuizen, Khaled Masoumifard

Stellenbosch University; MuViSU

Wednesday 25 November 2026, 15:40 – 16:00 · Weaver · Compositional data ↑ programme p. 27

Compositional responses arise when several interdependent components jointly represent a whole. Examples include species compositions, geochemical concentrations, land-cover proportions and soil particle-size fractions. Standard machine learning methods do not explicitly account for the constant-sum constraint and may consequently produce invalid or incoherent predictions. Log-ratio transformations are commonly used to address this problem, but they add complexity to model development and require predictions to be back-transformed to the original compositional scale.

We present the Dirichlet Random Forest (DirichletRF), a transformation-free random forest method for predicting compositional responses. The approach incorporates the Dirichlet distribution directly into the tree-building procedure. Candidate splits are evaluated according to improvements in the fit of the Dirichlet distribution within the resulting child nodes. The method therefore models all components jointly while preserving the geometry and constraints of the response.

The proposed framework is evaluated through a digital soil texture mapping case study using LUCAS topsoil observations from Germany. Clay, silt and sand proportions were predicted from a collection of environmental covariates. DirichletRF was compared with established transformation-based approaches in which conventional and multivariate random forests were fitted to isometric log-ratio coordinates. DirichletRF achieved the strongest overall predictive performance in the case study and generated coherent soil texture predictions without transformation or back-transformation. These findings illustrate how probability distributions defined on the simplex can be integrated into machine learning algorithms. The framework provides a practical and broadly applicable approach for predicting compositional responses across environmental and other scientific disciplines.

Keywords: Compositional data analysis; Dirichlet distribution; soil texture; statistical learning; tree-based models

A likelihood-based approach to compositional random forests

Khaled Masoumifard, Stephan van der Westhuizen, Sugnet Lubbe

Stellenbosch University

Wednesday 25 November 2026, 16:00 – 16:20 · Weaver · Compositional data ↑ programme p. 27

Compositional data are vectors of non-negative components constrained to sum to a constant. They arise across the geosciences, ecology, and the social sciences, and require methods that respect the geometry of the simplex. Applying random forests to such data poses a specific challenge in that a central step in growing a tree is the recursive splitting of a subset of the data into two child subsets, and the splitting rule must itself respect the constrained geometry of the simplex. Existing likelihood-based random forests for compositional data use the Dirichlet distribution to define the splitting criterion. This choice has two limitations. The Dirichlet is a rigid model, in that a single precision parameter governs the variability of every component and the induced correlations between components are necessarily negative, so it cannot represent the heterogeneous or positive associations that compositional data frequently exhibit. It is also expensive to evaluate because the maximum-likelihood estimates of the Dirichlet parameters have no closed-form solution and must be obtained iteratively for each candidate split.

We extend this framework in two directions. The first follows the log-ratio approach, using the multivariate logistic-normal likelihood as the splitting criterion. The second draws on directional statistics, in which the square-root transformation transfers a composition onto the unit hypersphere, where the Kent distribution provides an elliptically contoured model. Both families relax the Dirichlet's dependence structure. Unlike log-ratio methods, the square-root transformation is well defined on the boundary of the simplex, allowing compositions containing structural zeros to be handled without imputation.

In simulation studies, both methods substantially improve computational efficiency relative to the Dirichlet baseline while remaining competitive in predictive accuracy. The resulting forests support conditional mean prediction, estimation of full conditional distributions at a leaf, out-of-bag error estimation, and permutation-based variable importance.

Keywords: Compositional data analysis; Kent distribution; log-ratio transformation; logistic-normal distribution; random forests; variable importance

Fault detection and diagnostic analysis in multivariate compositional data

Mduduzi Maphosa, Roelof Coetzer, Charl Pretorius, Shawn Liebenberg

North-West University

Wednesday 25 November 2026, 16:20 – 16:40 · Weaver · Compositional data ↑ programme p. 27

Industrial processes frequently generate compositional data, in which variables represent proportions that sum to one. Compositional data are characterised by constrained, nonnegative proportions and interdependencies among components, posing unique challenges that require specialised methodologies. Data-driven fault detection procedures enable the identification of abnormal behaviour, process interruptions, and system breakdowns within industrial facilities. We propose retrospective Dirichlet-based likelihood ratio tests to detect process faults in multivariate compositional data streams. Two complementary test statistics are derived, each with associated asymptotic critical values, and their performance is evaluated using extensive Monte Carlo simulations. The results demonstrate that the tests have power to detect a single fault across different sample sizes and dimensions. In addition to detection, we develop a diagnostic procedure that attributes a detected fault to the responsible component. The proposed detection and diagnostic procedures are further validated on the Tennessee Eastman Process, a widely used benchmark for industrial process monitoring.

Keywords: Compositional data; Dirichlet distribution; Likelihood ratio

A Contaminated Model for Overdispersed Multinomial Microbiome Count Data

Ockert van Heerden, Andriëtte Bekker, Seite Makgai, Arno Otto, Antonio Punzo

University of Pretoria; University of Catania, Italy

Wednesday 25 November 2026, 16:40 – 17:00 · Weaver · Compositional data ↑ programme p. 27

Multinomial count data, such as microbial composition profiles derived from sequencing studies, frequently contain anomalous observations that distort parameter estimates. The Dirichlet-multinomial (DM) distribution is widely used in this setting but remains sensitive to such contamination. We propose the contaminated Dirichlet-multinomial (CDM) distribution, a two-component mixture in which the regular data come from a DM component with a lower dispersion and the irregular data come from a DM component with an inflated dispersion parameter. This construction accommodates anomalies without requiring their removal, and yields a natural rule for anomaly detection via posterior probabilities. Through sensitivity analyses involving both single-point anomalies and background noise, we demonstrate that the CDM distribution effectively downweights the influence of anomalous observations on the parameter estimates. The model is applied to gut microbiome data from a colorectal carcinogenesis study, where it consistently outperforms the DM distribution across all information criteria and identifies biologically plausible anomaly proportions in both the healthy and carcinoma subsets.

Keywords: Compositional data analysis; Dirichlet distribution; soil texture; statistical learning; tree-based models

Biostatistics 3

Wednesday, 25 November 2026, 15:40 – 16:40 · Yellowwood · Chair: Najmeh Nakhaeirad

Joint latent class modelling of time-to-event and discrete longitudinal data

Alizandre Klynsmith, Najmeh Nakhaeirad, Tristan J. E. Harris, Sphiwe B. Skhosana

University of Pretoria

Wednesday 25 November, 15:40 – 16:00 · Yellowwood · Biostatistics 3 ↑ programme p. 27

Discrete longitudinal data occurs frequently in clinical data, although, modelling these data jointly with a survival outcome has largely been restricted to shared random effects models. However, this approach leaves the clinician without model-based clustering solutions to account for underlying heterogeneity typically present in clinical data. Joint latent class modelling simultaneously models longitudinal and time to event outcomes by accounting for unobserved population heterogeneity through latent classes. While a Gaussian model is commonly used for the longitudinal component, many applications involve discrete or count valued outcomes, for which alternative approaches are required. This study proposes a latent class joint modelling framework incorporating a generalised linear mixed effects model (GLMM) for the longitudinal component and a Weibull model for the time to event component. This allows individuals with similar longitudinal trajectories and event risks to be grouped into latent classes, while accounting for the dependence between the longitudinal and time to event processes. Maximum likelihood estimation is performed using the expectation maximisation (EM) algorithm for parameter estimation. The effectiveness of the proposed approach will be assessed through simulated data and an appropriate application to real world Alzheimer's data.

Keywords: joint latent class models; time-to-event; discrete longitudinal data; mixture modelling; expectation-maximization

Deep learning enhanced non-linear repeated measurement modelling in latent class joint models of time-to-event and longitudinal data with application to psychometric scores in Alzheimer's patients

Tristan Harris, Najmeh Nakhaeirad, Sphiwe Skhosana

University of Pretoria

Wednesday 25 November, 16:00 – 16:20 · Yellowwood · Biostatistics 3 ↑ programme p. 27

Non-linear longitudinal data occurs frequently in pharmacological and clinical settings such as CD4 counts, psychometric scores, and pharmacokinetic processes. Often these longitudinal processes are linked with a corresponding event-time such as time-to-onset of AIDS, time-to-hospital readmission or time-to-disease progression. Further, these data can also exhibit sub-populations representing underlying population heterogeneity often present in clinical data. We can jointly model these features generally by using a joint latent class model of time-to-event and longitudinal data. This model will link the longitudinal and time-to-event processes by using model-based clustering. However, the model typically assumes a linear link between longitudinal responses and fixed-effect covariates, which makes the choice of the traditional model inadequate for non-linear longitudinal data. In this paper, we propose a neural-network based approach to handle non-linearities in the relationship between the longitudinal response and it's covariates. Additionally, capitalizing on recent advances in non-linear modelling for joint latent class models, we assume a non-linear flexible structure in the class-membership model by using a neural network to capture potential non-linearities. The model is flexibly estimated using a general expectation conditional maximization type algorithm. We illustrate the model's finite sample properties through an extensive simulation study, finding reduced prediction error under a longitudinal data generating process that is non-linear. Finally, the model is applied to an epidemiological study jointly modelling psychometric scores linked to onset of dementia.

Keywords: joint latent class models; time-to-event; non-linear longitudinal modelling; deep-learning

Tail-aware latent subgrouping for survival: Identifying patient subgroups by extreme-event risk rather than by median prognosis

Brenda Mac'Oduol, Mohammad Arashi

University of Pretoria

Wednesday 25 November, 16:20 – 16:40 · Yellowwood · Biostatistics 3 ↑ programme p. 27

Latent-class survival models separate patients by where their survival distributions sit, typically by shifting location across classes. Two subgroups can nonetheless share a median survival exactly at every covariate value while differing substantially in the probability of very early failure, because that probability is governed by the weight of the distributional tail rather than by its centre. Under a log-logistic accelerated failure-time model the class median does not involve the shape parameter, so classes with identical regression vectors and shape parameters 0.7 and 3.0 have identical medians but early-failure probabilities of 0.275 and 0.015. Subtyping methods in current use do not define their latent classes through this parameter. We define latent subgroups by the survival-tail index and ask when they can be recovered. Two results answer this, and they point in opposite directions. First, to leading order the tail of the observed mixture is governed entirely by the class attaining the smallest tail index, so the remaining classes leave no first-order trace and distinct parameter vectors are tail-equivalent. Second, using a uniqueness argument for Dirichlet series together with real analyticity of the conditional survival function, the whole parameter, every tail index included, is identified from an arbitrarily short window of follow-up. Subgroups defined by extreme-event risk must therefore be estimated from the body of the distribution rather than from its tail. Estimation proceeds by an EM algorithm over a compact shape set, which is required because the unconstrained mixture likelihood is unbounded above. In simulation the method recovers median-identical subgroups that a shared-shape latent-class model does not, although it yields no gain in concordance and its tail-index estimates are sensitive to misspecification of the component family. We illustrate the approach on the Rotterdam breast-cancer cohort, with external validation on the German Breast Cancer Study Group data.

Keywords: latent class model; survival-tail index; accelerated failure time; precision medicine

Model-Free Screening for Treatment-Interactive Variables with Multivariate Responses in High-Dimensional Data

Shiferaw Bizuayehu; Denekew Belay; Najmeh Nakhaeirad

University of Pretoria

Wednesday 25 November, 16:40 – 16:50 · Yellowwood · Biostatistics 3 ↑ programme p. 27

Variable screening for ultrahigh-dimensional data with multiple outcomes has received little attention, particularly in identifying treatment-interactive variables for individual treatment regimes and precision medicine. We proposed a novel multivariate conditional sure independent ranking and screening procedure based on the conditional cumulative distribution function. The proposed model-free framework introduces a multivariate conditional dependence measure that characterizes the interaction association between covariates, treatment, and multiple outcomes without imposing parametric model assumptions. By isolating covariates whose effects on the outcomes are modified by treatment, our proposed method distinguishes prescriptive variables from prognostic main effect, providing a unified framework for ranking treatment-relevant covariates in the presence of any number of treatments, outcome variables, and ultrahigh-dimensional predictors. We establish the ranking consistency and the sure screening property of the proposed method under mild regularity conditions. Extensive simulation studies demonstrate its excellent finite-sample performance under a variety of complex data generating mechanisms, including scenarios where interacting variables exhibit weak or null main effects. The practical applicability of the proposed procedure is further illustrated through the analysis of the AIDS Clinical Trials Group Protocol 175 (ACTG175) dataset.

Keywords: Variable Screening; Model-free; Multiple Outcome; Prescriptive Variable; High dimensional covariate; Multi-category treatment; Precision medicine

Multivariate statistics

Wednesday, 25 November 2026, 15:40 – 17:00 · Cedarwood · Chair: Ruan Buys

Dynamic visualisation: A bridge between multivariate data tours and GPAbin biplots

Johané Nienkemper-Swanepoel, Dianne Cook

Stellenbosch University; Monash University

Wednesday 25 November, 15:40 – 16:00 · Cedarwood · Multivariate statistics ↑ programme p. 27

Missing values are a reality and remain a persistent challenge in data-driven practice. Multiple imputation addresses this by generating a distribution of plausible responses for each missing value, which results in multiple completed data sets. For data visualisation, however, interpreting multiple figures that show subtle variations of the same underlying information quickly becomes cumbersome. The biplot is a well-established multivariate tool that simultaneously displays sample and feature information in a low-dimensional space. Biplot methodology has been extended to sensibly combine visualisations from multiple imputations into a single display, referred to as GPAbin biplots. However, in its current form, GPAbin biplots are static and restricted to two dimensions.

This presentation links this biplot methodology to data tours, animated sequences of low-dimensional projections using geodesic interpolation, which reveal the joint distribution of multivariate data through many smoothly changing views. Since a biplot is itself one projection, embedding it within a tour allows us to examine what that particular projection preserves or distorts relative to the full-dimensional structure.

Using simulated examples, we show that this combined approach reveals structural characteristics of the data before and after imputation that cannot be captured by a single biplot. The tours further highlight differences between multiple imputation techniques, which offers a visual basis to compare techniques and ultimately decide on a suitable imputation strategy.

Keywords: biplots; missing data; multiple imputation; tours

Tracking changing rainfall regimes through animated biplots

Raeesa Ganey

University of the Witwatersrand

Wednesday 25 November, 16:00 – 16:20 · Cedarwood · Multivariate statistics ↑ programme p. 27

Rainfall seasonality is inherently multivariate. It is described not by a single quantity, but by a combination of features such as wet-season onset and cessation, season length, total rainfall, number of rainy days, rainfall intensity, and the contrast between rainfall and temperature across seasons. These quantities are commonly examined separately, producing a collection of individual trends rather than a unified view of how rainfall seasonality changes. In this talk, rainfall seasonality is treated as an evolving multivariate system. Each station-year is represented as a point in a low-dimensional seasonality space and animated biplots, implemented in the moveEZ R package, are used to follow changes across 30 years of observations from South African rainfall stations. Two complementary approaches are considered. A global principal component analysis provides a fixed reference space in which stations can be tracked through time. This is compared with year-specific biplots, where both station positions and variable axes are allowed to evolve. The yearly configurations are aligned using generalised Procrustes analysis, making it possible to distinguish movement of stations from changes in the multivariate structure itself. The comparison provides more than a visualisation. Procrustes statistics, congruence coefficients, bias measures, and principal angles are used to assess whether apparent changes reflect genuine shifts in rainfall-seasonality structure or instability arising from repeated estimation. The resulting animations reveal patterns that are difficult to see through separate trend analyses, including redundancy among seasonality measures, dominant directions of variation, changing relationships between rainfall characteristics, and whether climatic regions are converging or separating over time.

Keywords: animated biplots; rainfall seasonality

Interpreting evidential deep learning through gradient-guided biplots

Ruan Buys, Sugnet Lubbe, Luca Steyn

Stellenbosch University, South Africa

Wednesday 25 November, 16:20 – 16:40 · Cedarwood · Multivariate statistics ↑ programme p. 27

Evidential deep learning (EDL) provides a single-network framework for classification in which model outputs are interpreted as evidence parameters of a Dirichlet distribution, allowing predictive class probabilities and an associated uncertainty mass, or vacuity, to be obtained simultaneously. However, understanding how this uncertainty varies throughout a high-dimensional predictor space remains challenging. This work proposes an uncertainty-aware biplot for exploring the vacuity structure of an EDL classifier. Rather than selecting a two-dimensional representation solely from variation in the observed predictors, as in ordinary principal component analysis (PCA), the proposed method incorporates gradients of the fitted vacuity function with respect to the inputs. These gradients identify directions in $\mathbb{R}^p$ along which the classifier's

uncertainty changes most rapidly. A two-dimensional affine subspace is constructed by balancing geometric representation of the predictor data with retention of vacuity-gradient information. A tuning parameter controls the trade-off between conventional variance preservation and representation of the local geometry of the model's uncertainty function. The fitted vacuity function is evaluated across the selected subspace and displayed as a continuous surface together with calibrated axes, allowing changes in model uncertainty to be interpreted in terms of the original variables. An empirical support rule is additionally used to distinguish model confidence from regions of the predictor space that are poorly supported by the observed data. The construction is assessed by quantifying the trade-off between two-dimensional geometric reconstruction and retention of uncertainty information, and is illustrated using simulated data with known geometry. The guided display can retain substantially more uncertainty-gradient information than ordinary PCA at a measurable cost in reconstruction quality, providing a model-interpretability tool for identifying directions and regions of predictor space associated with changes in classifier uncertainty.

Keywords: Evidential deep learning; biplots

Speed session 1

Wednesday, 25 November 2026, 15:40 – 16:40 · Cypress · Chair: Chantelle Clohessy

Anomaly identification and portfolio construction in equity markets: An unsupervised machine learning approach

Adrian Crisp, Fabio M. Correa

University of the Free State

Wednesday 25 November, 15:40 – 15:50 · Cypress · Speed session 1 ↑ programme p. 27

Financial markets exhibit heterogeneous responses to changing market conditions that may not be fully captured by conventional risk measures such as historical volatility and market beta. This study investigates whether anomaly-derived information can provide an alternative cross-sectional description of equity behaviour and a basis for market segmentation and portfolio construction. Daily adjusted closing prices for 39 large-cap United States equities from 2005 to 2025 were analysed using Isolation Forest and a Long Short-Term Memory autoencoder. Their signals were combined to capture both isolated point shocks and sustained temporal deviations. Five stock-level features were constructed: upward and downward anomaly frequencies, directional anomaly balance, mean rolling volatility, and average daily return. The standardised features were clustered using K-means, with HDBSCAN employed as a comparative density-based approach. K-means identified four economically interpretable behavioural regimes characterised as high-growth, large-cap growth, defensive, and cyclical. HDBSCAN produced three broader groups and did not retain all economically relevant separations. Assignment correspondence with the full-period K-means solution was 53.8% for 2015–2019 and 79.5% for 2020–2025. One regime, absent before COVID-19, achieved perfect correspondence after 2020, indicating a structural reorganisation of cyclical and financial equities. Equal-weighted cluster portfolios exhibited distinct risk–return profiles. The high-growth cluster generated the highest annualised returns but also the highest volatility and deepest drawdown, whereas the cyclical cluster recorded the weakest risk-adjusted performance.

Keywords: Anomaly detection; Isolation Forest; LSTM autoencoder; K-means; HDBSCAN; equity markets

Traffic data availability accessibility

Diamando Katsaitis

University of Pretoria

Wednesday 25 November, 15:50 – 16:00 · Cypress · Speed session 1 ↑ programme p. 27

Reliable traffic data is essential for effective transportation planning and accessibility analysis. However, its collection is characterised by logistical constraints, infrastructure bias, and persistent spatial gaps, as sensors are typically deployed on major arterial roads while secondary and residential streets remain undermonitored. Before spatial models are used to estimate movement, it is necessary to assess whether the underlying traffic data is complete and representative of the network as a whole. This study assesses the consequences of incomplete data across two accessibility models of increasing complexity: a baseline expected travel time model, and a homogeneous Markov chain model in which transition probabilities are derived from estimated travel time rather than geometric distance alone. Network-constrained spatial statistics, namely network Kernel Density Estimation and the linear network K-function, are used to detect clustering in traffic monitoring locations and map monitoring intensity across the road network. Preliminary results suggest that monitoring coverage, though generally high, is unevenly distributed and concentrated on national highway networks, with select corridors showing statistically significant sensor clustering. This method is applied to a case study of Amsterdam to illustrate how spatial gaps in traffic data availability can bias accessibility estimates, and to demonstrate the value of network-based spatial statistics in identifying where transport monitoring infrastructure is most needed.

Keywords: Network-constrained spatial statistics; Traffic monitoring coverage; Accessibility modelling; Markov chain travel time model; Kernel Density Estimation

Spatial dependency modelling of disjoint spatial areas: A SAPRIN urban node analysis

Ephent Selahle, Inger Fabris-Rotelli, Nada Abdelatif

University of Pretoria; South African Medical Research Council

Wednesday 25 November, 16:00 – 16:10 · Cypress · Speed session 1 ↑ programme p. 27

Spatial data analysis requires understanding dependencies between polygons to ensure accurate modelling. Spatial weight matrices quantify spatial and temporal relationships by describing how values at one location relate to those at another. Traditional approaches often rely on Euclidean distance to construct spatial weight matrices, assuming straight-line proximity between locations. However, in disjoint urban areas, Euclidean distance can misrepresent true spatial relationships. The new connectivity structure is derived from the underlying road network. This study compares spatial dependency modelling using Euclidean distance versus road network distance, focusing on three geographically separated areas in South Africa: Melusi, Atteridgeville, and Hillbrow. Road network distances were derived using shortest-path calculations, producing a more realistic representation of the connectivity structure. Local Indicators for Categorical Data (LICD) were applied to assess spatial autocorrelation in key categorical variables, revealing distinct pockets of significant clustering for factors such as seropositivity, exclusive toilet access, cleaning practices, and employment status. The univariate case of LICD was further extended to the bivariate case, capturing both co-location (where two variables share the same value) and divergence (where they differ) between spatial units. The findings demonstrate that network-based measures capture spatial structure more accurately in disjoint areas. Beyond spatial dependencies, our research delves into demographic disparities across disjoint urban areas, such as residential, commercial, and mixed-use zones, within a city, based on data collected by South African Population Research Infrastructure Network (SAPRIN). A particular focus is given to how gender influences these patterns, since we can assume that there is no spatial difference per gender. We aim to uncover the relationships between geographic features, urban development distribution, population density, and their correlations with health risk profiles and seropositivity probabilities. These insights are vital for guiding urban planning and public health policies towards creating more equitable and healthy urban environments.

Keywords: Spatial Autocorrelation; Euclidean Distance; Road Network; LICD; Public Health Policies

Traffic simulation on spatial linear networks

Dylan Dabbadie

University of Pretoria

Wednesday 25 November, 16:10 – 16:20 · Cypress · Speed session 1 ↑ programme p. 27

Traffic congestion is a persistent challenge in urban environments, with significant implications for economic productivity, quality of life, and environmental sustainability. In South Africa, these effects are particularly pronounced among commuters in low-income areas who depend on public transport and have few alternative routes available to them. Accurate traffic simulation models are essential for understanding how congestion originates, propagates, and dissipates through urban road networks. Mesoscopic models offer a promising middle ground between computational efficiency and behavioural accuracy, yet their performance under South African traffic conditions has never been rigorously evaluated. This study implements and evaluates the Link-Level Interrupted Flow Traffic Dynamics (LIFT) model, an event-based mesoscopic simulation framework, on the road network of Hatfield, Pretoria. LIFT explicitly tracks backward travelling spaces to detect queue spillback and enforces strict first-in-first-out ordering at diverging intersections. Its performance is benchmarked against macroscopic and mesoscopic competitors using SUMO microscopic simulation as ground truth, with model accuracy assessed through mean absolute error of link density and mean path speed. Results indicate that LIFT reproduces congestion propagation and queue spillback dynamics substantially more accurately than competing models, achieving mean absolute density errors well below those of store-and-forward and existing mesoscopic approaches. These findings establish LIFT as a viable and accurate tool for urban traffic simulation in South African cities, with direct implications for signal timing optimisation, road network planning, and congestion management in growing urban environments.

Keywords: Spatial linear networks; Queue spillback; Mesoscopic traffic simulation; Urban congestion; LIFT

Spatial modelling of wastewater data: The case of Ekurhuleni

Abongile Nombona, Inger Fabris-rotelli

University of Pretoria

Wednesday 25 November, 16:20 – 16:30 · Cypress · Speed session 1 ↑ programme p. 27

Traditional clinical surveillance systems may underestimate the true burden of infectious diseases because they capture those cases that are present for testing, while asymptomatic and undiagnosed infections may remain undetected. Wastewater-based epidemiology (WBE) addresses this limitation by analysing pathogen signals shed into wastewater, enabling non-invasive, population-level surveillance within connected wastewater networks. However, much of the existing WBE literature focuses more on temporal trend analysis while giving comparatively little attention to the spatial structure of wastewater signals even though wastewater catchments are inherently spatial units with distinct and often irregular geographic boundaries. This study addresses this gap by developing a spatial statistical framework for analysing and visualising wastewater surveillance data, using pathogen concentration measurements from two wastewater catchments namely, Olifantsfontein and Vlakplaats in Ekurhuleni, Guateng as a case study. The catchments are treated as areal spatial units, with spatial weight matrices constructed using contiguity-based and distance-based approaches to represent spatial relationships between catchments. The global Moran's I statistic is used to assess overall spatial autocorrelation, while Local Moran's I (LISA) is used to identify local clusters and spatial outliers that exist within the study area. The spatial statistical analysis is complemented by using various types of maps to show the spatial distribution and spatial patterns of wastewater signals. By integrating spatial statistical methods with wastewater surveillance, this research aims to demonstrate how incorporating spatial dependence can enhance the interpretation of WBE data, support the identification of potential disease hotspots, and contribute to more geographically targeted public health surveillance and response.

Keywords: Wastewater; pathogen concentrations; spatial autocorrelation; spatial analysis

Comparing three CNN architectures: Baseline compound scaling and residual learning for photovoltaic fault classification

Nico Bezuidenhout, Chantelle Clohessy, Warren Brettenny

Nelson Mandela University

Wednesday 25 November, 16:30 – 16:40 · Cypress · Speed session 1 ↑ programme p. 27

Photovoltaic (PV) systems have grown in capacity and use in South Africa due to the country's increased energy demands and the need for suitable energy sources. These systems are subject to degradation and anomalies negatively affecting energy output. Unmanned aerial vehicles (UAVs) paired with thermal imaging are currently being used to capture images to detect PV faults. Manual procedures of fault detection from the thermal images are time consuming and labour intensive therefore studies have made use of Convolutional Neural Networks (CNN) for more automated procedures to detect faults. This study focuses on comparing a baseline CNN with EfficientNetB0 and ResNet18 to determine whether added complexity provides a tangible benefit compared to using a simple baseline model. Model performance is assessed using a five-fold cross-validation approach on 80% with 20% held out as an independent test set, and evaluated using accuracy, macro-averaged F1-score and AUC. Findings show that the simple baseline CNN outperformed the ResNet18 and EfficientNetB0 models.

Keywords: photovoltaic fault classification; convolutional neural networks; ResNet18; EfficientNetB0; class imbalance; compound scaling; residual learning

A contaminated Kumaraswamy regression model for responses on the unit interval

Ronewa Mathuba, Johan Ferreira

University of the Witwatersrand

Wednesday 25 November, 16:40 – 16:50 · Cypress · Speed session 1 ↑ programme p. 27

A contaminated Kumaraswamy distribution is proposed for modelling heavy-tailed insurance loss data. Insurance losses are frequently heterogeneous and may include observations that deviate significantly from the main data, posing challenges for conventional models such as the gamma distribution, which is commonly used in generalised linear modelling. To address this challenge, a two-component mixture distribution is introduced, comprising a reference component for typical observations and a contaminant component for mild outliers. Parameter estimation for the contaminated Kumaraswamy model is conducted using the expectation-maximisation algorithm. The contaminated Kumaraswamy distribution is then extended to a regression framework, in which the median of the re-parameterised standard Kumaraswamy distribution serves as the location parameter and the response variable. Covariates representing policyholder risk factors are incorporated within this regression structure. The performance of the proposed model is evaluated through simulation studies, then applied to the Danish fire loss and Belgian motor insurance datasets, and compared with the standard beta and Kumaraswamy regression models.

Keywords: beta regression; contamination; finite mixtures; EM algorithm

ABSTRACTS

Thursday, 26 November 2026

Official and business statistics

Thursday, 26 November 2026, 09:40 – 10:40 · Weaver · Chair: Oabona Machete

Optimizing household energy production, distribution, and consumption for sustainability: A data-driven approach to renewable energy integration (net metering), grid optimization, and energy demand forecasting using time-series analysis and machine-learning models

Hausitoe Nare

National University of Science and Technology; Bulawayo; Zimbabwe

Thursday 26 November, 09:40 – 10:00 · Weaver · Official and business statistics ↑ programme p. 28

This research investigates optimizing energy consumption and production patterns in Zimbabwe, focusing on household energy usage, renewable energy integration through net metering, and energy demand forecasting. The study analyses household energy consumption patterns using time-series analysis, revealing seasonal fluctuations influenced by socioeconomic factors. These patterns offer opportunities for optimizing energy management, such as through load-shifting techniques. The effectiveness of net metering for solar energy integration was also assessed. Findings show that solar systems can provide significant financial savings for households within a reasonable payback period. Solar adoption is seen as a viable solution to Zimbabwe's energy deficit, though challenges like high initial costs and grid connection limitations remain. For energy demand forecasting, machine-learning models including Random Forest, Gradient Boosting, XGBoost, CatBoost, and stacked ensemble models were tested. The stacked ensemble model emerged as the best-performing model, showing its potential for improving grid management and reducing load shedding. The research concludes that optimizing energy consumption through demand-side management and promoting solar energy integration are key strategies for addressing Zimbabwe's energy challenges. Additionally, the use of the stacked ensemble model for accurate demand forecasting can improve grid performance. The study recommends further exploration of real-time forecasting and policy incentives to encourage solar adoption and enhance energy supply management.

Keywords: Demand forecasting; renewable energy integration

Women's labour-force participation in Botswana: A multivariable analysis of the 2022 Population and Housing Census

Oabona Machete

Statistics Botswana

Thursday 26 November, 10:00 – 10:20 · Weaver · Official and business statistics ↑programme p. 28

Women's participation in the labour force is an important dimension of economic inclusion and gender equality. This study examines demographic, socioeconomic and geographic factors associated with women's labour-force participation in Botswana using data from the 2022 Population and Housing Census.

Women aged 15 - 64 years were included. Labour-force participation was derived from census questions on main activity during the seven days preceding the census and activity among women who were not working. Women who were working or actively seeking work/business were classified as being in the labour force, while those not working and not actively seeking work were classified as outside the labour force. Descriptive analysis was followed by multivariable binary logistic regression to estimate adjusted associations with selected characteristics. Age was modelled nonlinearly using a natural spline after comparison with a linear specification.

Among 725,725 women with a stated labour-force status, 61.6% were in the labour force. Participation varied substantially by age, education, marital status and locality. The final multivariable model included 691,424 women. Compared with women with primary or below education, the odds of labour-force participation were higher among women with secondary (OR=1.70), certificate (OR=1.96), diploma (OR=2.90) and degree/postgraduate education (OR=1.97). Married women had lower odds of participation than never-married women (OR=0.71, 95% CI: 0.70–0.72). Compared with rural women, odds were higher among women in urban villages (OR=1.27, 95% CI: 1.26 - 1.29) and towns (OR=1.59). The nonlinear age specification provided substantially better fit than a linear age specification (Δdeviance=82,091; $p<0.001$). The findings demonstrate substantial differences in women's labour-force participation and provide nationally comprehensive evidence to inform policies promoting women's economic participation and gender equality in Botswana.

Keywords: Women's labour-force participation; Multivariable logistic regression; Official statistics; Population and Housing Census; Gender equality

Assessing the effect of sample size on supervised and voting ensemble learning models for binary classification of hiring decisions

Keketso Ashley Baboile, Tshegofatso Botlhoko

North-West University

Thursday 26 November, 10:20 – 10:40 · Weaver · Official and business statistics ↑programme p. 28

Binary classification problems continue to attract significant research attention, particularly in assessing whether ensemble learning techniques can enhance the performance of individual supervised learning models. This study compared the performance of supervised machine learning techniques such as Adaptive Linear Neuron (ADALINE) and Extreme Learning Machine (ELM) with voting ensemble learning techniques such as Hard Voting (HV) and Soft Voting (SV) in classifying hiring decisions using the recruitment dataset. The models were tested on four different sample sizes: 25%, 50%, 75%, and 100% of the recruitment dataset. ADALINE, ELM, and the ensemble learning techniques, HV and SV, were analysed using classification accuracy, recall, precision, and F1-score. The overall dataset consisted of 1 500 observations with 10 independent variables and a binary dependent variable of hiring decisions, indicating hired or not hired. The results of this study demonstrated that while both ADALINE and HV performed very well, SV was the overall best-performing model. ELM only offered maximum performance on the 25% sample size. From this study, it is recommended that organisations consider implementing the best-performing model as decision-support tool to enhance objectivity in hiring, while regularly monitoring model performance and fairness to ensure ethical and reliable outcomes.

Keywords: Adaptive Linear Neuron (ADALINE); Extreme Learning Machine (ELM); Hard Voting (HV); Soft Voting (SV)

Financial statistics 2

Wednesday, 25 November 2026, 15:40 – 17:00 · Yellowwood · Chair: Energy Sono

Redundancy and synergy in an emerging equity market: A mutual information and partial information decomposition approach to the JSE40

Palesa Kale, Fabio M. Correa

University of the Free State

Wednesday 25 November, 15:40 – 16:00 · Yellowwood · Financial statistics 2 ↑ programme p. 28

Financial markets exhibit complex dependence structures that may not be adequately captured by conventional linear correlation measures. This study investigates nonlinear and higher-order dependencies among constituents of the FTSE/JSE Top 40 index using a combined framework of Mutual Information (MI), Minimum Redundancy Maximum Relevance (mRMR), and Partial Information Decomposition (PID). Daily data for 34 JSE Top 40 stocks from 2018 to 2026 were analysed. Mutual information was estimated using a k-nearest-neighbour approach, while mRMR was used to identify informative and non-redundant predictors for XGBoost return forecasting. PID was applied to 17,952 stock triplets to distinguish between redundant and synergistic information and to examine whether these dependencies were concentrated within or across economic sectors. The results show that informational dependence in the JSE Top 40 is predominantly redundant, with 79.81% of the analysed triplets classified as redundancy-dominant. However, redundancy was not concentrated within individual sectors, while synergy was concentrated across sectors, suggesting the influence of market-wide common factors. Despite identifying substantial statistical dependence, neither the XGBoost models nor the AR(1) benchmark achieved directional forecasting accuracy meaningfully above chance at daily or weekly horizons. Furthermore, out-of-sample portfolio comparisons showed that the classical Markowitz minimum-variance portfolio outperformed the MI-based portfolio strategies in terms of Sharpe ratio and drawdown. These findings demonstrate that MI, mRMR, and PID provide a useful framework for characterising complex dependence structures in an emerging African equity market, while highlighting that statistical dependence does not necessarily translate into improved forecasting or portfolio performance.

Keywords: Mutual Information; mRMR; Partial Information Decomposition; Nonlinear Dependence; Financial Markets; Portfolio Optimisation

A regime-dependent structural-deep framework for probabilistic financial forecasting: Formulation and analysis

Energy Sonono

North-West University

Wednesday 25 November, 16:00 – 16:20 · Yellowwood · Financial statistics 2 ↑ programme p. 28

Financial returns often exhibit nonlinear, time-varying, and regime-dependent behaviour, particularly during periods of market stress, uncertainty, and structural change, which can challenge forecasting approaches based on time-invariant specifications. This study proposes a regime-dependent structural-deep framework for probabilistic financial forecasting that combines Bayesian structural time-series modelling with an LSTM using Monte Carlo dropout for uncertainty quantification. The framework is formulated in a one-step-ahead setting that avoids look-ahead bias and focuses on predictive distributions rather than point forecasts alone. The analysis investigates whether predictive estimates from the Bayesian structural time-series and LSTM components provide complementary information, and how the relative accuracy and dependence structure of their forecast errors influence potential gains from regime-dependent combination. Controlled simulation experiments are conducted under alternative data-generating processes representing changing dynamics, market regimes, and distributional conditions. Performance is assessed in terms of point-forecast accuracy, probabilistic forecast quality, calibration, and tail-risk representation.

Keywords: probabilistic financial forecasting; Bayesian structural time series; LSTM; regime-dependent forecasting; uncertainty quantification

Evaluating VaR for stock indices using a contaminated framework

Lehlonololo Mphanya, Johan Ferreira, Eden Gross

University of the Witwatersrand

Wednesday 25 November, 16:20 – 16:40 · Yellowwood · Financial statistics 2 ↑ programme p. 28

Risk measures are central to financial risk management, with Value-at-Risk (VaR) and Expected Shortfall (ES) among the most widely used. A key challenge in their estimation is that financial return data frequently exhibit outliers, which can undermine standard distributional assumptions such as normality and lead to unreliable risk estimates. This study addresses this issue by modelling returns using a two-component contamination mixture in which both components share a common mean but the second component has an inflated variance, allowing the model to accommodate mild outliers while retaining a parsimonious structure. VaR and ES are investigated under this framework, and their properties are examined relative to standard normal-based estimates. The approach is illustrated using real stock price data from the S&P 500 index, and the resulting risk estimates are interpreted in light of the Basel III regulatory framework, highlighting the practical implications of accounting for outlier contamination in capital adequacy calculations.

Keywords: contamination; S&P 500; risk measures

Environmental statistics

Thursday, 26 November 2026, 09:40 – 10:40 · Cedarwood · Chair: Tyriq Emmanuel Pillay

Modelling seasonal compound extreme risk using vine copula towards enhancement of renewable energy production

Maashele Kholofelo Metwane, Daniel Maposa, Caston Sigauke

University of Limpopo; University of Venda

Thursday 26 November, 09:40 – 10:00 · Cedarwood · Environmental statistics ↑ programme p. 28

Renewable energy sources, such as solar photovoltaic and wind, are heavily dependent on weather conditions and are greatly influenced by climate extremes and weather fluctuations. However, concurrent extreme events and seasonal variations constrain the ability of solar and wind systems to generate renewable energy independently, potentially disrupting production and increasing risk. Consequently, identifying periods when compound climate extreme events occur simultaneously and understanding their joint occurrence could inform strategies to boost renewable energy production and enhance operational planning, resource expansion, maintenance and storage scheduling, and grid integration. Traditional multivariate distributions, such as the elliptical and Archimedean families, assume limiting dependence structures for meteorological variables but often fail to capture the dependence structure of compound extremes and associated risk assessments. These dependencies often exhibit asymmetric dependence, non-linear relationships, and tail dependence. Vine copulas can break down complex high-dimensional dependence into flexible pairwise copulas. The study used a vine copula to model and evaluate the risk associated with seasonal fluctuations in solar and wind renewable energy production, using four meteorological variables: temperature, humidity, wind speed, and wind gust. The study developed two subsets of meteorological variables: Y1 (temperature, wind speed, relative humidity) for wind power, and Y2 (temperature, relative humidity, wind gust) for solar photovoltaic and grid resilience. These two subsets of meteorological variables are applied to two clusters in the Eastern Cape province of South Africa, distinguished by onshore and inland locations based on elevation. The analysis showed that the canonical vine (C-vine) copulas successfully captured the multivariate dependence structure in both solar and wind energy production scenarios. The results showed that solar photovoltaic systems generally generate more electricity during weather conditions or seasons of high solar irradiance, while wind turbines tend to produce more power during weather conditions or seasons of lower solar irradiance. The findings suggest that evaluating solar and wind energy resilience and operational risk necessitates considering compound meteorological variables rather than individual variables. The findings provide essential insights for energy policymakers, showing the importance of considering joint concurrent extremes and building hybrid infrastructure that can withstand concurrent extreme weather while still generating power.

Keywords: Climate extremes; compound extremes; meteorological variables; renewable energy; seasonal variation; vine copulas

A loggamma generalised linear model for NO2 emissions data from South Africa's Eskom's coal-fired power stations when the data are non-normal and the variance is non-constant

Mpendulo Wiseman Mamba, Delson Chikobvu

Central University of Technology; University of the Free State

Thursday 26 November, 10:00 – 10:20 · Cedarwood · Environmental statistics ↑ programme p. 28

The aim of this paper is to determine if the Loggamma distribution model in a Generalised Linear Model (GLM) setup is a better model than the traditional simple linear regression model and the Lognormal-based GLM when fitted to nitrogen dioxide (NO2) emissions data generated during the production of electricity from 13 Eskom's coal-fuelled power stations in South Africa. The variables explaining the NO2 emissions data are selected using backward stepwise variable selection techniques. The variables considered include the power station itself, the amount of electricity generated from the power station, the age in years of the power station, the abatement technology (filter) used at the particular power station, and the month of the year. Interaction terms between the variables are also considered. The maximum likelihood estimation (MLE) method is used to estimate parameters of the GLM, and ordinary least squares is used to estimate parameters for the regression model. The Normal, Lognormal, and Loggamma distribution models with identity link function are fitted to the NO2 emissions data. The variance of the NO2 emissions increases with mean emissions and the Loggamma model plots, and the explained variance metrics (the variance-function-based R^2 and adjusted R^2) confirm the best fit to the data over the Normally distributed regression model and Lognormal-based GLM. Thus, NO2 emissions at Eskom in South Africa can be explained and predicted by employing the Loggamma-based GLM model. The findings will assist in providing information for the development of effective strategies for mitigating air pollution and promoting sustainable practices within the energy sector in South Africa.

Keywords: NO2; air quality; generalised linear model; Loggamma distribution; linear regression

Performance of static and dynamic occupancy models

Tyriq Emmanuel Pillay, Greg Distiller, Res Altwegg

University of Cape Town

Thursday 26 November, 10:20 – 10:40 · Cedarwood · Environmental statistics ↑ programme p. 28

Occupancy models are used in ecology to model the distribution of species, accounting for imperfect detection. Static occupancy models analyse data from a single time period, whereas dynamic occupancy models consider multiple time periods by explicitly modelling colonisation and local extinction between seasons. Using simulated datasets, we compare the performance of static and dynamic occupancy models across three areas: ecological factors such as initial occupancy and detection probability, sampling conditions such as site imbalance and incomplete seasonal coverage, and environmental variables driving non-equilibrium occupancy dynamics. Dynamic occupancy models outperformed static occupancy models in both precision of estimates and prediction of true occupancy, showing less bias, particularly at lower detection probabilities. Dynamic occupancy models were also robust to imbalance in the number of site visits and tolerated skipped survey seasons provided enough seasons were sampled. Static occupancy models capture a "snapshot" of occupancy in an environment, whereas dynamic occupancy models, through their colonisation and extinction rates, allow us to examine how occupancy responds to changing environmental conditions; static models were consequently less reliable than dynamic models when species were not in equilibrium with their environment. We applied the insights gained from the simulations to real detection/non-detection data from the Southern African Bird Atlas Project (SABAP2), focusing on the Cape Rockjumper (Chaetops frenatus), a species with low detectability, comparing static and dynamic model estimates of its occupancy dynamics. These results provide guidance to ecologists and statisticians on when and how to apply static and dynamic occupancy models.

Keywords: Occupancy Modelling; Detection Probability; Species Distribution; Ecological Statistics

Sport statistics

Thursday, 26 November 2026, 09:40 – 10:40 · Cypress · Chair: Paul van Staden

A mathematical model to select medley relay teams in competitive swimming

Paul J van Staden

University of Pretoria

Thursday 26 November, 09:40 – 10:00 · Cypress · Sport statistics ↑ programme p. 28

This study extends research in the selection of relay teams in sport, with specific focus on identifying the fastest possible teams for 4 x 100 meter medley relays in competitive swimming. In a medley relay, each of the four swimmers is assigned one of the four different strokes in a fixed sequence: backstroke, breaststroke, butterfly, and finally freestyle. Traditional medley relay teams comprise either four female or four male swimmers, whereas a mixed-gender medley relay team has two female and two male swimmers, who may be assigned to the four strokes in any order. The proposed mathematical model uses an integer optimisation routine to select the medley relay teams based on the eligible swimmers' seasonal-best or personal-best times. The model is illustrated using data for South African swimmers who competed at the 2026 Commonwealth Games.

Keywords: Integer linear programming; Optimisation; Swimming; Team selection

An integer linear programming approach to team selection for the mixed medley relay

David Forbes, Gary Sharp

Nelson Mandela University

Thursday 26 November, 10:00 – 10:20 · Cypress · Sport statistics ↑ programme p. 28

South African Masters Swimming (SAMS) competitions require that relay teams be selected from swimmers with varying performance levels and ages while abiding by the gender, stroke and cumulative age requirements of the specific event. This research develops an integer linear programming approach for the selection of optimal teams by minimising the difference between team relay times and the relevant national records for a club competing in the mixed medley relay event at a SAMS competition. The model is formulated subject to the full set of event constraints and applied to SAMS national records and the best times of competitors from a South African club, providing a practical tool for relay team selection under these constraints.

Keywords: Linear Programming; Team Selection; Integer Programming; Masters Swimming

Sports analytics: Defining a performance measure for a cricket batsman

Justin Friend

Nelson Mandela University

Thursday 26 November, 10:20 – 10:40 · Cypress · Sport statistics ↑ programme p. 28

The International Cricket Council (ICC) uses a proprietary methodology for ranking T20I batsmen. These rankings are not always reflected in the ranked orderings produced by traditional batting performance metrics. This paper investigates which commonly reported measures of batting performance exhibit the strongest agreement with the ICC rankings. Various methods of ranking batsmen are proposed, including a newly defined metric, each demonstrating varying levels of agreement with the ICC rankings. The findings are believed to be useful in promoting the development of cricket at the school level and in lower-tier leagues, where detailed performance data may not be available.

Keywords: Sports; T20 Cricket; Performance Measurement; Player's Ranking

Spatial statistics 2

Thursday, 26 November 2026, 11:00 – 12:40 · Weaver · Chair: Raeesa Manjoo-Docrat

Assessing the impact of TB vaccination strategies in South Africa: A stochastic spatial compartmental modelling approach

Raeesa Manjoo-Docrat

University of the Witwatersrand

Thursday 26 November, 11:00 – 11:20 · Weaver · Spatial statistics 2 ↑ programme p. 29

South Africa's severe tuberculosis (TB) burden, compounded by co-endemic HIV and marked socio-economic and spatial inequality, motivates the need for statistically rigorous tools to guide the roll-out of emerging TB vaccines under tight resource constraints. This study develops a stochastic, age-stratified compartmental model of TB transmission, extended to incorporate spatial heterogeneity in disease risk and vaccine impact across South Africa.

Five candidate vaccination strategies are formulated as distinct spatial allocation rules: targeting HIV-positive individuals at testing sites, targeting occupational risk groups (mining and correctional facilities), vaccinating school-leavers, targeting geographic TB hotspots, and untargeted random allocation. Each strategy is simulated stochastically under varying assumptions of vaccine efficacy and durability, with vaccinator throughput constraints built into the transition rates. Model outputs are compared against a baseline of current preventative treatment regimens to quantify relative reductions in transmission and mortality.

The core statistical contribution lies in a spatial optimisation framework layered on top of the transmission model: a stochastic optimisation algorithm accounts for spatial dependence in vaccine effectiveness across regions of differing TB burden and vulnerability, while a hierarchical model incorporates this spatial heterogeneity into cost-effectiveness estimation. Sensitivity analyses characterise how uncertainty in efficacy, durability and cost parameters propagates through both the transmission dynamics and the optimised allocation. Together, these components identify the combination of strategies and locations that maximises epidemiological impact under fixed budget and workforce constraints.

This work demonstrates how stochastic compartmental modelling can be coupled with spatial statistical optimisation to move beyond simple strategy comparison toward quantitative, uncertainty-aware resource allocation. The resulting framework is generalisable to other vaccine-preventable diseases and resource-constrained settings, and contributes to the growing methodological toolkit for spatial and stochastic modelling in public health decision-support.

Keywords: spatial modelling; infectious diseases; applied statistics

The impact of spatial connectivity structures on model fit, inference, and data integration in spatial epidemiology and wastewater-based epidemiology: A systematic review

Claris Siyamayambo, Inger Fabris-Rotelli

University of Pretoria

Thursday 26 November, 11:20 – 11:40 · Weaver · Spatial statistics 2 ↑programme p. 29

Spatial connectivity structures define how geographic units are considered neighbours. Neighbourhood relationships may be based on contiguity, distance, k-nearest neighbours, or network-informed weights. These structures are fundamental to classical spatial models such as CAR/BYM, SAR, and GWR. Such models are widely used in spatial epidemiology and wastewater-based epidemiology (WBE). The choice of structure is rarely treated as a modelling decision requiring justification, comparison, or sensitivity testing. This review systematically searched peer-reviewed and grey literature (January 2000–May 2026) for studies applying CAR/BYM, SAR, GWR, or SPDE-based models. Eligible studies compared two or more distinct spatial connectivity structures within the same modelling framework. Findings were synthesised through quantitative frequency mapping, a logistic-regression-based effect-modification analysis, and a methodological-quality assessment. A total of 54 studies met eligibility. Contiguity/adjacency-based structures dominated (87.0% of studies), while distance-based, mobility-derived, and hybrid structures were used mainly as secondary comparators. Changing the connectivity structure altered the model fit in the large majority of studies that tested it (74.1% favoured a novel structure over its conventional comparator). Effect-modification analysis showed that studies outside epidemiology with public health and biostatistics were more likely to favour novel connectivity structures (OR = 8.22, 95% CI: 1.55–68.56, p = 0.024). However, the association did not remain significant after correction for multiple comparisons. Only 42.6% of studies quantified how connectivity choice affected parameter estimates, and just 14.8% explicitly resolved spatial data misalignment. Critically, none of the 54 studies integrated WBE data using a connectivity structure reflecting sewer or hydrological topology. Spatial connectivity structure choice is not a neutral technical detail which measurably affects model fit, parameter estimates, and inference, particularly with incomplete or misaligned data. The review proposes evidence-informed, quality-graded recommendations for selecting and reporting connectivity structures, and identifies wastewater-clinical data integration via topologically-informed connectivity as a substantially unaddressed methodological gap with direct relevance to public health surveillance.

Keywords: Spatial connectivity structures; spatial epidemiology; wastewater-based epidemiology; misaligned data; model fit; inference

Spatial Bayesian clustering of voting patterns

Nontembeko Dudeni-Tlhone, Jenny Holloway, Renee Koen, Tyrone Naidoo, Mpho Mokoatle

Council for Scientific and Industrial Research, Pretoria

Thursday 26 November, 11:40 – 12:00 · Weaver · Spatial statistics 2 ↑ programme p. 29

Voting patterns may show spatial variation, reflecting differences in the social, demographic and political characteristics of communities. This study develops and applies a Bayesian spatial clustering framework to identify geographically similar groups of voting districts (VDs), with similar party vote-share compositions. Voting outcomes are treated as compositional data, where the proportions allocated to multiple political parties are jointly modelled and constrained to sum to one. Empirical Bayes smoothing was applied to the party vote shares to stabilise party vote-share arising from small voting districts, and improve the reliability of subsequent clustering. The clustering framework is built on a Bayesian Dirichlet mixture model for compositional data, in which each VD is assumed to belong to an unobserved cluster characterised by a cluster-specific Dirichlet distribution. The model estimates the concentration parameters of these distributions while accounting for uncertainty in cluster membership. Markov Chain Monte Carlo methods, combining Gibbs sampling and Metropolis-Hastings updates, are used to obtain posterior distributions for the model parameters and cluster allocations. To incorporate geographic structure, a spatial dependence component based on Queen contiguity between neighbouring VDs is introduced into the cluster-allocation process. This enables geographically adjacent VDs with similar voting compositions to have a greater probability of belonging to the same cluster while retaining the compositional information contained in the voting behaviour. The approach provides both probabilistic and spatially interpretable clusters, allowing voting patterns to be examined beyond individual party performance. The resulting clusters can be characterised according to their posterior party compositions and spatial distribution, providing a basis for understanding the spatial structure of political support. The methodology demonstrates the potential of Bayesian spatial compositional clustering as a flexible framework for analysing complex, multi-party electoral patterns. The resulting clusters provide a basis for extrapolating observed voting patterns to predict the overall election outcome.

Keywords: Spatial clustering; Bayesian clustering; compositional data; spatial dependence; voting patterns; election prediction

Extending cross-type K-functions for polygonal and linear features: Applications to natural hydrogen in Mpumalanga, South Africa

Calvin Botha, Ansie Smit, Adam Bumby, Inger Fabris-Rotelli, Najmeh Nakhaeirad

University of Pretoria

Thursday 26 November, 12:00 – 12:20 · Weaver · Spatial statistics 2 ↑programme p. 29

Many spatial features are observed as extended geometries rather than as points, yet classical second-order spatial methods are primarily formulated for point patterns. This study develops K-CAP (*K*-functions with Controlled Allocation of Points), a framework for extending cross-type *K*-function methodology to interaction between bounded polygonal and linear spatial features. Each parent feature is represented using the same fixed number of geometrically determined points, retaining information on feature geometry while preventing larger polygons or longer linear features from contributing more heavily solely because of their size. Four complementary components are developed. K_{shape} measures global boundary-based polygon-line interaction, $K_{\text{direction}}$ uses an axis-based polygon representation to investigate directional structure, K_{polygon} decomposes the interaction according to individual polygonal features, and $K_{\text{lineament}}$ provides the corresponding decomposition for individual linear features. Together, these measures allow both global spatial structure and feature-level heterogeneity to be investigated across a range of distances. A simulation study is used to assess the sensitivity of the framework to the number of representative points allocated to each feature. The methodology is applied to mapped sub-circular depressions and geological lineaments in Mpumalanga, South Africa, motivated by the investigation of spatial structures associated with possible natural hydrogen occurrences. The application demonstrates how different geometric representations and parent-level decompositions can reveal complementary aspects of polygon-line interaction that would be lost by reducing extended features to single locations. K-CAP therefore provides a general framework for extending established second-order spatial methodology while retaining both the geometry and identity of the observed parent features.

Keywords: Cross-Type K-Functions; Natural Hydrogen; Polygon-Line Interaction; Spatial Statistics

The fuzzy lattice catchment areas for diarrhoea mapping

Amy Pieters, Inger Fabris-Rotelli, Renate Thiede, Rene Stander, Kabelo Mahloromela

University of Pretoria

Thursday 26 November, 12:20 – 12:40 · Weaver · Spatial statistics 2 ↑programme p. 29

One of the leading causes of death in developing countries throughout the world is diarrhoea. This burden is seen more so in women and children. Novel methodology to map disease origination when data exists in aggregated forms is presented. This methodology builds on the fuzzy lattice catchment area framework laid out. We develop a spatial approach for estimating diarrhoeal risk under the uncertainty that aggregated data creates by combining the fuzzy lattice catchment areas with kriging techniques. Then applying geographically weighted regression methodologies. Specific attention is given to how the risk changes over age categories, genders and wards in the city of Cape Town municipality.

Keywords: GWNBR; Kriging; diarrhoea

Emerging topics in biostatistics 4

Thursday, 26 November 2026, 11:00 – 12:40 · Yellowwood · Chair: Din-Geng Chen

ML-validated tensor RPCA for high-dimensional medical image data analysis

Ding-Geng Chen, Habte Tadesse Likassa

University of Pretoria; Arizona State University, Phoenix, AZ, USA

Thursday 26 November, 11:00 – 11:20 · Yellowwood · Emerging topics in biostatistics 4 ↑ programme p. 29

High-dimensional medical images contain complex anatomical structures, noise, artifacts, and strong inter-channel correlations that conventional matrix-based RPCA may fail to preserve. To address these challenges, we propose a Tensor Robust Principal Component Analysis (T-RPCA) method based on a Truncated Weighted Tensor Nuclear Norm (TWNN) for AI/ML-driven high-dimensional medical image enhancement and analysis. The proposed formulation is developed as an optimization problem that decomposes medical images into low-rank structural and sparse components while preserving multidimensional anatomical information. CLAHE and Tucker-SVD are incorporated to enhance image representation, while the resulting optimization problem is efficiently solved using the Alternating Direction Method of Multipliers (ADMM). Experiments on retinal and brain imaging datasets demonstrate improved reconstruction, structural preservation, and robustness under different levels of corruption compared with conventional RPCA and existing learning-based methods. Furthermore, T-RPCA-enhanced images provide improved feature representations for downstream machine learning, supporting more reliable classification and computer-aided diagnosis. These results demonstrate the potential of the proposed optimization based T-RPCA method as an effective and interpretable preprocessing and feature-enhancement approach for AI-driven medical image analysis.

Keywords: Intelligent Image Analysis; Medical Imaging enhancement; Principal Component Analysis; ML/AI

High-dimensional likelihood ratio tests for equality of means, outliers and profile analysis applied to the analysis of TB data

Carlos A. Coelho

NOVA University of Lisbon, Portugal

Thursday 26 November, 11:20 – 11:40 · Yellowwood · Emerging topics in biostatistics 4 ↑ programme p. 29

LRTs developed for the high-dimensional case, that is, for situations where the sample size is smaller than the number of variables, are applied to the analysis of TB data to test for the presence of outliers, for equality of mean vectors among populations and to make a full profile analysis, testing for profile parallelism, profile coincidence and possible flatness of the population profiles.

Keywords: High-dimensionality; Likelihood Ratio tests; Equality of mean vectors; Outliers; Profile Analysis

Bayesian assurance for diagnostic accuracy studies: An application to the SAHeart dataset

Lizelle Venter, Inger Fabris-Rotelli, Ding-Geng Chen

University of Pretoria

Thursday 26 November, 11:40 – 12:00 · Yellowwood · Emerging topics in biostatistics 4 ↑ programme p. 29

Determining an appropriate sample size for diagnostic accuracy studies is challenging because the precision of sensitivity and specificity estimates depends on several factors, including disease prevalence, prior information, and the choice of diagnostic threshold. Bayesian assurance provides a useful framework for study planning by expressing design performance as the probability that a study will achieve a prespecified level of precision before the study is conducted. In this work, Bayesian assurance methodology is applied to the SAHeart dataset to investigate the effect of threshold specification on the design and evaluation of a diagnostic accuracy study. Two approaches are considered: a fixed-threshold method, in which the diagnostic threshold is treated as known in advance, and a variable-threshold method, in which uncertainty associated with threshold selection is incorporated into the analysis. Sensitivity and specificity are used as the primary measures of diagnostic performance, with study success defined in terms of achieving prespecified posterior credible interval widths. The two approaches are applied to the SAHeart dataset to examine how accounting for threshold uncertainty influences assurance and the resulting sample size requirements. Sensitivity analyses are further conducted to assess the robustness of the results to assumptions regarding prior information, prevalence, threshold selection, and design criteria.

Keywords: Bayesian assurance; Diagnostic accuracy studies; Sample size determination; Variable threshold

A Bayesian joint framework for matrix-valued mediation: Application to breast cancer

Kasahun Takele, Ding-Geng Chen

University of Pretoria

Thursday 26 November, 12:00 – 12:20 · Yellowwood · Emerging topics in biostatistics 4 ↑ programme p. 29

High-dimensional radiomic features taken from medical images are being used more and more in the diagnosis of breast cancer. However, because of their dimensionality and correlation, these features, which are frequently naturally structured as matrices, present considerable difficulties for conventional mediation analysis. In this paper, we proposed a Bayesian joint framework for mediation analysis with matrix-valued mediators. The joint model combines a logit model for binary outcomes, a mediator model on the latent features, and probabilistic multilinear principal component analysis for dimension reduction. The identification of active mediation indicators and parameter estimation are done simultaneously using a Gibbs sampling algorithm with Varimax rotation. Simulation studies show that, especially in low-dimensional settings with small sample sizes, the joint Bayesian method produces more accurate and stable estimates of natural direct (NDE), indirect (NIE), and total effects (TE) than a two-step method. The joint model approach effectively identified important morphological characteristics, such as the worst and mean values of radius, perimeter, and area, as the most significant mediators in the pathway from tumour size to recurrence when applied to the Wisconsin Prognostic Breast Cancer (WPBC) dataset. Based on the results obtained from the Bayesian joint modeling, we recommend that the method provides a consistent and insightful approach for identifying complex causal relationships in high-dimensional biological data.

Keywords: Breast; Cancer; Joint; MPCA; Varimax

Testing before the plateau in mixture cure models

James Allison, I. Van Keilegom

University of South Africa; KU Leuven, Belgium

Thursday 26 November, 12:20 – 12:40 · Yellowwood · Emerging topics in biostatistics 4 ↑ programme p. 29

Mixture cure models allow for individuals who will never experience the event of interest. Checking the assumed lifetime distribution among susceptible individuals can be difficult when follow-up ends before a clear plateau appears in the survival curve. We propose a goodness-of-fit test that estimates the susceptible proportion and the distribution parameters jointly from the censored likelihood. The fitted event-time sub-distribution is then compared with its Kaplan–Meier estimate. The test assesses fit over the observed follow-up period and does not require a plateau to estimate the susceptible proportion. We discuss its asymptotic properties and bootstrap implementation. Monte Carlo comparisons examine size and power under both sufficient and insufficient follow-up. We apply the test to relapse data from 4,028 children in the third and fourth National Wilms Tumour Studies.

Statistics in education

Thursday, 26 November 2026, 11:00 – 12:40 · Cedarwood · Chair: Rene Stander

Early prediction of student success for a South African higher education institution

Tayla Wannenberg

University of the Western Cape

Thursday 26 November, 11:00 – 11:20 · Cedarwood · Statistics in education ↑ programme p. 29

Academic success remains a challenge for many students in South African universities, particularly in relation to passing modules, completing their studies on time, and progressing to graduation. An important step towards addressing these challenges is the timely identification of students who may be at risk of academic failure and an understanding of the factors contributing to that risk. This study explores whether academic success can be predicted before students begin university. More specifically, it aims to develop a statistical model using application data to estimate the likelihood of first-year academic success for students entering the Faculty of Natural Sciences at the University of the Western Cape. The study will use statistical learning techniques applied to historical student data, including high school marks, demographic characteristics, and registration information. As the target variable exhibits a slight imbalance between students who succeed and those who do not, techniques such as stratified sampling, and resampling methods including Synthetic Minority Over-sampling Technique (SMOTE) will be considered to reduce the potential effect of class imbalance on model performance. Model performance will be evaluated using classification metrics including but not limited to accuracy, recall, precision, F1-score, and the P4-metric. The early prediction of academic success may assist the faculties with admission decisions and, more importantly, enable earlier and more targeted support for students who are likely to experience academic difficulties. By identifying both students at risk and the factors associated with academic success or failure, the study aims to support a more proactive and data-driven approach to improving student success in South African higher education.

Keywords: Predicting academic outcomes; educational data analysis; demographic student profiling; early intervention strategies; improving student throughput

Efficacy of a student-led teaching intervention in mathematical statistics, framed by Fink's Taxonomy of Significant Learning

Morné Sjölander

University of the Free State

Thursday 26 November, 11:20 – 11:40 · Cedarwood · Statistics in education ↑ programme p. 29

This study investigates a novel pedagogical intervention implemented within an undergraduate Mathematical Statistics module (STSM3714: Statistical Inference) grounded in Fink's Taxonomy of Significant Learning. The intervention shifts from traditional lectures to student-led instruction by assigning randomized student groups to present core topics following lecturer orientations and tutorials.

Data collection utilizes a dual-instrument framework:

1. A pre- and post-module Likert scale questionnaire evaluating shifts in student perceptions across Fink's six relational dimensions (Foundational Knowledge, Application, Integration, Human Dimension, Caring, and Learning How to Learn), analysed via paired t -tests. 2. A formative "traffic light" self-assessment system (Green, Yellow, Red) completed on our online learning management system after each session to evaluate immediate comprehension and presentation quality, analysed using AI-assisted thematic analysis.

By examining how student-led presentations foster significant learning with regards to the specific taxonomic facets, this study aims to advance teaching practices in mathematical statistics education.

Keywords: Fink's Taxonomy; Significant Learning; Pedagogical Intervention; Student-Led Presentations

From AI use to intellectual independence: Rethinking PhD supervision in the generative AI era

Danielle Roberts, Michael von Maltitz, Inger Fabris-Rotelli, Ansie Smit, Sonali Das, Daniel Maposa

University of KwaZulu-Natal; University of the Free State; University of Pretoria; University of Limpopo

Thursday 26 November, 11:40 – 12:00 · Cedarwood · Statistics in education ↑ programme p. 29

The increasing attraction of industry careers, partly driven by the Fourth Industrial Revolution, has led many Statistics graduates to forgo postgraduate study. This has contributed to a shortage of PhD-level supervision expertise in Statistics departments across South Africa and reduced opportunities for mentorship from senior academics as many experienced professors approach retirement. In response, the authors developed a guiding rubric to support early-career PhD supervisors in Statistics. The rubric emerged from a multi-year collaborative programme informed by doctoral supervisors across South African Statistics departments and refined through scholarly engagement. The rubric is a structured, developmental framework guiding supervisors and students throughout the PhD journey. Rather than focusing only on the final thesis, it emphasises the development of an independent researcher through professional growth, research integrity, and critical thinking. It also promotes transparency and consistency in doctoral supervision practices. The emergence of Generative Artificial Intelligence (GenAI) has introduced new challenges and opportunities for postgraduate supervision. AI tools can now assist with literature reviews, coding, statistical analysis, and academic writing. While these technologies may enhance learning and productivity, they raise questions about cognitive engagement, ethical AI use, and the extent to which students genuinely understand and can defend their research. Central to these concerns is the question: Who is doing the thinking? This presentation describes how the original rubric has evolved to support the responsible and transparent integration of GenAI into PhD supervision. The revised rubric moves beyond AI detection or prohibition and instead places AI literacy, critical reflection, verification, and scholarly judgement at the centre of postgraduate research training. Although developed within the context of Statistics in South Africa, the rubric offers a broader framework for higher education and provides a practical model for navigating supervision and academic integrity in an increasingly AI-mediated academic environment.

Keywords: GenAI; PhD supervision; Guiding rubric; Independent researcher

Empowering learners to thrive in a data-driven future

Delia North

University of Kwa-Zulu Natal

Thursday 26 November, 12:00 – 12:20 · Cedarwood · Statistics in education ↑ programme p. 29

In September 2026, Department of Basic Education subject advisors and lead teachers from districts around the country, attended a ground breaking 2-day workshop, at the Garden Court Hotel in Johannesburg. The workshop aimed to shed light on strategies to enhance Statistics and Data Science education within school systems, even in schools where technological infrastructure is limited, or where these subjects are not yet embedded in the curriculum. The workshop was conceptualized by Prof Delia North (UKZN), materials and presentation were by two undisputed world experts in Statistics Education, ie Prof Pip Arnold (Auckland Univ, NZ) and Prof Christine Franklyn (Univ Georgia, USA). Key areas of focus included

• teacher development - identifying the types of support required to build both pedagogical and content knowledge in statistics and data science for educators working with school-aged learners; • student learning experiences - designing accessible and meaningful learning opportunities that help students grasp foundational concepts in statistics and data science, even in environments where these subjects are underrepresented or absent; • and engagement through data investigating practical and efficient methods for involving both teachers and students in data collection and visualization activities. These experiences should foster statistical reasoning and thinking, and data literacy

This talk aims to give an overview of the workshop, so as to share interesting ways to enhance Statistics Education amongst the youth of South Africa, that should be of interest to Statistics Departments from around the country

Keywords: education

Telling a better story about statistics: Tackling anxiety amongst business students

Rene Stander

University of Pretoria

Thursday 26 November, 12:20 – 12:40 · Cedarwood · Statistics in education ↑ programme p. 29

Statistics anxiety is a commonly observed challenge among undergraduate students, especially those encountering the subject for the first time. It is associated with low motivation, disengagement, and poor academic performance, and is often rooted in negative experiences from prior schooling. These early experiences can create a cycle of avoidance, and students' attitudes toward statistics are known to strongly predict their eventual achievement in the subject. One promising way to address this is through storytelling. Narratives can make abstract concepts more accessible and easier to remember, and have been used specifically to reduce anxiety in mathematically demanding subjects.

This research proposes to explore storytelling as a pedagogical tool to enhance statistical learning and make key concepts easier to understand and remember. Using generative AI, the intention is to develop a narrative theme running through each chapter, with contextualised stories built around individual statistical concepts. The study examines the extent to which this approach influences students' anxiety, attitudes, and academic performance in the subject.

Keywords: statistics anxiety; pedagogical storytelling; statistics education; undergraduate

Speed session 2

Thursday, 26 November 2026, 11:00 – 12:20 · Cypress · Chair: Inger Fabris-Rotelli

Spatial linear network regression of crime data

Karabo Matjila, Kabelo Mahloromela, Inger Fabris-Rotelli

University of Pretoria

Thursday 26 November, 11:00 – 11:10 · Cypress · Speed session 2 ↑ programme p. 29

Crime events often exhibit complex spatio-temporal patterns, whereby the occurrence of one incident may influence the likelihood of subsequent incidents. Moreover, the spatial structure of an urban environment may affect the distribution and dynamics of crime. Herein, we investigate the use of Hawkes processes to model crime dynamics within an urban road network in the Hatfield CID. Hawkes processes are particularly suited to this setting as they allow previous events to temporarily increase the likelihood of subsequent events, capturing the self-exciting nature of crime occurrences.

The analysis is developed progressively through three models. First, a temporal Hawkes process is used to investigate whether crime incidents exhibit temporal self-excitation, allowing previous incidents to influence the occurrence of subsequent events. Second, the model is extended to a spatio-temporal Hawkes process, incorporating network-based spatial distance instead of conventional Euclidean distance to quantify spatial relationships.

Finally, a spatio-temporal Hawkes process with covariates is developed to incorporate contextual factors that may contribute to variations in crime intensity.

The progressive modelling framework allows the contribution of temporal dependence, network-based spatial dependence, and contextual factors to be investigated separately and in combination. The approach may contribute to improved understanding of crime patterns and support more informed approaches to urban safety planning and resource allocation.

Keywords: Hawkes process; Crime modelling; Point processes; Spatial networks; Covariates

Spatial variable selection in mixed graphical models: Antimicrobial resistance in KwaZulu-Natal

Marquis Pretorius, Inger Fabris-Rotelli

University of Pretoria

Thursday 26 November, 11:10 – 11:20 · Cypress · Speed session 2 ↑ programme p. 29

COVID-19 has had a profound impact on the medical world's view on global outbreaks and the most appropriate tools to manage it. It is clear post pandemic that public health officials require a tool to react predictively rather than retrospectively. Thus, time is a valuable resource in any crisis, as any delay in reaction from the government and public health sector can lead to the death of many. For this exact reason we want to focus on a well-established tool within this field. Water-based epidemiology (WBE) is a surveillance tool that provides public health officials with the framework to both detect and track infectious markers that were shed by individual that are infected. This in turn allows public health officials the ability to monitor and manage the spread of disease within a country. Spatial modelling allows this information that was received from water-based epidemiology to both be visualised and quantified which will in turn assist medical officials to complete three tasks. These tasks include tracking the spread of a disease, managing that spread and allocating resources to the communities most in need. Furthermore, Mixed graphical models will allow us to visualise and understand the relationships between covariates. In this report we aim to identify and visualise the relationships between covariate through the application of mixed graphical models. Furthermore, we aim to investigate how these covariates are influenced by a spatial factor. This will be achieved through the use of spatial autocorrelation methods which will further be discussed later in the article. Through the rigorous application of all of these methods together and the analysis present in the application section we aim to analyses the impact of disease spread and how factors like geographic position and demographic groups can influence this.

Keywords: Wastewater-based epidemiology; Spatial modelling; Mixed graphical models; Spatial autocorrelation; Disease surveillance

Modelling spatially misaligned wastewater data in epidemiology

Esther Mazarura, Inger Fabris-Rotelli

University of Pretoria

Thursday 26 November, 11:20 – 11:30 · Cypress · Speed session 2 ↑ programme p. 29

A link between the present health status of a given population and the presence and concentrations of pathogens (disease-causing agents) in untreated wastewater has been established through many years of research and application. This relationship found particular application during the SARS-CoV-2 pandemic of 2020, where wastewater analysis revealed spread patterns and disease hotspots, and signaled new waves before case data was even available. The purpose of this research is to demonstrate the real-life applications of these techniques and clearly show the progression and spread of SARS-CoV-2 through a spatial model, using tools to measure spatial autocorrelation between case loads in Ekurhuleni subplaces.

Keywords: Spatial modelling; disease surveillance; SARS-CoV-2; Ekurhuleni

Investigating risk prediction models for intermittent disease progression using machine learning

Thlabane Mogashoa

University of Pretoria

Thursday 26 November, 11:30 – 11:40 · Cypress · Speed session 2 ↑ programme p. 29

Intermittent chronic diseases are characterised by phases of remission alternating with relapses or progression. These conditions can be challenging to predict. Patient outcomes differ irrespective of similar clinical profiles, making clinical monitoring and intervention timing difficult. This mini-dissertation examines the use of machine learning models to predict disease progression in chronic diseases by classifying patients as either progressing or in remission. This study compares three machine learning models, specifically the support vector machine (SVM), random forest (RF), and extreme gradient boosting (XGBoost) methods. They were selected for their ability to detect complex, non-linear relationships and interactions in patient data. The model performance was evaluated through various sensitivity, specificity and accuracy measures, together with other calibration and class separation indicators. This comparative application was illustrated through the use of real multiple sclerosis and gout datasets.

Keywords: Intermittent chronic diseases; Support vector machine; Random forest; Extreme gradient boosting; Risk prediction

Spatio-temporal modelling of extreme precipitation using INLA

Akani Msimeki, Ansie Smit, Janet van Niekerk, Inger Fabris-Rotelli

University of Pretoria

Thursday 26 November, 11:40 – 11:50 · Cypress · Speed session 2 ↑ programme p. 29

In recent years, understanding extreme rainfall events in South Africa has become increasingly important due to the potential economic and societal impact associated with such events. In April 2022, heavy rainfall in KwaZulu-Natal led to flooding, which claimed the lives of over four hundred people. Infrastructure, transportation and communication lines experienced heavy damage during that period. With the impact of climate change coupled with South Africa's diverse climatological landscape, such rainfall extremes are projected to increase in the coming years. For this reason, the development of a model that can provide inferences on where extreme rainfall is likely to occur and its behaviour over time is motivated. This study proposes a three-stage Bayesian hierarchical framework for modelling rainfall extremes across South Africa. At each stage, the Gamma, Bernoulli and Generalised Pareto distributions are fit to the data. The aim of this approach is to model average rainfall occurrence, the probability of threshold exceedance and the magnitude of those exceedances. To account for the spatial and temporal effect, an additive structure is implemented where an intercept, a spatial SPDE effect and a weekly temporal effect, modelled using a second-order random walk, are incorporated. To provide approximate inference for the posterior distributions of the mean rainfall level, probability of threshold exceedance and tail index, all varying across space and time, Integrated Nested Laplace Approximations (INLA) will be utilised via the R-INLA package in R. Heat maps, Intensity-Duration-Frequency (IDF) curves and maps displaying the spatial and temporal effect of extreme rainfall are produced. The aim is to contribute to the existing literature in Extreme Value Theory while also providing a framework that can assist financial industries with assessing and mitigating rainfall risk.

Keywords: Spatial; Temporal; Precipitation; INLA; SPDE

Non-linear association joint models for clinical studies

Sable Walters-Somers, Janet van Niekerk

University of Pretoria

Thursday 26 November, 11:50 – 12:00 · Cypress · Speed session 2 ↑ programme p. 29

Joint models in a clinical context aim to simultaneously model a patient's longitudinal (time-dependent) biomarker data and their Time-To-Event (TTE) survival data. For the sake of theoretical and practical simplicity, a linear linkage function between the two processes is usually assumed. However, this may easily fail to capture the complex and often non-linear processes and mechanisms that typically characterize the human body, potentially leading to medical decisions being made under inaccurate models.

Multiple linear and non-linear joint models are fitted through Bayesian methods to data from a Mayo Clinic primary biliary cirrhosis (PBC) study. By comparing the models, it is illustrated that for this data, the use of non-linear joint models leads to meaningfully different conclusions and decisions than in the case of a simple linear joint model.

Keywords: non-linear; joint model; survival; longitudinal; bayesian

Exploring an enriched innovation intensity INAR(1) model for count data

Muano Patrick Nemasea, Johan Ferreira

University of the Witwatersrand

Thursday 26 November, 12:00 – 12:10 · Cypress · Speed session 2 ↑ programme p. 29

Count data over time frequently exhibit characteristics that deviate from the norm, including departures from assumptions such as equi-dispersion. The literature suggests that the intensity of the underlying, or rather assumed, innovation process may be more complex in itself. While existing INAR models allow for their innovation intensity to depend on past observations, comparatively little attention has been given to account for this intensity when it could depend on past innovations themselves. This study propose an enriched innovation intensity INAR(1) model in which the conditional innovation intensity depends on both the previous observed count and the previous innovation. Theoretical properties of the model are derived, including moment structures, stationarity conditions, dispersion characteristics, and autocorrelation behaviour. Estimation is performed using a recursive conditional maximum-likelihood procedure, combined with innovation reconstruction, to accommodate the latent innovation process. The finite-sample behaviour of the model is investigated through an extensive Monte Carlo simulation study. The results show that innovation feedback generates stronger persistence, increased dispersion, and greater conditional heterogeneity than conventional count-driven specifications. An empirical application to real-world data illustrates the proposed model and provides an opportunity for reflection on its added value. Overall, the proposed framework extends conventional INAR models to accommodate dynamically dependent innovations by incorporating into an enriched innovation intensity and provides new insights into latent innovation recovery, parameter estimation, and the distinction between persistence detection and persistence attribution in count-valued time series.

Keywords: Poisson; intensity; count data

Virological failure and subsequent care outcomes among people living with HIV on a dolutegravir-based regimen in South Africa: A retrospective cohort study

Mbali P. Nkosi, Yukteshwar Sookrajh, Jennifer A. Brown, Maren Roehl, Johan S. van der Molen, Kwena Tlhaku, Thokozani Khubone, Richard Lessells, Sharana Mahomed, Nigel Garrett, Lara Lewis, Jienchi Dorward

Centre for the AIDS Programme of Research in South Africa (CAPRISA), University of KwaZulu-Natal, Durban, South Africa; eThekwini Municipality Primary Healthcare Services Directorate, Durban, South Africa; Institute of Medical Virology, University of Zurich, Zurich, Switzerland; Division of Clinical Epidemiology, Department of Clinical Research, University Hospital Basel, University of Basel, Basel, Switzerland; KwaZulu-Natal Research Innovation & Sequencing Platform, University of KwaZulu-Natal, Durban, South Africa; Desmond Tutu HIV Centre, University of Cape Town, Cape Town, South Africa; Nuffield Department of Primary Care Health Sciences, University of Oxford, Oxford, United Kingdom

Thursday 26 November, 12:10 – 12:20 · Cypress · Speed session 2 ↑ programme p. 29

Introduction: Dolutegravir (DTG)-based antiretroviral therapy (ART) is recommended for first- and second-line therapy by the World Health Organization and used by over 24 million people globally. However, evidence on virological failure (VF) on first-versus second-line DTG is limited. We compared VF between groups and described post-VF outcomes.

Methods: We conducted a retrospective cohort study of people living with HIV aged ≥10 years who started first- or second-line DTG-based ART from January 2020–December 2024, with ≥9 months potential follow-up through September 2025 cutoff. VF was defined as two consecutive viral loads ≥1,000 copies/mL ≥56 days apart. Follow-up began 60 days after DTG start and continued until VF, loss to follow-up, transfer out, death, switch off DTG or the cutoff, whichever came first. We estimated VF incidence and fitted a piecewise Cox model with facility-level gamma frailty, adjusted for age, sex, DTG start year and CD4 category, with a treatment-line-by-time interaction. Among those with VF through September 2024, subsequent 12-month outcomes were assessed.

Results: Among 1,790,185 people who started DTG, median age was 37 years (IQR 30–46), 67.8% were female and 5.0% received second-line DTG. There were 28,211 VF events over 5,076,975 person-years (incidence 4.61 vs 31.3 per 1,000 person-years for first- vs second-line DTG). Adjusted VF hazards for second- vs first-line DTG were 5.44 (95% CI 5.18–5.71), 3.54 (3.31–3.78), 3.08 (2.87–3.30) and 3.02 (2.77–3.28) at 0–6, 6–12, 12–24 and >24 months. Among 20,099 people with VF by September 2024, within 12 months, 47.0% resuppressed to <1,000 copies/mL, 30.9% had ongoing viraemia and 22.1% had no subsequent viral load.

Conclusion: People receiving second-line DTG had higher VF rates, with the greatest difference early in follow-up. Following VF, ongoing viraemia and lack of subsequent viral load monitoring were common, highlighting gaps in post-VF care.

Keywords: Survival analysis; Piecewise Cox regression; Shared frailty model; Time-varying effects; Virological failure

Sampling and survey methods and methodology

Thursday, 26 November 2026, 15:10 – 16:30 · Weaver · Chair: Thomas Farrar

Multiple inverse sampling without replacement for post-stratification

Thomas Farrar

Cape Peninsula University of Technology

Thursday 26 November, 15:10 – 15:30 · Weaver · Sampling and survey methods and methodology ↑programme p. 30

Stratified random sampling (StRS) allows for improved estimation of a population quantity of interest where the population of N units can be divided into H subpopulations ('strata') with known sizes $N_h, h = 1, \ldots, H$, according to a categorical label that is related to the quantity of interest. In instances where the stratum labels are unknown prior to sampling, StRS is infeasible but post-stratification is an alternative. Simple post-stratification (SPS) involves drawing a simple random sample (SRS) of size n and then weighting each stratum's contribution to the estimator according to that stratum's subpopulation size relative to subsample size, N_h/n_h. One drawback to SPS is that empty subsamples are possible, which introduces bias. Chang et. al (1998) thus introduced multiple inverse sampling (MIS), a sampling procedure that stops only when every subsample size n_h has reached a corresponding threshold m_h, $h = 1, \ldots, H$. Their method assumes that N is large enough and n/N small enough that the finite population correction can be ignored, or equivalently, the procedure modelled as sampling with replacement. However, the empty subsample problem with SPS is at least as serious when working with small populations. This study therefore investigates MIS without replacement (MISWOR). The joint distribution of the subsample sizes under MISWOR is derived and an R function written to compute it. The MISWOR estimator is shown to be unbiased and its variance is derived. The MISWOR estimator's performance is compared to SRS, StRS, SPS, and MISWR using a simulation experiment. Ideas for future extension of the research are explored.

Keywords: multiple inverse sampling; post-stratification; hypergeometric

Multi-objective stratified sampling: A GA for optimal allocation using the survey design effect

Georgi Borros, Sebnem Er, Sulaiman Salau

SALDRU, University of Cape Town

Thursday 26 November, 15:30 – 15:50 · Weaver · Sampling and survey methods and methodology ↑programme p. 30

Multivariate stratified sample allocation involves allocating sample size n across strata with the ultimate aim of estimating population parameters for multiple study variables. Methods used to optimise this allocation procedure form a substantial body of literature, with new approaches regularly introduced. Work to date, however, generally focuses on continuous stratifiers over a single estimator type. This is problematic for survey sampling using mixed data with different estimators of interest - such as the mean, total, proportion and categorical frequency. Our work addresses this gap by building on Kish's survey design effect to develop a fitness function across estimator type(s) and variables of interest. We develop a Genetic Algorithm (GA) for the optimisation and demonstrate its performance using the GHS 2024 dataset. To benchmark our method, we use Bethel-Chromy allocation from the SamplingStrata package - widely considered as the optimal method for multivariate stratified sample allocation despite its general use over continuous stratifiers. Here, we offer insight into how the Bethel-Chromy method may be applied to mixed data types. Finally, we demonstrate that the GA performs equivalently to Bethel-Chromy, with our design-effect approach offering extension to more complex sampling designs.

Keywords: Survey Methodology; Multivariate Allocation; Genetic Algorithm; Mixed Data

Transforming spatial data onto lattices to assess pathogen surveillance systems

Ciara Klinkert, Inger Fabris-Rotelli, Rene Stander, Renate Thiede, Kabelo Mahloromela

University of Pretoria

Thursday 26 November, 15:50 – 16:10 · Weaver · Sampling and survey methods and methodology ↑programme p. 30

Spatial data is often only available at units and resolutions that constrain analysis, comparability and interpretation. Although large geographic areas are intuitively conceptualised as collections of arbitrary but recognisable spatial units, there is no standardised methodology that can be used to transform spatial data into a target lattice structure. This research investigates alternative approaches for transforming POI spatial data onto a predefined, existing target lattice. The efficacy of the methodologies will be assessed through a simulation study and applied to a relevant practical research problem. This methodology aims to support an ongoing investigation into the use of wastewater pathogen detection systems, using COVID-19 data from Ekurhuleni wastewater plants. This research will enable COVID-19 cases recorded at hospitals in the area to be meaningfully related to wastewater catchment areas, so the wastewater data can be compared to actual cases and thus, assessed as a potential pathogen surveillance technique.

Keywords: Spatial data; Transformation; Lattices; POI; Disease Prevalence

Developing data-driven mental health risk indices using machine learning: A comparison with traditional index construction methods

Sbongiseni Mthethwa, Danielle Roberts

University of KwaZulu Natal

Thursday 26 November, 16:10 – 16:30 · Weaver · Sampling and survey methods and methodology ↑programme p. 30

Mental health disorders, particularly depression and anxiety, represent a growing public health challenge in South Africa. Effective monitoring of mental health risk is important for informing resource allocation and targeted interventions, yet dedicated mental health data are often limited, costly to collect, and not routinely available. An alternative approach is to exploit routinely collected survey data containing socioeconomic, demographic, household, community and environmental characteristics that may serve as predictors or proxies of mental health risk.

This study proposes a methodological framework for constructing outcome-specific Mental Health Risk Indices (MHRIs) using individual-level data from existing South African surveys. The primary focus is to investigate whether machine learning methods can provide a more predictive and flexible approach to index construction than traditional statistical techniques. Machine learning models will be used to identify and quantify potentially nonlinear and complex relationships between observed risk factors and specific mental health outcomes, such as depression and anxiety. The resulting machine learning-based indices will be compared with indices constructed using established dimension-reduction approaches, including principal component analysis (PCA) and factor analysis.

The different approaches will be evaluated according to their ability to predict observed mental health outcomes, including measures of predictive performance and discrimination. The study will also investigate the generalisability of the resulting indices across different datasets and population groups, and assess the contribution of individual predictors to the construction and interpretation of the indices.

The study aims to establish whether machine learning can provide a valid, interpretable and generalisable approach to constructing mental health risk indices from routinely collected survey data, and whether such approaches offer advantages over traditional index construction methods. The findings will contribute to methodological approaches for monitoring mental health risk in settings where direct mental health measures are limited, while providing a foundation for subsequent extensions incorporating additional contextual and spatial information.

Keywords: Mental health disorders; Mental Health Risk Indices; depression and anxiety

Theoretical statistics

Thursday, 26 November 2026, 15:10 – 16:30 · Yellowwood · Chair: Johan Ferreira

A flexible cosine-perturbed wrapped Lindley model

Delene van Wyk-de Ridder, Johan Ferreira, Janet van Niekerk

University of Pretoria; University of the Witwatersrand

Thursday 26 November, 15:10 – 15:30 · Yellowwood · Theoretical statistics ↑programme p. 30

Circular observations, common in fields such as animal movement and geological orientation, require probability models that respect the circle's periodic geometry; motivating flexible, interpretable, and computationally tractable circular distributions. We introduce and study the wrapped cos-Lindley (wc-L) distribution, a two-parameter model for circular data obtained from a Lindley-type construction. The model extends the classical wrapped Lindley distribution through an additional shape parameter, with the wrapped Lindley recovered as a special case. This additional parameter provides greater flexibility in the shape of the circular density, allowing the model to represent different modal structures. We investigate important distributional properties of the wc-L, including its trigonometric moments and modal behaviour, and identify parameter regimes associated with distinct density shapes. Maximum likelihood estimation is considered for parameter inference, and finite-sample performance is evaluated through simulation studies across a range of sample sizes and parameter settings. The usefulness of the model is demonstrated through applications to real-world circular datasets arising from transaction-time and animal-orientation data, illustrating how the additional shape parameter can provide greater flexibility than common competitors while retaining an arguably parsimonious two-parameter formulation.

Keywords: wrapping; cosine perturbations; trigonometric moments; maximum likelihood

Bayesian parametric modelling of credit risk using the three-parameter log-logistic distribution

Nthabiseng Khabele, Modisane Seitshero, Martin Chanza

North-West University

Thursday 26 November, 15:30 – 15:50 · Yellowwood · Theoretical statistics ↑programme p. 30

This paper proposes a unified parametric framework for credit risk modelling centred on the three-parameter log-logistic (LL3) distribution, whose heavy tail and flexible shape make it particularly well suited for capturing skewed loss-given-default and probability of default (PD) surfaces. Using the German Credit (Statlog) and Taiwan Credit Card Default datasets, we estimate LL3 parameters via classical maximum likelihood estimation (MLE) and a Markov chain Monte Carlo (MCMC) Bayesian framework and benchmark the results against Bayesian Additive Regression Trees (BART). Model adequacy is assessed through the probability integral transform (PIT histogram), the Kolmogorov-Smirnov (KS) test, the chi-squared (χ^2) goodness-of-fit test, and tail exceedance calibration. The LL3 distribution uniformly dominates nonparametric alternatives on all diagnostic criteria. In the second strand of the study, we couple a Bayesian ARIMA model with a Point-in-Time (PIT) adjusted logistic regression to construct forward-looking, scenario-conditional PD estimates consistent with IFRS 9 provisioning rules. Macroeconomic momentum is channelled into firm-level default probability through a two-stage propagation mechanism, and capital volatility metrics are introduced to link PD trajectories to Basel-compatible capital adequacy buffers. The resulting framework directly addresses the procyclicality inherent in Through-the-Cycle (TtC) PD models and provides a coherent basis for scenario-weighted expected credit loss (ECL) computation under the three-scenario (base, upside, downside) convention.

Keywords: Log-logistic distribution; Bayesian inference; Bayesian Additive Regression Trees; Bayesian ARIMA; Expected credit loss

Parameter estimation techniques for the hybrid Weibull–exponentiated Rayleigh distribution

Tolulope Adeniji, Akinwumi Odeyemi, Chioneso Marange

University of Fort Hare; University of Free State

Thursday 26 November, 15:50 – 16:10 · Yellowwood · Theoretical statistics ↑programme p. 30

The choice of an appropriate estimation method is as critical as the selection of the distribution itself, particularly for newly developed flexible lifetime models. The Hybrid Weibull–Exponentiated Rayleigh Distribution (HWERD) is a recently proposed three-parameter lifetime distribution that offers flexibility for modelling non-monotonic hazard functions and heavy-tailed data. However, the performance of different parameter estimation techniques for this distribution has not been systematically investigated. This study provides a comprehensive comparison of five classical estimation methods for the HWERD: maximum likelihood (ML), weighted least squares (WLS), Cramér–von Mises (CVM), maximum product of spacings (MPS), and Anderson–Darling (AD).

An extensive Monte Carlo simulation study is conducted across three distinct parameter configurations and sample sizes ranging from $n = 20$ to $n = 500$, with 1,000 replications per scenario. Performance is evaluated using bias, root mean square error (RMSE), and coverage probabilities of nominal 95% confidence intervals. The results reveal that all estimators are consistent, with bias and RMSE decreasing as sample size increases. The shape parameter α is generally the most challenging to estimate, particularly in small samples, while the parameter λ exhibits relatively stable behaviour across all scenarios.

Among the estimators, the AD method demonstrates superior point estimation performance, producing among the smallest RMSE values and coverage probabilities closest to the nominal 95% level. The MPS estimator achieves 100% convergence rates and the lowest RMSE for λ in challenging configurations (e.g., RMSE $= 0.0348$ at $n = 500$). The ML estimator exhibits strong asymptotic properties and provides the most reliable interval estimation. An application to guinea pig tuberculosis survival data confirms that ML yields the lowest Akaike Information Criterion (AIC) value of 193.25, closely followed by MPS and AD. This study provides practical guidance for practitioners applying the HWERD model in real-world lifetime data analysis.

Keywords: Quantile function; Hybrid Weibull-Exponentiated Rayleigh Distribution; parameter estimation; maximum likelihood estimation; Anderson-Darling estimator

Moment-weighting to improve the power of empirical likelihood ratio-based tests for normality under symmetric short-tailed alternatives

Lwando Dlembula, Chioneso S. Marange, Yichuan Zhao

University of the Free State; Georgia State University, Atlanta, GA, USA

Thursday 26 November, 16:10 – 16:30 · Yellowwood · Theoretical statistics ↑programme p. 30

The empirical likelihood (EL) technique is a nonparametric likelihood-based methodology with extensive theoretical and practical uses. Previous studies report that empirical likelihood ratio (ELR) tests for normality may demonstrate low power against symmetric short-tailed alternatives. Motivated by this limitation, we propose two moment-weighted modifications of existing ELR tests to improve sensitivity to symmetric short-tailed alternatives. The Type I error rates of proposed tests are assessed through Monte Carlo (MC) simulations. Power is examined under a range of symmetric short-tailed distributions, and their consistency is assessed across a broader range of both symmetric and asymmetric alternatives. For each moment k, the proposed moment weighted test statistics have chi-square limiting distributions. The proposed tests are compared against several studied classical tests and existing EL tests for normality. The simulation results show that the proposed tests have good Type I error rates, and exhibit increasing rejection probabilities with sample size under the alternatives considered, supporting their consistency. The MC simulations revealed the superiority of the proposed tests under symmetric short-tailed alternatives.

Keywords: Empirical likelihood; normality testing; moment weighting; symmetric short-tailed alternatives; Monte Carlo simulation

Biostatistics 5

Thursday, 26 November 2026, 15:00 – 16:20 · Cedarwood · Chair: Miguel Torres

Modelling seasonality of stroke types and associated risk factors in South Africa using a Poisson quantile regression approach: Hospital-based study 2014-2018

Lyness Matizirofa, Delson Chikobvu, Edmore Ranganai

University of South Africa; University of the Free State

Thursday 26 November, 15:10 – 15:30 · Cedarwood · Biostatistics 5 ↑ programme p. 30

Background: Globally, stroke is the leading cause of death and disability. In South Africa (SA), stroke is the second leading cause of death after HIV/AIDS and is among the top ten leading causes of disability. Although the seasonal variation in stroke cases has long been recognised in Africa, there is still limited published data on the seasonal variation of stroke type in the Sub-Saharan region. Understanding the seasonality of stroke type is important for SA planning and resource allocation. Few studies focus on the seasonality of stroke types in SA. This study aims to examine the seasonality of stroke types and the associated risk factors using a Poisson quantile regression model.

Methods: A cross-sectional study design was used to investigate the seasonality of stroke types in SA based on data collected between January 2014 and December 2018. A validated data retrieval form was used to retrieve data from the medical records of stroke patients. In Poisson regression, the dependent variable is count data, the number of strokes per month. Poisson quantile regression (PQR) models the conditional quantiles of a count variable. It accounts for seasonality by including time-based predictors in the model.

Results: Of the 35730 stroke patients, 27550 had ischemic stroke (IS), and the rest had hemorrhagic stroke (HS). The prevalence of IS was also highest in winter and lowest in summer. The risk of IS was significantly higher in patients with hypertension, diabetes, heart problems and hypercholesterolemia ($p < 0.0001$). The prevalence of modifiable factors and IS were higher in winter. The female gender, higher age groups, and black and Indian/Asian races were significantly associated with IS cases in winter.

Conclusion: The prevalence of modifiable risk factors and IS was higher in winter. Information about seasonal trends influenced by stroke predictors helps prepare awareness campaigns during periods/season of increased risk. The findings of this study call for programmes and policies to curb unhealthy lifestyles and promote physical activity, especially during winter.

Keywords: prevalence; stroke type; risk factors; seasonality; Poisson quantile regression; South Africa

Competing risks vs. censoring: Methodological insights from RR-TB trials

Lateef Babatunde Amusa, Rubeshan Perumal, Kogie Naidoo

CAPRISA; University of Ilorin, Nigeria

Thursday 26 November, 15:30 – 15:50 · Cedarwood · Biostatistics 5 ↑ programme p. 30

Tuberculosis (TB) treatment outcome studies commonly define a composite unfavourable outcome endpoint that combines death, treatment failure, and loss to follow-up (LTFU), analysed against treatment success. The use of a composite endpoint is a common strategy often employed to enhance statistical power and address the low incidence of individual outcomes. Using data from participants diagnosed with rifampicin-resistant TB enrolled in four separate clinical trials in KwaZulu-Natal, South Africa, we conduct a sensitivity analysis in which treatment success is treated as a competing risk rather than as censoring. Through comprehensive simulation studies, we further reevaluated the suitability of the Cox proportional hazards model in this context by demonstrating the methodological consequences of analysing composite unfavourable TB outcomes without accounting for competing risks.

Keywords: Competing Risks; Simulation; RR-TB; Cox Proportional Hazards; Clinical Trials

Statistical modelling of antimicrobial resistance using wastewater based epidemiology

Miguel Torres, Raeesa Docrat, Inger Fabris-Rotelli, Adam Sykulski

University of the Witwatersrand; University of Pretoria; Imperial College London

Thursday 26 November, 15:50 – 16:10 · Cedarwood · Biostatistics 5 ↑ programme p. 30

According to the World Health Organisation, antimicrobial resistance (AMR) is one of the most important threats to public health. In South Africa, AMR was associated with 35 054 deaths in 2021 alone. A system that can make real-time, spatially aware predictions to curb delays in intervention could save lives and allow efficient allocation of resources. Wastewater-based epidemiology (WBE) is an approach to disease monitoring that could enable real-time monitoring and modelling of AMR. WBE is inclusive, non-invasive and can account for asymptomatic individuals. Often, key epidemic parameters of interest such as true case numbers and biomarker concentrations before decay are unobservable. Observed wastewater signals and clinical data naturally follow a dual latent and observed state representation. State space models (SSMs) capture this structure by design, but have not been extended to include spatial information. Furthermore, SSMs model data sequentially, therefore they enable real-time predictions and disease outbreak detection through change-point detection. We consider the synthesis and application of statistical methods to address AMR modelling using WBE in four main steps: (1) a systematic review on statistical modelling of disease using WBE, (2) constructing a spatial SSM and verifying its properties via simulation, (3) applying the model to empirical AMR and clinical data, (4) comparison of change-point detection methods for detecting disease outbreaks. This framework provides a scalable tool for near real-time spatial surveillance and public health decision-making.

Keywords: antimicrobial resistance; wastewater-based epidemiology; spatial statistics; state space models; change-point detection

An automated analysis of a novel flow cytometry assay for TB vaccines

Sian Wood, Miguel Rodo, Elisa Nemes, Sebnem Er

University of Cape Town

Thursday 26 November, 16:10 – 16:30 · Cedarwood · Biostatistics 5 ↑ programme p. 30

Flow cytometry measures the expression of cellular markers on millions of cells within blood samples, enabling immune cell profiling for the investigation of vaccine-induced effects in clinical trials. Traditionally, this profiling is achieved by grouping cells into populations based on their marker expression in a process known as "manual gating". Although well-established, manual gating is not scalable or reproducible and introduces bias.

We explore automated clustering alternatives in the context of large TB vaccine clinical trials, which require approaches that are accurate and scalable. A manual-gating-based preliminary analysis indicated that the most informative phenotypes are expressed on rare cell subsets. However, an automated method for accurately clustering such rare subsets has not yet been demonstrated. We therefore prioritised theoretical strengths in rare subset identification when selecting four candidate approaches: FlowSOM, CHOIR, PARC and FAUST.

Each was evaluated on its ability to recover rare phenotypes identified as immunologically significant in the preliminary analysis. CHOIR and PARC overclustered and FlowSOM performed inconsistently. FAUST's discretisation of markers into distinct levels of expression per cell was found to be both accurate and interpretable. However, FAUST excluded approximately 10% of cells based on rarity, including subsets identified as important during the preliminary analysis. We therefore adapted FAUST's approach to force the annotation of every cell.

FAUST was used to generate an exhaustive set of phenotypes, which then required downselection. We constructed a graph connecting highly correlated phenotypes and partitioned it using the Leiden algorithm. A representative phenotype was selected from each community based on its median within-community correlation. The resulting pipeline minimises manual intervention, accurately reproduces manual gating results, and identifies rare phenotypes for differential analysis, including subsets highlighted in the preliminary analysis. This approach is scalable across large multi-sample datasets, and is currently being applied to data from two additional TB vaccine clinical trials.

Keywords: Cluster analysis; flow cytometry; tuberculosis; vaccine clinical trial; rare phenotype discovery

ABSTRACTS

Friday, 27 November 2026

Process monitoring

Friday, 27 November 2026, 09:10 – 10:10 · Weaver · Chair: Roelof Coetzer

A distribution-free method for anomaly detection in industrial processes

Praise Obanya, Roelof Coetzer, Shawn Liebenberg

North-West University

Friday 27 November, 09:10 – 09:30 · Weaver · Process monitoring ↑ programme p. 31

Statistical process monitoring (SPM) is essential for identifying process anomalies or performance deviations as quickly as possible. Most of the approaches applied in industry are based on assuming that the data follows some parametric distribution (e.g., normality). However, in industry, this assumption is not always feasible, thereby limiting the operational reliability of monitoring frameworks. In this paper, permutation entropy (PE), which is a time series analysis tool that identifies unusual patterns in a series, is used for univariate SPM. PE is distribution-free and robust to outliers. The power of PE is illustrated using a simulation study for fractional Brownian motion processes with different choices of Hurst parameter values. The effectiveness of PE is investigated using the Tennessee Eastman process as a case study for the detection of faults in an industrial process. In addition, PE is compared to two non-parametric CUSUM methods. PE is shown to be robust and effective, and outperforms the CUSUM methods in detecting various TEP process faults considered.

Keywords: CUSUM; permutation entropy; statistical process monitoring; Tennessee Eastman Process

Refining wastewater surveillance through regression kriging: A geostatistical approach to improve spatial prediction and public health applications

Odey Mofokeng, Raeesa Manjoo-Docrat, Inger Fabris-Rotelli, Pravesh Debba, Jenny Holloway, Nontembeko Dubeni-Tlhone, Jabulani Jele, Wouter le Roux, Lisa Schaefer, René Stander, David Rose, Nada Abdelatif

University of the Witwatersrand; University of Pretoria; Council for Scientific and Industrial Research; South African Medical Research Council

Friday 27 November, 09:30 – 09:50 · Weaver · Process monitoring ↑ programme p. 31

Wastewater-based epidemiology has emerged as a valuable complement to clinical surveillance for infectious disease monitoring, yet integrating environmental and clinical data across mismatched spatial scales remains a persistent methodological challenge. This study develops a Regression Kriging (RK) framework that combines multivariate linear regression with geostatistical interpolation of residuals to predict weekly COVID-19 incidence at the subplace level in the Ekurhuleni Metropolitan Municipality, South Africa, with a four-week early warning horizon. A spatial disaggregation framework was implemented to reconcile differences between COVID-19 case data (six health sub-districts), wastewater gene concentrations (eight treatment plants), and vulnerability covariates (673 subplaces). Population-weighted areal interpolation was used to disaggregate case data, while a hierarchical mapping approach linked subplaces to wastewater treatment plants. Variable selection using LASSO regularisation, variance inflation factor screening, and domain knowledge produced six predictors with low multicollinearity, and model performance was evaluated using a week-by-week spatial 10-fold cross-validation design.

The RK model achieved a mean R^2 of 0.759, a mean absolute error of 0.045 weekly cases per 1 000 population, and a symmetric mean absolute percentage error of 9.09%, outperforming a Regression-Only model ($R^2 = 0.683$) and demonstrating substantial gains from residual interpolation. Performance improved across successive pandemic waves, and Moran's I of 0.672 ($p < 0.001$) confirmed strong residual spatial dependence captured by the Kriging component. A Linear Mixed Model with wastewater treatment plant random effects performed poorly, while a Spatial Random Forest achieved higher temporal accuracy under less stringent validation. These results show that the predictive value of wastewater signals emerges through multivariate spatial modelling, and that geostatistical interpolation captures spatial dependence not accounted for by conventional approaches. The framework is generalisable to other pathogens and settings requiring integration of environmental and clinical data at fine spatial resolutions.

Keywords: Infectious disease prediction; Public health surveillance; Regression Kriging; Spatial linear mixed model; Wastewater modelling

Novel multivariate multiple regression synthetic EWMA control chart for monitoring general linear profiles

Maonatlala Thanwane, Jean-Claude Malela-Majika, Frans Kanfer, Kashinath Chatterjee, Saralees Nadarajah

University of Pretoria; Augusta University, Augusta, USA; University of Manchester, UK

Friday 27 November, 09:50 – 10:10 · Weaver · Process monitoring ↑ programme p. 31

Classical control charts are designed to monitor one or several quality characteristics that do not depend on other variables. When there is a functional relationship between the quality characteristic and one or more explanatory variables, the use of classical charts becomes inappropriate. In such case, the literature suggests using simple or general profiles based on the number of exogenous variables. Simple profiles are used when there is only one explanatory variable. In this paper, we introduce a new method for designing general linear profiles (GLPs). In addition, a novel synthetic multivariate multiple regression (MMR) exponentially weighted moving average (EWMA) chart for GLPs is proposed to demonstrate the design, implementation, and application of the new method. The performance of the new multivariate EWMA (MEWMA) chart is investigated using extensive simulations and they are compared to that of the existing MMR MEWMA chart for GLPs in terms of the run-length properties. It is found that the proposed synthetic MEWMA chart performs better than the competing MEWMA chart in most cases. A numerical real-life example is also provided to demonstrate the practicability of the novel synthetic MMR MEWMA chart.

Keywords: Conforming sub-chart; General linear profile; Multivariate multiple regression chart; Simple linear profile; Synthetic chart

Forecasting and timeseries

Friday, 27 November 2026, 09:10 – 10:10 · Yellowwood · Chair: Tendai Makoni

Regime-aware volatility forecasting of South African gold sales: A comparative analysis of Markov-switching and LSTM models

Tendai Makoni, Delson Chikobvu

University of the Free State

Friday 27 November, 09:10 – 09:30 · Yellowwood · Forecasting and timeseries ↑ programme p. 31

Gold markets are characterised by periods of shifting uncertainty and volatility, making accurate volatility forecasting important for risk management, investment decisions and market planning. This study examines the volatility dynamics and forecasting performance of monthly gold sales using a regime-aware modelling framework. Monthly gold sales data from January 2000 to July 2026 are analysed. Monthly gold sales growth rates are calculated as the first differences of the natural logarithm of gold sales and are subsequently examined for stationarity, volatility clustering and conditional heteroskedasticity. The study compares conventional, asymmetric, long-memory, regime-switching and nonlinear forecasting approaches. A GARCH(1,1) model serves as the baseline volatility model, while GJR-GARCH, EGARCH and APARCH models are used to capture asymmetric responses to shocks. FIGARCH is used to investigate long-memory behaviour in volatility. Markov-Switching models are employed to identify distinct volatility regimes, while a Markov-Switching GARCH (MS-GARCH) model integrates regime changes with conditional volatility dynamics. A Long Short-Term Memory (LSTM) neural network is also implemented as a nonlinear machine-learning benchmark. The models are evaluated using a chronological in-sample and out-of-sample forecasting framework to avoid look-ahead bias. Forecast performance is assessed using root mean squared error (RMSE), mean absolute error (MAE) and the quasi-likelihood (QLIKE) loss function. The empirical results indicate that MS-GARCH provides the most accurate out-of-sample volatility forecasts, outperforming conventional GARCH-family models and other regime-based specifications. LSTM records the second-best forecasting performance, demonstrating competitive predictive ability in capturing nonlinear volatility dynamics. The superior performance of MS-GARCH indicates that explicitly accounting for shifts in volatility regimes yields valuable information for forecasting monthly gold-sales volatility. The findings underscore the importance of regime awareness when modelling volatile gold markets and show that combining regime-switching mechanisms with GARCH dynamics can deliver substantial forecasting gains. The results have implications for risk management, market monitoring, and decision-making in environments characterised by shifting volatility.

Keywords: Gold sales; Volatility forecasting; MS-GARCH; LSTM; Regime-switching volatility

Analysing the efficiency of deseasonalisation methods in removing seasonal bias in outliers

Rofhiwa Netshiomvani, Shawn Liebenberg, Gerrit Grobler

North-West University

Friday 27 November, 09:30 – 09:50 · Yellowwood · Forecasting and timeseries ↑ programme p. 31

Deseasonalisation is a common preprocessing step in climate time series analysis, particularly when the focus is anomaly detection, where seasonal patterns can mask or bias extreme events. Most deseasonalisation approaches focus on removing the seasonal mean, implicitly assuming that this is sufficient to eliminate seasonal patterns. However, many meteorological variables exhibit seasonality in both mean and variance, which may bias the detection of outliers. In this talk, we will show whether deseasonalisation methods effectively remove seasonal bias in outliers from bivariate climate time series obtained from weather station and ERA5 data. Daily wind speed data from sixteen locations across all nine provinces of South Africa were analysed using paired observations from both sources. Standard deseasonalisation methods were used to remove seasonality in the mean, while less well-known approaches also account for seasonal variance. In addition, the effects of serial correlation and conditional heteroscedasticity were removed with an AR–ARCH model. To assess the presence of seasonal bias, the monthly distribution of detected outliers was evaluated using a chi-square goodness of fit test. The results show that deseasonalisation can either remove or introduce seasonal bias in detected outliers, depending on both the deseasonalisation approach and the outlier detection method used.

Keywords: Deseasonalisation; Seasonal bias; Outlier detection; Bivariate climate time series; Wind speed

The path to the 2030 SDG target: Statistical forecasting of maternal mortality in South Africa using Box-Jenkins time series analysis

Gomolemo Rakale, Sam Thembelihle Ntuli, Tshepo Ramarumo, Solly Matshonisa Seeletse

Sefako Makgatho Health Sciences University; University of South Africa

Friday 27 November, 09:50 – 10:10 · Yellowwood · Forecasting and timeseries ↑ programme p. 31

Maternal mortality remains a critical global health challenge. Sustainable Development Goal (SDG) Target 3.1 aims to reduce the global Maternal Mortality Ratio (MMR) to below 70 per 100,000 live births by 2030. Assessing the likelihood of achieving this target in high-burden settings like South Africa requires robust, forward-looking analyses. This study aimed to model and forecast trends in institutional maternal mortality in South Africa using time-series analysis to evaluate the nation's projected progress toward SDG Target 3.1.

Methods: We employed a Box-Jenkins (ARIMA) methodology. Data were extracted from the triennial Saving Mothers reports (2014–2022) published by South Africa's National Committee for Confidential Enquiries into Maternal Deaths (NCCEMD). The best-fitting model was selected based on the Akaike and Bayesian Information Criteria (AIC/BIC). Residual diagnostics were performed to validate model adequacy. The fitted model was used to generate forecasts for the period 2024–2030.

Results: The optimal model was an ARIMA (1,1,1). Diagnostics confirmed the residuals were white noise (Ljung-Box Q-test p=0.098), indicating a well-specified model. The forecast predicts a continuing decline in the institutional MMR, with a point estimate of 94.2 (95% CI: 72.5 – 130.6) per 100,000 live births for the year 2030.

Conclusion: While the model confirms a significant positive downward trajectory in maternal mortality, the forecast suggests that at the current pace, South Africa is unlikely to achieve SDG Target 3.1 by 2030. The projected 2030 value of 94.2 substantially exceeds the target of 70. This quantitative finding highlights the need for accelerated, transformative interventions beyond current efforts to alter the trajectory and meet this critical global health goal.

Keywords: Maternal health; Maternal mortality; Time-series analysis; ARIMA; Sustainable Development Goals

Stochastic processes

Friday, 27 November 2026, 09:10 – 10:10 · Cedarwood · Chair: Etienne Pienaar

Self-normalised sequential monitoring and break estimation for panel data

Charl Pretorius, Marie Hušková

North-West University; Charles University, Czechia

Friday 27 November, 09:10 – 09:30 · Cedarwood · Stochastic processes ↑programme p. 31

We consider sequential monitoring for changes in the means of a high-dimensional panel, where observations arrive over time and a decision to stop must be made without access to the full sample. CUSUM-type detectors are proposed and their limit behaviour established under weak dependence conditions on the errors and regressors. We then allow strong cross-sectional dependence, modelled through a common factor structure, and show how the detectors behave once the factors are accounted for. A self-normalised version avoids estimation of long-run variances, which matters in practice: the usual estimators are unreliable at the moderate sample sizes typical of applications, and the resulting size distortion is substantial. A simulation study supports the theory for small to moderate panels. We close with an illustrative application, and with the question that follows an alarm: which of the units actually changed?

Keywords: Break detection; Common factor model; High-dimensional panel data; Online monitoring; Self-normalisation

Assessing the impact of screening sensitivity and follow-up duration on parameter identifiability in a five-state hidden continuous-time Markov model

Tshepo Maja

University of Limpopo

Friday 27 November, 09:30 – 09:50 · Cedarwood · Stochastic processes ↑programme p. 31

Hidden continuous-time Markov models provide a flexible framework for modelling disease progression when the underlying disease state cannot be observed directly and individuals are detected through imperfect screening or clinical diagnosis. This study develops and evaluates a five-state hidden continuous-time Markov model representing progression from a healthy state through early preclinical, late preclinical, early clinical, and late clinical disease states. The model incorporates imperfect screening at two scheduled screening times, with separate sensitivities for the early and late preclinical states, followed by clinical detection during follow-up. Transition probabilities are obtained using matrix exponentials, and the six model parameters are estimated by maximum likelihood. A simulation study was conducted to investigate the effects of screening sensitivity and follow-up duration on parameter recovery and practical identifiability. Data were generated under three screening designs with increasing follow-up durations and four combinations of screening sensitivities. For each scenario, repeated simulations were used to assess parameter bias, relative standard error (RSE), and root mean squared error (RMSE). Practical identifiability was assessed using a 20% RSE criterion. The results indicate that both screening sensitivity and follow-up duration influence the accuracy and precision of parameter estimation. Longer follow-up generally reduces censoring and improves parameter recovery, resulting in lower bias, RSE, and RMSE. The disease-onset intensity was the most difficult parameter to estimate, particularly under shorter follow-up designs, whereas the downstream transition intensities showed improved precision under longer follow-up and higher screening sensitivity. The findings suggest that extending follow-up may be particularly important for improving practical identifiability of early disease-onset parameters. Overall, the study demonstrates the importance of jointly considering screening sensitivity and observation duration when designing studies based on hidden disease-progression models.

Keywords: Sensitivity; Follow-up duration; Identifiability; Hidden continuous-time Markov models

Evaluating first passage for Brownian motion with jumps

Etienne Pienaar, Iain MacDonald

University of Cape Town

Friday 27 November, 09:50 – 10:10 · Cedarwood · Stochastic processes ↑programme p. 31

The probabilistic evolution of stochastic differential equations with jumps are governed by partial integro-differential equations. Despite the numerous fields of applications for these models, the lack of closed-form solutions to such PDEs limits the set of 'tractable' models to less than a handful of specifications. Indeed, the transitional density can be found for a few examples if we allow solutions up to specification of the characteristic function. For first passage time distributions there exists a single specification with a closed form solution. (Well, sort of.) In this talk, we demonstrate a numerical technique for finding the first passage distribution (due to jumps) for Brownian motion plus jumps that follow an arbitrary jump distribution.

Keywords: Brownian Motion; First Passage Time; Stochastic Differential Equation; Poisson Jump Process

GARCH modelling

Friday, 27 November 2026, 10:30 – 11:30 · Weaver · Chair: Lerato Mahlangu

Stochastic aspects of market risk under different Basel Accord regimes

Sarah L. Mahlangu, Sandile C. Shongwe, Maxim Finkelstein

University of the Free State

Friday 27 November, 10:30 – 10:50 · Weaver · GARCH modelling ↑programme p. 32

The global financial crisis prompted the Basel Committee to replace Basel 2.5 with the Fundamental Review of the Trading Book (FRTB) which was implemented in South Africa in July 2025. While the Basel 2.5 Internal Models Approach relied on Value-at-Risk (VaR) and stressed VaR, FRTB introduces expected shortfall and a redesigned Standardised Approach. Reported market risk capital nonetheless fell at the changeover, raising the question of what drives it. This study addresses that question in two stages. First, an interrupted time-series analysis of 45 quarterly market risk risk-weighted assets observations (2014 Q4 to 2025 Q4) for a global bank operating in South Africa: an ARIMA(5,1,0) fitted to the 43 pre-transition quarters on the logarithmic scale (AIC = 31.43) projects a Basel 2.5 counterfactual of ZAR8.04 billion and ZAR5.63 billion for 2025 Q3 and Q4, against realised FRTB figures of ZAR3.57 billion and ZAR3.06 billion, which are 51.5% below the counterfactual and outside the 80% prediction interval at both horizons. Second, a scenario and sensitivity experiment on the bank's 33-bucket FX book tests whether concentration in G10 currencies explains the gap. Holding total gross sensitivity constant and sweeping the G10 share from 0% to 100%, delta capital rises 82.0%, from ZAR8.47 million to ZAR15.41 million. Neutralising individual portfolio features isolates three channels: the risk-weight relief under Basel Market Risk paragraph MAR21.88 reduces capital by 17.0%, the loss of offsetting raises it by 125.6%, and concentration contributes 2.1%. The composition curve is non-monotonic and the binding correlation scenario switches between G10 shares. The bank's actual book, at a 75.9% G10 share, sits on the steep upward segment, where the offsetting effect exceeds the risk-weight relief roughly sevenfold. Lower FRTB capital therefore reflects the recognition of offsetting and a changed risk measure rather than reduced risk and comparisons across regimes require caution.

Keywords: Fundamental Review of the Trading Book (FRTB); Basel 2.5; market risk capital; Standardised Approach; Internal Models Approach

Dynamic prices and volatility linkages across sectoral stock markets in South Africa: A cointegration and GARCH- based approach

Pepenene Mofolo, Sandile C. Shongwe

University of Free State

Friday 27 November, 10:50 – 11:10 · Weaver · GARCH modelling ↑ programme p. 32

While cointegration techniques provide a framework for examining whether non-stationary time series have long-run equilibrium relationships; the generalized autoregressive conditional heteroskedasticity (GARCH) models provide insights into financial risk dynamics. This study investigates dynamic price and volatility linkages across sectoral stock markets on the Johannesburg Stock Exchange (JSE), by focusing on long-run equilibrium relationships, causality, dependence and volatility transmissions. With daily closing prices from January 2015 to June 2025, unit root tests are first applied to check the order of integration of the stock price series. The analysis is conducted under the bivariate and multivariate frameworks. In the bivariate analysis, the Engle-Granger, Johansen and Gregory-Hansen cointegration tests are used to identify long-term relationships and structural breaks. Copula models are then used to examine dependence between selected pairs, while causality analysis checks the direction of information transmission. Within the multivariate framework, the Johansen and Maki cointegration tests are employed to check long run relationships while accounting for structural breaks, impulse response analysis and forecast error variance decomposition are then used to access the dynamic responses to shocks and contribution of individual markets to movements in other sectors within their portfolio. Multivariate causality analysis is also conducted to establish the direction of information transmission. Under both the bivariate and multivariate case studies, cointegrated stocks are further modelled using GARCH-family type models to examine volatility persistence, leverage effects, volatility dynamics and risk spill overs. The findings suggest that the South African equity market is integrated, with evidence of cross sectoral linkages and transmission of shocks and volatility between markets. These findings have practical implications for portfolio diversification, risk management, asset allocation, par trading strategies particularly for strategic investment decision-making and where investors seek to understand cross sectoral dependencies and manage exposure to transmitted market risk.

Keywords: Cointegration; Generalized autoregressive conditional heteroskedasticity (GARCH); Volatility clustering; Johannesburg Stock Exchange (JSE); Stationarity

Volatility and dependence in COVID 19 mortality across South African epicenters: A bivariate copula based NB INGARCH approach

Thembhani Hlayisani Chavalala, Retius Chifurira, Knowledge Chinhamu, Jacob Majakwara

University of Limpopo; University of Kwazulu-Natal; University of the Witwatersrand

Friday 27 November, 11:10 – 11:30 · Weaver · GARCH modelling ↑ programme p. 32

Daily COVID 19 death counts in South Africa's three largest provinces: Western Cape, KwaZulu Natal, and Gauteng, exhibit synchronised mortality patterns characterised by volatility, overdispersion, heavy tails, and autocorrelation. The univariate NB-INGARCH(1,1) model effectively addressed conditional heteroscedasticity and overdispersion. Granger causality tests on standardised residuals revealed strong interaction between Gauteng and KwaZulu-Natal, with bidirectional causality, while Western Cape was influenced only by KwaZulu-Natal. These cross provincial linkages motivated a bivariate copula extension. Gaussian, Student-t, Clayton, and Gumbel copulas were compared using AIC and BIC, with the Student t copula providing the best fit by capturing moderate tail dependence. Spearman's rank correlations and scatterplots confirmed satisfactory calibration, while Fisher-z test showed no significant differences between empirical and simulated pseudo observations. Overall, copula based NB INGARCH models offer a robust framework for jointly modelling marginal dynamics and inter provincial dependence, enhancing understanding of extreme co movements and spillover risks critical for public health planning and resource allocation.

Keywords: COVID-19; Volatility; NB-INGARCH; Student-t copula; inter-provincial dependence

Applied statistics

Friday, 27 November 2026, 10:30 – 11:30 · Yellowwood · Chair: Stefan Britz

Compression-based distance measures in text classification

Keara-Linn Stapelberg, Surette Bierman

Stellenbosch University

Friday 27 November, 10:30 – 10:50 · Yellowwood · Applied statistics ↑ programme p. 32

A relevant problem in the classification of text documents is quantifying distance measures between documents. An attractive parameter-free approach to this problem is text compression. This approach is particularly attractive in settings where labelled documents are scarce. Text classification using compression-based distance measures compares query texts to training texts in terms of how well they compress together. More specifically, the change in compressed size when a query text is concatenated to a training text, is used to classify the query text. The training text in the concatenation can either be a single document (instance-based classification), or a class prototype consisting of several documents from the same class (profile-based classification). The most widely used compression-based distance measure is the 'Normalised Compression Distance' (NCD), but it behaves inconsistently between instance-based and profile-based classification settings. In this paper, we identify two sources of size disparity that degrade the performance of the NCD. An alternative compression-based distance measure is proposed, and its properties analysed, showing invariance to size disparities. Empirical results support the use of the proposed measure compared to the NCD in profile-based classification.

Keywords: distance measures; document compression; NCD; text classification

Structural connectivity in the African air transport network

Stefan Britz, David Snaddon, Sebnem Er

University of Cape Town

Friday 27 November, 10:50 – 11:10 · Yellowwood · Applied statistics ↑ programme p. 32

The African Air Transport Network (AATN) plays a critical role in enabling connectivity across a geographically diverse and infrastructurally fragmented continent. However, the airports and routes carrying the greatest scheduled capacity do not necessarily occupy correspondingly central structural positions. This study analyses scheduled passenger routes in 2025 to investigate how operational capacity and structural connectivity differ across the AATN, and how these differences relate to its geographical organisation.

The data form an undirected network of 381 airports and 1184 intra-African routes, with scheduled seat capacity retained as an edge weight. Centrality measures, capacity-weighted multi-resolution Louvain community detection and K-core decomposition are used to compare operating scale, brokerage, modularity and embeddedness. Airport capacity is strongly associated with structural measures overall (Spearman's $\rho = 0.671$-0.745), but the leading roles differ: OR Tambo International has the greatest capacity, whereas Addis Ababa has the highest degree and betweenness. At route level, capacity and edge betweenness are only weakly associated ($\rho = 0.236$), and their top-20 sets do not overlap. Community structure aligns most strongly with African subregions at intermediate resolutions and with countries or areas at finer resolutions, while the deepest core spans all five African subregions.

The results show that capacity, brokerage, geographical modularity and core embeddedness capture distinct dimensions of African air connectivity.

Keywords: African Air Transport Network; network analysis; structural connectivity; community detection

Intervention modelling of time series tourist arrivals data from the non-SADC African and unspecified group from July 2010 to February 2024: In the context of the COVID-19 pandemic

Musara Chipumuro, Delson Chikobvu, Tendai Makoni

University of the Free State

Friday 27 November, 11:10 – 11:30 · Yellowwood · Applied statistics ↑ programme p. 32

This study quantifies the impact of the COVID-19 pandemic on monthly tourist arrivals to South Africa (SA) from non-SADC African countries and the unspecified category over the period July 2010 to February 2024. Employing a Seasonal Autoregressive Integrated Moving Average with Exogenous Factor (SARIMAX) model within the Box-Tiao intervention framework, we incorporate a pulse intervention function to capture both immediate and decaying effects of the pandemic. The application of the log transform is borrowed from the pre-intervention model's Box-Cox transformation output. The preferred model SARIMAX$(2,1,0)(1,0,0)_{12}$ was selected based on error evaluation metrics such as the MAPE, MASE and RMSE. Given the need to estimating the structural intervention effects in relation to purely forecasting future values, model validation relied on in-sample residual diagnostics and parameter significance. Results reveal an immediate decline of 45.82% in March 2020, followed by a near-total collapse of 99.96% in April 2020 with a 95% confidence interval (CI) of 99.94% to 99.98%. The estimated decay parameter δ=0.8475 indicates high persistence with arrivals remaining at approximately 0.34% below baseline levels by February 2024. These results indicates the prolonged recovery trajectory for this under-researched market segment as well as highlighting the necessity for targeted policy interventions. Diagnostics checks confirm model adequacy as evidenced by the Ljung-Box test's p-value of 0.088.

Keywords: Tourist arrivals; Box-Tiao methodology; non-SADC African countries; Unspecified group; SARIMAX

Spatial statistics 3

Friday, 27 November 2026, 10:30 – 11:30 · Cedarwood · Chair: Kabelo Mahloromela

Exploratory investigation of Discrete Pulse Transform features for roof-covering-material classification

Mark de Lancey, Inger Fabris-Rotelli

University of Pretoria

Friday 27 November, 10:30 – 10:50 · Cedarwood · Spatial statistics 3 ↑ programme p. 32

In many formally regulated building contexts, such as parts of Europe, roof construction and covering materials are relatively standardised and often supported by established building codes and records. In contrast, informal and rural settlements, for example in South Africa, may include more heterogeneous, incrementally constructed or undocumented roof structures, increasing uncertainty when inferring roof-covering material from aerial imagery. Reliable information on roof-covering material supports risk assessment, infrastructure planning and building-stock characterisation. This study investigates whether features derived from the Discrete Pulse Transform (DPT) improve roof-covering-material classification relative to non-DPT baselines. The Discrete Pulse Transform (DPT) is a nonlinear multiscale image representation that decomposes an image into structures, called pulses, at different spatial scales.

The experiments used roof crops from the Netherlands-based RoofSense dataset from Mozambique-based Nacala dataset, each set spanning multiple roof material classes. The decompositions were summarised using features describing pulse abundance, scale distributions, pulse magnitudes and spatial structure. Models using these features were evaluated within a fixed random forest framework and compared with corresponding non-DPT baselines based on crop dimensions or basic roof geometry. Feature screening and fold-wise residualisation were used to assess whether any improvement persisted after accounting for associations with image crop size, a feature that may be spuriously associated with material class.

Across both datasets, adding DPT-derived features improved roof-material classification over the corresponding non-DPT baselines, with useful predictive information remaining after adjustment for size-related effects. Roof material and slope are useful building attributes for building-stock characterisation, risk and infrastructure assessment, thermal and energy studies, and rooftop renewable-energy planning. At settlement scale, these attributes may also inform housing-quality, climate-exposure and broader public-health assessments. The persistence of predictive gains across these contrasting datasets supports the DPT as a useful, engineered representation of roof-material characteristics across substantially different building contexts.

Keywords: Roof-covering material classification; aerial imagery; texture analysis; statistical learning; Discrete Pulse Transform

Defining homogeneity for spatial linear networks

Arthur Antonio, Inger Fabris-Rotelli, Renate Thiede, Rene Stander

University of Pretoria

Friday 27 November, 10:50 – 11:10 · Cedarwood · Spatial statistics 3 ↑ programme p. 32

Homogeneity is formally defined in spatial statistics for point patterns as constant first-order intensity. This study introduces a novel definition homogeneity for spatial linear networks through two complementary components. The locational component records where line segments are concentrated; the directional component records how those segments are oriented. Herein, intensity-based concepts for homogeneity testing are extended to cover both segment counts per spatial unit and mean directional orientation. Each network is partitioned into a grid of quadrats. A modified χ^2 test on segment counts per unit area assesses the locational component, while the directional component is assessed by a non-parametric bootstrap test on differences in circular means across quadrats. Conclusions drawn from one fixed grid may be an artefact of grid placement, so the procedure is repeated across several grid resolutions and under controlled grid jittering. Size and power of both tests are calibrated in a road network simulation study.

Keywords: spatial linear networks; homogeneity testing; quadrat counts; circular statistics; sensitivity analysis

Quantifying local anisotropy in point pattern analysis

Kabelo Mahloromela, Inger Fabris-Rotelli

University of Pretoria

Friday 27 November, 11:10 – 11:30 · Cedarwood · Spatial statistics 3 ↑ programme p. 32

A spatial point pattern is termed anisotropic if its statistical properties exhibit directional variation. Global anisotropy, which occurs when a spatial point pattern exhibits uniform directional characteristics across the study area, has been extensively studied. Local anisotropy analysis for point patterns is less commonly addressed in the literature and often developed for particular applications. In many real-world applications, anisotropy may vary spatially due to environmental, structural, or anthropogenic factors. Thus, accurately characterising point patterns with heterogeneous directional behaviour requires understanding and quantifying local anisotropy. Herein, we develop methodology to quantify local anisotropy, providing spatially resolved measures of direction and strength across the point pattern domain.

Keywords: point pattern; anisotropy; first-order characteristics; second-order characteristics

ABSTRACTS

Posters

The effect of including direct-travel neighbourhood matrices in hierarchical Bayesian models for disease mapping

Johannes Petrus Meyer Moller, Janet van Niekerk

University of Pretoria

Wednesday 25 November, 17:30 – 18:45 · board P01 · Poster session ↑ programme p. 33

Disease mapping in spatial epidemiology quantifies relative risk across geographical areas by smoothing random noise and borrowing strength from neighbouring regions. Classical hierarchical Bayesian frameworks, such as Conditional Autoregressive (CAR) models and the scaled Besag-York-Mollie (BYM2) specification, traditionally rely on a spatial neighbourhood matrix constructed strictly from physical boundary contiguity. However, modern human mobility and transport infrastructure create high-volume interaction corridors between non-adjacent regions, decoupling true social interaction from immediate geographical contiguity.

This study outlines a methodological framework for augmenting the spatial neighbourhood graph by combining physical contiguity with mobility networks. Rather than replacing boundary adjacencies, existing border-sharing neighbours are maintained, while non-contiguous direct-travel links, such as major transit routes and commuter corridors, are incorporated as additional first-order neighbours into a single unified matrix. Because adding non-contiguous connections alters the precision matrix Q, the generalized marginal variance within the BYM2 framework is dynamically rescaled to maintain consistent hyperprior interpretation. The computational setup is formulated via Integrated Nested Laplace Approximations (INLA) to evaluate parameter sensitivity across structured epidemiological networks.

By explicitly combining physical and travel-based adjacencies into a unified first-order neighbourhood structure, this proposed model addresses the limitations of strict boundary restrictions. The framework aims to prevent artificial over-smoothing between physically adjacent yet functionally isolated regions while accurately capturing risk transmission along transport corridors.

Keywords: Spatial disease mapping; BYM2 model; Human mobility; Hierarchical Bayesian modeling

Modelling extreme maximum temperatures using block maxima and peaks-over-threshold approaches

Anna M Seimela

University of Limpopo

Wednesday 25 November, 17:30 – 18:45 · board P02 · Poster session ↑ programme p. 33

Climate extremes such as heat waves pose serious risks to society and the environment, particularly in developing countries such as South Africa. This study analyses maximum temperatures in Limpopo Province using a comparative analysis of return levels from the Block Maxima (BM) and Peaks-Over-Threshold (POT) approaches. The BM approach uses the Generalised Extreme Value (GEV) distribution, while the POT approach uses the Poisson point process and Generalised Pareto Distribution (GPD). The profile likelihood is used to construct 95% confidence intervals. Four stations, namely Mara, Messina, Polokwane and Thabazimbi, are considered. The results show comparable return-level estimates across the three methods for most stations, except Polokwane, where the Poisson point process and GPD produce higher estimates than the GEV method. At Thabazimbi, a temperature of 40°C is estimated to be exceeded at least once in 500 years. Overall, the BM and POT approaches produce comparable estimates for longer return periods. Future research may consider other extreme-value models, such as the r-largest order statistics approach.

Keywords: Block maxima; GEV; maximum temperatures; Poisson point process; GPD

Comparative modelling of tail risk in global equity markets: Evidence from heavy-tailed distributions and back-testing

Sthembiso Dlamini, Sandile C. Shongwe

University of the Free State

Wednesday 25 November, 17:30 – 18:45 · board P03 · Poster session ↑ programme p. 33

In financial risk modelling, the ability to accurately predict future market movements (especially losses), rather than over-under-estimating is vital for proper risk management. In recent times, the global financial crisis (2007–2008), COVID-19 pandemic (2019–2021) and the most recent wars in the Middle-East have been amongst the most notable rare events that have had major impacts on stock markets globally. The economic instability created by these events has over the years strengthened the call for the use of more robust models together with stronger model evaluation mechanisms. This research explores the application of heavy-tailed distributions in quantifying tail risks associated with these events across different equity markets, with a specific focus on the Nairobi Securities Exchange (NSE20), the South African Equity Market (FTSE/JSE Top40), and the US Equity Index (S&P500). The study aims to identify the most suitable distribution between the Generalized Extreme Value Distribution (GEVD), the Generalized Pareto Distribution (GPD), Skewed t, Generalized Gamma, and Burr distribution for modelling the distributional behavior of our financial returns. To address volatility clustering, four Generalized Autoregressive Conditional Heteroskedasticity (GARCH) models are first fitted to the return series. Subsequently, the five distributions are then fitted to the standardized residuals, where for each index, the models are ranked using graphical diagnostics, goodness-of-fit statistics. To assess the predictive performance of our models, Value-at-Risk back-testing is conducted using Kupiec's unconditional coverage test and Christoffersen's conditional coverage test at the 99th confidence level. Overall, the back-testing results demonstrate that EVT models, particularly the GPD, remain highly effective for modelling extreme market risk, while alternative heavy-tailed distributions provide useful flexibility under certain market conditions. Beyond its technical contributions, this study has broader implications for building sustainable and resilient financial systems in both emerging and developed markets.

Keywords: Back-testing; Generalized extreme value distribution; Generalized Pareto distribution; Nairobi Securities Exchange; South African equity market

Investigating the strengths and weaknesses of the SARIMAX intervention model in time series intervention analysis

Thabiso Masena, Sandile Shongwe

University of the Free State

Wednesday 25 November, 17:30 – 18:45 · board P04 · Poster session ↑ programme p. 33

A few recent studies have shown that the seasonal autoregressive integrated moving average model with a pulse function covariate vector as an exogenous variable (SARIMAX) is an ideal model in assessing, modelling and quantifying the impact of the pandemic on the South African economy. This study aims to investigate the strengths and weaknesses of the recently modified SARIMAX intervention model by comparing its adherence to the Box-Jenkins model assumptions of white noise residuals before and after fitting a pulse function covariate vector. The analysis was based on four datasets containing monthly sales from the wholesale, retail, tourism accommodation and the food and beverages sectors. March 2020 was used as the intervention point. The results of the study showed that the sales for the retail and wholesale datasets had a quick short-term recovery of 12 and 13 months, with total estimated losses of ZAR 97580 million and ZAR 306850 million, respectively. However, the tourist accommodation and food and beverage sales datasets had long-term recovery of 64 and 70 months, with total estimated losses of ZAR 102461 million and ZAR141855 million, respectively. The residuals from the SARIMAX intervention models fitted on the retail and wholesale datasets with short-term recovery periods were white noise before and after fitting a pulse function covariate vector. On the contrary, the residuals from the SARIMAX intervention model fitted on the tourism and food and beverages datasets with long-term recovery periods were white noise before fitting the pulse function covariate vector only. The analysis indicate that the SARIMAX intervention model seems to be more appropriate in quantifying and modelling interrupted time series datasets with short-term recovery periods as compared to the datasets with gradual and long-term recovery periods in this study.

Keywords: Intervention analysis; pulse function; SARIMAX intervention model; covariate vector; white noise

Through-the-cycle to point-in-time probability of default conversion under IFRS 9: Integrating macroeconomic variables into credit risk modelling

Hope Alum, Kgotso Moremoholo, Sandile Shongwe

University of the Free State

Wednesday 25 November, 17:30 – 18:45 · board P05 · Poster session ↑ programme p. 33

Under International Financial Reporting Standard 9 (IFRS 9), financial institutions are required to estimate expected credit losses (ECL) using forward-looking information, replacing the incurred-loss approach of International Accounting Standard 39 (IAS 39). This requires converting through-the-cycle (TTC) PD estimates, which reflect long-term borrower-specific default patterns, into point-in-time (PIT) PD estimates that incorporate current and forecast macroeconomic conditions. However, existing literature remains fragmented, with the selection of macroeconomic variables (MEVs), the modelling of default risk, and the TTC-to-PIT conversion commonly treated as separate problems. Few studies integrate these components within a single empirical framework. Using the United States' Freddie Mac single-family mortgage data, this research first compares Pearson and Spearman correlation methods for MEV selection, finding that Spearman's correlation better captures the monotonic, non-linear relationships between MEVs and default rates. Default rates are then modelled using ordinary least squares, an autoregressive distributed lag model, gradient boosting, and random forests. Results show that gradient boosting yields the lowest prediction error at some cost to interpretability. Finally, the additive scalar, multiplicative scalar, and Vasicek-based TTC-to-PIT conversion methods are evaluated for responsiveness, stability, and interpretability. The results show that no single method dominates across all criteria: the multiplicative scalar approach responds strongly to deteriorating economic conditions while preserving probability bounds; the Vasicek approach provides a theoretically grounded framework but exhibits a known internal inconsistency; and the additive scalar approach offers greater stability but is less responsive to economic changes. By integrating MEV selection, default-rate modelling, and TTC-to-PIT conversion within a unified empirical framework, this study provides an evidence-based comparison of alternative modelling choices. Additionally, it offers practical guidance for financial institutions in developing and justifying defensible PD estimates under IFRS 9.

Keywords: International Financial Reporting Standard 9 (IFRS 9); expected credit loss (ECL); probability of default (PD); through-the-cycle (TTC); point-in-time (PIT)

A cointegration-GARCH–EVT–copula framework for modelling tail risk in equity markets

Moses Mokhatholane, Sandile C. Shongwe

University of the Free State

Wednesday 25 November, 17:30 – 18:45 · board P06 · Poster session ↑ programme p. 33

This study aims to develop a financial risk modelling framework for cointegrated stocks by integrating Extreme Value Theory (EVT) and copula models to leverage the strengths of each approach in estimating Value at Risk (VaR) and Expected Shortfall (ES). EVT is particularly effective in modelling rare, extreme movements that can lead to substantial losses, while copula models provide a flexible framework for capturing complex and nonlinear dependence structures among random variables. Combining cointegration analysis with EVT and copula methodologies, the proposed framework seeks to enhance the modelling of long-run relationships and joint tail behaviour among stock returns. This integrated approach offers a comprehensive tool for analysing co-movements in financial markets, particularly during periods of heightened volatility, market stress, and structural change, thereby improving the accuracy, reliability of risk measurement and portfolio risk management. The research explores three pairs of highly correlated and cointegrated S&P500 stocks that operate in the United States economy. The stocks considered in this study are Home Depot, Sherwin Williams, Cintas Corporation, Costco Wholesale Corporation, and S&P Global Inc. Generalized Autoregressive Conditional Heteroskedasticity (GARCH) models were employed to capture volatility clustering commonly observed in financial return series. EVT is then applied to the standardised residuals of the GARCH models to estimate VaR and ES faced by the stocks. To model the marginal distributions of the stocks, several candidate distributions including the log-normal, Student-t, gamma, Weibull, and exponential distributions were fitted and evaluated to identify the most appropriate distribution for each asset. The selected marginal distributions were subsequently combined using copula models to construct joint distributions that capture the dependence structure and joint downside risk among financial assets. This framework provides a foundation for the development and enhancement of pair-trading strategies by explicitly incorporating the risk dynamics and dependence structures associated with cointegrated assets.

Keywords: Cointegration; Copula; Extreme value theory (EVT); Generalized Autoregressive Conditional Heteroskedasticity (GARCH); Risk measures

Analysis of statistical, deep learning and hybrid models for foreign exchange rates

Kuhle Mxotwa, Seoka Bolalane, Sandile Shongwe, Nokuthula Mokoena

University of the Free State

Wednesday 25 November, 17:30 – 18:45 · board P07 · Poster session ↑ programme p. 33

The autoregressive integrated moving average (ARIMA) model is a standard benchmark for univariate time series forecasting, but it cannot incorporate engineered features, capture non-linear and complex patterns. We investigate the effectiveness of the deep learning (DL) methods: recurrent neural network (RNN), gated recurrent unit (GRU), and long short-term memory (LSTM) network, alongside their hybrids (ARIMA-RNN, ARIMA-GRU, ARIMA-LSTM) against the benchmark ARIMA model. The task is one-day-ahead forecasting of five major South African Rand (ZAR) denominated foreign exchange (FOREX) rates, from 1 January 2006 to 18 March 2026: United States Dollar (USD/ZAR), Euro (EUR/ZAR), British Pound Sterling (GBP/ZAR), Japanese Yen (JPY/ZAR), and Chinese Yuan (CNY/ZAR). The DL models utilize a sixteen-dimensional feature vector, including daily percentage change, log returns, 50-day simple moving average,14-day relative strength index, and commodity prices (gold, platinum and crude oil). To ensure robust evaluation without look-ahead bias, a strictly chronological expanding-window cross-validation scheme is employed (5 folds, 0.60 initial training fraction) with training-set-only MinMax scaling. The forecasting accuracy is assessed using error metrics, coefficient of determination and directional accuracy. We initially hypothesized that hybrid models would outperform both standalone architectures by combining ARIMA's linear structure with the DL models' capacity to learn non-linear residual dynamics. However, empirical testing revealed that ARIMA accounted for the vast majority of predictable variance, resulting in negligible error correction by the DL component. To determine whether this lack of improvement is an architectural flaw or a unique characteristic of the FOREX data, we extended the evaluation to a secondary, highly non-linear dataset as a structural stress test. By providing one of the few systematic comparisons of statistical, deep learning, and hybrid forecasting approaches applied to ZAR-denominated FOREX pairs, this study clarifies whether hybrids offer measurable improvements for FOREX risk management, interest rate decisions, and hedging strategies.

Keywords: Recurrent Neural Network (RNN); Gated Recurrent Unit (GRU); Long Short-term Memory (LSTM); Hybrid Models; Foreign Exchange (FOREX)

AI techniques for risk analysis in banking sector: A review and perspective with emphasis on credit risk modelling

Themba Khohlakala, Sandile C. Shongwe

University of the Free State

Wednesday 25 November, 17:30 – 18:45 · board P08 · Poster session ↑ programme p. 33

Effective credit risk assessment is essential to the stability and long-term sustainability of banking institutions, which have historically relied on statistical methods such as logistic regression, credit scoring systems, and linear discriminant analysis, valued for their transparency and alignment with regulatory expectations. This study examines six models to improve upon the traditional methods (i.e., logistic regression, random forest, gradient boosting methods, linear support vector machine (LSVM), neural network) on the Give Me Some Credit dataset, a real-world unsecured consumer credit dataset comprising 104,805 borrowers, evaluated using Area Under the Receiver Operating Characteristic (AUROC) curve, F1-score, precision, recall, accuracy, the Kolmogorov-Smirnov statistic and Gini coefficient. Contrary to expectations drawn from the wider credit-scoring literature, in which gradient boosting and ensemble methods typically outperform linear approaches, logistic regression achieved the highest AUROC (0.86), narrowly ahead of boosting methods and LSVM (0.85 to 0.86), random forest (0.84) and a neural network (0.77). This result is attributed principally to the use of fixed, literature-informed hyperparameters rather than cross-validated tuning a methodological choice made for computational tractability and is discussed as a limitation of the study design rather than as evidence of logistic regression's inherent superiority. The study further considers broader implications for model interpretability, algorithmic bias, data governance, and regulatory compliance under frameworks such as Basel III, and discusses the potential role of generative artificial intelligence (AI) and large language models (LLMs) in supporting future credit risk workflows, though these were not empirically tested in the present study but were administer through a survey. The findings underscore that predictive performance in credit risk modelling cannot be assessed in isolation from the methodological choices underpinning model development, and that transparent, interpretable models retain practical relevance even amid the growing sophistication of AI-based alternatives.

Keywords: Artificial Intelligence (AI); Banking; Credit Risk; Credit Scoring; Extreme Gradient Boosting (XGBoost); Generative AI (GenAI); Large Language Models (LLMs); Logistic Regression

Empowering potential of the smartphone in South Africa

Fallo Happy Khanye, Renette Blignaut, Julia Keddie, Koen Ponnet, Lieven De Marez

University of the Western Cape; Ghent University

Wednesday 25 November, 17:30 – 18:45 · board P09 · Poster session ↑ programme p. 33

South Africa is widely recognised as a mobile-first country, yet whether smartphones alone enable meaningful digital participation remains an important policy question. Using data from the nationally representative Digital Engagement South Africa (DESA) survey, collected through a stratified multistage cluster sampling design, this study examines whether exclusive smart-phone ownership provides sufficient access to digitally empowering opportunities. Smartphone-only users were compared with individuals owning multiple smart devices across a range of resource-enhancing digital activities.

The findings indicate that, despite high levels of smartphone ownership, substantial digital inequalities persist. Compared with users who own multiple smart devices, smartphone-only users are significantly less likely to engage with artificial intelligence applications, digital health services, e-government platforms, online financial transactions, and online learning. Exclusive smartphone ownership is also disproportionately concentrated among socioeconomically disadvantaged groups, placing them at greater risk of exclusion from meaningful participation in the digital economy and society.

These findings challenge the assumption that increasing smartphone ownership alone is sufficient to achieve digital inclusion. The study argues that digital inclusion policies should extend beyond expanding internet connectivity and smartphone access to promote affordable access to a broader ecosystem of digital devices, including computers and tablets. By demonstrating the limitations of smartphone-only access, this research provides evidence to inform policies aimed at reducing digital inequalities and fostering more equitable digital participation in South Africa.

Keywords: digital access; internet; smartphone; digital empowerment; meaningful

Human activity recognition: Personalisation through user calibration

Thutego Potsane, Stefan Janse Van Rensburg, Chantelle Clohessy

Nelson Mandela University

Wednesday 25 November, 17:30 – 18:45 · board P10 · Poster session ↑ programme p. 33

An activity recognition model works best when it has seen labelled examples from the user. However, asking new users to label many examples of their activities is inconvenient. A short calibration phase, where the user provides only a few labelled examples, may be a useful compromise. We investigate how much labelled data from an unseen user is needed to calibrate an impersonal human activity recognition model. We also consider whether that calibration significantly improves recognition performance. An ensemble model is trained using data from a group of subjects. It is then calibrated by progressively adding more labelled examples from a new unseen user. Recognition performance is evaluated throughout the calibration process to examine the relationship between the amount of user-specific data and model performance. Learning curves assess the accuracy of calibration and identify which activities benefit most from calibration. The results will show whether a short calibration phase delivers enough improvement to justify asking users for labelled examples.

Keywords: Human Activity Recognition; User Calibration; Personalization; Machine Learning; Learning Curves

Comparison of convolutional neural network architectures for photovoltaic fault classification using infrared imagery

Justin Marshall, Chantelle Clohessy, Warren Brettenny

Nelson Mandela University

Wednesday 25 November, 17:30 – 18:45 · board P11 · Poster session ↑ programme p. 33

Infrared thermography images of photovoltaic modules have been used to detect and classify faults. This approach has gained traction because traditional methods rely on manual inspection, which is labour-intensive and time-consuming given the scale of large PV plants. Previous studies have proposed convolutional neural networks (CNNs) for detecting and classifying PV faults. These methods perform well; however, more recent CNN architectures have not been tested in this domain. This study trains and evaluates DenseNet121 and EfficientNetB0 on infrared images of PV modules using five-fold cross-validation and a held-out test set, and compares them against a custom CNN developed for this study. The custom CNN achieved the highest test accuracy (91.9%), matching or exceeding both pretrained architectures (89.2% and 87.8%), with no statistically significant difference between models.

Keywords: convolutional neural networks; photovoltaic fault classification; infrared thermography; transfer learning; image classification

Bayesian modelling of extreme rainfall in flood-prone coastal and escarpment regions of KwaZulu-Natal

Felicia Sethosa

University of the Free State

Wednesday 25 November, 17:30 – 18:45 · board P12 · Poster session ↑ programme p. 33

Severe rainfall is a major challenge for flood management in KwaZulu-Natal. The estimation of rainfall extremes is challenging due to uncertainties and fluctuations over time and space. This research applied Bayesian extreme value models to enhance estimation accuracy and quantify uncertainty. Daily rainfall data from 1986 to 2026 were examined using two traditional approaches: annual maxima with the Generalized Extreme Value distribution and threshold exceedances with the Generalized Pareto Distribution. Initially, classical stationary and non-stationary models were fitted to establish baseline results. Following this, Bayesian hierarchical GEV and GPD models were formulated utilizing the Integrated Nested Laplace Approximation (INLA) and the Stochastic Partial Differential Equation (SPDE) framework to incorporate spatial dependence. As a further extension, a joint Bayesian spatio-temporal model was created to capture both spatial and temporal variations in annual maximum rainfall. The Bayesian models generated posterior estimates with credible intervals, and the spatio-temporal analysis revealed significant spatial differences and a small decline in extremes over time. These results indicate that Bayesian extreme value modelling with INLA-SPDE provides a more robust and dependable method for analysing rainfall extremes, facilitating flood-risk assessment and climate adaptation in KwaZulu-Natal.

Keywords: Generalized Extreme Value (GEV); Generalized Pareto Distribution (GPD); Stochastic Partial Differential Equation (SPDE); Stochastic Partial Differential Equation (SPDE); Rainfall extremes

Clustering daily solar irradiance patterns

Wehan Smith, Stefan Janse Van Rensburg, Chantelle Clohessy

Nelson Mandela University

Wednesday 25 November, 17:30 – 18:45 · board P13 · Poster session ↑ programme p. 33

The increasing use of solar energy in South Africa highlights the need for reliable methods to characterize solar irradiance variability. Understanding these patterns is important for solar-energy planning, forecasting and investment decision-making. Clustering supports these tasks by reducing a large set of daily profiles to a small number of representative day types. Clustering methods for identifying daily solar irradiance patterns have been widely investigated internationally. However, comparatively limited research has applied these approaches to South African locations. This project aims to apply and evaluate an established clear-sky-index-based clustering methodology. The methodology will be used to identify robust and physically meaningful daily solar irradiance patterns for selected South African locations using CAMS data.

Keywords: Solar irradiance; Clustering; Renewable energy; Clear-sky-index; Daily irradiance profiles

Forecasting of daily electricity demand in South Africa using machine learning models

Hodisane Clement Mamaila, Sisa Pazi, Siphumlile Mangisa

Nelson Mandela University

Wednesday 25 November, 17:30 – 18:45 · board P14 · Poster session ↑ programme p. 33

South Africa has historically experienced electricity generation capacity constraints, underscoring the importance of accurate electricity demand forecasting for effective system planning and operations. Hence this study aims to develop a machine learning-based statistical model to forecast electricity demand. The data used in this study are sourced from Eskom, spanning five years from 1 April 2021 to 31 March 2026. The Eskom dataset is supplemented with temperature data from the South African Weather Services, and feature engineering is conducted to derive calendar, lag, and rolling statistics features, which are used alongside temperature as predictors for electricity demand forecasting. Machine learning models used in the study are Extreme Gradient Boosting (XGBoost), Random Forests (RF), Long Short-Term Memory (LSTM), and an XGBoost-RF-LSTM ensemble. A walk-forward approach is employed to assess and compare the models using metrics such as the root mean squared error (RMSE), mean absolute error (MAE), and mean absolute percentage error (MAPE).

Keywords: Machine learning; Feature engineering; Electricity demand forecasting

Modelling relative humidity with upper boundary observations: A comparison of alternative beta distributions

Mbikazi Ketelo, Siphumlile Mangisa, Sonali Das

iYunivesithi Walter Sisulu; Nelson Mandela University; University of Pretoria

Wednesday 25 November, 17:30 – 18:45 · board P15 · Poster session ↑ programme p. 33

Relative humidity is an important environmental variable for understanding atmospheric moisture conditions and plays a significant role in processes such as cloud formation, precipitation, evaporation, and thermal comfort. It is a bounded, continuous variable taking values on the closed interval $[0, 100]$. The presence of observations at the upper boundary presents a modelling challenge for the standard beta distribution, whose support is restricted to the open interval $(0,1)$. This study considers alternative distributional approaches for modelling relative humidity data containing boundary observations. Four models are investigated: a standard beta distribution with boundary observations excluded, a beta distribution with boundary observations transformed, a One-Inflated Beta distribution, and a Scale-Location-Truncated Beta distribution. These models provide alternative approaches to accommodating the bounded nature and upper boundary behaviour of relative humidity. The parameters of the respective models are estimated using the method of maximum likelihood, and the models are compared using the Akaike Information Criterion and Bayesian Information Criterion.

Keywords: Relative humidity; bounded data; beta distribution; One-Inflated Beta distribution; Scale-Location-Truncated Beta distribution

Classification of statistical methodologies in South African medical literature: A review of SAMJ and SAJCC

Kutloano Mavovana, Erin Bromley-Gans

Nelson Mandela University

Wednesday 25 November, 17:30 – 18:45 · board P16 · Poster session ↑ programme p. 33

Clinicians and other medical professionals need to understand the statistical methods used in medical research in order to trust and apply findings. However, it has been reported in literature that the complexity of statistical methods is increasing over time. Consequently, the accurate interpretation of medical literature has become more challenging. To assess this for the South African landscape, this study evaluates the evolution of statistical complexity in the South African Medical Journal (SAMJ) and the South African Journal of Critical Care (SAJCC) between 2010 and 2025. Articles were randomly selected from both journals and the statistical methods used were recorded. The methods were then classified as Basic, Intermediate, or Advanced using a custom Statistical Complexity Framework built for this study and trends over time were assessed. The findings of this study may be used to evaluate the scope of statistical training required in medical education.

Keywords: Statistical complexity; Medical literature; Biostatistics education; South Africa; Statistical literacy

Robust calibration-based imputation scheme and estimators for finite population mean

Maria Lekganyane, Ahmed Audu, Sam Ntuli

Sefako Makgatho Health Sciences University; Busman Danfodiyo University, Nigeria

Wednesday 25 November, 17:30 – 18:45 · board P17 · Poster session ↑ programme p. 33

Missing data remain a major source of bias and efficiency loss in sample surveys, necessitating the development of robust imputation procedures that effectively utilize auxiliary information. This study proposes two robust and improved imputation estimators based on the calibration approach for estimating the finite population mean under missing observations. The performance of the proposed estimators was evaluated through an empirical study and compared with the conventional sample mean and several well-established imputation estimators available in the literature. The comparison was based on bias, Mean Squared Error (MSE), and Percentage Relative Efficiency (PRE) under two sampling scenarios. The empirical results reveal that both proposed estimators consistently outperform the existing methods in terms of estimation efficiency. Both estimators also exhibited negligible biases comparable to or smaller than those of the competing methods. The findings demonstrate that the proposed estimators effectively exploit auxiliary information to substantially reduce estimation error while maintaining high accuracy. Their consistent superiority across different sample sizes confirms their robustness, reliability, and practical suitability for imputing missing data in finite population surveys.

Keywords: Robust; Imputation; Estimator; Calibration; Population mean

Comparison of Cox proportional hazards and Fine–Gray competing risk models for loss to follow-up among tuberculosis patients in Kenya

Nelson Ouma, Hillary Katsabola, Moses Ngari, Innocent Maposa

Stellenbosch University; Kenya Medical Research Institute (KEMRI)-Wellcome Trust Research Programme, Kenya

Wednesday 25 November, 17:30 – 18:45 · board P18 · Poster session ↑ programme p. 33

Background: Tuberculosis (TB) is a preventable and curable infectious disease; however, it remains a major cause of morbidity and mortality globally. Beyond mortality and treatment interruption, loss to follow-up (LTFU) continue to pose major challenges to TB control programmes and the achievement of successful treatment outcomes. Terminating TB treatment outcomes especially in survival data setup is due to one of several possible causes. However, experiencing a LTFU event usually precludes an individual from experiencing any other type of events such as death, such type of events is typically referred to as competing risks. In this study, therefore, we aim to compare and evaluate the performance of conventional Cox-PH and the Fine-Gray models in estimating the cumulative incidences based on simulated and secondary TB data from Kenya.

Methods: Using a retrospective study design, Bayesian cause-specific hazard and Fine-Gray subdistribution Stan models will be fitted on simulated and secondary data with the primary estimand; cumulative incidence function (CIF) for LTFU in the presence of death as a competing event. Model performance under varying event rates, and percent censoring will be evaluated using bias, mean squared error (MSE), and coverage. We will additionally compare cause-specific hazard ratios (HRs) and the subdistribution hazard ratios (sHRs) and assess the consistency of covariate effects across modelling approaches.

Results: Preliminary simulation results indicate good recovery of prespecified parameters under the data-generating mechanism, with estimated cumulative incidence and regression parameters showing good agreement with their corresponding true values. Further simulations scenarios are being conducted to evaluate model performance across varying sample sizes, competing-event rates, censoring mechanisms, and effect sizes.

Conclusion: This study will generate evidence on how different analytical approaches influence estimates of incidence and risk factors for LTFU while contributing to methodological best practice in TB research. The results will also support more reliable identification of patients at risk of treatment disengagement and contribute to improved TB treatment retention strategies.

Keywords: Tuberculosis; Loss to follow-up; Bayesian; Cumulative Incidence Function

CRISP-NAM for survival analysis under competing risks using risk set sampling

Michael Falk, Justine Nasejje, Steven James

University of the Witwatersrand

Wednesday 25 November, 17:30 – 18:45 · board P19 · Poster session ↑ programme p. 33

Deep survival models for competing risks face severe computational bottlenecks due to exhaustive risk-set evaluations. To resolve this, this study integrates a novel dynamic, entropy-guided risk-set sampling strategy into an interpretable Neural Additive Model (CRISPNAM). Evaluated across synthetic and clinical datasets (PBC, Framingham), traditional sampling methods like Nested Case-Control and Static Case-Cohort have novel perfomace gain along with some calibration failures under high compression. Conversely, the Entropy sampler intelligently targets high-variance subjects. By leveraging stabilized Inverse Probability Weighting, the optimal entropy configuration acts as a data-level regularizer. It successfully isolated ground-truth biostatistical features and matched the Full Cohort baseline predictive performance (achieving a C-index of 0.851 on synthetic data) while reducing computational training time by over 50%. Ultimately, this variance-driven dynamic sampling provides a computationally scalable, highly calibrated, and clinically transparent framework for large-scale competing risk analysis.

Keywords: Deep Survival Analysis; Competing Risks; Fine-Gray Model; Risk-Set Sampling; Neural Networks

From risk factors to risk trajectories: Integrating survival analysis, machine learning and explainable AI for first exposure to intimate partner violence in South Africa

Tshiamo Medupe, Inoccent Maposa, Tonya Esterhuizen, Nwabisa Shai

Stellenbosch University; University of the Witwatersrand; South African Medical Research Council

Wednesday 25 November, 17:30 – 18:45 · board P20 · Poster session ↑ programme p. 33

Intimate partner violence (IPV) remains a major public health and social challenge, yet statistical analyses of IPV often focus on identifying associations between individual risk factors and violence exposure. While such analyses provide population-level insights, they may overlook the temporal nature of first exposure and the possibility that risk emerges through complex, nonlinear and interacting patterns. This study explores how survival analysis, statistical learning and explainable artificial intelligence can be integrated to move beyond conventional risk-factor analysis towards a dynamic and predictive understanding of first IPV exposure. Using data from the 2016 South Africa Demographic and Health Survey, we investigate first exposure to IPV among women aged 15–49 years. A person-period dataset is constructed to represent the time-to-first-event process, enabling discrete-time survival modelling within a nationally representative survey framework. Penalised regression and machine-learning approaches are incorporated to investigate whether nonlinearities, interactions and higher-order predictor patterns can improve prediction beyond conventional statistical modelling. This brings together three complementary questions: Which factors are associated with IPV? When does risk emerge? And how accurately can that risk be predicted and explained? Model discrimination, calibration and predictive performance will be evaluated, alongside explainability methods to examine how behavioural, demographic and socioeconomic characteristics contribute to predicted risk. Rather than treating statistical inference and machine learning as competing paradigms, we position them as complementary tools. Survival analysis- principled framework for modelling time to first exposure, machine learning- complex predictive structure, and explainability connects prediction back to interpretable population-health insights. By integrating these perspectives, we seek to demonstrate how population-based epidemiological research moves from asking "who is associated with IPV?" towards "who is at elevated risk, when does that risk emerge, and what drives that prediction?" The framework may provide a generalisable approach for integrating inference, prediction and explainability in complex time-to-event outcomes in population health.

Keywords: Discrete-time survival analysis; Machine learning; Explainable artificial intelligence; Intimate partner violence; Population health

Modelling and forecasting of daily gold prices using an ARIMA–GARCH hybrid model

Happy Maluleke

University of Limpopo

Wednesday 25 November, 17:30 – 18:45 · board P21 · Poster session ↑ programme p. 33

Gold is an important precious metal that serves as jewelry, a commodity, and an investment asset. However, gold prices can fluctuate considerably and exhibit periods of high and low volatility, making accurate forecasting challenging. This study investigates the application of a hybrid ARIMA–GARCH model for forecasting daily gold prices. Secondary daily gold-price data were obtained from Yahoo Finance and the World Bank for the period 2 January 2020 to 31 December 2024. The ARIMA component was used to model the mean structure and dependence in the series. Since the daily gold-price series was initially non-stationary, first-order differencing was applied to achieve stationarity. Residual analysis subsequently revealed evidence of time-varying volatility and volatility clustering, motivating the use of GARCH to account for conditional heteroskedasticity. Two competing models, ARIMA(1,1,1)–GARCH(1,1) and ARIMA(1,1,0)–GARCH(1,1), were fitted and evaluated using the Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC/SBC), Mean Absolute Percentage Error (MAPE), parameter significance, and residual diagnostics, including the Shapiro–Wilk test. The results favoured the ARIMA(1,1,0)–GARCH(1,1) model, which recorded an AIC of 10,932.435, BIC/SBC of 10,937.560, and MAPE of 105.918%. The model provided a more parsimonious representation while maintaining competitive forecasting performance. Residual diagnostics indicated no significant remaining autocorrelation, suggesting that the dependence structure was adequately captured. The ARCH(1) and GARCH(1) coefficients were significant at 0.0525 ($p = 0.0007$) and 0.9280 ($p < 0.0001$), respectively, indicating strong persistence in gold-price volatility. Overall, the findings demonstrate that the ARIMA(1,1,0)–GARCH(1,1) model can effectively account for both the mean behaviour and time-varying volatility of daily gold prices, providing useful information for investment decision-making.

Keywords: daily gold price; hybrid ARIMA-GARCH; forecasting; investment

Dynamic prediction of disease progression using longitudinal biomarkers: Joint modeling and landmarking

Melkamu M. Ferede, Ding-Geng Chen

University of Pretoria

Wednesday 25 November, 17:30 – 18:45 · board P22 · Poster session ↑ programme p. 33

The increasing availability of biomedical follow-up data with repeated longitudinal biomarker measurements requires statistical methods for accurate individualized prediction of disease progression. Unlike conventional risk models based on baseline information, dynamic prediction methods update risk as new longitudinal data become available. Several approaches have been proposed, including landmarking and joint modeling (JM). However, limited evidence compares these methods regarding predictive performance, interpretability, and computational efficiency. Landmarking is widely used for its simplicity, whereas JM models longitudinal biomarker trajectories and time-to-event outcomes while accounting for their association and measurement error. Consequently, clinicians and researchers lack clear guidance on selecting the most appropriate method. This study addresses this gap by comparing landmarking (two-stage) and joint modeling approaches through simulation studies, with applications to colorectal cancer and/or chronic kidney disease data. The findings provide guidance on method selection, demonstrate how integrating longitudinal biomarkers, baseline covariates, and survival outcomes improves personalized prediction of disease progression, supporting disease monitoring, timely intervention, and informed clinical decision-making.

Keywords: Dynamic risk prediction; Longitudinal biomarkers; Survival analysis; Joint modeling; Landmarking

Factors associated with consistent condom use among sexually active HIV-negative PrEP users in South Africa: Results from the 2022 South African HIV household survey

Goitseone Maseko, Lesiba Molopa, Adlai Davids, Thembelihle Ginyana, Lehlogonolo Makola, Sean Jooste, Sbonelo Chamane, Musawenkosi Mabaso, Sizulu Moyo, Nompumelelo Zungu, Khangelani Zuma

Human Sciences Research Council, Pretoria, South Africa

Wednesday 25 November, 17:30 – 18:45 · board P23 · Poster session ↑ programme p. 34

Aim: Pre-exposure Prophylaxis (PrEP) is a biomedical strategy to prevent new HIV infections that also requires the combined use of condoms to prevent sexually transmitted infections and unplanned pregnancies. Despite efforts to expand PrEP uptake, its combined use with condoms remains low. This analysis examines factors that influence the combined use of condoms and PrEP in South Africa.

Methods: Data from the 2022 South African National HIV Prevalence, Incidence, and Behaviour Survey were analysed among sexually active, HIV-negative persons aged ≥15 years currently using PrEP. Consistent condom use with the most recent sexual partner was the primary outcome. Pearson's chi-squared tests were used for descriptive analysis. Variables associated with the outcome at $p<0.20$ in bivariate analysis were included in a multivariable modified Poisson regression model with robust standard errors to estimate adjusted risk ratios (ARRs) and 95% confidence intervals (CIs).

Results: Among 298 PrEP users, consistent condom use was low (28.8%; 95% CI: 19.9–39.8). PrEP users with correct HIV knowledge and who rejected HIV myths were less likely to report consistent condom use (ARR=0.43, 95% CI: 0.22–0.85; p=0.016). Single PrEP users were more likely to report consistent condom use than married users (ARR=7.52, 95% CI: 1.28–44.13; p=0.025). Female users had a marginally lower likelihood of consistent condom use than males (ARR=0.59, 95% CI: 0.34–1.01; p=0.056). Age, geographic location, number of sexual partners, and self-perceived HIV risk were not significantly associated with consistent condom use.

Conclusion: PrEP use and consistent condom use were low, highlighting gaps in combination HIV prevention. While PrEP protects against HIV, condoms remain important for preventing other sexually transmitted infections and unintended pregnancies. Targeted interventions should therefore promote PrEP uptake alongside consistent condom use to strengthen dual protection.

Keywords: Consistent condom use; PrEP; household survey; combination prevention therapy

Assessment and comparison of error metrics used for photovoltaic energy forecasting

Tatenda Nyoni, Chantelle Clohessy, Stefan Janse van Rensburg

Nelson Mandela University

Wednesday 25 November, 17:30 – 18:45 · board P24 · Poster session ↑ programme p. 34

Accurate energy forecasting is essential for integrating photovoltaic (PV) systems into the power grid while maintaining stability and economic efficiency. This project investigates current statistical techniques used in forecasting power output across various time horizons for PV systems. Historical power data, combined with meteorological variables such as solar irradiance and temperature, will be analysed to identify underlying seasonal patterns and irregular fluctuations. In addition to examining forecasting approaches, the study places particular emphasis on the evaluation of error performance metrics. These metrics include Mean Absolute Error (MAE), Root Mean Square Error (RMSE), Mean Absolute Percentage Error (MAPE), and Mean Bias Error (MBE). These metrics will be used to assess the accuracy, bias, and overall reliability of the forecasts under different weather conditions. By analysing both the forecasting techniques and their associated error measures, the research aims to provide a clearer understanding of model performance and limitations. The findings support improved grid management, operational planning, and the effective integration of solar energy into the power system.

Keywords: Photovoltaic power forecasting; forecast error metrics; forecast evaluation

Prevalence and risk factors associated with HIV infection among pregnant antenatal care attendees in Limpopo Province, South Africa

Itumeleng Tebele Boshomane, Kontswi Donald Tshabalala, Oratilwe Penwell Mokoena, Thembelihle Sam Ntuli

Sefako Makgatho Health Science University

Wednesday 25 November, 17:30 – 18:45 · board P25 · Poster session ↑ programme p. 34

Early screening for HIV infection provides an opportunity for mother-to-child transmission and optimizes the care of HIV-infected mothers and unborn babies to improve clinical outcomes. This study aimed to determine the prevalence and the risk factors associated with HIV infection among pregnant women attending antenatal care at the District Hospital and its feeder community health center of the Limpopo Province (LP), South Africa. The study was a cross-sectional descriptive study carried out over 2-months from 01 May 2019 to 30 June 2019. A consecutive sample of pregnant women who attended antenatal care during the study period was asked to participate. In total, 211 pregnant women participated in this study. Their mean age was 28.4±5.7 years, ranging from 18 to 41 years. More than half (56.4%) were aged <30 years old, 51.7% had secondary education, 71.1% were unmarried, and 72.0% were unemployed. The majority (66.4%) of pregnant women had multiple pregnancies and 70.6% were in the third trimester. Few (0.95%, n=2) had a history of alcohol use. The HIV prevalence was 15.2%, and significantly high in illiterate, elementary school educated and multiparous women. The HIV infection rate in this setting is relatively associated with the level of education and parity. The social risk factors of health in each municipality should be considered when local health authorities implement policies. Women should be continually provided with health education about modes of transmission of HIV prevention particularly those with lower levels of education and reproductive age.

Keywords: South Africa; Limpopo Province; Capricorn District; HIV prevalence; Pregnant Women

Calibrating Basel III capital adequacy for South Africa: A vector error correction and copula-based analysis of the trade-off between bank safety and lending capacity

Keneuoe Rasupu, Kojo Essael-Mensah

University of the Free State

Wednesday 25 November, 17:30 – 18:45 · board P26 · Poster session ↑ programme p. 34

South Africa applied a capital adequacy requirement that is substantially above the Basel III minimum once the South African Reserve Bank (SARB)'s Pillar 2A add-on and Domestic Systemically Imported Bank buffer are included. This requirement has not previously been evaluated through a locally calibrated analysis balancing bank safety against lending capacity. This study addresses this gap by jointly estimating the relationship between the capital adequacy ratio (CAR), bank safety and lending capacity for the South African banking sector, using monthly sector-wide aggregate data from January 2013-July 2025 and an explicit calibration framework incorporating macroeconomic stress conditions. The relationship between CAR, a capital-exclusive systemic Z-score (safety), loan growth (lending) and macroeconomic drivers was estimated using a Vector Error Correction Model (VECM), following a Johansen cointegration test that identified two long-run equilibrium relationships among the variables. The joint distribution of safety and lending shocks was modelled using a copula, where Clayton copula outperformed the Gaussian, Student-t and Gumbel alternatives. Counterfactual safety and lending outcomes across a range of capital adequacy ratios were generated using Monte Carlo simulation, solving the estimated long-run equilibrium directly for each capital linked residual variation under baseline, adverse and severely adverse macroeconomic scenarios. The results show bank safety and lending capacity decline together as CAR rises rather than trading off across the tested range and in every scenario. In conclusion, the findings indicate that higher CARs are associated with reduced lending and lower equilibrium values of the capital-exclusive systemic safety across estimated scenarios. Under severely adverse scenarios, none of the CAR levels between 10% and 16% satisfy the specified safety-floor criterion

Keywords: capital adequacy ratio; bank lending; Vector Error Correction Model; Monte Carlo simulation; macroprudential regulation

Comparative analysis of the performance of the standard SIR model across COVID-19 waves in South Africa

Olaoluwa Odeyemi, Tolulope Adeniji

University of Fort Hare

Wednesday 25 November, 17:30 – 18:45 · board P27 · Poster session ↑ programme p. 34

This study assesses the performance of the standard Susceptible-Infected-Recovered (SIR) model across the Beta, Delta, and Omicron COVID-19 waves in South Africa. Secondary daily COVID-19 data was used to construct the three compartments of the standard SIR model which consists of confirmed cases, recoveries, deaths, and population sizes. The model was fitted to each variant-driven wave and then using numerical integration and nonlinear least-squares estimations the transmission and recovery rates, as well as basic reproduction number for each wave were obtained. Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), coefficient of determination (R^2) were the metrics used to assess the performance of the standard SIR model. Observed-versus-predicted epidemic curves and residual plots were also evaluated.

The estimated transmission rates were 0.49331, 0.50948, and 0.86467 for the Beta, Delta, and Omicron waves. The R_0 basic reproduction number values following each wave were then 1.07037, 1.11217 and 1.14064. The Omicron wave exhibited the highest transmission rate, while the Delta wave produced the highest R^2 value with the lowest RMSE and MAE values. The strongest overall fit was found on the Delta wave. The model reproduced general epidemic trends, however systematic patterns in the residuals showed that the temporal dynamics of the three waves were not fully captured. The results showed that standard SIR model is most reliable when depicting the general behavior of an infectious disease such as the COVD-19 virus, as it is limited due to it simplified assumptions.

Keywords: SIR model; COVID-19; Parameter estimation; Epidemic waves

Evaluating growth trajectories in HIV-exposed uninfected children using a calibrated simulated HIV-unexposed comparator population

Mthobisi M Zondi, Henry G Mwambi, Sileshi Fanta Melesse, Nonhlanhla Yende-Zuma

University of KwaZulu-Natal

Wednesday 25 November, 17:30 – 18:45 · board P28 · Poster session ↑ programme p. 34

Background: HIV-exposed uninfected (HEU) children often show poorer linear growth than HIV-unexposed uninfected (HUU) children, but real-world HUU comparators are frequently unavailable. Simulation-based counterfactuals offer an alternative, yet their plausibility requires rigorous evaluation.
Methods: We simulated a HUU reference population using published growth parameters calibrated to local characteristics from a HEU cohort across four African countries. Latent class growth mixture models were fitted for height-for-age (HAZ) and weight-for-age (WAZ) z-scores from 0 to 60 months in both the observed HEU cohort and the simulated HUU population. Country effects on class membership were compared using multinomial logistic regression.
Results: Entropy values were modest (0.52-0.67). In the simulated HUU population, South Africa (OR 1.86, 95% CI 1.17-2.95) and Zimbabwe (OR 1.45, 95% CI 1.04-2.03) showed higher odds of favourable HAZ class membership than Malawi. In the HEU cohort, these countries showed lower odds of the most unfavourable HAZ class (South Africa: OR 0.13, 95% CI 0.07-0.24; Zimbabwe: OR 0.17, 95% CI 0.10-0.27). Strong positive correlations were observed between HEU and HUU high-growth trajectories (r = 0.96 for WAZ). WAZ country effects were absent in the simulated population but emerged in the HEU cohort (South Africa Class 2 OR 5.14, 95% CI 2.07-12.74).
Conclusion: A calibrated simulated HUU population provides a plausible counterfactual. HIV exposure modifies geographic growth patterns in complex ways, supporting simulation-based approaches when real comparators are unavailable.

Keywords: SIR model; COVID-19; Parameter estimation; Epidemic waves

Assessing copula symmetry and tail dependence in multi-source temperature data from South Africa

Vusi Masingi, Gerrit Grobler, Shawn Liebenberg

North-West University

Wednesday 25 November, 17:30 – 18:45 · board P29 · Poster session ↑ programme p. 34

Reliable meteorological data is essential for agricultural risk management. This is critical in South Africa, where weather station records in agriculturally active regions are often incomplete and have limited historical coverage. The aim of this project is to apply copula models to the imputation of incomplete weather station records, with particular emphasis on tail behaviour, and to the bias correction of ERA5 reanalysis data from the European Centre for Medium-Range Weather Forecasts (ECMWF). To achieve these objectives, we investigate the dependence structure between reference temperatures at weather stations (WS) and auxiliary ERA5 and neighbouring WS observations. The analysis focuses on departures from radial and reflection symmetry, as well as on tail dependence. To capture this dependence, we fitted and compared six copulas: two symmetric copulas, two that are reflection symmetric but radially asymmetric, and two capable of capturing both radial and reflection asymmetry. The results revealed that the Khoudraji Clayton-Gumbel copula provided a better fit in most cases, consistent with the detected asymmetry in the maximum temperature pairs. However, for minimum temperatures, the Student-t copula yielded a slightly better fit, reflecting the more symmetric dependence structures observed for this pair. The results further show stronger upper-tail dependence for maximum temperature in the WS–ERA5 comparison, whereas the WS–WS comparison exhibits a more symmetric dependence structure. In the WS-ERA5 comparison, lower tail dependence is weaker for minimum temperature. This suggests that ERA5 is well suited for modelling heat stress, as it captures extreme high temperatures reliably, but is less suitable for modelling cold stress because it does not represent extreme low temperatures as accurately.

Keywords: Tail dependence; Copulas; temperature extremes; bivariate symmetry; multi-source climate data

Notes