# 61ST ANNUAL CONFERENCE OF THE SOUTH AFRICAN STATISTICAL ASSOCIATION

25 - 29 November 2019
Nelson Mandela University
Port Elizabeth, South Africa

## PROGRAMME & ABSTRACTS

SASA
2019
the 61st annual conference of SASA

SAS® Certified Industry Professor

Show industry, faculty and students you have the SAS® credentials.

To find out more, contact us at: intouch@sas.com

THE POWER TO KNOW.

# INTRODUCTION

The South African Statistical Association (SASA) and the Nelson Mandela University Department of Statistics are proud to host the 61st annual SASA conference. The SASA conference will be held from 27 November to 29 November 2019 on the South Campus of Nelson Mandela University in Summerstrand, Port Elizabeth.

We welcome all to the Nelson Mandela University in the beautiful coastal city of Port Elizabeth in the Eastern Cape. Nelson Mandela University is the largest higher education institution in the Eastern Cape, with approximately 25 000 students enrolled across six different campuses. The Summerstrand South Campus, the hosting site for the SASA 2019 conference is one of the six campuses across the Eastern and Western Cape and is situated on a nature reserve. The nature reserve conserves fynbos unique to the area and is home to wildlife including springbok, zebra and red hartebeest.

# REGISTRATION

Registration for the conference with take place in the following venues, at the times specified.

| | | |
|---|---|---|
| Monday & Tuesday (25-26 November) | 08h00 – 09h00 | Foyer, Conference Centre, Summerstrand North Campus |
| Wednesday (27 November) | 07h30 – 12h00 | Foyer, Building 123, Summerstrand South Campus |
| Thursday (28 November) | 07h45 – 10h00 | Foyer, Building 123, Summerstrand South Campus |

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

# PARKING

There is sufficient parking for delegates close to the venues. Follow the signs once you enter the Nelson Mandela University campuses.

# NAME TAGS

Delegates should wear their name tags at all times to gain access to the lecture halls, tea breaks, lunches and social functions.

# TEA & MEALS

Teas will be served in the North Campus Conference Centre during the workshops, and in Building 123 during the conference in the allocated sessions in the programme.

Lunch meals will be served in the North Campus Conference Centre during the workshops, and in the Chancellor's Room in building 24 (the indoor sports centre), a seven-minute walk from building 123, during the allocated sessions in the programme. Please take note that there is a premium on space in the lunch venue and delegates are asked to move through the system in and orderly and prompt manner to simplify the space restriction.

Cover photographs by *Aviwe Gqwaka Photography*.

## INTERNET

Wireless internet will be made available to the delegates on campus during the conference. Delegates can choose to us the "eduroam" Wi-Fi network, provided that they have access. Other Wi-Fi networks, and the login details, will be communicated to the delegates at the conference.

## SESSION VENUES

All session venues used during the conference can be found on the maps on pages 67 and 68. Use the following table to identify the locations of the venues on the maps.

| Venue | Map Location |
|---|---|
| Conference Centre (North C) | North Campus, Building 235 |
| Building 123 | South Campus, alongside parking lot in Aloe road |
| Building 35 | South Campus, alongside Building 123 |
| Building 24 | South Campus, corner of Aloe and Protea road. |
| Building 7 (Venues 0246 and 0248) | South Campus, alongside parking lot in Springbok road. |
| Auditorium (South C) | South Campus, alongside Main Building |

## MEETINGS AND SOCIAL FUNCTIONS

**Meet and Greet**

Monday 25 November at 17h00

Venue: Conference Centre, North Campus

**SASA Executive meeting**

Tuesday 26 November at 15h45 – 17h00

Venue: Conference Centre, North Campus, Theatre

**MDAG Annual General Meeting**

Tuesday 26 November at 15h45 – 17h00

Venue: Conference Centre, North Campus, Venue 3

**Meet and Greet**

Tuesday 26 November at 17h00

Venue: Conference Centre, North Campus

**Opening Ceremony**

Wednesday 27 November at 09h15 – 10h30

Venue: South Campus Auditorium, Main Building

**SASA AGM**

Wednesday 27 November at 15h00 – 15h40

Venue: Building 123, Venue; Lecture Theatre no. 1

**Welcoming Function**

Wednesday 27 November at 17h30 - 19h30

Venue: Building 123, Venue; Foyer

**Young Statisticians' Function**

Wednesday 27 November at 19h15 – 21h30

Venue: Bridge Street Brewery, Lower Baakens Valley

**IBS Group SA Meeting**

Thursday 28 November at 16h20 – 16h50

Venue: Building 123, Venue; Lecture Theatre no. 3

**Bayes Group Meeting**

Thursday 28 November at 16h55 – 17h25

Venue: Building 123, Venue; Lecture Theatre no. 3

**Data Science Group Meeting**

Thursday 28 November at 16h30 for 17h30

Venue: Building 35, Venue; 35 00 40

**Gala Dinner**

Thursday 28 November at 18h30 for 19h00

Venue: The Willows, Marine Drive

# SASA 2019 ORGANISING COMMITTEES

## Local Organising Committee

Chair: Prof Gary Sharp

Department of Statistics

Nelson Mandela University

Port Elizabeth

South Africa

6019

Members

Ms Rae Vincent le-Roux

Dr Chantelle Clohessy

Mr Sisa Pazi

Dr Warren Brettenny

Prof Lizanne Raubenheimer (Finances)

## Scientific Committee

Chair: Prof Gary Sharp

Department of Statistics

Nelson Mandela University

Port Elizabeth

South Africa

Members

Dr Johan Hugo

Dr Chantelle Clohessy

Dr Warren Brettenny

Prof Leonard Santana (Proceedings)

Dr Humphrey Brydon (Young Statisticians Competition)

Prof Delia North, Dr Lizelle Fletcher and Ms Yoko Chhana (Industry and Education Joint Session)

# GUIDELINES TO SPEAKERS AND CHAIRPERSONS

*Speakers*

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

*Chairpersons of sessions*

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

*The above instructions are intended as a basic guideline for the sessions.*

*Please use your own initiative in the sessions to keep them running smoothly.*

# GUIDELINES TO POSTER PRESENTERS

**Venue and Time:** All posters will be viewed on Wednesday 27 November from 17h00 – 18h30.

*Poster presenters must setup their posters during the lunch break on Wednesday*. As per the information circulated to delegates prior to the conference, boards which can take posters of size A0 in portrait format will be available for each delegate. Boards will be demarcated with a delegate's name to assist with the poster competition assessment. Please use the board that you have been allocated and do not remove names from the boards.

# PROGRAMME

| MONDAY: 25 NOVEMBER 2019 | |
|---|---|
| | |
| *08h00 - 09h00* | **Registration**<br>**Registration venue: Conference Centre; North campus**<br>Venue: Theatre, North Campus Conference Centre |
| *09h00 - 11h00* | *Introduction to the use of SAS for analysing longitudinal data with mixed models*<br>Workshop 1, Session 1<br>Prof Emmanuel Lesaffre |
| *11h00 - 11h30* | **Tea**<br>Venue: Conference Centre |
| *11h30 - 13h00* | *Introduction to the use of SAS for analysing longitudinal data with mixed models*<br>Workshop 1, Session 2<br>Prof Emmanuel Lesaffre |
| *13h00 - 14h00* | **Lunch**<br>Venue: Conference Centre |
| *14h00 - 15h30* | *Introduction to the use of SAS for analysing longitudinal data with mixed models*<br>Workshop 1, Session 3<br>Prof Emmanuel Lesaffre |
| *15h30 - 15h45* | **Tea**<br>Venue: Conference Centre |
| *15h45 - 16h45* | *Introduction to the use of SAS for analysing longitudinal data with mixed models*<br>Workshop 1, Session 4<br>Prof Emmanuel Lesaffre |
| *17h00* | **Meet and Greet**<br>Venue: Conference Centre, North Campus |

| TUESDAY: 26 NOVEMBER 2019 | | | | |
|---|---|---|---|---|
| | | | | |
| *08h00 - 09h00* | **Registration** | | | |
| | **Venue: Conference Centre, North Campus** | | | |
| | **WORKSHOP 2** | | | |
| | **Getting Started in Bayesian Statistics** | | **WORKSHOP 3** | **WORKSHOP 2** |
| | **Prof Tim Swartz** | | **Ideas for Bringing Data Science into Statistics Education** | **Multivariate Data Analysis Group** |
| | Venue: Theatre, Conference Centre | | **Prof Deborah Nolan** | **Various Presenters** |
| *08h30 - 10h30* | *Session 1* | | Venue: D101, North Campus | Venue: Venue 2, Conference Centre |
| | | *09h00 - 11h00* | *Session 1* | *Session 1* |
| *10h30 - 11h00* | **Tea** - Conference Centre Foyer | | | |
| *11h00 - 12h45* | *Session 2* | *11h00 - 11h30* | **Tea** - Conference Centre Foyer | |
| | | *11h30 - 13h15* | *Session 2* | *Session 2* |
| *12h45 - 13h30* | **Lunch**<br>Conference Centre Foyer | | | |
| | | *13h15 - 14h00* | **Lunch**<br>Conference Centre Foyer | |
| *14h00 - 15h45* | Vacant | *14h00 - 15h30* | *Session 3* | *Session 3* |
| | | *15h30 - 15h45* | **Tea** - Conference Centre Foyer | |
| *15h45 - 17h00* | **SASA Exec meeting** | *15h45 - close* | *Session 4* | **Multivariate Data Analysis Group**<br>**AGM** |
| *17h00* | **Meet and Greet** | | | |
| | Venue: Conference Centre Foyer | | | |

| WEDNESDAY: 27 NOVEMBER 2019 | | | | | |
|---|---|---|---|---|---|
| | | | | | |
| 08h00 - 14h30 | **Registration** | | | | |
| | Venue: Foyer, Building 123 (South Campus) | | | | |
| 09h15 - 10h30 | **Opening Ceremony** | | | | |
| | Venue: Auditorium | | | | |
| | 09h15 - 09h20 | MC: SASA Past-President: Prof Danie Uys | | | |
| | 09h20 - 09h30 | Welcoming: Nelson Mandela University, Dean of Science: Prof Azwinndini Muronga | | | |
| | 09h30 - 10h00 | Presidential address: SASA President: Prof Maseka Lesaoana | | | |
| | 10h00 - 10h20 | Awards: SAS awards for best honours projects: Ms Nombuso Zondo (hand over by Mr Murray de Villiers) | | | |
| | | Award: Post Graduate paper competition winner: Dr Lizelle Fletcher (o.b.o. Fransonet Reynecke) | | | |
| | | Award: Sichel medal: Prof Paul Fatti | | | |
| | | Awards: Fellowship and Honorary Members: Prof Danie Uys | | | |
| | | Award: Thought leader: Prof Danie Uys | | | |
| | 10h20 - 10h30 | Platinum Sponsor Address: SAS, Mr Murray de Villiers | | | |
| 10h30 - 11h00 | **Tea** | | | | |
| | Venue: Foyer, Building 123 | | | | |
| 11h00 - 11h55 | **Plenary Session** | | | | |
| | Venue: Lecture Theatre 1, Building 123 | | | | |
| | 11h00 - 11h45 | Guest and Title | Prof Tim Swartz: Sports Analytics: Reflections and some current projects | | |
| | | Chair | Dr Warren Brettenny | | |
| | 11h45 - 11h55 | Questions | Audience | | |
| 11h55 - 12h00 | **Parallel Sessions** | | | | |
| 12h00 - 13h00 | Stream | Young Statisticians | Biostatistics | Bayesian Statistics | Young Statisticians |
| | Chair | Johane Nienkemper-Swanepoel | Danielle Roberts | Allan Clark | David Hofmeyr |
| | Venue | Lecture Theatre 1 | Lecture Theatre 2 | Lecture Theatre 3 | 35 00 40 |
| 12h00 - 12h15 | | Deep Learning for Photovoltaic Defect Classification through Thermal Infrared Imaging | Joint Modelling of Anaemia and Malaria in Young Children using Data from Complex Survey Designs | Bayesian Semi-Parametric Linear Mixed Model using Smoothing Spline: Application to Longitudinally Measured Fasting Blood Sugar Level Data | Statistical Modelling of Decomposed Rainfall Time Series and Generalised Extreme Value Distribution |
| | | *Christopher Dunderdale* | *Danielle Roberts* | *Tafere Aniley* | *Willard Zvarevashe* |

| WEDNESDAY: 27 NOVEMBER 2019 | | | | | |
|---|---|---|---|---|---|
| 12h20 - 12h35 | | Machine Learning ALSI 40 Index Futures Option Prices | Survival Analysis of First to Second Childbirth Interval among Women in South Africa: Rural-Urban Differential | Estimation of the Degrees of Freedom for the Student t-Distribution using a Bayesian Procedure | A Goodness-of-Fit Test for the Rayleigh Distribution Based on the Mellin Transform |
| | | *Duncan Saffy* | *Rotimi Afolabi* | *Abrie van der Merwe* | *Shawn Liebenberg* |
| 12h40 - 12h55 | | Extended Applications of GPAbin Biplots | Joint Model of Longitudinal and Survival Data for Covariates Measures with Correlated Error | A Gibbs Sampler for Multi-Species Site Occupancy Models | Factorisation Machines for Recommender Systems |
| | | *Johane Nienkemper-Swanepoel* | *Adeboye Azeez* | *Allan Clark* | *Bronwyn Dumbleton* |
| 13h00 - 14h10 | **Lunch** | | | | |
| | Venue: Chancellor's Room, Sport Centre | | | | |
| | **NOTE: All Posters must be set up prior to the end of lunch** | | | | |
| 14h15 - 14h20 | **Parrallel Sessions** | | | | |
| 14h20 - 15h00 | Stream | Development of Spatial Statistics | Statistics in Sport | Applied Statistics | Quality Control |
| | Chair | Michaela Ritchie | Paul van Staden | Esta Bekker | Jean Claude Malela-Majika |
| | Venue | Lecture Theatre 1 | Lecture Theatre 2 | Lecture Theatre 3 | 35 00 40 |
| 14h20 - 15h40 | 14h20 - 14h35 | New Local Measures for Geostatistical and Lattice Data | Multiple Two-Sample Comparisons Based on a Gradual Change Model | Usefulness of Applying Power Allocation in Survey Sampling in South Africa | Distribution-Free Precedence Schemes with a Generalized Runs-Rule for Monitoring Unknown Location |
| | | *Christine Kraamwinkel* | *Zdenek Hlavka* | *Esta Bekker* | *Jean Claude Malela-Majika* |
| | 14h40 - 14h55 | Sub-Pixel Land Cover Classification in a Resource Constrained Environment: One Study Area, Three Algorithms and Seven Images - What Can We Learn? | Points and Rating Systems in Professional Tennis | Asymptotic Tail Probability of the Discounted Aggregate Claims under Homogenous, Non-Homogenous and Mixed Poisson Risk Model | A New Double Sampling Control Chart for Monitoring an Abrupt Change in the Process Location |
| | | *Michaela Ritchie* | *Paul van Staden* | *Franck Adekambi* | *Collen Motsepa* |
| 15h00 - 15h40 | **SASA AGM**<br>Venue: Lecture Theatre 1 | | | | |

| WEDNESDAY: 27 NOVEMBER 2019 | | | | | |
|---|---|---|---|---|---|
| 15h40 - 16h00 | **Tea** | | | | |
| | Venue: Foyer, Building 123 | | | | |
| 16h00 - 17h20 | **Parallel Sessions (including posters)** | | | | |
| | Stream | Young Statisticians | Exp Design & Biostatistics | Stochastic Processes | Young Statisticians |
| | Chair | Saadiyah Mayet | Francesca Little | Daniel Mashishi | Alpheus Mahoya |
| | Venue | Lecture Theatre 1 | Lecture Theatre 2 | Lecture Theatre 3 | 35 00 40 |
| 16h00 - 17h20 | 16h00 - 16h15 | Models for Early Identification of Students at Risk of Failing | Experimental Design Criteria and Optimization for Chemical Process Modelling | Hidden Markov Models Based on Truncated Weibull Distributions | Ruin Probability in the Delayed Poisson Renewal Risk Model Perturbed by Diffusion Process |
| | | *Saadiyah Mayet* | *Willem van Deventer* | *Iain MacDonald* | *Essodina Takouda* |
| | 16h20 - 16h35 | Multi- Level Modelling of Associations Between Inflammation, Smoke Exposure and Pneumococcal Carriage in a Gambian Birth Cohort Study | Classifying Gene Expression Data with Mixture Models | Comparative Analysis of the 100-Year Return Level of the Average Monthly Rainfall for South Africa: Parent Distribution Versus Extreme Value Distributions | Ruin Formulas for Delay Renewal Risk Model with General Dependence |
| | | *Raymond Nhapi* | *Michelle de Klerk* | *Daniel Mashishi* | *Kokou Essiomole* |
| | 16h40 - 16h55 | Identifying Immunological Risk Factors for TB Progression using a Hidden Markov Model | Destructive COM-Poisson Cure Rate Model and Likelihood Inference with Lognormal Lifetime | Modelling the Sporadic Behaviour of Rainfall Time Series using ETS State Space and SARIMA Models in the Limpopo Province, South Africa | A Deep Learning Framework for Individual Clanwilliam Cedar Tree-Crown Detection using High Resolution Aerial Imagery |
| | | *Miguel Rodo* | *Jacob Majakwara* | *Selokela Molautsi* | *Blessings Hadebe* |
| | 17h00 - 17h15 | Statistical Accuracy of a Linear Object Extraction Algorithm for Greyscale Images | Models for Analysing Growth for a South African Cohort of Infants | Using Jump Models for Fire Detection in NDVI Data | Optimal Smoothing of Cycles using Sinusoidal waves |
| | | *Renate Thiede* | *Francesca Little* | *Etienne Pienaar* | *Alpheus Mahoya* |
| 16h45 - 18h30 | **Poster Session** | | | | |
| | Venue: Foyer, Building 123 | | | | |
| | **Viewing posters from 17h00 to 18h30** | | | | |
| | | | | | |
| 17h30 - 19h30 | **Welcoming Function** | | | | |
| | Venue: Foyer, Building 123 | | | | |

| THURSDAY: 28 NOVEMBER 2019 | | | | | |
|---|---|---|---|---|---|
| | | | | | |
| 07h45 - 10h00 | **Registration** | | | | |
| | Venue: Foyer, Building 123 | | | | |
| 09h00 - 13h00 | **Special meeting of HoDs and other stake holders to discuss UCDP future planning and feedback on DST scientometric report** | | | | |
| | Chair: Prof Freedom Gumedze | | | | |
| | Venue: 35 00 01 | | | | |
| 08h30 - 10h30 | **Parallel Sessions** | | | | |
| | Stream | Young Statisticians | Education and Industry | Methods and Theory | Biostats |
| | Chair | Carmen Stindt | Delia North | Daniel Maposa | Chris Muller |
| | Venue | Lecture Theatre 1 | Lecture Theatre 2 | Lecture Theatre 3 | 35 00 40 |
| 08h30 - 08h45 | | Artificial Neural Networks in Precipitation-Stream Flow Modeling | An Industry Perspective on the Relevance of Statistics Programs at Universities | An Introduction to Gaussian Belief Propagation with Exploratory Remarks | Selecting Variables for Predicting the Hazard of Alcohol Intake Initiation among Tertiary Students in Thohoyandou using Selected Penalized Likelihood and Gradient Boosting Approaches |
| | | *Romelon Chetty* | *Murray de Villiers* | *Francois Kamper* | *Alphonce Bere* |
| 08h50 - 09h05 | | Longitudinal Analysis of Brain Metabolite Levels for HIV Infected Children from Ages Five to Eleven Children using Multivariate Approaches | Innovative Linkages Between Universities and Organizations | Spatial Statistics of Extremes with a View Towards Application to Extreme Weather Events in South Africa and Other Neighbouring Countries in the SADC Region | A System of Longitudinal Regression Approach for Investigation of HBV Vaccination Rates |
| | | *Noelle van Biljon* | *Jennifer Priestley* | *Daniel Maposa* | *Tanita Cronje* |
| 09h10 - 09h25 | | Weighted Poisson Regression using Polynomial Weight Functions | Statistics Skills Development Through Collaborative Projects at UKZN | Change-Point Detection in Panel Data with Stationary Regressors | An Early Infant HIV Risk Score for Targeted HIV Testing at Birth |
| | | *Nicola Gawler* | *Delia North, Temesgen Zewotir* | *Charl Pretorius* | *Chris Muller* |
| 09h30 - 09h45 | | Modelling Malaria Incidence in the Limpopo Province, South Africa: Comparison of Classical and Bayesian Methods of Estimation | Industry Aligned Degrees | Bootstrap Based Test for Two Independent Time Series | Pregnancy Incidence and Risk Factors among Women Participating in Vaginal Microbicide Trials for HIV Prevention: Systematic Review and Meta-Analysis |
| | | *Makwekantle Sehlabana* | *Renette Blignaut* | *Modisane Seitshiro* | *Alfred Musekiwa* |

| **THURSDAY: 28 NOVEMBER 2019** | | | | | |
|---|---|---|---|---|---|
| 09h50 - 10h05 | | Structural Equation Models (SEM): Size Matters, for Now | Challenges, Opportunities and Success Factors for Statistics and Statisticians Within the Greater Data Science Field | New Tests for Exponentiality Based on the Interarrival Times of the Poisson Process | Non Parametric Techniques for Multilevel Discrete Survival Data |
| | | *Carmen Stindt* | *Roelof Coetzer* | *Jaco Visagie* | *Thambeleni Nevhungoni* |
| 10h10 - 10h25 | | Determination of Factors Associated with Time to Recovery from Pneumonia using Cox Proportional Hazard Model | Doctoral Supervision in Statistics in South Africa – Perspectives from the DIES/CREST Doctoral Supervision Training Course | Robust Estimation of Pareto-Type Tail Index Through an Exponential Regression Model | Gaussian Mixture of Expert Model for Censored Data |
| | | *Masimthembe Lala* | *Inger Fabris-Rotelli* | *Richard Minkah* | *Elham Mirfarah* |
| 10h30 - 11h00 | **Tea** | | | | |
| | Venue: Foyer, Building 123 | | | | |
| 11h00 - 11h55 | **Plenary Session** | | | | |
| | Venue: Lecture Theatre 1, Building 123 | | | | |
| | 11h00 - 11h45 | Guest and Title | Prof Deborah Nolan: How can data science improve statistics education? | | |
| | | Chair | Dr Frans Kanfer | | |
| | 11h45 - 11h55 | Questions | Audience | | |
| 12h00 - 13h00 | **Parallel Session** | | | | |
| | Stream | Young Statisticians | Education and Industry | Interesting Topics | Data Science and Big Data |
| | Chair | Carel van der Merwe | Delia North | Trudie Sandrock | Ariane Neethling |
| | Venue | Lecture Theatre 1 | Lecture Theatre 2 | Lecture Theatre 3 | 35 00 40 |
| 12h00 - 12h15 | | Modelling Diabetes in South Africa | Robust Prediction Interval Modelling of Hourly Electricity Demand Forecasts | Cluster Analysis Methods for Homogeneous Groupings in Efficiency Benchmarking | Tree-Based Ensemble Methods for Classification |
| | | *Nina Grundlingh* | *Caston Sigauke* | *Aviwe Gqwaka* | *Danie Uys* |
| 12h20 - 12h35 | | Classifying Yield Spread Movements Through Triplots: A South African Application | The Role of Statisticians in Harnessing the 4th Industrial Revolution | Genetic Algorithms (GA) for Feature Selection | Statistics Behind Big Data Analysis: Bridging the Gap Between Satellite Imagery and Business Intelligence |
| | | *Carel van der Merwe* | *Mark Nasila* | *Edward Jones* | *Ariane Neethling* |

| THURSDAY: 28 NOVEMBER 2019 | | | | | |
|---|---|---|---|---|---|
| 12h40 - 12h55 | | Skew Generalised Normal Innovations for the AR(1) Model | Fundamentals of Geostatistics | Classification of Musical Instruments in Audio Samples | An Efficient Kernel Quantile Estimator |
| | | *Ane Neethling* | *David Rose* | *Trudie Sandrock* | *Francois van Graan* |
| 13h00 - 14h10 | Lunch | | | | |
| | Venue: Chancellor's Room, Sports Centre | | | | |
| 14h20 - 16h00 | Parallel Sessions | | | | |
| | Stream | Young Statisticians | Statistics Education | Multivariate Data Analysis | General Statistics |
| | Chair | Stefan Janse van Rensburg | Jeanette Pauw | Sugnet Lubbe | Siphumlile Mangisa |
| | Venue | Lecture Theatre 1 | Lecture Theatre 2 | Lecture Theatre 3 | 35 00 40 |
| 14h20 - 14h35 | | SARIMA Modelling and Forecasting the Monthly Rainfall of the City of Cape Town, South Africa | CASIO Technology in the Classroom | A Novel Application of Survival and Competing Risk Models in Agriculture | Points of Impact Analysis in Functional Linear Regression Setting: A Case Study |
| | | *Rambuwani Kenneth* | *Lauren Izaaks* | *Sugnet Lubbe* | *Siphumlile Mangisa* |
| 14h40 - 14h55 | | Nested EM Algorithms for Estimation of Hierarchical Wind Speed Distributions | The Power of Markdown for Teaching, Research, and Consultation | Levels and Determinants of Knowledge of HIV/AIDS among Women in South Africa | Estimation of the Frequency-Size Power Law Slope Parameter without Knowledge of the Time-Varying Level of Completeness of the Dataset |
| | | *Michaela Laidlaw* | *Sean van der Merwe* | *Tshepo Matlwa* | *Ansie Smit* |
| 15h00 - 15h15 | | Machine Learning Techniques to Assess the Volatility of Equity Returns | Application of Discrete-Time Survival Analysis Models in Modelling Student Dropout: A Case of Engineering Students at Tshwane University of Technology, South Africa | Principal Component Analysis Data Reduction Procedure for Body Shape Classification for South African Men | On the Maximum Likelihood Parameter Estimation of the Bimodal Skew-Normal Distribution |
| | | *Nikita Gorlach* | *Princess Ramokolo* | *Busisiwe Tabo* | *Mehrdad Naderi* |
| 15h20 - 15h35 | | Imputation; How Good is it? | The Impact of Assessment on Student Learning | An R Package for Heteroskedasticity Diagnostics | Open Set Recognition with the Generalised Pareto Distribution |
| | | *Nyiko Khoza* | *Dries Naude* | *Thomas Farrar* | *Luca Steyn* |

| THURSDAY: 28 NOVEMBER 2019 | | | | | |
|---|---|---|---|---|---|
| 15h40 - 15h55 | | Skewed-t Score Driven Volatility Models Applied to FTSE/JSE ALSI Returns | Does Training in Statistics Make Me a Good Intuitive Statistician? Some Thoughts Based on "Thinking, Fast and Slow" by Daniel Kahneman | Speeding Up Projection Pursuit using Fast Kernel Computations | A Different Approach for Choosing a Threshold in POT |
| | | *Stefan Janse van Rensburg* | *Jeanette Pauw* | *David Hofmeyr* | *Andrehette Verster* |
| 16h00 - 16h30 | **Tea** | | | | |
| | Venue: Foyer, Building 123 | | | | |
| 16h30 - 17h30 | **Parrallel Sessions** | | | | |
| | Stream | Young Statisticians | Applied Statistics | Meeting | Meeting |
| | Chair | Priyanka Nagar | Kajingulu Malandala | Various | Various |
| | Venue | Lecture Theatre 1 | Lecture Theatre 2 | Lecture Theatre 3 | 35 00 40 |
| 16h30 - 16h45 | | Regressions in the Sandbox: Projection Pursuit Regression in Comparison to Other Regression Techniques in Digital Soil Mapping | Quantile Regression for Count Data using Delaporte Distribution | IBS Group South Africa Meeting (16h20 - 16h50) | Data Science Group Meeting 16h30 - 17h30 |
| | | *Stephan van der Westhuizen* | *Kajingulu Malandala* | | |
| 16h50 - 17h05 | | New Contributions to Möbius Transformation-Induced Distributions on the Disc | Parameter Estimation with Mixture Cure Models: A Simulation Comparisons Methods | Bayes Group Meeting (16h55 - 17h25) | |
| | | *Priyanka Nagar* | *Tolulope Adeniji* | | |
| 17h10 - 17h25 | | A Dual-Stress log-Normal Accelerated Life Testing Model | Some Useful Real Life Data Applications of Diao et al. (2013)'s Accurate Confidence Intervals | | |
| | | *Neill Smit* | *Yegnanew Shiferaw* | | |
| | | | | | |
| 18h30 for 19h00 | **Gala Dinner** | | | | |
| | Venue: The Willows | | | | |

| FRIDAY: 29 NOVEMBER 2019 | | | | | |
|---|---|---|---|---|---|
| 08h30 - 09h30 | **Registration** | | | | |
| | Venue: Foyer, Building 123 | | | | |
| 08h30 - 10h30 | **Parallel Sessions** | | | | |
| | Stream | Statistical Ecology | Applied Statistics | Time Series | General |
| | Chair | Morries Chauke | Temesgen Zewotir | Igor Litvine | Yegnanew Shiferaw |
| | Venue | Lecture Theatre 1 | Lecture Theatre 2 | Lecture Theatre 3 | 35 00 40 |
| 08h30 - 08h45 | | Stand Height Growth Model Conditioned to Changes in Rainfall for Eucalyptus Pulpwood in Mondi South Africa | Comparisons of Quality of Life in Haemodialysis and Peritoneal Dialysis Patients: A Case Study of Alice and King Williams's Town, South Africa | Time Series Analysis of Two Different Sources of Financial Statistics Data: Data from Administrative Sources and Survey Data | Understanding the Dynamic Dependence Between Oil, Mineral Commodities and USD-ZAR Exchange Rate: Evidence from South Africa |
| | | *Morries Chauke* | *Lekhoele L Moleleki* | *Sagaren Pillay* | *Yegnanew Shiferaw* |
| 08h50 - 09h05 | | Using Ecological Distance with Continuous-Time Spatial Capture-Recapture Models | Comparative Assessment of Machine Learning Models in the Ranking of Child Mortality Data | Factors of the Term Structure of Sovereign Yield Spreads and the Effect on the Uncovered Interest Rate Parity Model for Exchange Rate Forecasting | Modelling Dependence Structures of Extreme Wind Speed using Bivariate Distribution: A Bayesian Approach |
| | | *Greg Distiller* | *Samuel Oduse* | *Chun-Sung Huang* | *Tadele Diriba* |
| 09h10 - 09h25 | | An Investigation into Tree Growth Modeling.  A Bayesian Approach | Determinants of Fertility Intentions among Women of Reproductive Age in South Africa: Evidence from the 2016 Demographic and Health Survey | Models for Cycles in Financial Time Series | Modelling Time-Varying Temperature Extremes in Kwazulu-Natal using Generalized Extreme Value Distribution |
| | | *Lulama Kepe* | *Olusegun Ewemooje* | *Igor Litvine* | *Murendeni Nemukula* |
| 09h30 - 09h45 | | Fishing Vessel Anomaly Detection using AIS Within South African Waters | Factors Responsible for Safe Sex Practice among Female Adolescent: A Case Study of the University of Fort Hare Students | Regime Switching Models in Time Series | Hierarchical Forecasting of the Zimbabwe International Tourist Arrivals |
| | | *Nosihle Dlamini* | *Makhadimola Leshabane* | *Henri Moolman* | *Tendai Makoni* |

| | | | | | |
|---|---|---|---|---|---|
| **FRIDAY: 29 NOVEMBER 2019** | | | | | |
| 09h50 - 10h05 | | Modelling Malaria Time to Re-Infection with Time Varying Covariates Effects: A Case Study of Outpatients in DR Congo | Modelling Determinants of Birth Interval among Women of Reproductive Age in Hadiya Zone, Southern Ethiopia: A Facility Based Cross-Sectional Study | Modelling Extreme Co-Movement Between Oil Prices and Economic Growth | A Formulated Combined Model in Forecasting Long-Term Energy Consumption in South Africa |
| | | *Ruffin Mutambayi* | *Ashenafi Abebe* | *Katleho Makatjane* | *Livhuwani Nedzingahe* |
| 10h10 - 10h25 | | Forecasting Extreme Conditional Quantiles of Electricity Demand in South Africa | Spatial Analysis of Tuberculosis-HIV in Ethiopia. Bayesian Multilevel and Generalized Linear Mixed Models | Stochastic Modelling of Inflation and Interest Rates for Defined Benefit Pension Plan Projections in Ghana | |
| | | *Norman Maswanganyi* | *Leta Lencha Gemechu* | *Ezekiel Nortey* | |
| 10h30 - 10h55 | **Tea** | | | | |
| | Venue: Foyer, Building 123 | | | | |
| 11h00 - 11h55 | **Plenary Session** | | | | |
| | Venue: Lecture Theatre 1, Building 123 | | | | |
| | 11h00 - 11h45 | Guest and Title | Prof Jennifer Priestley: Ethics of Data | | |
| | | Chair | Dr Lizelle Fletcher: Chair of SASA Education Committee | | |
| | 11h45 - 11h55 | Questions | Audience | | |
| 12h00 - 12h30 | **Closing Ceremony: SASA 2019** | | | | |
| | Prof Maseka Lesaoana | | | | |
| | Venue: Lecture Theatre 1, Building 123 | | | | |

| **POSTERS** | | |
|---|---|---|
| **WEDNESDAY, 26 NOVEMBER 2019, 17h00 - 18h30** | | |
| Venue: Foyer, Building 123 | | |
| **Surname** | **Name** | **Title** |
| Bredenkamp | Deidre | Wind Speed and Power Modelling using Mixture of Life Distributions |
| Fazzini | Chiara | Using Circular Statistics to Analyse Aoristic Data |
| Fehr | Fabio | Text Content Classification on News Articles |
| Gilfillan | Michelle | Clustering: An Introduction |
| Ivey | Peter | Changes in Rainfall Seasonality in the Western Cape |
| le Roux | Enrike | Bayesian Estimation for the Ratio of Two Exponential Parameters |
| Lloyd | Tessa | Estimating Survival of an Enigmatic Frog from Capture-Mark-Recapture Data |
| Makgolane | Kgethego Sharina | Determinants of Under-Five Mortality in South Africa using Poisson Regression Model |
| Makulube | Mzamo | The Use of Repeat LIDAR Flights to Model Dominant Height |
| Marques | Ricardo | Gaussian Process Regression and Classification: Elliptical Slice Sampling |
| Mashau | Thanyani | Mixtures of Regression: Expectation Maximisation Type Algorithms |
| Mawonike | Romeo | Kalman Filtering of the Generalized Vasicek Term Structure Models with Infinite Maturity |
| McCready | Carlyle | Latent Variable models for Longitudinal Outcomes from a Parenting Intervention Study |
| Naidoo | Tristan | A Comparison of Modern Dimensionality Reduction Techniques Through the Classification of Extragalactic Objects |
| Osuji | Georgeleen | A Bayesian Mixture Modelling for Zero-Inflated Multivariate Data |
| Potgieter | Arminn | Spatial Modelling of the Association Between Crime and Weather |
| Pretorius | Wilben | Music Generation with Neural Networks |
| Ramkilawon | Gopika | Computational Features for Efficient Estimation of Some Zero-Inflated Models |
| Ravele | Thakhani | Medium Term Load Forecasting in South Africa using Generalized Additive Model with Tensor Product Interactions |
| Saben | Vaughn | Car Accident Feature Extraction from a Drone-Based Video Feed |

**POSTERS**

**WEDNESDAY, 26 NOVEMBER 2019, 17h00 - 18h30**

Venue: Foyer, Building 123

| Surname | Name | Title |
|---|---|---|
| Seidu | Issah | Simulation Studies in Stochastic Ergodicity |
| Seimela | Anna | An Investigation of Parent Distributions and Long-Term Trends of Average Maximum and Minimum Temperature in the Limpopo Province of South Africa |
| Tloubatla | Moyagabo Danny | Wavelet-based Time Series Analysis of South African Financial Data |
| van Heerden | Carl | A New Characterisation-Based Test for Symmetry |
| van Wyk | Delene | Spherical (skew-) Normal Distributions Under a Geodesic Distance Measure |
| von Schoultz | Kaamilah | Statistical Modelling of Degradation Rates of Photovoltaic Modules |
| Watchurst | Lee | Cluster Analysis for Group Selection in Launch Sale Predictions |
| Westraadt | Edward James | Assessment of Photovoltaic Software Using Modelled and Measured Energy Yields |
| Wilson | Mampane Phokwane | Basis Beyond Linear Regression using Wavelets, Splines and Locally Weighted Polynomial Regression |
| Zondo | Nombuso | The Level of Difficulty and Discrimination Power of the NSC Mathematics Examination Questions |
| Zondo | Mothibi | The Spatio-Temporal Analysis of Under-Five Mortality in Lesotho |

# ABSTRACTS: Oral Presentations

### *Modelling Determinants of Birth Interval among Women of Reproductive Age in Hadiya Zone, Southern Ethiopia: A Facility Based Cross-Sectional Study*

*Ashenafi Abebe*
*Wachemo University*
*Abebe, ND (PhD candidate in Epidemiology), Ermias, A (Resident of General Surgery)*
*ashujajura@gmail.com*

Background: The study of the dynamics of timing and spacing of births is important for several reasons, including an understanding of completed family size as well as maternal and child mortality (1). Modeling fertility data is one of the greatest interests in population economic studies. Several indicators are used to measure fertility patterns, such as the first birth interval after marriage (2). In recent years, the study of birth intervals has been a main determinant of the levels of fertility in the populations, as it is associated with rates of fertility and population growth. Objective: The purpose of this study was to apply survival analysis for modeling of birth interval through exploring its determinants. Materials and Methods: In this facility based cross sectional study, the fertility history of 654 women was collected and extracted in five districts of Hadiya Zone from 2011-2017. We used the survival analysis such as Cox regression and alternative parametric models to evaluate the socio-economic and prognostic factors of birth interval. Results: Among the explanatory variables of interest, age at marriage, level of mother's education, menstrual status had highly significant effects on the duration of birth interval after marriage (p<0.01). Conclusion: It is concluded that the suitable parametric models would be a useful tool for fitting the birth interval data, the fact that has been less paid attention to in various researches so far.

### *Asymptotic Tail Probability of the Discounted Aggregate Claims under Homogenous, Non-Homogenous and Mixed Poisson Risk Model*

*Franck Adekambi*
*University of Johannesburg*
*Kokou, E*

In this paper, we derive a closed form-expression of the tail probability of the aggregate discounted claims under homogenous, non-homogenous and mixed Poison risk model with constant force of interest using a general dependence structure between the inter-arrival claims times and the claim amounts. This dependence is relevant since it is well known that under catastrophic or extreme events the inter-claim times and the claim severities are dependence.

### *Parameter Estimation with Mixture Cure Models: A Simulation Comparisons Methods*

*Tolulope Adeniji*
*University of Fort Hare*
*Azeez, A (Department of Statistics, University of Fort Hare)*
*tolbum17@gmail.com*

Parameter Estimation with Mixture Cure Models: A Simulation Comparisons Methods Adeniji Tolulope*, Azeez Adeboye* Department of Statistics, University of Fort Hare, 5700, Alice. Department of Statistics, University of Fort Hare, 5700, Alice. Cure fraction models are popular concept for analyzing biostatistics and medical data. Many patients with cancer can be long-term survivors of their disease, and cure models can be a useful tool to analyze and describe cancer survival data. In Survival analysis, Cox proportional hazard model typically assumed that every individual will eventually experience the event of interest with adequate follow-up time. But there are some occasions in which fractional part of the population of interest may not experience the event of interest. However, a cured fraction can be incorporated in the statistical model for analysis. In this study, we comprehensively evaluated mixture cure model compared with standard Cox model for best performance. Cancer is generally considered as a disease with low cure rate, and we are using cure models compared with standard Cox proportional hazards model, to evaluate whether there is an evidence that treatment induce a proportion of patients to be long-term survivors. We comprehensively simulated data for the analysis. The results from the simulated models indicated that cure models fit better and can prolong the survival rate of the patients.

### Survival Analysis of First to Second Childbirth Interval among Women in South Africa: Rural-Urban Differential

Rotimi Felix Afolabi
North West University and University of Ibadan
Palamuleni, ME (Population Studies & Demography Programme and Population and Health Focus Area, North-West University, South Africa)
rotimifelix@yahoo.com

South Africa fertility rate remains the lowest in sub-Saharan Africa but unevenly distributed by residence type. Even though fertility transition is characterised by inter-birth intervals lengthening, information on interval between first and second birth (SbI) that could impact on maternal and child health is rare. We examined the correlates of SbI by residence type among South Africa women. We analysed 2016 South Africa Demographic and Health Survey data on 6124 women aged 15-49yrs who had reported at least one childbirth as of the survey date. Survival analysis methods were employed at 5% significance level. The SbI was significantly longer among urban (76mths) relative to rural (66mths) women. After controlling for other variables, urban women who used contraceptive were less likely to delay second birth in the first 4yrs (HR=10.67; CI: 9.11-12.49) but were more likely to postpone it as from =16.5yrs (HR=0.08; CI: 0.05-0.14). Rural women who used contraceptive had a 23% increased hazard to shorten SbI. Women's ethnicity, wealth-quintile, marital status at first birth, desire more children and employment were other significant predictors of SbI irrespective of residence. Having higher education and age at first birth respectively were protective of second birth in rural and urban. Women residing in rural have higher tendency to shorten SbI than those in urban. These correlates as revealed in this study could form the basis for health education interventions in South Africa.

### Bayesian Semi-Parametric Linear Mixed Model using Smoothing Spline: Application to Longitudinally Measured Fasting Blood Sugar Level Data

Tafere Tilahun Aniley
Department of Statistics, University of South Africa
Debusho, LK, Diriba, TA (Department of Statistics, University of South Africa)
67146929@mylife.unisa.ac.za

In the linear mixed effect model, the continuously measured longitudinal data is modeled as a function of fixed linear predictors and random components. However, the individual and average profile plots sometimes might exhibit a polynomial function which violates the linearity assumptions. Linear mixed-effects models also accounts random effects for capturing inter-and intra-individual variability. In this study, we propose a Bayesian semi-parametric linear mixed model, in particular using Spline smoothing method to model and analyze longitudinally measured fasting blood sugar level data. The Bayesian semi-parametric linear mixed model shows a better result than the linear mixed effects models. This is due to the fact that the longitudinal change of the fasting blood sugar level data of the diabetic patients is well captured and estimated non-parametrically using smoothing spline method. The study also revealed that the rate of changes in fasting blood sugar level of diabetic patients changes with follow-up time and weights. The proposed model shows better fit to the data. The individual profile curve is used to follow patient's specific FBS level trends over time. Hence the method can be used to monitor the individual-level trends of patients over time.

### Joint Model of Longitudinal and Survival Data for Covariates Measures with Correlated Error

Adeboye Azeez
Department of Statistics, University of Fort Hare
Mutambayi, RM, Ndege, J, Qin, Y (Department of Statistics, University of Fort Hare)
azizadeboye@gmail.com

The study investigates the association between longitudinal covariates and time-to-event process of examining the within-subject measurement error that could influence estimation when the assumption of normality and mutual independence are violated by using the conditional score approach. In view of assumption violation, we proposed an estimating equation approach based on the generalized conditional score to relax parametric distributional assumptions for repeated measures random effects. We jointly model the time-dependent biomarkers and event times using the Cox model with intermittent time-dependent covariates measure. Estimates of the parameters are obtained by a restricted maximum likelihood estimate (REML). A modified Cholesky decomposition method is used to capture the within-subject covariance for a positive definite and symmetric matrix, assumed that observed data from different subjects are independent. We conducted simulation studies to compare the proposed method with other existing methods and illustrated the proposed method by a real data set for TB patients with impaired renal function.

### *Usefulness of Applying Power Allocation in Survey Sampling in South Africa*

*Esta Bekker*
*Department of Mathematical Statistics and Actuarial Science, University of the Free State*
*Neethling, A (Department of Mathematical Statistics and Actuarial Science, University of the Free State)*
*esta30312@gmail.com*

In many surveys, reliable estimates are required both at the national level (entire population) and for subnational areas (strata in the population). When the subnational areas vary considerably in population size or importance, problems can arise in terms of obtaining useful representation from all areas when standard allocation techniques are applied. In many South African surveys, variables such as province and race group are used as stratification variables in country wide surveys. Taking into account e.g., the differences in size between the resulting strata using these variables, it can easily happen that some strata are not adequately represented in the sample when using a standard allocation technique. Power allocation will be discussed as a simple allocation method, which allows for a balancing of the stratum population size, the stratum variability and a desirable 'spread' of sample allocation across strata. Medium and small sized strata are allocated adequate observations (elements) to enable accurate estimation on the stratum level, while large strata are still adequately covered to also ensure reliable estimation for the entire population being studied. The use of this allocation method as well as the effect on survey weights and the calculation of estimates will be discussed during this presentation.

### *Selecting Variables for Predicting the Hazard of Alcohol Intake Initiation among Tertiary Students in Thohoyandou using Selected Penalized Likelihood and Gradient Boosting Approaches*

*Alphonce Bere*
*University of Venda*
*Sithuba, G, Mashabela, R, Kyei, K, Sigauke, C (Department of Statistics, University of Venda)*
*alphonce.bere@univen.ac.za*

In the study, we use variable selection methods to build discrete survival models for age at first alcohol intake among students at two tertiary institutions in Thohoyandou, South Africa. Variable selection in discrete survival modelling is complicated by the need employ two penalties; one for variable selection and the other for stabilization of the baseline hazard parameter estimates. The study compares the results obtained through a recently developed penalized likelihood approach and a gradient boosting approach. The results show that the methods do not select the same set of variables.

### *Industry Aligned Degrees*

*Renette Blignaut*
*University of the Western Cape*

The University of the Western Cape aligned undergraduate and honours modules to better prepare students for the world of work as well as for a masters in the field of Data Science. Our unique collaborative master's programme with the Centre for Business, Mathematics and Informatics (BMI) at Northwest University will be discussed. The BMI programme provides students with business analytics skills and prepares students to complete projects in a business environment. The development of an interdepartmental postgraduate diploma in Data Analytics and Business Intelligence has been a very interesting experience. The process of industry involvement and buy-in will be shared.

### *Stand Height Growth Model Conditioned to Changes in Rainfall for Eucalyptus Pulpwood in Mondi South Africa*

*Morries Chauke*
*University of KwaZulu-Natal*
*Chauke, M, Mwambi, H (School of Mathematics, Statistics and Computer Science, University of KwaZulu-Natal, PMB)*
*morries.chauke@mondigroup.com*

Dominant height is commonly used as a measure of site productivity in forestry and it is usually related to age only. This study investigated the effect of incorporating rainfall as a covariate of the dominant height model in Eucalyptus grandis x urophylla. The dominant height values used in the study were obtained from plots aged between 0.6 to 9.4 years located in Coastal Zululand of KwaZulu-Natal. The rainfall data were obtained from four rainfall stations. The Chapman-Richards model has been widely used in forestry to represent dominant height as a function of age. However the model does not account for plot to plot variability. A random effect and mean monthly precipitation modifier (as a covariate) was included in the asymptote of the model to account for plot to plot heterogeneity and rainfall effect. The asymptote was selected since it is a parameter that governs the maximum value of the dominant height that a plot can attain. The resultant nonlinear mixed effects model was more precise, with a gain in precision in terms of standard error of estimate of about 87%. Possible applications of this model include adjustment of yield estimates on measured stands as well as stands with rainfall data but lacking inventory data. Extensions will include the inclusion of heterogeneity in other growth parameters via random effects, which are themselves bound to be correlated.

### *Artificial Neural Networks in Precipitation-Stream Flow Modelling*

*Romelon Chetty*
*University of Cape Town*
*Ernie, B, De Vries, M (Department of Statistics, University of Cape Town)*
*romelon.chetty@gmail.com*

Precipitation-stream flow modelling is one of the key challenges in the field of hydrology. Many approaches exist ranging from physically-based to fully data-driven models. In this project, we proposed a novel data-driven approach, investigating the utility of Artificial Neural Networks (ANNs) for short term forecasting of stream flow in the Jonkershoek catchment area. Research into the use of general Feed Forward Neural Networks (FFNNs), Recurrent Neural Networks (RNNs), and in particular, the Long Short-Term Memory (LSTM) model in order to forecast stream flow from precipitation, was conducted. The aim of this project was to build a model linking rainfall to stream flow which could be used to predict how stream flow is likely to be affected by projected changes in precipitation seasonality. The work considers historical data of precipitation measured in millimetres and stream flow measured in cubic metres from the Jonkershoek catchment area from the year 1940 until 1991. Two experiments were conducted differing in the resolution of the data used (daily and weekly time intervals). ANNs were seen to effectively predict stream flow from rainfall data alone. In both experiments the LSTM model outperformed the other models in predictive capabilities, suggesting the potential of the LSTM for hydrological modelling applications.

### *A Gibbs Sampler for Multi-Species Site Occupancy Models*

*Allan Clark*
*University of Cape Town*
*Altwegg, R (Department of Statistical Sciences, University of Cape Town)*
*allan.clark@uct.ac.za*

Occupancy models (MacKenzie et al, 2002) are an important statistical technique that has been developed to infer the probability that a species under investigation occupies a region. Bayesian analysis of these models can be undertaken using statistical packages such WinBUGS, OpenBUGS, JAGS and more recently Stan, however, since these packages were not developed specifically to fit occupancy models, one often experiences long run times when undertaking an analysis. We develop a simple Gibbs sampler to obtain posterior samples from the posterior distribution of the parameters of this type of model when logit link functions are used to model the regression effects of the detection and occupancy processes. As an application of the methods developed, we analyze data collected from a camera-trapping study in the Karoo, South Africa. The study area consisted of two regions in South Africa namely, the Anysberg Nature Reserve and a group of 22 neighbouring sheep farms in the Koup. Reversible jump MCMC was used in order to undertake model selection. We found that the species richness between the two study sites were practically the same and that elevation was an important covariate that explained the occupancy status for most of the mammals in the region.

### *Challenges, Opportunities and Success Factors for Statistics and Statisticians Within the Greater Data Science Field*

*Roelof Coetzer*
*University of the Free State*
*roelof.coetzer@gmail.com*

Data Science and the demand thereof have increased significantly over the last number of years. However, Data Science as a field entails many different disciplines of which statistics is only one aspect. In industry, statisticians need to cope with the interest, the demand and many times the misunderstandings of what Data Science is and can or should be delivering. In this paper I will discuss the many challenges that statisticians in industry face and will face, but also the many opportunities that exist. I will discuss a case study which illustrates the journey of scientific learning and discovery within an industrial setting, which will provide some insights into the factors required for being successful in industry, and a few other examples of applying statistical methods in an innovative way to deliver value for the business. The examples should provide insight into a training curriculum at University.

### *A System of Longitudinal Regression Approach for Investigation of HBV Vaccination Rates*

*Tanita Cronje*
*Department of Statistics, University of Pretoria*
*Ferreira, JT, Arashi, M, Mirfarah, E (Department of Statistics, University of Pretoria)*
*tanitacronje@gmail.com*

Delay of childhood vaccines has increased in recent years and is believed to cluster in some communities. Such clusters could pose public health risks and barriers to achieving immunization quality benchmarks. Providing hepatitis B vaccine to all neonates within 24 hours of birth is the key preventative measure to control perinatal hepatitis B virus infection. In this talk, a system of longitudinal regression model is proposed to potentially identify geographical clusters of vaccination behaviour with regards to percentage of infants vaccinated against hepatitis B across different facilities in counties of New York State between 2012 and 2018. Possible socio-economic effects of results will be briefly discussed.

### *Classifying Gene Expression Data with Mixture Models*

*Michelle De Klerk*
*University of Pretoria*
*Kanfer, F, Millard, S (Department of Statistics, University of Pretoria)*
*michelle.deklerk@up.ac.za*

The study of gene expression is an active area of research as the understanding of genomics can be used to identify, diagnose and develop treatments for genetic diseases. The sequencing of cellular RNA is a modern alternative to microarray with many advantages such as measuring gene expression data with no prior knowledge of the genome sequence. Mixtures of distributions is considered to determine differentially expressed gene profiles. RNA-seq read counts are generated by mapping short reads to a reference gene for which discrete component distributions can be identified using a mixture model. Differentially expressed genes can be identified by comparing reads mapped to each gene for the treatment and control group in a finite mixtures of Poisson or Negative Binomial distributions model.

### *An Industry Perspective on the Relevance of Statistics Programs at Universities*

*Murray De Villiers*
*SAS Institute*
*zafmdf@zaf.sas.com*

The advent of the fourth Industrial Revolution has introduced previously unheralded challenges to companies and governments. These include deepened business complexity, rapid changes in business competitiveness factors and the enhanced availability of structured and unstructured data and associated analytical capabilities to leverage these for business decision-making. These developments emphasise the importance of problem-solving, data savvy, solution structuring Statisticians, with the commensurate adaptation of university programs. This presentation explores the dynamics of the analytics skills-related challenges faced by industry and academia, providing some strategic and tactical insights towards successful and jointly profitable collaboration between industry and academia.

### *Modelling Dependence Structures of Extreme Wind Speed using Bivariate Distribution: A Bayesian Approach*

*Tadele Akeba Diriba*
*University of South Africa*
*Legesse Kassa Debusho (Department of Statistics, UNISA), Joel Botai (South African Weather Service)*
*diribtd@unisa.ac.za*

When investigating extremes of weather variables, it is often not just a single station which determines the damage caused, but in turn extremes may be caused from the combined behaviour of several weather stations. In order to investigate the joint dependence of extreme wind speed, a bivariate generalised extreme value distribution (BGEVD) have been considered from the frequentist and Bayesian approaches to analyse the extremes of component wise monthly maximum wind speed data. In the frequentist approach, the parameters of EVDs were estimated using maximum likelihood, whereas in the Bayesian approach the MCMC technique was used. The results show that the BGEVD fitted to component wise maxima of extreme weather variables provide apparent benefits over the univariate method, which allows information to be pooled across stations and resulted improved precision of the estimate for the parameters as well as return levels of the distributions. The paper also discusses a method to construct informative priors empirically using historical data of the underlying process from weather characteristics of surrounding four pairs of weather stations at various distances. The results from the Bayesian analysis show that posterior inference might be effected by the choice of priors used to formulate the informative priors. From the results, it can be inferred that the Bayesian approach provide satisfactory estimation strategy in terms of precision compared to the frequentist approach.

### *Using Ecological Distance with Continuous-Time Spatial Capture-Recapture Models*

*Greg Distiller*
*University of Cape Town*
*Borchers, D (Department of Mathematics and Statistics, University of St Andrews)*
*greg.distiller@uct.ac.za*

The distance between a particular detector and an individual's activity centre plays a part in the detection function of Spatial Capture Recapture (SCR) models whereby detectability declines with increasing distance. SCR models have been extended to take account of salient landscape features by using a more realistic distance metric ( "ecological distance" ) compared to straight line Euclidean distance. While the SCR framework was developed with the primary aim of estimating density, incorporating ecological distance means that other processes like space usage and landscape connectivity can be explored. Continuous-time (CT) SCR models include a temporal dimension and hence extend the types of inferences that can be made by estimating usage patterns at different times of the day. These models can be used to explore nuanced spatio-temporal processes that can lead to interesting analyses on predator-prey dynamics, sympatry and interspecific competition.

### *Fishing Vessel Anomaly Detection using AIS Within South African Waters*

*Nosihle Dlamini*
*Dlamini, NP, Krishnannair, S (Department of Mathematical Sciences, University of Zululand),*
*Meyer, R (ICT for Earth Observation, Next Generation Enterprises and Institutions, CSIR)*
*nosihledlamini62@gmail.com*

Since South Africa occupies an Exclusive Economic Zone greater than its own land, its maritime safety, security and economy becomes a major priority. The significance of Illegal, Unreported and Unregulated fishing and other activities in the developing countries requires more attention. Yet anomaly detection has the capability of translating the available raw data into actionable information that is critical for decision-making. The approach to anomaly identification and analysis of a vessel behavior is based on generating the model of normalcy from the patterns formed by the vessels and consider everything that deviate from the model as anomalous. This paper make uses the Automatic Identity System data within the South African waters to detect fishing vessel anomalies. The fishing vessel data is classified according to their speed over ground into two categories, fishing vessels at the fishing zone and the fishing vessels at travel. The applied method make use of the Gaussian Processes to fit the trained model and test the sampled data points then flag its anomalies based on the 95% confident interval.

### *Factorisation Machines for Recommender Systems*

*Bronwyn Dumbleton*
*Stellenbosch University*
*Bierman, S (Department of Statistics and Actuarial Science, Stellenbosch University)*
*dumbleton@sun.ac.za*

Recommender systems are information filtering systems used to propose relevant items to users. Their successful application, especially in online retail, has proven their worth in terms of increased customer satisfaction and sales revenue. Hence it may be argued that recommender systems currently present some of the most successful and widely used machine learning algorithms in business. Latent factor models for collaborative filtering (such as matrix factorisation) have been used as part of many effective recommender systems. However, a limitation of standard latent factor models is that their input is typically restricted to a set of item ratings. In contrast, general purpose supervised learning algorithms allow more flexible inputs, but are typically not able to handle the degree of data sparsity prevalent in recommendation problems. Factorisation machines were intended to bridge this gap (Rendle, 2010). Being supervised learning methods, they are able to incorporate more flexible inputs, but they are also well suited to deal with the effects of data sparsity. We provide an overview of the use of latent factor models, and of factorisation machines in the recommender setting. Several extensions to factorisation machines are also discussed. We conclude with the results from an empirical comparison of the extent to which standard latent factor models and factorisation machines are able to reduce the adverse effects of data sparsity.

### *Deep Learning for Photovoltaic Defect Classification Through Thermal Infrared Imaging*

*Christopher Dunderdale*
*Department of Statistics, Nelson Mandela University*
*Brettenny, W. Clohessy, C (Department of Statistics, Nelson Mandela University),*
*van Dyk, EE (Department of Physics, Nelson Mandela University)*
*s214046117@mandela.ac.za*

As the global energy demand continues to soar, solar energy has become an attractive and environmentally conscious method to meet this demand. This study examines the use of machine learning techniques for defect classification in photovoltaic systems using thermal infrared images. A deep learning approach is investigated for the purpose of detecting and classifying defective photovoltaic modules. The VGG-16 and MobileNet deep learning models are shown to provide good performance for the classification of defects. Furthermore, these models are also shown to discriminate between defective and non-defective photovoltaic modules with a high accuracy. The successful implementation of this approach has significant potential for cost reduction in defect classification over currently available methods.

### *Ruin Formulas for Delay Renewal Risk Model with General Dependence*

*Kokou Essiomle*
*School of Economics and Econometrics, University of Johannesburg*
*Franck Adekambi (School of Economics and Econometrics, University of Johannesburg)*
*essiomle16@gmail.com*

In this paper, we derive explicit expressions and an upper bound of the ruin probability for the compound discounted delay renewal aggregate risk model when taking into account dependence. We therefore specify different sources of dependence: among the claims, the subsequent inter-claims, dependence between the subsequent inter-claims and the claims, between the first occurrence time and the forthcoming occurrence times, dependence between the first claim and the first occurrence time, between the first occurrence times and the forthcoming claims, in order to account for the heterogeneity of the risk process and the environment of the market (climate conditions, ages, gender…). Using a specific mixture of exponential distributions of the claims and the occurrence times we derive a closed-form expression and the upper bound of the ruin probability for an ordinary and a delay renewal risk model. Numerical examples are given using Clayton copula to analyze the impact of the dependency on the ruin probability as well as the impact of the delay on the ruin probability.

### *Determinants of Fertility Intentions among Women of Reproductive Age in South Africa: Evidence from the 2016 Demographic and Health Survey*

*Olusegun S. Ewemooje*
*Department of Statistics, Federal University of Technology, Akure, Nigeria. and Population and Health Research Entity, North-West University, South Africa.*
*Biney, E. (Population and Health Research Entity, North-West University, South Africa),*
*Amoateng, AY (Population and Health Research Entity, North-West University, South Africa)*
*olusegunewemooje@gmail.com*

Against the background of the fertility decline in South Africa in recent years, the present study aimed at understanding the factors that shape South African women's childbearing aspirations. Specifically, the study sought to examine the effect of selected sociodemographic factors on the fertility aspirations of two groups of reproductive-aged women, 15-29 years (younger women) and 30-49 years (older women), in South Africa using multinomial logistics regression analysis. The results showed that being black African, increased education, urban residence, relationship status and increased parity significantly determine women's intentions for more children, regardless of the age cohort. Employment, increased wealth and increased household size significantly determine fertility intention for younger women, while contraceptive use significantly determines fertility intention of older women. The study also found that contraceptive use and increased parity significantly reduces fertility indecision for all women. The implications for policy interventions are discussed.

### *Doctoral Supervision in Statistics in South Africa – Perspectives from the DIES/CREST Doctoral Supervision Training Course*

*Inger Fabris-rotelli*
*University of Pretoria*
*inger.fabris-rotelli@up.ac.za*

The issue of ageing Statistics academia causes a lack of guidance to young supervisors. The DIES/CREST Online Training Course for Supervisors of Doctoral Candidates at African Universities was presented mid-2019, which I partook in. This course provided excellent context and the appropriate setting to grow further as a supervisor. I propose and provide a portfolio arrangement to use as guidance for a supervisor(s) and the doctoral candidate using the concepts and outcomes of this DIES/CREST course, while adding the necessary adaptions to Mathematical Sciences, specifically Statistics. The young supervisor can thus learn through the process and the candidate has structure and understanding of the journey. This portfolio will hopefully assist the process of developing supervisors as well as graduating doctoral candidates quicker and with more direction in order to develop Africa's Statistics supervisors. The portfolio is structured in two parts. The first part as guidelines to a young supervisor and the second, a conversation between supervisor(s) and the doctoral candidate for the full journey of the doctorate. This latter conversation consists of two steps in each time frame, the first requires the candidate to reflect on their understanding of that element of the journey, and second provides concrete advise from the supervisor after the reflection.

### *An R Package for Heteroskedasticity Diagnostics*

*Thomas Farrar*
*University of the Western Cape; Cape Peninsula University of Technology*
*Blignaut, R, Luus, R, Steel, S (Department of Statistics and Population Studies, University of the Western Cape)*
*farrart@cput.ac.za*

Since the 1960s there has been a steady stream of new proposals for the testing of heteroskedasticity in the linear regression model. Google Scholar citation counts suggest that newer methods have gained little traction among practitioners, who tend to use classical heteroskedasticity tests such as those of Breusch and Pagan (1979) and White (1980). A possible reason for the lack of uptake of the newer methods is that the classical tests are implemented as built-in functions in standard statistical software (e.g., SAS, Eviews, Stata, R [lmtest package]), whereas newer tests are not. Consequently, an R package has been developed specifically for the purpose of making accessible to practitioners the full suite of heteroskedasticity testing methods that are proposed in the literature. An overview of the package functions and documentation is presented, together with practical insights gained from the experience of developing an R package in RStudio.

### *Modelling Time-Varying Temperature Extremes in Kwazulu-Natal using Generalized Extreme Value Distribution*

*Nceba Gagaza*
*University of KwaZulu-Natal*
*Nemukula, MM, Chifurira, R, Roberts, DJ (School of Mathematics, Statistics and Computer Science, University of KwaZulu-Natal)*

The use of extreme value theory (EVT) is usually aimed at quantifying and modelling the asymptotic behaviour of extreme values. In this paper we discuss the use of the block-maxima (BM) approach of EVT in modelling temperature extremes in rural KwaZulu-Natal (KZN) province of South Africa. The purpose of this paper is to model the asymptotic behaviour of extreme high temperature in rural KZN. This is accomplished using EVT by applying the generalized extreme value distribution (GEVD) with a BM approach. The data that is used in this paper is a time series of maximum daily temperatures (MDT) that is obtained from the South African Weather Service over the period 01 January 2007 to 31 December 2018. The MDT data is then divided into the winter and non-winter seasonal versions, where non-winter is the period from 1 September to 30 April of each year and the rest of the data is for the winter season. Only the data for non-winter season is used in this paper for the purpose of modelling the occurrence of maximum temperatures. The GEVD was then modified to take into account the temporal non-stationary trend in the annual maxima. The Weibull class of distributions is established as the appropriate model for modelling MDT in rural KZN. It is also revealed that extreme temperatures in rural KZN are associated to increasing trend over the last decade and also that the inclusion of the non-stationary trend with respect to time significantly improves the modelling of MDT in rural KZN. The return levels for specific return periods are calculated and discussed. The highest temperature of 34.07 degrees Celsius will be exceeded on average, once in 10 years in Emerald weather station, whereas the highest temperature of 37.58 degrees Celsius is expected to be exceeded on average once in 100 years in Ixopo weather station.

### *Weighted Poisson Regression using Polynomial Weight Functions*

*Nicola Gawler*
*Department of Statistics, University of Pretoria*
*Visagie, IJH, Santana, L (Department of Statistics, North-West University)*
*visagiejaco3@gmail.com*

When performing regression analysis in the case where the response is a count variable, the response is often assumed to be realised from a Poisson distribution. The weighted Poisson distributions make up a class of distributions containing the Poisson as a special case. This class of distributions is obtained when the Poisson probability mass function is adjusted by multiplication with a suitable weight function. This talk proposes the use of the weighted Poisson distribution with a polynomial weight function as the model for the response in count regression. A practical application is included, wherein it is shown that this more flexible model outperforms Poisson regression.

### *Spatial Analysis of Tuberculosis-HIV in Ethiopia. Bayesian Multilevel and Generalized Linear Mixed Models*

*Leta Lencha Gemechu*
*University of South Africa*
*Debusho, LK (Department of Statistics, University of South Africa)*
*letit_98@yahoo.com*

Tuberculosis(TB) has claimed many lives throughout the history of mankind and it continues to be a global threat in the coming decades, especially in developing countries like Ethiopia. The incidence and mortality due to TB cases is not equally distributed across the globe; they vary by geographic region, subpopulation, and spread by close and prolonged contact with an infected individual. Over the past few decades, Ethiopia has been implementing different prevention and controlling strategies with objective of achieving the TB free society. However, none of the studies done in Ethiopia considered the characteristics of TB distribution at national level and take in to account the possibility of correlation due to aggregation of data in different scales or hierarchy. Further, no study considered the joint modelling of TB and HIV, which could likely have strong correlation and may share common risk factors. The general objective of this study is to assess spatio distribution of TB, detect clustering and identifying major risk factors of TB/HIV distribution in Ethiopia For detecting the Spatial autocorrelation both global and local spatial test statistics (univariate and bivariate) Moran’s I and Geary’s were used. Bayesian multilevel modelling and general linear mixed effect model which incorporate spatial component as combination of both random and fixed effect were fitted for the data. Test results of spatial autocorrelations showed strong positive correlation of TB/HIV

### *Machine Learning Techniques to Assess the Volatility of Equity Returns*

*Nikita Gorlach*
*Nelson Mandela University*
*Sharp, G, Brettenny, W (Department of Statistics, Nelson Mandela University)*
*ngorlach@gmail.com*

The study undertakes an investigation into the assessment of asymmetric distributional modelling. The necessity for this is based on two supporting theories, which explain the behaviour and volatility of equity returns. Firstly, it has been shown that equity returns are asymmetric. Secondly, there is evidence to suggest that equity returns experience two states of volatility over time. This study aims to investigate these two theories simultaneously. A survey of the literature has found that research into the joint theories is absent and this research attempts to support the joint theories empirically. Several machine learning algorithms are used to classify simulated equity return data into two volatility groups. Some of the methods that are explored include support vector machines, decision trees, artificial neural networks, k-nearest neighbours and logistic regression. Furthermore, real-world data in the form of global equity indices are assessed as a case study. There is evidence to suggest that the volatility of equity indices differs between emerging and developed markets. This study compares the volatility of various indices across different markets. The results of this study may prove useful for investment decisions concerning South Africa, an emerging market. Research of this nature is particularly important at a time when the South African government is devoted to encouraging investment in the country.

### *Cluster Analysis Methods for Homogeneous Groupings in Efficiency Benchmarking*

*Aviwe Gqwaka*
*Nelson Mandela University*
*Brettenny, W, Tshangela, Z (Department of Statistics, Nelson Mandela University)*
*aviwe.gqwaka@mandela.ac.za*

Efficiency benchmarking is conducted for the assessment and comparison of performances of decision making units (DMUs), such as firms or organisations, relative to each other or to some ideal frontier. These DMUs may vary in certain dimensions such as structure, size, level of investment etc., where comparisons of the implementation of their processes may yield adverse findings if DMUs are assumed to operate under the same technology. Perhaps clustering DMUs into homogeneous groupings and assessing levels of efficiencies according to these clusters may provide equitable insights into performance. This study thus assesses a centre-based as well as hierarchical clustering method to group institutions in the South African education sector. The efficiency levels of these homogeneous groups are then determined using data envelopment analysis (DEA).

### *Modelling Diabetes in South Africa*

*Nina Grundlingh*
*University of KwaZulu-Natal*
*Zewotir, T, Roberts, D (Department of Statistics, University of KwaZulu-Natal),*
*Manda, S (Biostatistics Research Unit, South African Medical Research Council)*
*ninagrundlingh30@gmail.com*

Diabetes is a metabolic disorder associated with high blood sugar or glucose levels due to the absence or insufficient production of insulin. It was estimated that, of the 1.8 million people between 20- and 79 years old with diabetes in South Africa (SA) in 2017, 84.8% were undiagnosed. This study utilised data from the South African Demographic and Health Survey (SADHS) done in 2016 where sampled individuals aged 15 years and older had their glycated haemoglobin level (HbA1c) tested. Individuals were classified as non-diabetic, prediabetic or diabetic according to the American Diabetics Association guidelines. The potential risk factors considered comprised of a range of demographic, socio-economic and anthropometric variables. A multinomial survey logistic regression model was fitted however, the residuals indicated spatial autocorrelation. Therefore, a multinomial generalized additive mixed model (GAMM) with longitude and latitude as fixed spline effects was fitted in order to adjust for surface correlation in the residuals. The observed prevalence of diabetes and pre-diabetes from the sampled and tested 6442 individuals was 21.9% and 66.6%, respectively. From the GAMM, females were more likely than males to be non-diabetic rather than pre-diabetic. Black/Africans were more likely to be diabetic rather than pre-diabetic. Diabetes is of major concern in SA. It does not receive the attention it deserves where policy makers are generally unaware of its current prevalence.

### *Multiple Two-Sample Comparisons Based on a Gradual Change Model*

*Zdenek Hlavka*
*Charles University, Department of Statistics*
*Huskova, M (Department of Statistics, Charles University)*
*hlavka@karlin.mff.cuni.cz*

We propose two-sample gradual-change analysis and show that it can be more powerful than usual two-sample t-tests both with and without multiple testing corrections. The real life motivation comes from a study comparing observed maximum jumping speeds of approximately 800 boys and girls: standard two-sample t-tests suggest that jumping speed does not depend on gender until 10 years and that boys' jumping speed is clearly higher from 13 years on (with a p-value 0.047 in the age category 12-13 years). We apply a simple two-sample gradual change model and, by estimating a single change-point estimator, we are able to detect statistically significant difference already at 11.26 years. Under homoscedasticity, we establish asymptotic normality of the change-point estimator leading to simple formulas for p-values and confidence intervals. For heteroscedastic observations, the distribution of the change-point estimator is approximated by wild bootstrap. We discuss also several practical issues such as multiple testing and the relationship between p-values and one-sided confidence intervals.

### *Speeding Up Projection Pursuit using Fast Kernel Computations*

*David Hofmeyr*
*Stellenbosch University*
*dhofmeyr@sun.ac.za*

Projection pursuit refers broadly to the class of dimension reduction problems in which a chosen index is optimised over all linear projections of a data set. The most well known projection pursuit model is that of Principal Component Analysis (PCA), in which the projection index may be formulated as the variance of the projected data. More complex projection indices, however, may require estimation of more than just the moments of the projected data. Independent Component Analaysis (ICA) is based on the assumption that the observed data represent realisations of a linear combination of independent sources. The corresponding projection pursuit is based on finding the projection which maximises the likelihood under this independence assumption. Since the true likelihood is not known, pseudo-likelihoods based on kernel estimators are used instead. Basing the projection index on exact evaluations of the pseudo-likelihood results in high computation cost, however, and so simple surrogates for an equivalent problem based on the sample entropy are popular. Recent advances in fast kernel computations have allowed the exact pseudo-likelihood approach to be realised for larger data sets than was previously possible. This talk will discuss the application of these fast and exact kernel methods to solving the ICA problem, as well as other problems with computationally expensive projection indices based on non-parametric estimates of the projected data distribution.

### *Skewed-t Score Driven Volatility Models Applied to FTSE/JSE ALSI Returns*

*Stefan Janse Van Rensburg*
*Nelson Mandela University*
*Sharp, GD (Department of Statistics, Nelson Mandela University)*
*sjansevanrensburg@gmail.com*

This study examines the ability of score driven (SD) models based on skewed-t distributions to model and forecast various aspects of Johannesburg Stock Exchange (JSE) All Share Index (ALSI) daily returns. It is shown that these models are competitive with various GARCH specifications in terms of risk forecasts and density forecasts. For both GARCH models and SD models, there is evidence to suggest that the conditional distribution of ALSI daily returns is negatively skewed.

### *Genetic Algorithms (GA) for Feature Selection*

*Edward R. Jones*
*Texas A&M University*
*Zhang, H (Department of Sociology, Texas A&M University)*
*ejones@tamu.edu*

This paper proposes an approach for feature selection using genetic algorithms (GA). Although the focus is on using this approach in linear regression, it can be extended to other machine learning methods. The GA approach is tailored to regression models and then compared to traditional feature selection using stepwise and lasso. In this research, emphasis is placed on finding the best feature subset among all possible combinations. The approach is illustrated using a case study from fracking oil wells and simulations. The observation is that GA selection has great benefit for applying machine learning in applications with many nuisance features, some of which may be highly colinear. GA selection is more likely to find the best model among all possible subsets. Constraints from model restrictions, data transformations, data encoding are naturally incorporated into the algorithm. Although the time needed to find the best solution is higher than shrinkage methods, in most cases it is acceptable when compared to the improved selection and confidence in the selected features. The case study and simulations used SAS® Enterprise Miner™ and python.

### *An Introduction to Gaussian Belief Propagation with Exploratory Remarks*

*Francois Kamper*
*Department of Statistics and Actuarial Science, Stellenbosch University*
*francoisk@sun.ac.za*

Belief propagation (BP) has been applied as an approximation tool in a variety of inference problems. The purpose of this talk is to provide an introduction to a specific application of BP and discuss some potential novel applications. BP is applied to a Markov graph (MG) constructed from a Gaussian distribution in canonical parameterisation, this is known as Gaussian belief propagation (GaBP). The main drawback is that GaBP is not guaranteed to converge when applied to MGs of arbitrary topology. This drawback can be overcome by applying a regularisation scheme to the message-passing. Two potential novel applications of

GaBP will be considered. In the first, a method of augmenting GaBP to allow for solving the lasso in linear regression is discussed. The advantage of this augmentation over other methods is that the algorithm can be easily distributed. The second potential application is the use of GaBP to post-process similarity measures between elements of a set of objects. The idea is to construct a precision matrix, based on the similarity scores, apply GaBP to this precision matrix and use the strength of the messages between objects as new similarity measures.

### *An Investigation into Tree Growth Modelling. a Bayesian Approach*

*Lulama Kepe*
*Nelson Mandela University*
*Little, K (Department of Forestry, Nelson Mandela University), Hugo, J (Department of Statistics, Nelson Mandela University)*
*lulama.kepe@mandela.ac.za*

In commercial forest production, predicted yields based on mensuration recommendations are seldom realized. Moreover, minimal silvicultural standards based on results from trial blocks do not account for the relationship between initial planting density and final stocking. Growth models designed to investigate management options must therefore employ competition indices to provide reliable predictions under extreme silviculture. To investigate competition amongst individual trees, a Bayesian mixed effects model, similar in characteristics to a Sire model used in animal breeding, will be proposed. In animal breeding models, the Sire Model allows for the inclusion of a numerator relationship matrix containing inbreeding coefficients. In a similar way, documented competition indices will be included in the model in an attempt to estimate posterior probabilities of individual trees being the strongest growers during different stages of growth.

### *SARIMA Modelling and Forecasting the Monthly Rainfall of the City of Cape Town, South Africa*

*Rambuwani Khathutshelo Kenneth*
*Statistics South Africa*
*Prof Ndlovu (University of South Africa)*
*kennethram@statssa.gov.za*

The city of Cape Town is one of the eight metropolitan municipalities in South Africa, it receive most of its rain in the opposite season as the rest of the country. The main objective of the study was to find the model that best describes the pattern of rainfall data in the city of Cape Town. Using univariate Box-Jenkins methods, a Seasonal Auto Regressive Moving Average (SARIMA) model was developed to forecast monthly rainfall using the monthly average rainfall data (1997 to 2017) collected from seven weather stations situated in the city of Cape Town. Using R-Studio the SARIMA model $(2;0;0)\times(2;1;2)_{12}$ with Akaike Information Criteria (AIC) = 2288.691 was found to be the best t model to forecast the average rainfall in the city of Cape Town.

### *Imputation; How Good is it?*

*Nyiko Khoza*
*University of South Africa*
*Bhekisipho Twala*
*ngobenh@unisa.ac.za*

This study focused on comparing different techniques when handling missing observations in data to assess if imputation techniques are good or not? We looked at techniques that manage missing observations by deleting missing observations when there is missing observations in data and techniques which imputes observations. Looking at consequences each technique encounters when managing the problem of missing observations in a data, this study allowed us to reach to a better conclusion. This study was therefore able to propose a model and a good technique to be used when handling missing observations in data to correctly address the problem of missing observations in data and measure reliable statistical estimates. No matter how many times the same analysis is conducted, or different users conduct the analysis on the same data without knowing the seed run; they will be able to get the same solution when handling missing observations in data. This study has found that the multiple values imputation technique is not as good as it is said to be when compared to an exceptionally good way of handling missing observations in data. A model to describing this phenomenon is outlined here.

### *New Local Measures for Geostatistical and Lattice Data*

*Christine Kraamwinkel*
*University of Pretoria*
*Fabris-Rotelli, IN, Magagula, E (Department of Statistics, University of Pretoria)*
*christine.kraamwinkel@gmail.com*

There are three types of spatial data namely geostatistical data, lattice or regional data, and point patterns. These three spatial data types differ based on the properties of the domain and values observed. It is therefore important to understand the nature of the spatial data before making model building decisions. Current exploratory techniques for geostatistical and lattice data indicate a global strength of spatial dependence, but provide little to no information on the nature of this relationship, namely whether values are clustered or regular. We propose here an extension of the point pattern summary functions K and L for use on geostatistical and lattice data. We focus only on these two summary functions as the functions F,G and J depend on an estimated intensity $\lambda$ of the process which cannot be kept constant in our approach. The K-function, however, is widely used and most often the first choice.

### *Nested EM Algorithms for Estimation of Hierarchical Wind Speed Distributions*

*Michaela Laidlaw*
*University of Pretoria*
*Bekker, A, Ferreira, JT (Department of Statistics, University of Pretoria)*
*michaelalaidlaw@gmail.com*

Generating power by harnessing natural phenomena such as wind is one of the most common and successful solutions to the need for sustainable energy. In order for a wind farm to be successful in generating power the behaviour of wind speed at the site being considered is essential. Common methods implemented for modelling wind speed focus on unimodal, simple distributions; these are not always sufficient as data is often multimodal. In order to model more complex wind speed behaviour, mixtures of distributions have been introduced and parameters obtained via the EM algorithm. This presentation will investigate the EM algorithm for mixtures for some familiar wind speed distributions as well as an extension to the Birnbaum-Saunders distribution. The validity of hierarchical mixture distributions, with nested EM algorithms for parameter estimation, is discussed. Model fit is evaluated using a Kolmogorov-Smirnov measure and coefficient of determination, as behaviour at different sites affect the success of the goodness of fit process.

### *Determination of Factors Associated with Time to Recovery from Pneumonia using Cox Proportional Hazard Model*

*Masimthembe Lala*
*University of Fort Hare*
*Mutambayi, R, Chiruka, R (Department of Statistics, University of Fort Hare)*
*mlucalala@gmail.com*

Background: Pneumonia is the leading cause of hospitalization and death in children as well as adults in South Africa. Tuberculosis and HIV-infected individuals of all age groups have a substantially elevated risk of hospitalization and death due to pneumonia. We used survival analysis techniques to determine the recovery time and the effects of the associated factors and comorbidity conditions on recovery of pneumonia. Methods: Over a period of 4 years, we prospectively collected patients' information on factors of pneumonia in a selected hospital in King Williams Town, Eastern Cape to observe recovery time and define the long-term clinical outcomes of pneumonia patients using Cox Proportional Hazard model. Results: the results from the Cox PH model indicated that three covariates had higher significant relative risk. Having a diabetes (p = 0.002; RR = 3.443; CI = 0.606:2.597), exposed to toxic fumes (p = 0.007; RR = 3.666; CI = 0.911:1.790) and patients with high alcohol consumption (p = 0.048; RR = 0.687; CI = -.702: -0.15) were significantly associated with prolonged or poor recovery over time. Conclusion: The effects of prolonged recovery time, leads to continued morbidity and death in those infected with pneumonia. Therefore, future research should be aimed at developing effective novel therapeutics to minimize morbidity, mortality, and long-term complications from infection.

### *Factors Responsible for Safe Sex Practice among Female Adolescent: A Case Study of the University of Fort Hare Students*

*Makhadimola Rosa Leshabane*
*University of Fort Hare*
*Jubane, I, Ndege, J*

HIV/AIDS is still affecting large populations globally, female youths are at a greater risk than males according to World Health Organization (WHO). The aim is to investigate factors that influence safe sex practices among females. A cross sectional study was conducted on 200 females (18 - 39 Years), data was collected through Google forms. SPSS descriptive analysis was presented and dimension reduction to determine factors, SPSS Amos to determine relationships between variables using SEM, logistic regression to build a statistical model.

### *A Goodness-of-Fit Test for the Rayleigh Distribution Based on the Mellin Transform*

*Shawn Liebenberg*
*North-West University*
*Allison, JS (North-West University)*
*shawn.liebenberg@nwu.ac.za*

We develop a new goodness-of-fit test for the Rayleigh distribution based on the Mellin transform. The finite-sample performance of the tests is studied and it is found that the test performs well compared to competitor tests.

### *Models for Analysing Growth for a South African Cohort of Infants*

*Francesca Little*
*University of Cape Town*
*Jacqui Niehaus (University of Cape Town)*
*francesca.little@uct.ac.za*

Given data from a population birth study of mother-baby pairs, followed up from birth to at least 5 years, we look at models to characterise and compare growth measurement profiles. Several candidate models for the nonlinear growth profiles are compared, including linear and nonlinear parametric formulations and nonparametric spline-based models. These are fitted within a mixed effect modelling framework to account for the repeated measures. The mixed effect modelling approach is compared to a neural network approach. We assess the interpretability of the different approaches. A second class of models look at identifying predictors of stunting and wasting and compare logistic regression models to random forest classification trees.

### *Models for Cycles in Financial Time Series*

*Igor Litvine*
*Nelson Mandela University*
*igor.litvine@mandela.ac.za*

We introduce several definitions for cycles in time series: shape-cycles, peak-trough cycles, speculative cycles. The practical uses of these definitions are demonstrated on real asset prices and simulated series. Several smoothing techniques, based on the concept of differentiable waves are suggested for cyclical financial time series. A comparison of the smoothing techniques is conducted.

### *A Novel Application of Survival and Competing Risk Models in Agriculture*

*Sugnet Lubbe*
*Stellenbosch University*
*Daniels, A, Nieuwoudt,H (Institute of Wine Biotechnology, Stellenbosch University),*
*le Roux, N (Department of Statistics and Actuarial Science, Stellenbosch University)*
*slubbe@sun.ac.za*

Survival analysis is typically associated with biostatistics. Although no new theory is developed, the existing theory is applied in a novel application – the survival of table grapes in cold storage. Grapes from the first and second harvest of different cultivars and production areas were analysed in both 2016 and 2017. At weekly intervals, two boxes of each were inspected – one grape at a time – to determine whether there were any of 27 defects present. The weight of the grapes with each defect was determined, opening two different boxes each week. Unlike patients in a clinical trial, there are no repeated measurements and whether each grape's survived t weeks is only observed at a single time point t. This means each observation is either left or right censored. It will be shown how standard survival analyses have been adapted for this so-called current state data. The number of patients that survive in clinical trials are whole numbers, here a continuous value in grams is observed of those grapes that did not survive until time t. Furthermore, since there are 27 different types of defects, a grape is subject to not surviving due to competing risks. It is illustrated how a competing risks model is applied to the survival functions based on the current state data.

### *Hidden Markov Models Based on Truncated Weibull Distributions*

*Iain Macdonald*
*University of Cape Town*
*iain.macdonald@uct.ac.za*

There already exist models for time series of counts which can cope with under- or overdispersion in the counts. These are hidden Markov models (HMMs) having Conway-Maxwell-Poisson (CMP) distributions as the state-dependent distributions. Although such models seem flexible and show promise, they have some computational disadvantages and can be slow to fit. Here I explore the use of truncated Weibull distributions as an alternative to CMPs in such models. The probability mass function of a truncated Weibull is much simpler to evaluate than that of a CMP, and the resulting HMMs are therefore easier to fit. I apply such models to several illustrative data-sets and compare the adequacy of truncated Weibull HMMs with that of Conway-Maxwell-Poisson HMMs.

### *Optimal Smoothing of Cycles using Sinusoidal Waves*

*Alpheus Mahoya*
*Nelson Mandela University*
*Litvine, IN (Department of Statistics, Nelson Mandela University)*

Financial time series data is characterised by high volatility which makes forecasting difficult. This study used hierarchical methods to separate segments of overall downward and upward trends in the cycles and derive peaks and troughs (buy-sell points) between these segments. Sinusoidal waves because of their flexibility and differentiability power, were used to fit buy/sell points, extracting a new smoothed series. Daily share price data of several JSE listed companies was collected and analysed. The fitted wave performed well with for example an adjusted R-squared (> 0.94) was observed across the entire series. Successful implementation of this method helps investors and management with decision making and avoid losses.

### *Destructive COM-Poisson Cure Rate Model and Likelihood Inference with Lognormal Lifetime*

*Jacob Majakwara*
*University of the Witwatersrand*
*jacob.majakwara@wits.ac.za*

In this paper, we consider a flexible cure rate model by assuming the initial risk factors to undergo a destructive process so that what is recorded is only from the undamaged portion of the original number of risks factors. A Conway-Maxwell (COM) Poisson distribution is assumed for the initial risk factors and the lognormal distribution for the lifetimes of the noncured individuals, and the steps of the expectation maximisation algorithm are developed in detail to estimate the model parameters. An extensive simulation study is carried out to demonstrate the performance of the proposed estimation method. The flexibility of the COM-Poisson family is utilized to carry out a model discrimination using the likelihood ratio test. Finally, a melanoma data is analysed for illustrative purpose.

### *Modelling Extreme Co-Movement Between Oil Prices and Economic Growth*

*Katleho Makatjane*
*North West University*
*Moroke, N, Ncube, B (Department of Statistics and Operations Research, North West University)*
*katleho.makatjane@yahoo.com*

The current study investigated an extreme co-movements between oil prices and economic growth in South Africa (SA). To achieve our objective, a stationary hybrid threshold vector error correction model multivariategeneralised hyperbolic (MGH)-skew-Student's t-distribution-copulas (TVECM-MGH-skew-Student's t-distribution-copulas) was estimated. The MGH-skew-Student's t-distribution affine-linearly transformed random vectors with stochastically independent and generalized hyperbolic marginals while Archimax Copulas identify extreme dependence and co-movement structures between oil prices and economic growth in SA. Our results indicates oil prices are moving together in the long-run according to the observered cointergrating parameter di,j. Nevertheless, we then discover that our MGH distribution gave the parameter of the lower tail has an exponentially decaying behaviour while the upper tail has a polynomial decaying bahivour. This shows that the moments at the tails of the distributions are higher than that of a normal distribution implying extreme movement of 80.8%. To check for the extreme dependence, the results of the Archimax Copulas indicate that showed the moderate dependence with oil price and economic growth. However, negative news have a larger impact on the degree of dependence than positive news. Contagion effect is observed in both the oil price and economic growth in SA.

### *Hierarchical Forecasting of the Zimbabwe International Tourist Arrivals*

*Tendai Makoni*
*University of the Free State*
*Chikobvu, D (Department of Mathematical Statistics and Actuarial Science University of the Free State),*
*Sigauke, C (Department of Statistics, University of Venda)*
*tpmakoni@gmail.com*

Forecasting international tourist arrivals using hierarchical forecasting methods is important since it provides critical information for effective planning and other strategic decisions. Business people, investors, the government and tourism stakeholders are among beneficiaries of accurate forecasting of tourist arrivals. Monthly Zimbabwe international tourist arrivals data provided by the Zimbabwe Statistics Agency (ZIMSTAT) from years 2002 to 2018 is disaggregated according to Purpose Of Visit (POV). A single hierarchy with 6 nodes of level 0 and level 1 series is constructed. Bottom-up, top-down and the optimal combination approaches are adopted in forecasting international tourist arrivals to Zimbabwe. The three approaches are used in coming up with point forecasts prior to forecasting performance evaluation. According to the mean absolute percentage error (MAPE), the bottom-up approach gives the most accurate forecasts. Sixty month step-ahead point forecasts are produced using the bottom-up approach. Hierarchical forecasting allows the identification of areas that need special attention. Forecasts indicate a general increase in aggregate series. Disaggregated forecasts indicate no change in shopping and with slow increase in educational tourists. Results also indicate the need for educational reforms and local companies should diversify their products and market them internationally to attract international tourists. Holiday and in-transit tourists forecasts indicate a

### *Quantile Regression for Count Data using Delaporte Distribution*

*Kajingulu Malandala*
*University of South Africa*
*Ranganai, E*

Quantile regression estimates the relationship between explanatory variables and quartile points of the response variable (Koenker, 2009). Most researches on Quantile regression have been focusing on theory and applications in various areas such as econometrics, marketing, ecology, and finance (Neelon et al., 2015; Davino et al., 2013; Hao and Naiman, 2007). Quantile regression has been developed and well documented for continuous responses. However, for discrete data, much focus has been placed on Poisson and Negative binomial to model the conditional mean of the response measurements. Nonetheless, Poisson and Negative Binomial distributions are not void of limitations. Count data are non-negative, integer value and often their variance is larger than the mean. Using Poisson and negative Binomial distributions for modeling count data may mislead the inference for under/overspread data. To overcome these limitations, this paper proposes a more flexible frequentist and robust approach using Delaporte distribution. The results demonstrated that the proposed approach is a robust and flexible quantile regression model.

### *Distribution-Free Precedence Schemes with a Generalized Runs-Rule for Monitoring Unknown Location*

*Jean-claude Malela-majika*
*University of South Africa*
*Rapoo, EM (Department of Statistics, University of South Africa),*
*Mukherjee, A (Department of Production, Operations and Decision Sciences, XLRI-Xavier School of Management, XLRI-Jamshedpur, India),*
*M.A. Graham (Department of Science, Mathematics & Technology Education, University of Pretoria)*
*malelm@unisa.ac.za*

Nonparametric statistical process monitoring schemes are robust alternatives to traditional parametric process monitoring schemes, especially when the assumption of normality is invalid or when we do not have enough information about the underlying process distribution. In this paper, we propose to improve the well-known precedence scheme using the 2-of-(h+1) supplementary runs-rules (where h is a nonzero positive integer). The in-control and out-of-control performances of the proposed control schemes are thoroughly investigated using both Markov chain and simulation based approaches. We find that the proposed schemes outperform their competitors in many cases. A real-life example is given to illustrate the design and implementation of the proposed schemes.

### *Points of Impact Analysis in Functional Linear Regression Setting: A Case Study*

*Siphumlile Mangisa*
*Nelson Mandela University*
*Das, S (Department of Business Management, University of Pretoria)*
*siphumlile.mangisa2@nmmu.ac.za*

The prediction of scalar outcomes using functional predictors is a frequent problem in functional Data Analysis. Standard linear functional regression model has been successfully applied to many problems. However, results are often difficult to analyse and interpret since the model is a weighted average of the whole trajectory of the functional predictors, which make it difficult to assess specific local characteristics in the underlying process. In some applications only specific locations (points of impact) in the functional predictors have an impact on the outcome (i.e. points that are 'most influential'). Such points of impact are usually unknown. We use a case study of the Global Crises Index as a scalar response of interest, versus the US and UK stock market rolling window correlations as the functional predictor. The US and UK stock market are chosen because they are important and the major drivers of the global economy, and they also provide a long time series of data spanning about two centuries. Findings from this investigation will be shared and implications discussed.

### *Spatial Statistics of Extrees with a View Towards Application to Extreme Weather Events in South Africa and Other Neighbouring Countries in the SADC Region*

*Daniel Maposa*
*University of Limpopo*
*Maposa, D (Department of Statistics and Operations Research, University of Limpopo, South Africa),*
*Cochran, JJ (Department of Information Systems, Statistics and Management Sciences, University of Alabama, USA)*
*danmaposa@gmail.com*

We look at the possibility of application of spatial extreme value statistics to the extreme weather events such as precipitation (or flood heights) and temperature in South Africa and other neighbouring countries in the Southern African Development Community (SADC) region. This study is motivated by the existence of spatio-temporal variability of extreme weather events in South Africa and other neighbouring countries in the SADC region revealed from the previous studies, particularly with reference to the 1991/1992 severe droughts in the region, the year 2000 disastrous floods in the region and the most recent 2019 disastrous floods caused by cyclone Idai which affected parts of Malawi, Mozambique, South Africa and Zimbabwe.

### *Comparative Analysis of the 100-Year Return Level of the Average Monthly Rainfall for South Africa: Parent Distribution Versus Extreme Value Distributions*

*Daniel Mashishi*
*University of Limpopo*
*Maposa, D, Lesaoana, M (Department of Statistics and Operations Research, University of Limpopo)*
*daniel.mashishi@ul.ac.za*

In this paper we model average monthly rainfall for South Africa using the parent distribution and extreme value theory (EVT). The 100-year return level plays an important role to hydrologists, meteorologists and civil engineers. Hence our attention in this study is on modelling the 100-year return level of average monthly rainfall for South Africa using the parent distribution and EVT. The main purpose of this paper is to compare the extreme quantile estimates of the EVT and parent distributions as well as to reveal the risk brought by heavy rainfall in South Africa. The method of maximum likelihood was used to estimate unknown parameters. In this paper, we first investigate the parent distribution of the average monthly rainfall for South Africa. The results showed that the Weibull domain of attraction which include the two-parameter Weibull distribution is the appropriate parent distribution to model the data. We then perform a comparative analysis of the 100-year return level using the two-parameter Weibull distribution, the generalised extreme value distribution (GEVD) and the Poisson point process. The findings revealed that the 100-year return level of the two-parameter Weibull distribution was lower as compared to that of the GEVD and Poisson point process model. The 100-year return level of the GEVD was equal to that of the observed maximum for the series, whereas that of the Poisson point process was slightly higher than the observed maximum average monthly rainfall.

### *Forecasting Extreme Conditional Quantiles of Electricity Demand in South Africa*

*Norman Maswanganyi*
*University of Limpopo*
*Sigauke, C (Department of Statistics, University of Venda), Ranganai E (Department of Statistics, University of South Africa)*
*nmaswanganyi72@gmail.com*

The study is focused on ensuring access to affordable, reliable, sustainable and modern energy for the developed and developing countries. In achieving the national goals relevant to establishing renewal energy policy objectives, South Africa needs to identify the key goals for the nations and how the electricity sector fits among its priorities. The study proposes the method on detecting the estimates of the probabilities of rare events and extreme quantiles between 0.95 and 0.9999. The additive quantile regression (AQR) model is compared with extreme conditional quantile based approach. The probabilistic accuracy measures for both models are also calculated and compared. The comparisons are carried out using daily peak electricity demand (DPED) data specifically from South Africa ranging from January 1997 to March 2014. Based on semi-parametric mixture model, kernel density is fitted to bulk model and tail model for the observations above the threshold. In this study, the methods that show the best results are presented. This work is motivated by one-sage and two-stages extreme conditional quantile procedures used to illustrate the estimation of the proposed approach.

### *Levels and Determinants of Knowledge of HIV/AIDS among Women in South Africa*

*Tshepho Matlwa*
*Statistics South Africa*
*brianmatlwa@gmail.com*

In South Africa, prevalence of HIV/AIDS is high and on the increase. However, there is paucity of studies on knowledge of HIV/AIDS especially among women in the country, as such it is a concern. Adapting a quantitative analysis approach, the study use the South African Demographic Health Survey (SADHS 2016) data to explore the knowledge of HIV/AIDS among women aged 15-49 in S.A. The SADHS data is a cross-sectional data collected by Stats SA from a total of 15 292 households across the country. Analysis was carried out at the bivariate and multivariate levels. The binary logistics regression was used at the multivariate level. Finding were expressed using charts and tables. Findings shows that 22.2% of women in Gauteng have heard of HIV/AIDS. Also, about 67% of women who are not working have knowledge, while 63% of women in rural areas have not heard of HIV/AIDS. In addition, 6.4% of women with no education knew or have heard of HIV/AIDS and 48% of never married women have heard. At bivariate level all tested variables were found to be associated. However, at multivariate level women who are working are more likely to have the knowledge, with coloured having higher likelihood irrespective of the employment status. Also, E.C and KZN were found not significant, while Gauteng was found to be more significant. The study concludes by recommending that findings from the study be considered for all policy and programme development around knowledge and awareness of HIV/AIDS in

### *Models for Early Identification of Students at Risk of Failing*

*Saadiyah Mayet*
*University of Cape Town*
*Singo, U, Scott, L (Department of Statistics, University of Cape Town)*
*saadiyah.mayet@uct.ac.za*

Though the historic 2015 and 2016 calls for free education brought the value of tertiary education to the forefront of South African dialogue in an unprecedented way, student performance at university has long been a research topic of interest globally. Prior academic factors have been used to predict student performance with much success in previous work; however, such research is lacking in the South African context. Moreover, the extent to which a student's socio-economic background affects the likelihood of success at a tertiary level remains under-explored across the world. This project investigated the factors contributing to student academic performance in introductory statistics and mathematics courses at the University of Cape Town, and built models with the goal of early identification of students at risk of failing. An array of demographic and academic features were used to build logistic regression, multiple linear regression, and support vector machine models to predict student performance. It was found that students' mathematics marks at the end of high school and, where applicable, in prerequisite courses were reasonably strong indicators of their likelihood of passing the course. Additionally, there was evidence that students with a wealthier background were more likely to pass. That said, it was apparent that there were other factors at play which models could not capture, possibly relating to the emotional and mental well-being of students.

### *Robust Estimation of Pareto-Type Tail Index Through an Exponential Regression Model*

*Richard Minkah*
*University of Ghana*
*de Wet, T (Stellenbosch University)*
*rminkah@ug.edu.gh*

In this paper, we introduce a robust estimator of the tail index of a Pareto-type distribution. The estimator is obtained through the use of the minimum density power divergence with an exponential regression model for log-spacings of top order statistics. The proposed estimator is compared to an existing estimator for Pareto-type tail index based on fitting an extended Pareto distribution with minimum density power divergence. A simulation study is conducted to assess the performance of the estimators under different contaminated samples from different distributions. The results show that the proposed estimator has better mean square errors and less sensitivity to an increase in the number of top order statistics. In addition, the estimation of the exponential regression model yields estimates of second-order parameters that can be used for estimation of extreme events such as quantiles and exceedance probabilities. The estimators are illustrated with a practical dataset on insurance claims.

### *Gaussian Mixture of Expert Model for Censored Data*

*Elham Mirfarah*
*university of Pretoria*
*Chen, D, Naderi, M (Department of Statistics, University of Pretoria)*
*elham_mirfarah@yahoo.com*

Mixture of Linear Experts (MoLE) model is a promising statistical tool to simultaneously investigate the linear relationship of random phenomena under study and classify the set of complete heterogeneity data. However, in some practical application (especially in the analysis of lifetime data such as accelerated failure time model) censoring occur due to the nature of experiment. In this work, we propose MoLE to model the censored data under the Gaussian distribution for the error term. This approach helps us to regress as well as to cluster the heteroscedastic and multi-modal censored data. A feasible EM-type algorithm is implemented to obtain the maximum likelihood estimate of the parameters. The methodology performance is illustrated by simulation and a real-world data analysis.

### *Modelling the Sporadic Behaviour of Rainfall Time Series using ETS State Space and SARIMA Models in the Limpopo Province, South Africa*

*Selokela Victoria Molautsi*
*University of Limpopo*
*Lasisi, TA (Department of Statistics and Operation Research, University of Limpopo, South Africa),*
*Moeletsi, ME (Agricultural Research Council), Boateng, A (University of Cape Coast, Ghana)*
*victoria.molautsi@ul.ac.za*

Abstract: The effects of ozone depletion on climate change has, in recent years, become a reality, impacting on changes in rainfall patterns and severity of extreme floods or extreme droughts. The majority of people across the entire African continent live in semi-arid and drought-prone areas. Extreme droughts are prevalent in Somalia and eastern Africa, while life-threatening floods are common in Mozambique and some parts of other SADC countries. Research has cautioned that climate change in South Africa might lead to increased temperatures, reduced amounts of rainfall, thereby altering their timing and putting more pressure on the country's scarce water resources, with implications for agriculture, employment and food security. The average annual rainfall for South Africa is about 464mm, falling below the average annual global rainfall of 860mm. The Limpopo Province, one of the nine provinces in South Africa, and of interest to this study, is predominantly agrarian, basically relying on availability of water, with rainfall being the major source for water supply. It is therefore, pertinent that the rainfall pattern in the province be monitored effectively to ascertain the rainy period for farming activities and other uses. This study employed the SARIMA and ETS State Space models to capture the sporadic behaviour of rainfall data. These two models have been widely applied to climatic data by many scholars and adjudged to perform creditably well. In an attempt to find a suitable prediction model for monthly rainfall patterns in Limpopo Province, data from Macuville and Marnits Stations, ranging from January 1900 to December 2016 were analysed. SARIMA$(1,0,1)\times(1,1,1)_{12}$ and ETS(A,N,A) models based on exponential smoothing were built and the results showed that the two models were adequate.

### *Comparisons of Quality of Life in Haemodialysis and Peritoneal Dialysis Patients: A Case Study of Alice and King Williams's Town, South Africa*

*Lekhoele L Moleleki*
*University of Fort Hare*

Our study was designed to compare QOL (Quality of Life) in chronic kidney failure dialysis between PD (Peritoneal dialysis) and HD (Haemodialysis) patients from National Renal Care Unit. QOL was compared using a Medical Outcomes Short Form (SF-36). Our study showed that there was a significant difference between HD and PD patients for total SF-36 score. PD patients reported significantly greater overall discomfort from bodily pain than HD. There was a significant difference between HD and PD patients for total SF-36 score. These results could become in handy mostly in the development of health care policies.

### *Regime Switching Models in Time Series*

*Wessel Hendrik (Henri) Moolman*
*Walter Sisulu University*
*moolman.henri@gmail.com*

Regime Switching models are used when there are changes in a Time Series over different time intervals. These differences can be due to different parameters, error variances or error distributions. The theory and practical applications of such models will be discussed.

### *A New Double Sampling $\bar{X}$ Control Chart for Monitoring an Abrupt Change in the Process Location*

*Collen Motsepa*
*Statistics South Africa*
*Malela-Majika, JC (Department of Statistics, University of South Africa), Graham, MA (Department of Statistics, University of Pretoria)*
*cmmotsepa@gmail.com*

This paper develops a new double sampling (DS) monitoring scheme, namely, the side-sensitive DS $\bar{X}$ chart, to monitor the process mean. The operational procedure is presented first followed by the exact form of the probability of the in-control process under the normality assumption. Finally, the performance of the new scheme is investigated by minimizing the out-of-control average run-length and extra quadratic loss function. It was observed that the proposed chart presents a better overall performance than the existing DS $\bar{X}$ chart. An illustrative example is given to facilitate the design and implementation of the new chart.

### *An Early Infant HIV Risk Score for Targeted HIV Testing at Birth*

*Chris Muller*
*Stellenbosch University*
*du Plessis, NM, Avenant, T (Department of Paediatrics, University of Pretoria),*
*Pepper, MS (Department of Immunology, SAMRC Extramural Unit for Stem Cell Research and Therapy, University of Pretoria),*
*Goga, AE (Health Systems Research Unit, South African Medical Research Council)*
*cmuller@sun.ac.za*

Early HIV testing is needed for treatment success in young infants, but universal abstract testing is expensive. In this study, we examined the feasibility of early infant HIV risk scores for targeted polymerase chain reaction (PCR) testing and early HIV diagnosis. SAS Enterprise Miner was used to analyse patient data from between August 2014 and December 2016. 15 175 live infants were born at KPTH, 3356 (22.12%) of these to mothers infected with HIV. Informed consent was obtained from 1759 of 1911 (92.05%) eligible patients. Patients with birth HIV PCR test results were included.

### *Pregnancy Incidence and Risk Factors among Women Participating in Vaginal Microbicide Trials for HIV Prevention: Systematic Review and Meta-Analysis*

*Alfred Musekiwa*
*University of Pretoria*
*Muchiri, E (Aurum Institute) Manda, S (Biostatistics Unit, South African MRC),*
*Mwambi, HG (School of Mathematics, Statistics, and Computer Science, University of KwaZulu-Natal)*
*alfred.musekiwa@gmail.com*

INTRODUCTION: Pregnancy is contraindicated in vaginal microbicide trials for HIV prevention in women due to the unknown maternal and fetal safety of the microbicides. Women who become pregnant are taken off the microbicide but this reduces statistical power. This systematic review estimates the overall incidence rate of pregnancy in microbicide trials and describes associated risk factors. METHODS: A comprehensive literature search identified eligible studies from electronic databases. Two review authors independently selected studies and extracted relevant data from included studies. Meta-analysis of incidence rates of pregnancy was carried out and risk factors of pregnancy were reported narratively. RESULTS: Fifteen studies reporting data from 10 microbicide trials (N=27,384 participants) were included. A total of 4,107 participants (15.0%) fell pregnant and a meta-analysis of incidence rates of pregnancy from 8 microbicide trials (N=25,551) yielded an overall incidence rate of 23.37 (95%CI: 17.78 to 28.96) pregnancies per 100 woman-years. Hormonal injectable, intra-uterine device or implants or sterilization, older age, higher education and condom use were associated with lower pregnancy. Living with a man, history of pregnancy, desire for future baby, oral contraceptive use, increased number of unprotected sex acts and inconsistent condom use influenced pregnancy. CONCLUSIONS: Pregnancy incidence in microbide trials is high and strategies for its reduction are required.

### *Modelling Malaria Time to Re-Infection with Time Varying Covariates Effects: A Case Study of Outpatients in DR Congo*

*Ruffin Mutambayi*
*Department of Statistics, University of Fort Hare*
*Adeboye, A*

Patient's living conditions that could influence graft survival may also exhibit non-constant effects over time; therefore, violating the important assumption of the Cox proportional hazard (PH) model. The researchers describe the effects of some selected malaria covariates on the hazard of graft failure in the presence of re-infection based on 109 malaria outpatients in Lubumbashi Congo Hospital, who had re-infection status after six months of follow-up. The survival status of re-infected patients was based on the effect of various factors. The study was carried out using Cox PH, a variation of the Aalen additive hazard and Accelerated failure time (AFT) models. The selection of important factors was based on the purposeful method of variable selection. It was found that modelling the covariates using the Cox PH concept without adequate assessment of the model fit could under-estimate significant covariates. The additive hazard and AFT models offer more flexibility in understanding covariates with non-constant effects on survival. The results also suggest that the follow-up period till re-infection time when using the time-varying concept is accommodating more factors that are statistically linked to the re-infection of malaria patients in DR Congo.

### *On the Maximum Likelihood Parameter Estimation of the Bimodal Skew-Normal Distribution*

*Mehrdad Naderi*
*Department of Statistics, Faculty of Natural and Industrial sciences, University of Pretoria*
*Bekker, A (Department of Statistics, University of Pretoria),*
*Jamalizadeh, A (Department of Statistics, Shahid Bahonar University of Kerman, Iran)*
*mehrdad.naderi@ymail.com*

Modeling random phenomena based on the normal distribution is one the most widely used technique for the investigators. However, several practical datasets exhibit non-normal features, including asymmetry and bimodality, which are criticized the assumptions of the normal model. This paper presents a nice stochastic representation of bimodal skew-normal distribution which is useful for implementing the expectation-maximization (EM) algorithm to compute maximum likelihood estimate (MLE) of unknown parameters. Consequently, the superiority of the methodology is illustrated by the real data example.

### *New Contributions to Möbius Transformation-Induced Distributions on the Disc*

*Priyanka Nagar*
*University of Pretoria*
*Bekker, A (Department of Statistics, University of Pretoria), Arashi, M (Department of Statistics, Shahrood University of Technology)*
*priyanka.nagar@up.ac.za*

The joint modelling of angular and linear observations is crucial as data of this nature are prevalent in multiple disciplines, for example the joint modelling of wind direction and another climatological variable such as wind speed or air temperature, the direction an animal moves and the distance moved, or wave direction and wave height. Hence, there is a need for developing flexible distributions on the hyper-disc, which has support of the interior of the hyper-sphere, as it allows for modelling the combination of angular and linear observations. A new class of bivariate distributions is proposed which has support on the unit disc in two dimensions that includes, as a special case, the existing Möbius distribution on the disc. This new class of distributions for the disc have the ability to capture any inherent bimodality present in the data. The flexible behaviour of the proposed models in terms of bimodality and skewness is graphically demonstrated. The fit of the proposed models, which account for bimodality, to the Marion Island wind data were evaluated analytically and visually.

### *The Role of Statisticians in Harnessing the 4th Industrial Revolution*

*Mark Nasila*
*First National Bank*
*mnasila@fnb.co.za*

The 4th Industrial Revolution (4IR) is characterised by the fusion of technologies between cyber physical, digital, and biological spheres. This revolution represents a fundamental change in the way we live, work and relate to one another. It is a new chapter in human development enabled by these advances that have raised expectations from society for businesses to provide increasingly enhanced, customized offerings to help meet the needs of individuals and organizations as they evolve over time. Despite so much investment into making sure organisations are data driven, Forbes reports that only 12% of executives surveyed say they are executing a company-wide data strategy centred at what companies offer. In this talk I discuss the roles of statisticians in helping organisations Harness the 4th Industrial Revolution specifically around transforming their data strategies [Omni -Channel to Ubiquitous data strategies] using 4IR technological advancements (Artificial Intelligence, Machine Learning, Block-chain, Internet of Things) to strategically align their value proposition to society domains such as: Education, Security and Justice, Health and Hunger, Crisis Response etc.

### *The Impact of Assessment on Student Learning*

*Dries Naude*
*University of the Free State*
*naudeam@ufs.ac.za*

Research in the field of assessment indicates a shift in focus from traditional testing practices to a mere constructive assessment approach that aims to enhance learning. Assessment is an integral part of learning and should be planned and conducted in a constructive way, and not an add-on to teaching and learning. The aim is to address the paradigm shift that has occurred in assessment, and to provide generic guidelines on how to plan and conduct the process of assessment of learning.

### *A Formulated Combined Model in Forecasting Long-Term Energy Consumption in South Africa*

*Livhuwani Nedzingahe*
*ESKOM*

In planning, forecasting is an integral part that aid management to cope with the level of uncertainly into the future. This is done mainly by relying on the past, present data, management experience, knowledge and judgement. Long term energy forecasting is a necessity for Power Utilities and Government strategic planning from medium to long-term system adequacy outlook. The purpose of the study is to investigate a method to forecast long-term energy consumption in South Africa using a combined parametric method. Results shows that using Regression- Seasonal Autoregressive Moving Average (SARIMA) model in forecasting long-term energy consumption reduces the level of error terms and improves accuracy of results considerably. The study recommended Regression

### Statistics Behind Big Data Analysis: Bridging the Gap Between Satellite Imagery and Business Intelligence

Ariane Neethling
University of the Free State
Teessen, M (GeoTerraImage)
ariane_neethling@yahoo.com

In the digital age, increased availability of and accessibility to satellite imagery combined with the application of IR4.0 tools such as IoT, AI and Advanced Sensor Technology present increased opportunities for the capturing, processing, analysis and interpretation of large-scale data (Big Data) in creative ways to find relationships between subjects which previously would have seemed to be irrelevant. This enables the public and private sector to make a difference at all levels of society, resulting in the improved living standards of ordinary people, facilitating the smart cities initiative and supporting the protection of the natural environment. Amidst this opportunity lies the challenge to transform large amount of data, such as those from satellite, into timeous, relevant and reliable Business Intelligence. By referring to various real-life applications, the paper will demonstrate how the era of IR4.0, which is characterised by immense data sets represented as information in real-time, requires tools that can capture, process, analyse and present the results thereof in the shortest time possible – a function that defines the role of statistics in the era of the digital age.

### Skew Generalised Normal Innovations for the AR(1) Model

Ané Neethling
Department of Statistics, University of Pretoria
Ferreira, JT, Naderi, M (Department of Statistics, University of Pretoria)
ane.neethling@up.ac.za

In many real life scenarios and statistical applications, the assumption of symmetry is often violated – in fact, the demand for modelling asymmetry has been increasing. This implies a departure from the well-known assumption of normality defined for the innovations. The autoregressive process of order one is of particular interest in this talk, which is a popular model in time series and regression contexts. A skew generalised normal distribution is investigated as possible innovation structure for the AR(1) process. A simulation study illustrates the behaviour and statistical properties of the parameters and estimation methods. A real time series dataset is analysed and results are compared to previously proposed models.

### Non Parametric Techniques for Multilevel Discrete Survival Data

Thambeleni Nevhungoni
University of Venda
Bere, A (Department of Statistics, University of Venda), Manda, S (SAMRC)
tpnevhungoni@gmail.com

In discrete survival modelling the parametric approach has been receiving a lot of attention due to its computational ease. However, recently attention has shifted to nonparametric methods due to lack of flexibility with parametric approach. A lot of work has been done on nonparametric modelling of the baseline hazard, link function and the random effect separately. This study seek to build a discrete survival model where the base line hazard, link function and the random effect are simultaneously modelled nonparametrically.

### Multi- Level Modelling of Associations Between Inflammation, Smoke Exposure and Pneumococcal Carriage in a Gambian Birth Cohort Study

Raymond Nhapi
Division of Epidemiology and Biostatistics, University of Cape Town
Lesosky, M (Division of Epidemiology & Biostatistics, University of Cape Town),
Kwambana, B (Division of Infection and Immunity, University College London)
nhpray001@myuct.ac.za

Recent literature points out the role of air pollution as a major contributor to the risk of pneumonia acquisition especially in infants. Amongst these pollutants, biomass smoke and tobacco smoke exposures tend to be highly prevalent especially in low- to middle-income countries. This study aims to investigate the associations between smoke exposure and infant pneumococcal carriage specifically focusing on the role of inflammation in this 'potentially causal' pathway. A cohort of 120 mother-infant pairs recruited at birth in Gambia between March 2013 and September 2015. They were followed up monthly in the first 12 months then

quarterly in the following year. Medical data were collected from infant health cards and prescriptions, nasopharyngeal swabs, blood and serum samples were collected at a number of visits and anthropometric and demographic data was collected by the field team at each visit. Pneumococcal carriage was measured by qPCR focussing on the detection of the lyta gene whilst inflammation was measured by using the Alpha-1 glycoprotein and C-reactive protein. Variable selection is done by making use of a combination of variable importance statistics from random forests, penalized generalized estimation equations and theory from a literature search. Directed acyclic graphs are then used to obtain the minimal confounding sets. We compare Bayesian Multi-Level Models to Generalized Linear Mixed-effects Models in terms strengths of the hypothesized associations.

### *Extended Applications of GPAbin Biplots*

*Johané Nienkemper-Swanepoel*
*Stellenbosch University*
*le Roux, NJ, Lubbe, S (Department of Statistics and Actuarial Science, Stellenbosch University)*
*jnienkie@gmail.com*

Multiple imputation (MI) is a well-established technique for analysing missing data. Multiple imputed data sets are obtained and analysed separately using standard complete data techniques. The estimates from the separate analyses are then combined for inference. However, the exploratory analysis options of multiple imputed data sets are limited. Biplots are regarded as generalised scatterplots which provide a simultaneous configuration of both samples and variables. Therefore, a visualisation for each of the multiple imputed data sets can be constructed and interpreted individually, but in order to formulate an unbiased conclusion, the visualisations have to be appropriately combined for a unified interpretation. The GPAbin technique has been developed to address this problem. Generalised orthogonal Procrustes analysis (GPA) is used to align the biplots before combining them in a mean coordinate matrix. The name GPAbin is derived from the amalgamation of GPA and Rubin's rules, which are the combining steps used after MI. Simulation studies have confirmed the usefulness of the GPAbin method for categorical data in the context of multiple correspondence analysis based biplots. The GPAbin methodology is now extended to multivariate continuous data by using principal component analysis biplots and constructing log-ratio biplots for compositional data. The extended GPAbin method is illustrated by creating artificial missingness in complete data sets.

### *Stochastic Modelling of Inflation and Interest Rates for Defined Benefit Pension Plan Projections in Ghana*

*Ezekiel Nii Noye Nortey*
*University of Ghana, College of Basic and Applied Sciences*
*Quarshie, HD, Doku-Amponsah, K (University of Ghana, College of Basic and Applied Sciences)*
*ennnortey@ug.edu.gh*

Pension plan administrators, employers and managers in exchange for service provided currently by employees' pledges stated benefits in the prospective future. For this expense to be budgeted for the future, a pension cost method is used by the plan administrator to establish a form of warranty for the member. A fraction of the future liability to the present year known as the normal cost is allocated by a cost method. Two methods were used to calculate the normal cost, that is, the total and projected unit credit cost using different interest, inflation assumptions and constant single life annuity. The economic variables inflation and interest rate were modeled based on data from the Bank of Ghana. Several time series models were considered, SARIMA(3,1,0)x(2,0,0)12 was the appropriate time series model for inflation whereas ARIMA(1,1,0) was the best model for interest rate based on the selection criterion among the different ARIMA models fitted to the data. Based on the final models selected for the variables, 30 years ahead was forecasted, 100 simulations were carried out on inflation and interest rate variables for the stochastic scenarios, the difference in the annual averages were used as they could be either a decrease or increase for the inflation and interest rate under study. Numerous economic scenarios were generated, 5th, 25th, 50th, 75th and 95th percentiles of probabilities associated with the values were obtained from the cost. The study revealed at age 59, the cost under the total unit cost of allocation method had a 0.05 probability of been less than 1.694 and a 0.95 probability that the cost would be lesser than 1.859 and under projected unit cost of allocation method, the cost had a 0.05 probability of been less than 37.284 while 0.95 probability of the cost been less than 45.408 at age 59.

### *Statistics Skills Development Through Collaborative Projects at UKZN*

*Delia North, Temesgen Zewotir*
*University of KwaZulu-Natal*

The ability to analyse data in real time and do predictive modelling enables faster, more informed decision-making. In-house data analytics skills are consequentially in high demand in the business world. Universities are challenged to feed this demand and

also to bring their curriculums in line with industry needs, which are continuously evolving. This talk highlights the UKZN experience in aiming to graduate more job ready statisticians.

### *Comparative Assessment of Machine Learning Models in the Ranking of Child Mortality Data*

*Samuel Oduse*
*University of KwaZulu-Natal*
*Zewotir, T, North, D (Department of Statistics, University of KwaZulu-Natal)*
*babasegy@gmail.com*

Many of the studies on estimation of risk factors uses conventional statistical methods such as linear regression. However, in real life many variables are not linear, and the variance distributions are non-constant. Therefore, these linear models leave a good portion of the variation unexplained. In recent years, the introduction of more sophisticated machine learning techniques offered possibilities to overcome these constraints. However, these techniques also have drawbacks, and some have been labeled "blackbox," because they cannot provide appropriate explanations how the prediction results are derived. This makes the ranking of the most important variables in the model a challenging task. Different technique uses different approach to estimate variable importance. However, the Random Forest technique come with a mechanism that gives variable importance. This can be a unifying point where other techniques can serve as surrogate models in obtaining the variable importance. The objective of this study is to compare the ranking of child mortality risk factors using four machine learning techniques namely; Random Forest, Artificial Neural Network, Logistic Regression and Support Vector Machine. A demographic health survey data with 247 232 birth records and 42 potential risk factors was used. Non-parametric methods were used to compare variable rankings by the techniques. Result shows all techniques produced similar rankings of child mortality risk factors.

### *Does Training in Statistics Make Me a Good Intuitive Statistician? Some Thoughts Based on "Thinking, Fast and Slow" by Daniel Kahneman*

*Jeanette Pauw*
*Nelson Mandela University*
*Prof Smit, CF*
*jeanette.pauw@mandela.ac.za*

In our everyday lives we are all faced with decisions based on our beliefs concerning the likelihood of uncertain events. Statements such as "I think that…", "it is likely that…", "there is not a big chance that…" reflect our beliefs. The relevant questions to which we give these answers are often complex and the data available to us is limited. How is it then possible that we almost always come up with an intuitive answer to these questions? When we present intuitive answers to difficult questions, we often substitute the original question with an easier alternative. We are very often insensitive to situations in which we heuristically employ these alternatives and therefore unaware of the resultant systematic errors caused by these heuristics. The troubling thought for a statisticians is whether we are also prone to these biases, or are we truly better intuitive statisticians than the layman at large? This talk is based on the book "Thinking, Fast and Slow" written by Daniel Kahneman. The book is a comprehensive account of the work on judgement and decision making that Kahneman over many years did in collaboration with Amos Tversky. Six years after the death of Tversky, Kahneman received a Nobel Prize for this work.

### *Using Jump Models for Fire Detection in NDVI Data*

*Etienne Pienaar*
*University of Cape Town*
*Melvin Varughese*
*etienne.pienaar@uct.ac.za*

Jump diffusion processes are class of continuous-time stochastic processes that often find application in financial contexts. Specifically, as a modelling tool, the class permits conducting useful inference on extreme events observed in financial time series. These models can also easily be adapted for application in other fields of science. For purposes of an Ecological application, we demonstrate how to detect jumps under a jump-diffusion model of an ecological time serioes and show how the analysis can be used as a technique for detecting fires from a series of normalized difference vegetation index (NDVI) observations retreived from satelite data.

### *Time Series Analysis of Two Different Sources of Financial Statistics Data: Data from Administrative Sources and Survey Data*

*Sagaren Pillay*
*Statistics South Africa*
*sagarenp@statssa.gov.za*

There is a growing trend in national statistical organisations to research the use of data from administrative records maintained by programme administration agencies and government institutions in order to save costs in statistical production and reduce response burden. These organisations are increasingly making use of administrative and other secondary data sources for the production of statistics. Linkages of administrative data sets with survey data promise a new frontier in production of official statistics. This study examines the potential for the use of administrative data for collecting and compiling financial statistics that are currently collected by means of a survey. The research presents an analysis of data on turnover and other key variables collected from the annual financial statistics survey and the administrative data obtained from the Companies and Intellectual Property Commission.

### *Change-Point Detection in Panel Data with Stationary Regressors*

*Charl Pretorius*
*North-West University*
*Hušková, M, Láf, A (Department of Probability and Mathematical Statistics, Charles University)*
*charl.pretorius@nwu.ac.za*

We consider a panel regression model with cross-sectional dimension N. The aim is to test, based on T observations, whether the intercept in the model remains unchanged throughout the observation period. The test procedure involves the use of a CUSUM-type statistic derived from a quasi-likelihood argument. We provide the limit null distribution of our test under strong mixing and stationarity conditions on the errors and regressors, and show that the results remain valid if the unknown scaling parameters appearing in the test are appropriately estimated. The results are extended to the case where errors are allowed to be dependent across panels via a common factor error structure. Theoretical results are supplemented by a simulation study that indicates that the test works well in the case of small to moderate sample sizes. An illustrative application of the procedure to US mutual fund data demonstrates the relevance of the proposed procedure in a financial setting.

### *Innovative Linkages Between Universities and Organizations*

*Jennifer Priestley*
*Kennesaw State University*

How many .edu addresses are in your inbox right now? As organizations pursue digital transformation strategies, challenges related to finding and retaining analytical talent, objectively assessing the relevance of new, and emerging technology and engaging in deep and meaningful innovation with eventual ayback are common to all sectors of the economy. Deep, collaborative partnerships with universities can help mitigate many of these challenges. This is all the more true because data science itself has given rise to a new “entrepreneurial university” paradigm. Dr. Priestley is an academic Associate Dean, who worked for organizations like Accenture and VISA EU, and now manages corporate partnerships with the likes of Blue Cross Blue Shield, Emerson, Equifax, and GE, as well as fire departments and law enforcement. She will discuss the ways that organizations should be thinking about working with universities, but typically don’t – including research, innovation, “externships,” training options, recruitment, and other strategic relationships. After this session, you will never look at universities the same way again.

### *Application of Discrete-Time Survival Analysis Models in Modelling Student Dropout: A Case of Engineering Students at Tshwane University of Technology, South Africa*

*Princess Ramokolo*
*University of Limpopo*
*Maposa, D, Lesaoana, M (Department of Statistics and Operations Research, University of Limpopo)*
*princymasondo@gmail.com*

This study uses discrete-time survival analysis techniques to analyse the timing of dropout of undergraduate engineering students at Tshwane University of Technology. A discrete-time single risk model of dropout is used to investigate both the time-invariant and time-varying factors associated with dropout. The time-varying effects of the time-invariant factors are also analysed. In addition to the single-risk model of dropout, a discrete-time competing risks model of dropout is also estimated in order to account for graduation as a competing event. The results of the two models are compared to determine the adequacy of the single risk specification. The two models are also extended to account for unobserved heterogeneity by including a frailty term.

### *Factors of the Term Structure of Sovereign Yield Spreads and the Effect on the Uncovered Interest Rate Parity Model for Exchange Rate Forecasting*

*Chun-Sung Huang*
*Department of Finance and Tax, University of Cape Town*
*Desigan Reddy (Department of Finance and Tax, University of Cape Town)*

Using a Principal Component Analysis (PCA) approach, we investigate the sovereign yield spread term structure of the BRICS economies against that of the U.S. We show that the term structure for these markets are primarily driven by three latent factors which can be classified as the spread level, slope and curvature factors. We further postulate that a country's yield curve contains valuable information about its future economic state and as such the PCA derived spread factors, which are based on the differences between sovereign yield curves, encapsulates material macro-economic information between the countries. In light of this, we show that augmenting the traditional Uncovered Interest Rate Parity model (UIRP) with these factors improves the model's accuracy when forecasting exchange rate movements.

### *Sub-Pixel Land Cover Classification in a Resource Constrained Environment: One Study Area, Three Algorithms and Seven Images - What can We Learn?*

*Michaela Ritchie*
*Next Generation Enterprises and Institutions, Council for Scientific and Industrial Research and Department of Statistics, Nelson Mandela University*
*Debba, P. (Smart Places, Council for Scientific and Industrial Research and School of Statistics and Actuarial Science, University of Witwatersrand), Luck-Vogel, M. (Smart Places, Council for Scientific and Industrial Research and Department of Geography and Environmental Studies, University of Stellenbosch), Goodall, V. (Department of Statistics and Centre for African Conservation Ecology, Zoology Department, Nelson Mandela University)*
*michaelabeckley@gmail.com*

Statistical approaches can help environmental managers in resource constrained environments to better manage the areas for which they are responsible by enabling improved decision making through the use of sub-pixel classification algorithms of easier, cheaper and more frequently available Sentinel-2 data. For such managers, it is important to know if the same method would yield the best classifications for all seasons and years for their study area. It is known from literature that for different studies, the best algorithm varies, however, this study considers if the same holds true for the same study area across different seasons and years. Seven Sentinel-2 images across two seasons and four years are classified using three algorithms which area both seasonally and own date trained. Results show that the wet season imagery classifications are more accurate than those from the dry season for both seasonal and own date classifications. It is also shown that seasonal classifications are more acceptable for the wet season classifications while for the dry season there are larger differences between the accuracies for seasonally and own date trained algorithms. Finally, it is determined that even for the same study area and season, the best algorithm varies for different images.

### *Joint Modelling of Anaemia and Malaria in Young Children using Data from Complex Survey Designs*

*Danielle Roberts*
*University of KwaZulu-Natal*
*Zewotir, T (School of Mathematics, Statistics and Computer Science, University of KwaZulu-Natal)*
*danjader@gmail.com*

Anaemia and malaria are major contributors of childhood morbidity and mortality, particularly in sub-Saharan Africa. The causes of anaemia in children are multifactorial and include malaria. In malaria endemic regions, malaria is one of the most common causes of childhood anaemia, however severe anaemia can exacerbate malaria in these areas. Young children are particularly vulnerable to malaria as they have not yet developed an immunity to the disease. This study aims to investigate the joint effects of socio-economic, demographic and environmental factors on anaemia and malaria in young children in Kenya, Malawi, Tanzania and Uganda. Jointly modeling correlated outcomes can improve the efficiency of parameter estimates compared to fitting separate models for each outcome, as joint models have better control over type I error rates in multiple tests. Jointly modeling anaemia and malaria also allows the correlation between the two outcomes to be investigated and accounted for. This study makes use of data collected from demographic and health surveys, which utilize complex survey designs. Such surveys consist of multistage cluster sampling designs where samples are not collected in proportion to the population. This requires the design of the study to be accounted for in the analysis in order to avoid overestimation of standard errors and thus misleading results.

### *Identifying Immunological Risk Factors for TB Progression using a Hidden Markov Model*

*Miguel Rodo*
*University of Cape Town*
*Scriba, T (Department of Pathology, University of Cape Town), Little, F (Department of Statistics, University of Cape Town),*
*Rozot, V (Department of Pathology, University of Cape Town)*
*rdxmig002@myuct.ac.za*

Tuberculosis (TB) is the world's leading cause of death due to infectious disease. Despite much research effort, a poor understanding of the immunological determinants of host control of Mycobacterium tuberculosis (M.tb) remains a significant hurdle in developing the improved drugs and vaccines required to curb the epidemic. To better understand successful host-M.tb interaction, we investigated the phenotypic and functional immune responses from a two-year observational study of 6363 adolescents. From this cohort we analysed 37 individuals who developed TB and 37 matched controls who did not. An individual once infected with M.tb would either have eliminated the bacterium, or else had some degree of control over it. Since immunological risk factors for TB may depend on the underlying state of infection, we use a hidden Markov model to model risk of TB. We hypothesise that the latent state is correlated with inflammation and may be identified using markers of inflammation. Immune responses influencing state transition rates would be risk factors for TB. We suggest that this accurately identifies immunological risk factors for progression to disease.

### *Fundamentals of Geostatistics*

*David Rose*
*University of the Witwatersrand*

One way of distinguishing Geostatistics from other branches of Spatial Statistics is through the mathematical characterisation of its spatial domain, $\mathcal{D}$. Let $\{Z(\boldsymbol{x}); \boldsymbol{x} \in \mathcal{D} \subseteq \mathbb{R}^d, d \in \mathbb{N}\}$, where $Z$ is a variable of interest, and $Z(\boldsymbol{x})$ is the corresponding random function of a $d$-dimensional vector, $\boldsymbol{x}$. The data and the techniques used to analyse this data are classified as geostatistical if $\mathcal{D}$ is continuous but not random. Real-world problems have provided momentum to the development of Geostatistics, though its origins are largely attributed to the Mining Industry. Danie Krige, the late South African mining engineer and professor at Wits University, applied Mathematical Statistics to address the challenge of ore resource-reserve evaluation. He collaborated with Georges Matheron who formalised Krige's work in the 1960s, and who established a Centre de Geostatistique in Fontainebleau, outside of Paris. While Matheron is acknowledged as the Father of Geostatistics, the linear interpolation method of kriging takes its name from Professor Krige in recognition of his contribution to the field. The presentation aims to provide an overview of Geostatistics, its history and personalities. Naturally, attention will also be given in the talk to more technical aspects of the geostatistical toolbox. Variants of kriging and simulations are underpinned inter alia by stationarity requirements and the modelling of variograms, not unlike techniques encountered in other areas of Statistics.

### *Machine Learning ALSI 40 Index Futures Option Prices.*

*Duncan Saffy*
*University of Cape Town*
*Gebbie, T (Department of Statistical Sciences, University of Cape Town)*
*sffdun001@myuct.ac.za*

Option pricing models are a popular research area in financial mathematics and are of practical significance as market makers want efficient and accurate pricing methods. Often these models are developed on numerous assumptions that do not hold in the real world. To avoid these assumption we implement a model free option pricing approach through the use of a neural network to price European options of the ASLI 40 Index futures. Where the only constraint or assumption we will enforce is a put-call parity as a on the neural network to ensure a no-arbitrage market. We implement this no-arbitrage constraint in three ways: Pricing calls only and using put-call parity to price puts. The next two methods involve learning two networks at the same time, one to price calls and the other pricing puts, where the one attempts to enforce put-call parity through the use of a soft constraint. The second method enforces put-call parity as a hard constraint through the modification of the bias in the output layer. We compare the three models against each other as well as against Black's Futures option pricing model. The two main comparison metrics used are pricing accuracy as well as hedging error.

### *Classification of Musical Instruments in Audio Samples*

*Trudie Sandrock*
*Stellenbosch University*
*trudies@sun.ac.za*

Music information retrieval (MIR) is primarily concerned with the reduction of music to a workable data format and then extracting meaningful information from the data. A very topical field of research in MIR is musical instrument recognition. The goal of (machine) musical instrument recognition is to automatically determine the instrument or instruments playing in a given audio signal. There can be one instrument playing at a time, in which case the problem is referred to as monophonic, or there can be two or more instruments playing simultaneously, in which case the problem is called polyphonic. While the monophonic case is to a large extent considered solved, this is mostly in a Western context; that is, for instruments used in Western music traditions. Furthermore, the polyphonic case presents significant challenges, even for Western instruments. I will discuss some of the challenges faced in performing automatic musical instrument recognition and present results from some preliminary work using multilabel classification techniques and feature selection to tackle these challenges.

### *Modelling Malaria Incidence in the Limpopo Province, South Africa: Comparison of Classical and Bayesian Methods of Estimation*

*Makwelantle Asnath Sehlabana*
*University of Limpopo*
*Daniel Maposa (Department of Statistics and Operations Research, University of Limpopo, South Africa),*
*Alexander Boating (Department of Statistics and Actuarial Science, University of Cape Coast, Ghana),*
*Christel Faes (Centre for Statistics and Department of Mathematics and Statistics, University of Hasselt, Belgium)*
*asnathmakwelantle@gmail.com*

Malaria has infected and killed millions of people in Africa, mainly in hot regions where temperatures during the day and night are usually high. In South Africa, Limpopo is the hottest province in the country and therefore prone to malaria incidence. Bayesian and classical methods of estimation have been applied and compared on the effect of climatic factors on malaria incidence. Credible and confidence intervals from a negative binomial model estimated via Bayesian estimation – Markov chain Monte-Carlo process and maximum likelihood respectively, were utilised in the comparison process. Overall assumptions underpinning each method were given. The Bayesian method appeared more robust than the classical method in analysing malaria incidence in the Limpopo province of South Africa. The classical method identified rainfall and temperature during the night to be the significant predictors of malaria incidence in Mopani, Vhembe and Waterberg districts of Limpopo province. However, the Bayesian method identified rainfall, normalised difference vegetation index, elevation, temperature during the day and temperature during the night to be significant predictors of malaria incidence in Mopani, Sekhukhune, Vhembe and Waterberg districts. Both methods also affirmed that Vhembe district is more susceptible to malaria incidence, followed by Mopani district. We recommend that the Department of Health and Malaria Control Programme allocate more resources for malaria control to Limpopo.

### *Bootstrap Based Test for Two Independent Time Series*

*Modisane Seitshiro*
*North-West University*
*modisane.seitshiro@nwu.ac.za*

This paper is concerned with the derivative of a new bootstrap based test for the equality of the means for two independent stationary time series. The required properties of the test include satisfactory probability of Type I errors and high power. It is shown how critical points for various sample sizes and significance levels can be obtained by applying the parametric bootstrap method. A limited Monte-Carlo simulation study is conducted to illustrate the validity of the bootstrap approximation for the exact critical values, by producing satisfactory probability of Type I errors. It also shows that the newly proposed test compare favourably with standard two sample test in the absence of serial correlation, but are more powerful than the well-known t-test if small and moderate correlation structures are present, for a wide range of parameter values.

### *Some Useful Real Life Data Applications of Diao et al. (2013)'s Accurate Confidence Intervals*

*Yegnanew A. Shiferaw*
*Department of Statistics, University of Johannesburg*
*yegnanews@uj.ac.za*

Diao et al. (2013) proposed corrected confidence intervals for small area parameters based on the empirical best linear unbiased predictor (EBLUP) under the Fay-Herriot model using the Taylor series expansion. Their confidence intervals are accurate to terms $O(m^{-2})$ under unequal sampling variances. They conducted an extensive simulation study to illustrate their findings and demonstrate the superiority of their method. However, they could not demonstrate their proposed method using real life data. Having this in mind, the main aim of this paper is to apply their method with two different applications. Application 1 uses the 2016 Community Survey (CS) and the 2011 Population Census data to estimate the percentage of food insecurity at the local municipality levels of South Africa. Application 2 uses the 2011 Household Consumption Expenditure Survey (HCES) and the 2007 Population Census data to estimate the percentage of food expenditure at the zones and regional towns of Ethiopia.

### *Understanding the Dynamic Dependence Between Oil, Mineral Commodities and USD-ZAR Exchange Rate: Evidence from South Africa*

*Yegnanew A. Shiferaw*
*Department of Statistics, University of Johannesburg*
*Quarshie, HD*
*yegnanews@uj.ac.za*

The multivariate generalized autoregressive score (GAS) with Student-t model along with the Neural Granger causality approach was applied to examine the dynamic relationships among changes in Brent oil price, mineral commodity prices, and USD-ZAR exchange rate. Firstly, Brent oil price Granger causes mineral commodity prices and the other way around too. Brent oil price Granger causes USD-ZAR exchange rate, but not the other way around. Mineral commodity prices Granger cause USD-ZAR exchange rate, but not the other way around. Secondly, there are significant correlations between Brent oil price, mineral commodity prices and USD-ZAR exchange rate over time. In general, all the dynamic correlations are positive, which indicates that an increase in volatility of global crude oil price may lead to an increase in volatility for mineral commodity prices and USD-ZAR exchange rate. Empirical results in this study may help investors in the mining sector to explore alternative investment sets and monitor their investment risks. This may also help policymakers to manage their decision-making strategies about the global crude oil prices and the USD-ZAR exchange rate.

### *Robust Prediction Interval Modelling of Hourly Electricity Demand Forecasts*

*Caston Sigauke*
*University of Venda*
*caston.sigauke@univen.ac.za*

Uncertainty modelling of hourly load forecasts is important to system operators who have schedule and dispatch electrical energy on an hourly basis. The paper discusses short-term hourly load forecasting using additive quantile regression models. A comparative analysis is done using generalised additive models. Variable selection is done using least absolute shrinkage and selection operator via hierarchical interactions. The forecasts from the developed models models are then combined using quantile regression averaging (QRA). A comparative analysis of the developed models shows that the QRA model has the smallest prediction interval normalised average width and prediction interval normalised average deviation. The modelling framework discussed in this paper has established that going beyond summary performance statistics in forecasting has merit as it gives more insight into the developed forecasting models.

### *A Dual-Stress log-Normal Accelerated Life Testing Model*

*Neill Smit*
*North-West University*
*Raubenheimer, L (School of Mathematical and Statistical Sciences, North-West University)*
*neillsmit1@gmail.com*

Obtaining sufficient failure data to quantify the life characteristics of high-reliability items may not be viable in terms of financial and time constraints. A possible solution in this case is the use of accelerated life tests (ALTs). In ALTs, items are tested in more severe than their normal use environments, in order to induce early failures. The accelerated failure data can then be

extrapolated to estimate the reliability of the items at normal operating conditions. A functional relationship, known as a time transformation function (TTF), is assumed between the parameters of the life distribution and the accelerated stressors. In this paper, a Bayesian approach to an ALT with two stress variables is presented. The log-normal distribution is used as the life distribution and the generalised Eyring model as the TTF. This model allows for the use of one thermal stressor and one non-thermal stressor. Various priors are imposed on the model parameters and these models are compared via a sensitivity analysis for a data set obtained from an electronics epoxy packaging ALT. Markov chain Monte Carlo (MCMC) methods are used to obtain posterior samples due to the mathematically intractable posteriors. The log-concavity of the full conditional posterior distributions are assessed in order to determine an appropriate MCMC sampling method.

### *Estimation of the Frequency-Size Power Law Slope Parameter without Knowledge of the Time-Varying Level of Completeness of the Dataset*

*Ansie Smit*
*University of Pretoria Natural Hazard Centre, University of Pretoria*
*Kijko, A (UP Natural Hazard Centre, University of Pretoria)*
*ansie.smit@up.ac.za*

In natural hazard assessments, power laws are often used to represent the relationship between the frequency and the event-sizes of natural hazards. One of the disadvantages of the frequency-size power law, it the dependence of the slope parameter on the applied level of completeness m_c of the dataset. Although several techniques have been developed for its assessment, the determination of m_c remains problematic. In an attempt to provide a simple estimate for the power law slope that is free from an assumed level of completeness, both the Method of Moments (MM) and the Maximum Likelihood Estimation (ML) are applied to the most commonly observed shape of the apparent event-size distributions. This form usually follows a straight line that curves gradually at the lower tail of the distribution. The methodology is further extended to take into account an event-size dataset that exhibits time-varying levels of completeness. The updated procedure is not restricted to any particular shape of the apparent frequency-magnitude distribution. Additionally, independent information can be incorporated using the Bayesian formalism. The inclusion of weaker events, and accounting for time variation in the data, can provide more reliable input parameters for natural hazard and risk assessments. The procedure is applied to the earthquake dataset for the Ceres–Tulbagh area, which is the most seismic active region of South Africa.

### *Open Set Recognition with the Generalised Pareto Distribution*

*Luca Steyn*
*Stellenbosch University*
*de Wet, T (Department of Statistics and Actuarial Science, Stellenbosch University)*
*lucasteyn@sun.ac.za*

Classification is a subfield of pattern recognition where a model is built to classify observations into two or more categories. However, in some cases complete knowledge of the set of class labels is incomplete. In other words, not all classes are known during training. Consequently, test observations from a new class not seen during training will be misclassified as one of the known classes. Open set recognition methods generalise classification algorithms to detect these new classes during testing. Similar to anomaly detection, unknown classes are considered anomalous with respect to the known classes. Therefore, open set recognition methods perform two tasks. The model must accurately classify the observations from known classes. This model is then extended to also detect observations that do not belong to any of the classes seen by the model. In this talk we propose a method utilising extreme value theory and deep learning to perform open set recognition. A similarity score based on a ratio of distances is defined. It is then shown that this score can be accurately modelled with an extreme value distribution. Consequently, the probability that an observation is from a new class is estimated with this distribution. The method is applied to various image recognition datasets to demonstrate the predictive power and interpretability of the proposed model.

### *Structural Equation Models (SEM): Size Matters, for Now*

*Carmen Stindt*
*Nelson Mandela University*
*Sharp, GD (Department of Statistics, Nelson Mandela University)*
*carmen.stindt@mandela.ac.za*

Studies that make use of surveys as the method of data collection often make use of multivariate statistical techniques such as structural equation modelling (SEM). These techniques often require larger sample sizes, while survey-based studies are often riddled with logistical difficulties and there is a struggle to obtain the minimum required sample sizes (Bentler, P.M. and Yuan,

K.H., 1999). This often leads to a disjoint between the statistical techniques used and the data obtained. In practice, it is common to obtain sample sizes of 150 or less, which can result in issues when interpreting model fit due to the nature of the current methods of assessing model fit within SEM. This study makes use of a Monte Carlo simulation study to investigate a sample-size adjusted goodness-of-fit index as a practical alternative to allow for appropriate assessment of model fit.

### *Principal Component Analysis Data Reduction Procedure for Body Shape Classification for South African Men*

*Busisiwe Tabo*
*University of South Africa*
*Njuho, P (Department of Statistics, University of South Africa), Pandarum, K (Department of Consumer Science, University of South Africa)*
*bucitabo@yahoo.com*

This study focuses on development of body shape classification for men in South Africa. A questionnaire was administered on a sample of 300 men from Johannesburg South Africa, and the body measurements were taken using a 3D full body scanner found at UNISA Florida campus. About 14 body volume measurements (BVI) were obtained from the scanner and 21 variables including BMI, weight and height were obtained through the questionnaire. The 20 variables were tested for effect on BVI and BMI together. Only 6 variables had effect on BVI and BMI. The new set of variables were then used for body shape analysis. PCA was employed for data reduction with Promax rotation. Four components were found, the first accounted for Torso shape, the second for Leg shape, third for Demography and the fourth for Height. The components are to be used for further analysis to obtain body shape analysis.

### *Ruin Probability in the Delayed Poisson Renewal Risk Model Perturbed by Diffusion Process*

*Essodina Takouda*
*School of Economics, University of Johannesburg*
*Franck Adekambi (School of Economics, University of Johannesburg)*
*etakouda85fr@gmail.com*

In this paper, we consider the risk model perturbed by an independent diffusion process with a time delay in the arrival of the first claim. We derive the distribution of the delayed renewal process, the intego-differential equations of the ruin probabilities and generalize its defective renewal equations. With claim amounts following Exponential and Mixed Exponential distributions, an explicit expressions and asymptotic properties of the ruin probabilities are derived. Numerical illustrations of the ruin probabilities are proposed when claim amounts are exponentially and mixed exponentially distributed. As further extension, we consider the case of the delayed renewal risk model with exchangeable risks and derive the ruin probabilities.

### *Statistical Accuracy of a Linear Object Extraction Algorithm for Greyscale Images*

*Renate Thiede*
*University of Pretoria*
*Fabris-Rotelli, IN (Department of Statistics, University of Pretoria), Stein, A (ITC, University of Twente, Netherlands),*
*Debba, P (Spatial Planning and Systems, CSIR Built Environment, Pretoria)*
*renate.thiede@gmail.com*

Informal unpaved roads in developing countries arise naturally through human movement and informal housing setups. These roads are not authorised nor maintained by council, nor recorded in official databases nor online maps. Mapping such roads from satellite images is a common problem, as information on these roads is critical for sustainable city growth. Information on their location and extent may be gleaned from spatial big data. Attempts to do so are sparse, and no automatic or semi-automatic approach is freely available. This research develops a novel algorithm for extracting informal roads from multispectral satellite images, using physical road characteristics. These include near-infrared reflectance, addressed via the NDVI index, shape, addressed via the shape measures compactness and elongation, and grey-value intensity. The crux of the algorithm is the Discrete Pulse Transform, implemented via the Roadmaker's Pavage. The algorithm provides a classification of road objects, along with an associated uncertainty measure per road object. Accuracy is assessed using per-pixel assessment metrics and metrics based on road characteristics, including completeness, correctness, and Pratt's Figure of Merit, which is applied to road extraction accuracy for the first time. The algorithm is applied to areas in Gauteng and North West Provinces, South Africa. Sources of uncertainty and error are discussed, such as indefinite boundaries, surface type heterogeneity, trees and shadows

### *Tree-Based Ensemble Methods for Classification*

*Daniel Uys*
*Stellenbosch University*
*dwu@sun.ac.za*

Ensemble methods combine a large number of simpler base learners to form a collective model that can be used for classification. Learning methods such as bagging and random forests can be regarded as tree-based ensemble methods. In these methods, the standard approach is to express the model as a linear combination of the base learners where the coefficients, associated with the base learners, are all equal, i.e., the base learners are equally weighted. An alternative approach would be to assign to those base learners that are considered important, larger weights. Within a regression context, this has been done by estimating the coefficients of the base learners using least squares. Since a large number of base learners is involved, the residual sum of squares of the linear combination of base learners has to be penalised by, for example, the lasso penalty. However, the large number of base learners also complicates the minimisation of the coefficients in the penalised residual sum of squares criterion. By using the iterative forward stagewise linear regression algorithm for ensemble methods, estimators of the coefficients of the base learners can be obtained. In this talk, the principles of the forward stagewise linear regression algorithm are considered within a classification context by assigning different weights to different classification base learners. Various tree-based ensemble methods will be evaluated by applying the weighted classification techniques to simulated, as well as to real life data sets.

### *Longitudinal Analysis of Brain Metabolite Levels for HIV Infected Children from Ages Five to Eleven Children using Multivariate Approaches*

*Noëlle Van Biljon*
*University of Cape Town*
*Little, F (Department of Statistical Sciences, University of Cape Town)*
*anbnoe001@myuct.ac.za*

Under new treatment guidelines, HIV infected children initiate antiretroviral therapy (ART) early in life and remain on it lifelong. However, the long-term impact of ART and HIV on the maturing brain is not well documented and longitudinal neuroimaging studies are rare, especially in developing countries most heavily impacted by HIV where access to imaging resources are limited. As part of a longitudinal study, we aim to examine HIV related changes in metabolite level trajectories from 5-11 years in three brain regions using Magnetic Resonance Spectroscopy. We are focusing on the concentration of seven metabolites which are measures of cellular activity and health. Hence, this technique allows for a non-invasive investigation of brain health. Univariate mixed effect modelling approaches may be used to analyse this data, however this involves the creation and interpretation of 24 separate models that will not give any insight to the interactions between metabolite concentrations and regions of the brain. As we know the body is a connected system and we cannot assume these factors are independent. We have used various multivariate longitudinal approaches to identify the longitudinal profiles of these metabolites in different regions and to combine this information to see if these changes in metabolites associated with HIV are related. The preliminary models show that there is a significant region-metabolite interaction, confirming the need for multivariate modelling approaches.

### *Classifying Yield Spread Movements Through Triplots: A South African Application*

*Carel Van Der Merwe*
*Stellenbosch University*
*De Wet, T (Department of Statistics and Actuarial Science, Stellenbosch University)*
*cjvdmerwe@sun.ac.za*

Significant movements in yield spreads from a sparse data environment are classified using various share, interest rate, financial ratio, and economic type covariates in a visually interpretive manner. This allows for a better understanding of how various factors drive the changes in yield spreads. Additionally, this visualisation technique provides the ability to classify whether an unlisted debt instrument's yield spread had significantly changed or stayed stable during a specific observation period. The analysis was implemented in a web-based application as well.

### *The Power of Markdown for Teaching, Research, and Consultation*

*Sean Van Der Merwe*
*University of the Free State*
*vandermerwes@ufs.ac.za*

Are you tired of marking copied assignments? Do you wish you could remember what the code you wrote last year does and why you typed it that way? Are you repeatedly copy-pasting graphs into documents? Wouldn't it be nice if you could refit a complex model or redraw a complicated graph with just one click? Well I have the solution to all your problems and I'll give it you for free!* *Obviously not solutions **to ALL your problems.** Solutions may include new problems.

### *Estimation of the Degrees of Freedom for the Student t-Distribution using a Bayesian Procedure*

*Abrie Van Der Merwe*
*University of the Free State*
*Groenewald, PCN, Voges, JL (Department of Mathematical Statistics and Actuarial Science, University of the Free State)*
*matheeme@ufs.ac.za*

In most applied as well as theoretical research, the residual terms in linear models are assumed to be normally and independently distributed. However, such assumptions may not be appropriate in many practical situations. Many economic and business data, for example stock return data, exhibit heavy (fat) tail distributions and cannot be effectively modelled by the normal distribution. The use of the student-t distribution reduces the influence of outliers and thus makes the statistical analysis more robust. Inference about the number of degrees of freedom in the case of the t-distribution, represents an important task, as the degrees of freedom parameter governs the heaviness of the tails. The Bayesian framework requires a prior distribution to be assigned for the parameter, representing the initial uncertainty about its true value. In this talk, reference and probability-matching priors will be derived for the unknown parameters and will be compared with other objective priors. A real example on the daily log-returns of the IBM Company for the period 1989 to 1998 will also be considered.

### *Regressions in the Sandbox: Projection Pursuit Regression in Comparison to Other Regression Techniques in Digital Soil Mapping*

*Stephan Van Der Westhuizen*
*Stellenbosch University*
*Hofmeyr, D (Department of Statistics and Actuarial Sciences, Stellenbosch University),*
*Heuvelink, GBM (Department of Environmental Sciences, Wageningen University)*
*stephanvdw89@gmail.com*

Regression is a class of statistical techniques which models the relationship between a numeric output and several predictors or covariates. In digital soil mapping (DSM) regression-kriging is a commonly used technique that also addresses spatial correlation and combines a multiple linear regression model of the response on the covariates with kriging of the prediction residuals. However, there are situations when linear regression and regression-kriging are sub-optimal. Such situations include the absence of normality of the residuals, a lack of linearity between the response and the covariates, heteroscedasticity of the residuals or severe multicollinearity. Friedman and Stuetzle (1981) introduced projection pursuit regression (PPR), a regression technique which is not limited by such stringent assumptions. PPR projects the data matrix of covariates in an optimal direction before applying smoothing functions to the projected data. Even though the theory of PPR is well developed, not much research has gone into using it in DSM. We present a brief overview of PPR and its application and performance in DSM. We will compare PPR to regression-kriging and to other popular regression techniques in DSM such as random forests. A case study will be used to compare the regression techniques.

### *Experimental Design Criteria and Optimization for Chemical Process Modelling*

*Willem Adriaan Van Deventer*
*University of the Free-State*
*R.L.J. Coetzer (University of the Free-State, South Africa)*
*2014217968@ufs4life.ac.za*

Modelling a chemical process involving non-ideal and/or reactive mixtures which are influenced by process-, mixture- and qualitative variables, is a complex process. Prediction errors associated with any of the interacting models can potentially be amplified to the point that predictions made by the overall model are not plausible at all, only acceptable in a very small experimental region, or no predictions can be made due to simulator convergence issues. For a model of a chemical system to be successful, experimental designs should be generated for any number and combination of variables of different types, inside arbitrary, potentially dynamic, mutually exclusive, multimodal, nonlinear, feasible regions in addition to regions described by a convex hull of available measurements. Multi-phase chemical systems require experimental designs consisting of multiple simultaneous process-mixture designs which simultaneously apply. In this paper we will present an experimental design methodology and new design criteria for addressing many different types of constraints of various complexities in the modelling of chemical processes. The experimental designs generated were found to have similar, and in some cases higher efficiencies compared to designs generated with commercial experimental design software. The methodology developed can also be applied to optimization problems in general.

### *An Efficient Kernel Quantile Estimator*

*Francois Van Graan*
*North West University*

For an estimator of quantiles , the efficiency of the sample quantile can be improved by considering linear combinations of order statistics. A number of such estimators are available in the literature. A different route to estimate a population quantile is to propose an estimator for the underlying distribution function. In this presentation, a smooth nonparametric kernel distribution function estimator will be proposed to estimate the population quantile function. The proposed estimator will be compared to existing estimators in a small scale simulation study.

### *Points and Rating Systems in Professional Tennis*

*Paul Van Staden*
*Department of Statistics, University of Pretoria*
*Sajiwan, M (Department of Statistics, University of Pretoria)*
*paul.vanstaden@up.ac.za*

The Association of Tennis Professionals (ATP) for men's tennis and the Women's Tennis Association (WTA) use rolling 52-week points-accumulation rating systems. These systems rank tennis players according to points awarded for their performances in tournaments without taking strength of opponents or score differences into account. This paper develops a points-exchange rating system for tennis with the rating points of the competing players adjusted by equal, opposite amounts after each tennis match based on the margin of victory in that match. Since margin of victory cannot be directly calculated from traditional tennis scores, we also present a points system for tennis matches in which tennis scores are converted into performance points. The proposed points and rating systems are illustrated with the 2019 Wimbledon Championships.

### *A Different Approach for Choosing a Threshold in POT*

*Andrehette Verster*
*University of the Free State*

The peaks over threshold approach, where appropriate Extreme Value models are fitted to either absolute or relative excesses, has become very popular over recent decades. Applications of these can be found in Finance, Hydrology, Seismology, etc. One of the challenges when applying the peaks over threshold approach is the choice of threshold. Various difficulties arise in choosing the threshold either too low or too high, including biased estimates and uncontrolled variances. In this study a generalization of the Topp-Leone (GTL) distribution is considered to assist in the choice of threshold when the extreme value index is positive. The GTL Pareto distribution can be expressed as a Pareto type distribution with a slowly varying part and a generalization parameter that tends to 1 as n tends to infinity. A Bayesian approach is considered in this study for parameter estimation. It can also be shown, through simulation studies, that the GTL Pareto is less sensitive to the choice of threshold.

### *New Tests for Exponentiality Based on the Interarrival Times of the Poisson Process*

*Jaco Visagie*
*Department of Statistics, North-West University*
*Oosthuisen, AB (Department of Statistics, North-West University) Pretorius, C (Department of Statistics, University of Pretoria)*
*jaco.visagie@nwu.ac.za*

In this talk, we propose new classes of goodness-of-fit tests for the exponential distribution. Given a dataset containing only positive observations, we construct a counting process by sampling interarrival times from the observed data and test whether or not this process is Poisson by considering the distribution of the increments. This is achieved by calculating various distance measures between the empirical mass function of the sample increments and the corresponding Poisson mass function. A Monte Carlo study demonstrates that the finite sample power performance of the newly proposed tests is competitive when compared to existing tests for exponentiality.

### *A Deep Learning Framework for Individual Clanwilliam Cedar Tree-Crown Detection using High Resolution Aerial Imagery*

*Lionel Yelibi*
*Department of Statistical Sciences, University of Cape Town*
*Britz, S (Department of Statistical Sciences, University of Cape Town),*
*Moncrieff, G, Slingsby, J (Fynbos Node, South African Environmental Observation Network)*
*wzvarevashe@gmail.com*

The critically endangered Clanwilliam cedar, Widdringtonia wallichii, is an iconic tree species endemic to the Cederberg mountains in the Fynbos Biome. Consistent declines in its populations have been noted across its range primarily due to the impact of fire and climate change. Mapping the occurrences of this species over its range is key to the monitoring of surviving individuals and is important for the management of biodiversity in the region. Recent efforts have focused on the use of freely available Google EarthTM imagery to manually map the species across its global native distribution. This study advances this work by developing a workflow to automate the process of tree detection using a deep-learning based approach. The approach involves using sets of high-resolution red, green, blue (RGB) imagery to train artificial neural networks for the task of tree-crown detection. Additional models are trained on colour-infrared (CIR) imagery, since live vegetation has a red tone on the near-infrared (NIR) spectrum. Preliminary results show that using an intersection-over-union threshold of 0.5 yields an average tree-crown recall of 0.67 with a precision of 0.53, and that the addition of the NIR spectral band does not result in improved performance. The feasibility of using this approach to regularly update maps of the Clanwilliam Cedar and monitor population trends in the Cederberg is assessed.

### *Statistical Modelling of Decomposed Rainfall Time Series and Generalised Extreme Value Distribution*

*Willard Zvarevashe*
*University of Zululand*
*Krishnannair, S (Department of Mathematical Science, University of Zululand)*
*wzvarevashe@gmail.com*

The extreme rainfall patterns have a direct and indirect effect on all earth spheres particularly the hydrosphere, biosphere and lithosphere. Therefore, an understanding of the extreme rainfall patterns is very important for future planning and management. In this study, the rainfall time series is decomposed into intrinsic mode functions (IMFs) using a data-adaptive method, ensemble empirical mode decomposition. The IMFs are modelled using generalised extreme value distribution (GEVD). The model diagnosis and selection using QQ-Plot, PP-Plot and Akaike information criterion show that the decomposed IMFs have better models than the original rainfall data. Rainfall modelling using decomposed data may assist in future planning and further research by providing better predictions.

# ABSTRACTS: Poster Presentations

### *Wind Speed and Power Modelling using Mixture of Life Distributions*

*Deidre Bredenkamp*
*University of Pretoria*
*Naderi, M, Bekker, A*

Accurate modelling of wind speed and power densities is fundamental in the estimation of local wind turbine power. The Weibull distribution is the most commonly used distribution to model wind speed and power distributions. However, the used Weibull distribution may lead to unreliable results when it comes to modelling wind speed and power distributions, especially for cases with peaking wind speeds or data that contains outliers. This leads to an underestimation of the power produced by the wind. Therefore, alternative approaches are required to model wind speed and power. Contamination is a promising approach in order to model data containing outliers. Data from a certain distribution that is contaminated by “bad-points” or outliers from another distribution can be modelled by using contaminated distributions. In addition to the parameters of the fitted distribution, our contaminated distributions have for each cluster (outlier or not), a proportion controlling the portion of bad points and one specifying the level of contamination. Essentially these parameters do not have to be specified a priori, resulting in a more flexible model. In addition, each observation is given a posterior probability of belonging to a certain cluster, this allows for automatic outlier detection. In this paper, we will focus on contaminated lifetime distributions such as the Beta-Weibull (BW) and the Weibull (W) distribution. These distributions were evaluated against four other well-known distributions. Data from locations in Canada was observed. In order to demonstrate the effectiveness of the model, the maximum likelihood (ML) was used for parameter estimation of various distributions. Different criteria and goodness-of-fit tests were implemented to compare the suitability of the tested distributions. It was found that the contaminated Beta-Weibull distribution provided the best fit to model the wind speed distribution for all the observed locations. However, the Weibull distribution still proved to be the best model to use to model wind power density.

### *Using Circular Statistics to Analyse Aoristic Data*

*Chiara Fazzini*
*University of Pretoria*
*Nagar, P, Bekker, A*

Aoristic data encompasses data relating to crime against property like theft and burglaries. This data often provides information on the range or time interval in which some event could have occurred, but it usually lacks the exact time of occurrence. For example, an individual may leave their car unattended at a certain time and return later to find it stolen. The time interval in which the car was left unattended is known, but the exact time it was stolen remains unknown. This information is of great interest to researchers and law enforcement. There has been minimal research done on determining the exact time of events, but some methods proposed include known-time, start- and end-time, midpoint, random, aoristic and frequency domain time series methods. In this research report, circular statistics is used to estimate the most likely time of crime instances by assuming that the twelve months of the year follow a cyclic pattern. More specifically, the aoristic fraction method, combined with circular statistics, is used to obtain an aoristic distribution of bicycle thefts in the City of York, UK. There is limited research on the two techniques of circular statistics and aoristic analysis being used in conjunction with each other. This approach of circular statistics will enable researchers to analyse aoristic data more accurately through better analysis of the variations and distributions of temporal crime cycles.

### *Text Content Classification on News Articles*

*Fabio Fehr*
*University of Cape Town*
*Soutar, S*

The increased volume of electronic journalism and other text media has driven the processing of the natural language to gather meaning from large bodies of text. Text classification is the process of classifying a body of text into a predefined set of classes. This process can be computationally expensive. This research provides a case study of classifying online News24 news articles into a variety of single label classes. Machine learning techniques are used to classify the news articles with the objective to balance classification accuracy and computational efficiency. The text vectorization techniques Term Frequency Inverse

Document Frequency and Naive Bayes are contrasted while the classifier models: Naive Bayes classifier, k-Nearest Neighbours, Support Vector Machines and decision tree forests are compared by computation time and accuracy. It was found that the Naive Bayes vectorisation method produced models that were the most accurate and computationally efficient. The k-NN, SVM and decision tree forest models trained on Naive Bayes vectorised data were deemed to be good choices for text classification.

### *Clustering: An Introduction*

*Michelle Gilfillan*
*University of Pretoria*
*Millard, SM, Kanfer, FHJ*

What is clustering? Clustering forms part of unsupervised statistical learning where unlabelled observations are grouped according to similarity measures to identify the true structure of a data set. How was this vast field studied? Three clustering algorithms were studied from first principles, then the identification of the optimal number of clusters, followed by the application and evaluation of these clustering algorithms. Before applying an algorithm to predict the three true clusters of this data, an assumption must be made about the nature of the clusters.

### *Changes in Rainfall Seasonality in the Western Cape*

*Peter Ivey*
*University of Cape Town*
*Erni, B (Department of Statistics, University of Cape Town)*

Not much is known about the rainfall patterns within the last 100 years in the Western Cape. The more knowledge there is about how rainfall patterns have shifted over time, the better the region can be prepared for periods of flooding and drought. More optimal decisions can be made regarding agriculture if there is a better understanding of rainfall patterns. 100 years worth of data has been acquired from 30 weather stations within in the Western Cape. The Western Cape has three known different climate regions – the Mediterranean (winter rainfall), the Karoo (summer rainfall) and the South Coast (summer rainfall). The weather stations are grouped into five clusters using K-means clustering based on their average monthly rainfall throughout the year. Start and end of season days are computed for the 100 years using thresholds, criteria and Generalized Additive Models. These days are then compared over time to extract any patterns. Linear regression models and GAMs are fit to the data to pick up on any trend over the 100 years. Conflicting results are produced for individual clusters when using different methods. Some suggest a linear increase in start/end of season whilst others suggest the opposite. Many clusters appear to have some underlying cyclical nature.

### *Bayesian Estimation for the Ratio of Two Exponential Parameters*

*Enrike Le Roux*
*North-West University*
*Raubenheimer, L (School of Mathematical and Statistical Sciences, North-West University)*

In this paper the maximal data information prior and the probability matching prior for the ratio of two exponential parameters will be derived. The method by Datta & Ghosh (1995) will be used to derive the probability matching prior and the method proposed by Zellner (1971) will be used to derive the maximal data information prior. Simulation studies will be done to compare and evaluate the performance of the following five priors: the Jeffreys, uniform, probability matching, maximal data information priors and a prior suggested by Ghosh, Mergel and Liu (2011). We will investigate the performance of the credibility intervals for the ratio of two exponential parameters. These intervals will be compared with each other in terms of coverage rates and average interval lengths. It seems that if inference is made on the ratio of two exponential parameters, the Jeffreys prior performs better in terms of coverage rates, but the maximal data information prior performs better in terms of average interval lengths. References Datta, G. S. and Ghosh, J. K. (1995). On priors providing frequentist validity for Bayesian inference. Biometrika, 82(1), 37 – 45. Ghosh, M., Mergel, V. and Liu, R. (2011). A general divergence criterion for prior selection. Annals of the Institute of Statistical Mathematics, 63(1), 43–58. Zellner, A. (1971). An Introduction to Bayesian inference in Econometrics. New York: Wiley, 1st edition.

### *Estimating Survival of an Enigmatic Frog from Capture-Mark-Recapture Data*

*Tessa Lloyd*
*University of Cape Town*
*Njati, J (Department of Statistics, University of Cape Town), Altwegg, R (Department of Statistics, University of Cape Town)*

Capensibufo rosei (C. rosei) or Rose's mountain toad, is a small species of frog, in both size and population, that is uniquely found in two known locations as of date, with three populations, occupying sites within South Africa's Cape Peninsula. Due to the small number of populations, C. rosei has been listed by the International Union for Conservation of Nature as an endangered species. Using Capture-Mark-Recapture (CMR) data from two of these populations provided by the South African National Biodiversity Institute, this study endeavours to estimate the survival of the toad. Two models specific to CMR data, namely the Cormack-Jolly-Seber model implemented in the program MARK and JAGS through R, and the Jolly- Seber only implemented in MARK, population dynamic parameters that provide a wealth of information for conservation purposes were estimated. This report aimed to test results found by Becker and his colleagues from the recently published article. Becker hypothesized that the survival of C. rosei was inversely correlated to rainfall, therefore the severe drought experienced by the city of Cape Town over the past three years provided a unique opportunity to test this hypothesis. Contrary to the hypothesis, this study found that as rainfall reduced, survival rates dropped. We concluded that, as an amphibian, C. rosei relies heavily on water, especially the breeding puddles formed from winter rainfall, for reproduction and metamorphosis.

### *Determinants of Under-Five Mortality in South Africa using Poisson Regression Model*

*Kgethego Sharina Makgolane*
*University of Limpopo*
*Moloi, KD, Tessera, A (Department of Statistics and Operations Research, University of Limpopo)*

The under-five mortality rate (U5MR) is the probability of dying between birth and exactly five years of age expressed per 1000 live births. The rate is a leading indicator of the child's health and survival, it reflects the development of the country's economy. Sub Saharan countries including South Africa, are still leading in the U5MRs with one in every nine children dying before the age of five years. However, South Africa has made significant progress towards the reduction of U5MR but has not achieved its Millennium Development Goal 4 in 2015. To ensure a significant decline in U5MR the focus should be on the associated factors. Hence, this study attempts to apply Poisson regression model to identify the determinants of under-five mortality utilising the South African Demographic and Health Survey 2016 children data. Descriptive analysis and Poisson regression model were used for data analysis. Descriptive analysis has shown that 11.27% mothers has experienced at least one under-five death. Based on the results of Poison regression model, duration of breastfeeding, source of water, region, maternal occupation, maternal age group, paternal age and paternal education were found to be significantly associated with U5M in South Africa. Further, there was evidence of over-dispersion which was shown by the value of Pearson Chi-square. Therefore, other modelling techniques such as negative binomial regression model should be considered to address the issue of over-dispersion.

### *The Use of Repeat LIDAR Flights to Model Dominant Height*

*Mzamo Makulube*
*Mondi South Africa (Pty) Ltd*
*Chauke, M, Kotze, H (Mondi South Africa, Growth and Yield Research Department, Pietermaritzburg)*

In forestry, future yields are predicted using stand growth models. One of the main components of the stand growth model is the dominant height, which is defined as the average height associated with the 20% thickest trees in a sample. The dominant height over time is often modelled using Chapman Richards and Hossfeld models in guide curve form. However, the use of the guide curve approach makes it difficult to project and to calibrate the height growth. Alternatively, a guide curve can be transformed into a dynamic/difference form which can be used to calibrate and project dominant height growth. The guide curve and the difference form models are fitted simultaneously in SAS to estimate the set of parameters. This kind of model fitting requires interval data which can be obtained from permanent sample plots and long term trials. Moreover, data from repeat LIDAR flights can be used. The objective of this exercise is fit the difference form model using data from repeat LIDAR flights. The results indicate that this fitted model is more flexible and accounts for all age classes.

### *Gaussian Process Regression and Classification: Elliptical Slice Sampling*

*Ricardo Marques*
*University of Pretoria*
*de Waal*

The use of Gaussian processes to perform classification and regression is considered, however Bayesian inference can become difficult to perform, as Bayes theorem requires a closed form solution of an integral which is often intractable. This issue is addressed through the use of the probabilistic approach in sampling, particularly Markov chain Monte Carlo inference. Within Markov chain Monte Carlo inference, an explanation as to why this is a suitable approach, such that the difficulties arising from Bayes theorem and it's implementation are avoided. In addition to this, a comparison of the degree to which several Markov chain Monte Carlo techniques address this obstacle is provided.

### *Mixtures of Regression: Expectation Maximisation Type Algorithms*

*Thanyani Mashau*
*University of Pretoria*
*Millard, SM, Kanfer, FHJ*

As mixture models gain popularity in modelling heterogeneous data, different methods for fitting these models are studied. The expectation maximisation (EM) algorithm is a procedure for finding maximum likelihood estimates for mixtures of regression models. Since the formulation of the EM algorithm, many adaptions have been considered such as the classification EM (CEM) and stochastic EM (SEM) algorithms. This study comprises of a comparison of the EM, CEM and SEM algorithms based on observed data.

### *Kalman Filtering of the Generalized Vasicek Term Structure Models with Infinite Maturity*

*Romeo Mawonike*
*Department of Mathematics and Computer Science, Great Zimbabwe University*
*Ikpe, D (Department of Applied Mathematics, University of South Africa)*

In this paper, we give a brief introduction to the Kalman filter, and the generalized Vasicek models of the term structure of interest rates with special focus to the application of the Kalman filter equations to estimate one-and two-factor models. The model is expressed in a state space form and the Kalman filter is then used to estimate the unobserved state variables and the parameters of the model. The state space formulation has the advantages of taking into account both the cross-sectional and time series restrictions on the data and measurement errors in the observed yield curve. We finally derive the yield on a zero coupon bond with finite and infinite maturities and the Kalman filter equations of the state space formulation of the generalized Vasicek models. We perform simulations and illustrate how the Kalman filter works and the major weakness of the Vasicek model.

### *Latent Variable Models for Longitudinal Outcomes from a Parenting Intervention Study*

*Carlyle McCready*
*University of Cape Town*
*Little, F (Department of Statistical Sciences, University of Cape Town)*

The Sinovuyo Caring Families Programme (SCFP) is a low cost intervention program for primary caregivers of 2- to 9-year-old children to reduce harsh parenting practices and child behavioral problems in high-risk South African families. Child and parent behavior were assessed using likert scale instruments that together measure underlying behavioral traits, reported child behavior problems, and reported positive parenting, reported harsh parenting. To make recommendations on the success or failure of the SCFP program, we examine the use of Structural Equation modelling for longitudinal profiles. We create second order latent class growth models which impose a spline component on the time trend and compare the intervention group to a control group. This study examined the measurements taken at baseline, directly after the intervention period, and at a 12-month follow-up. The model estimates show improved behavior in terms of reported child behavior problems and reported harsh parenting with no differences between intervention and control groups. It was concluded for both reported child behavior problems and reported harsh parenting that subjects with higher initial scores decrease at a faster rate. The model estimates show improved behavior in terms of reported positive parenting with differences between intervention and control groups. Further tests indicated that the success of the intervention program is, in some cases, dependent on the community in which it was implemented.

### *A Comparison of Modern Dimensionality Reduction Techniques Through the Classification of Extragalactic Objects*

*Tristan Naidoo*
*University of Cape Town*
*Mayet, S, Arendse, J*

The Sloan Digital Sky Survey (SDSS) can be thought of as one of the biggest and most successful sky surveys in the history of astronomy. The unprecedented depth and complexity of the resulting dataset makes it an excellent case study for the statistical needs of modern day astronomy. Classification of extragalactic objects is one of the most relevant of these applications, but typically can only be conducted after the dimensionality of the data is significantly reduced. We applied a number of modern dimensionality reduction techniques to the SDSS data, namely independent component analysis, Isomaps, local linear embedding, and t-Distributed Stochastic Neighbour Embedding, as well as the more common and traditional principal component analysis. The reduced components from each of these techniques were used for classification of objects into galaxy, star, and quasar categories. The classification performance was used to compare the quality of the different dimensionality reduction techniques. PCA and t-SNE emerged as the most promising techniques. When compared to classification accuracy using all of the predictors, the former could produce similar accuracy using 2.8% of the original dimensionality, while the latter improved classification accuracy using only 0.03% of the original dimensionality.

### *A Bayesian Mixture Modelling for Zero-Inflated Multivariate Data*

*Georgeleen Osuji*
*University of Fort Hare*
*Mutambayi, R, Ndege, J, Qin, Y, Azeez, A (Department of Statistics, University of Fort Hare)*

Ordinal responses from multivariate data are frequent in many aspects of biomedical studies and public health research. If the multivariate features in the ordinal data are ignored by not taking into account the correlated errors, this may lead to substantial biased estimates and inference. However, such ordinal outcome may often demonstrate a high proportion of zero inflation at the tail end of the ordinal scale data. Thus, this zero-inflated multivariate data will be difficult to analyze. Several methods have been developed to address these challenges in zero-inflated data. However, in this study, we suggest a Bayesian mixture multivariate zero-inflated ordinal model to reduce and relax the estimate bias and inference. A Bayesian Markov Chain Monte Carlo (MCMC) approach was used to estimate the parameter. We conducted a simulation study to compare the performance of the proposed model with other two existing models. The simulation results show that the proposed model performed better than the other two models in terms of bias reduction and higher accuracy from Root Mean Square Error (RMSE).

### *Spatial Modelling of the Association Between Crime and Weather*

*Arminn Potgieter*
*University of Pretoria*
*Fabris-Rotelli, I, Breetzke, G, Wright, C*

In a recent paper Schinasi et al. utilized distributed lag non-linear modelling to model relationships between weather conditions and daily counts for different types of crime, such as robbery, drunk driving and disorderly conduct, that are non-linear in both the space of the predictor as well as the lag-space. In their analysis specific attention was paid to the daily mean heat index, a measure that uses temperature and relative humidity to produce an indication of how individuals truly experience the temperature in their environment. Their findings suggest that higher temperatures, especially during colder months, are associated with increased rates of violent crime. This is in line with Temperature Aggression theory which states that higher temperatures are associated with increased levels of discomfort and aggression in individuals, which in turn increases the likelihood of them seeking out violent confrontations. Such theories, which suggest that there exists a correlation between weather conditions and the propensity for violent crime, have been proposed numerous times by various scholars and philosophers throughout human history. The relationship between crime and weather has been investigated several times across the world, however most of the existing theory and application was developed in countries situated within the Northern Hemisphere. This study aims to expand the currently limited understanding of this relationship in a unique South African setting by investigating the relationship between crime and weather over a 10-year period in Khayelitsha, Cape Town using the same techniques employed by Schinasi et al. and supplemented with the inclusion of elements of spatial analysis. This is the first application of this methodology on data of this particular nature and origin. Distributed lag non-linear models are frequently used in epidemiological investigations of the effect of temperature variables on human health and mortality. This modelling framework allows the user to specify non-linear functions to model the relationship in the predictor- and lag-space and is a generalization of distributed lag modelling. In this paper we used distributed lag non-linear models to investigate the relationship between weather variables such as temperature, relative humidity and rainfall with crime counts for different categories of crime and using different intervals of time. Next we calculated empirical estimates for various distance-based functions such as Ripley's K function to determine whether the spatial distribution of crime events exhibits evidence of clustering across different years. These functions calculate the probability of a crime event occurring within a given distance of the location where another crime event occurred. Our goal was to determine whether crime events tend to occur randomly over the region of interest or if offenders tend to focus their attention on only a few select areas, so-called "crime hotspots". Lastly, we investigated whether the spatial distribution of crime events has changed over the course of the 10-year period under investigation. This is particularly relevant since Khayelitsha is known to be home to a large number of informal settlements or "squatter camps" that can potentially appear (or disappear) virtually overnight. The erection of more informal housing ("shacks") increases the number of potential victims and offenders. To achieve this we used Andresen's spatial point pattern test which uses Monte Carlo sampling to assess whether two sets of spatial point patterns are significantly different by aggregating the number of points within defined spatial units. We also utilized a recently developed method proposed by Fuentes-Santos et al. to achieve this same purpose by testing whether the density at the points of two spatial point patterns are equal. This second test overcomes some of the weaknesses of Andresen's test and can potentially yield more accurate results through the use of bootstrapping.

### *Music Generation with Neural Networks*

*Wilben Pretorius*
*University of the Free State*
*Ludick, Z (Department of Statistics, University of the Free State)*

Recent research has compared neural network types in the context of image generation. Two of these networks are the generative adversarial network (GAN), an implicit-density type network, and the variational autoencoder (VAE), an explicit-density type network. It is suggested that VAE generates better image data than GAN. This research seeks to compare the two network types within the task of music generation. In order to assess the abilities of each network type, we will use Zipf's Law and artificial art critics. Finally, we hope to compare the music generated from the two networks to music generated from other methods including Markov chains and hierarchical variational autoencoders.

### *Computational Features for Efficient Estimation of Some Zero-Inflated Models*

*Gopika Ramkilawon*
*University of Pretoria*

The performance of the estimates of Poisson, zero-inflated Poisson, zero-inflated negative binomial and zero-altered Poisson models are evaluated using simulated zero-inflated count data generated from the zero-inflated Poisson model. Indices, namely the zero-inflation index, dispersion index and heavy tail index are also evaluated for each model. Method of moments estimators were derived for each model in closed form. Results from the simulation study conducted using method of moments indicated that the zero-inflated negative binomial model performed better than the other models included in the study when using zero-inflated, overdispersed count data.

### *Medium Term Load Forecasting in South Africa using Generalized Additive Model with Tensor Product Interactions*

*Thakhani Ravele*
*University of Venda*
*Sigauke, C, Bere, A (Department of Statistics, University of Venda)*

The study develop medium term load forecasting models which will help decision makers in Eskom for planning of the operations of the utility company. Most peak loads occur at hours 19:00 and 20:00, during 2009 to 2013. Generalised additive models with and without tensor product interactions were used to forecast electricity demand at 19:00 and 20:00 including daily peak electricity demand. Least absolute shrinkage and selection operator (Lasso) and Lasso via hierarchical interactions were used for variable selection. The parameters of the developed models were estimated using restricted maximum likelihood and penalized regression. The best models were selected based on smallest values of the Akaike information criterion, Bayesian information criterion and Generalized cross validation along with the highest Adjusted $R^2$. Forecasts from best models with and without tensor product interactions were evaluated using mean absolute percentage error, mean absolute error and root mean square error. Operational forecasting was proposed to forecast the demand at hour 19:00 with unknown predictor variables. Empirical results show that modelling hours individually during the peak period results in more accurate peak forecasts compared to forecasting daily peak electricity demand. The performance of the proposed models for hour 19:00 were compared and the generalized additive model with tensor product interactions was found to be the best fitting model.

### *Car Accident Feature Extraction from a Drone-Based Video Feed*

*Vaughn Saben*
*University of Cape Town*
*Britz, S (Department of Statistics, University of Cape Town)*

This thesis designs and implements, in part, an automated Drone Response Service (DRS) to respond to specified events. For a test case, the specified event is restricted to represent urban car accidents only. A DRS, within this context, is tasked with identifying visual car damage of a target vehicle to infer the type of car accident that has occurred. A drone's visual sensory output (i.e. image and /or video) is analysed to answer 3 primary research objectives: (1) Identify whether there is at least 1 car within an image frame. (2) Classify and place a bounding box about damaged and non-damaged cars within an image frame. (3) Classify and locate visual car damage on a damaged vehicle within an image frame. Camera-mounted drones produce aerial points-of-views of cars at varied angles. Convolutional Neural Network (CNN) based architectures are state-of-the-art at classification (Objective 1), object detection (Objective 2) and instance segmentation (Objective 3) tasks. Training of these CNN-based architectures requires a lot of relevant, annotated imagery data. This thesis interacts with the Personal Computer (PC) game, Grand Theft Auto V's, Rockstar Advanced Game Engine (RAGE) to create annotated image data relevant to each objective. Base keras CNN classification architectures, trained on ImageNet, have been used to predict the existence of a car. Additional transfer learning model configurations are currently being tested.

### *Simulation Studies in Stochastic Ergodicity*

*Issah Seidu*
*Department of Statistics and Actuarial Science, University of Ghana.*
*Mettle, FO, Quaye, ENB, Aidoo, EK, Boateng, LP (Department of Statistics and Actuarial Science, University of Ghana)*

Irrespective of whether the test of for homogeneity is significant or not, most researchers assume time-homogeneity in analysing Markov chains due to scanty literature on analysis of time-inhomogeneous Markov chains. Based on the assumption that, for each point in time in the future, a stochastic process will be subjected to a randomly selected transition matrix from an ergodic set of transition matrices the process was subjected to in the recent past, a methodology was proposed for analysing the long-run behaviours of a time-inhomogeneous Markov chains. The proposed model was implemented to historical data consisting of exchange rate of cedi-dollar, cedi-pound and cedi-euro spanning over 6 years (January 2012 to December 2017). The results show that under certain "closeness" conditions, the long-run behaviours of the time-inhomogeneous case are almost identical to as those of the time-homogeneous case. The paper asserted that even if the Markov chain exhibit time-inhomogeneity, analysing the Markov chain under the assumption of time-homogeneity is a step-in the right direction under certain "closeness" conditions, otherwise the proposed method is recommended. It was also found that investing in dollars yields better returns than the other currencies in Ghana.

### *An Investigation of Parent Distributions and Long-Term Trends of Average Maximum and Minimum Temperature in the Limpopo Province of South Africa*

*Anna Seimela*
*University of Limpopo*
*Maposa, D (Department of Statistics & Operations Research, University of Limpopo)*

In studying natural hazards or disasters that occur due to temperature extremes such as heat waves and cold waves it is crucial to understand the underlying distributions of the maximum and minimum temperatures at a particular site or region. The present study intends to investigate the parent distributions of maximum and minimum temperatures at various sites in the Limpopo province of South Africa. Four candidate parent distributions; normal, lognormal, gamma and Weibull distributions, were fitted to the average monthly maximum and minimum daily temperatures. Akaike information criterion (AIC) and Bayesian information criterion (BIC) were used to select the best fitting distribution at a particular site. The parent distribution with the lowest value of AIC and BIC was chosen as the best fitting distribution for the data. The findings revealed that light-tailed distributions in the Weibull domain of attraction, which include the Weibull distribution, are the best fitting parent distributions for both maximum and minimum temperatures at all the stations, except for Thabazimbi and Polokwane, where the best fitting parent distributions for the minimum temperature were found to be in the normal and log-normal domain of attraction, respectively. The Mann-Kendall test and time series plots trend analysis findings showed that there is a downward and upward long-term trend in minimum and maximum temperature data, respectively.

### *Wavelet-Based Time Series Analysis of South African Financial Data*

*Moyagabo Danny Tloubatla*
*University of the Witwatersrand*
*Mlambo, F (University of the Witwatersrand)*

In this paper an efficient time series forecasting model for foreign exchange rates is proposed. Previous literature reveals that Multilayer Perceptron is very effective in financial time series forecasting involving less computational load and fast forecasting capability. Autoregressive Integrated Moving Average (ARIMA) models are well known for their remarkable forecasting accuracy. In this literature, we have used Discrete Wavelet Transform (DWT) to decompose the in-sample training data into linear (detailed) and nonlinear (approximate) components. The researcher then left out the highest frequency component when applying the Inverse Discrete Wavelet Transform, then applied the appropriate model to forecast the respective components. The proposed model show accuracy improvement in the Multilayer Perceptron, but not in the ARIMA model. A reason for the improvement in the Multilayer Perceptron (MLP) model accuracy might be, after Smoothing the time series the Multilayer Perceptron (MLP) was more able to capture the non linear relationships with a better precision. ARIMA models model the linear components of the time series. The USD/ZAR data had an increasing trend which is also influenced by the high frequency component, hence by filtering the time series, the ARIMA model predicted values that are slightly lower than the average trend because of no influence of high frequency component on the trend.

### *A New Characterisation-Based Test for Symmetry*

*Carl Van Heerden*
*North-West University*
*Pretorius, C (Department of Statistics, North-West University)*

A new test for symmetry is proposed based on a lesser-known characterisation of symmetric distributions. We derive the limiting null distribution of the test and show that the test is consistent against general alternatives. The performance of the new test is evaluated and compared to that of existing tests by means of a Monte Carlo study. It is found that the newly proposed test performs favourably compared to the other tests.

### *Spherical (Skew-) Normal Distributions under a Geodesic Distance Measure*

*Delene Van Wyk*
*University of Pretoria*
*Bekker, A, Ferreira, JT*

A noteworthy problem with regards to directional statistics is the fact that many models neglect to address the curvature of underlying sample spaces. Although these models have been efficient when it comes to computation, results are influenced by this oversight with respect to shape. To address this problem, the spherical normal (SN) distribution was introduced by Hauberg in 2018. This distribution is curvature-aware and inference techniques can be developed towards increasing efficiency and minimising calculation errors. Results and expressions for the isotropic as well as the anisotropic cases of the SN distribution is explored, and it is shown how algorithms can aid in simulating and understanding the distribution. Borrowing from this approach, the spherical skew-normal (SSN) distribution is proposed to accommodate skewness and is implemented with a simulation to aid the understanding of this new model. This model is derived by applying the principle of rewriting the multivariate mean mixture normal to a skew-normal distribution in the spherical environment. The differences between the two simulation approaches are discussed for the SN and the SSN distributions. Output is graphically recorded, and the effect of a skewness parameter is explored in terms of how data is generated before being rescaled to the spherical environment.

### *Statistical Modelling of Degradation Rates of Photovoltaic Modules*

*Kaamilah Von Schoultz*
*Nelson Mandela University*
*Brettenny, WJ, Clohessy, CM (Department of Statistics, Nelson Mandela University),*
*van Dyk, EE (Department of Physics, Nelson Mandela University*

South Africa has a high energy demand which has typically been met by fossil fuels. This is unsustainable since fossil fuels are finite and have an extremely negative effect of the environment. Due to this, there has been an increase in the use of renewable energy resources such as solar, wind, biomass, hydroelectric and geothermal. This study focusses on solar energy, in particular, photovoltaic (PV) systems The increased use of PV systems in South Africa, has brought about the requirement to adequately assess and monitor the performance of such. This project will seek to shed light on the sustainability of PV installations by providing a statistical method to assess the degradation rates of PV modules in a South African setting. This information will be critical to investors and developers as they plan the long-term viability of a project. This project will make use of outdoor weather data to adapt the degradation model mentioned by Pan et al. (2011) to included multiple weather factors such as temperature, humidity and irradiance. Time series seasonal autoregressive moving average (SARIMA) models are developed to forecast the outdoor weather conditions. The forecasted data together with the developed degradation model is used to predict future degradation for the specific site. References Pan, R., Kuitche, J., & Tamizhmani, G. (2011). Degradation Analysis of Solar Photovoltaic Modules: Influence of Environmental Factor. In 2011 Proceedings-Annual Reliability and Maintainability Symposium (pp. 1–5).: IEEE.

### *Cluster Analysis for Group Selection in Launch Sale Predictions*

*Lee Watchurst*
*Investec and Nelson Mandela University*
*Clohessy, CM, Brettenny, WJ (Department of Statistics, Nelson Mandela University)*

One way for business to stay ahead in a competitive market is through the launch of new products and planning for these launches optimally. This includes ordering the correct quantity of stock in advance as well as maintaining these stock levels while the item launches. However, holding too much stock can affect the business costs adversely. This research proposes the use of cluster analysis techniques to determine the up-front purchase quantity by identifying similar items and using their initial quantities sold. Products will be grouped based on their numerical and categorical attributes. Once the data is clustered, the Bass model will be used to obtain a sales profile for the new item. The Bass model is a popular choice for product life cycle planning due to the emphasis placed on the timing of adoption. The study will make use of data from a retail and wholesale company that sells, in part, single use items. With the planning for new launches being a key problem point in many companies, this research aims to optimize the planning process and ensure product launch success across stores.

### *Assessment of Photovoltaic Software using Modelled and Measured Energy Yields*

*Edward James Westraadt*
*Nelson Mandela University*
*Clohessy, CM, Brettenny, WJ (Department of Statistics, Nelson Mandela University),*
*van Dyk, EE (Department of Physics, Nelson Mandela University*

The potential for solar power production in South Africa is extremely promising, with regions receiving solar irradiance intense enough to produce large quantities of power. The use of photovoltaic (PV) systems in South Africa is growing rapidly, since the equipment and installation costs are being reduced. For PV assessment, energy yield data is required by developers and investors to determine the PV viability. Energy yield data is often predicted by specialised PV software. This software is costly and requires licencing. This study investigates the use of solaR, an R-package, for the prediction of energy yield data. This package runs in the statistical software, R, and is free and open-source. This study compares the measured energy yield data of the 1.3kW mono-crystalline silicon system located at Nelson Mandela University's Outdoor Research Facility, and the modelled energy yield data which is produced using the solaR R-package. A two-sample profile analysis is used to conduct this comparison.

### *Basis Beyond Linear Regression using Wavelets, Splines and Locally Weighted Polynomial Regression*

*Mampane Phokwane Wilson*
*University of the Witwatersrand*
*Mlambo, F (University of the Witwatersrand)*

Here we present techniques that deal with non linear data structures in regression. The techniques form a pool of growing techniques that aim to address the need to get better estimates of non linear patterns in regression. The techniques focused on in this paper are based on wavelets, splines and locally weighted polynomial regression. The following research investigates the effect or power of non linear techniques in regression. The study will look into the comparison of the three non linear regression techniques. The paper will also discuss why the techniques are better suited than other comparable standard techniques which will not be studied in detail in this paper like normal polynomial regression. This will mainly be done by arguing points why the research based techniques are better and also highlight cases where they may have short falls in comparison to the other standard techniques in existence. In the current era, the idea of big data is the talk of the moment. This simply means large data sets than those seen before in past decades are readily available and accessible for use. These large data sets come with their own complexities such as variables relationship structures that may not be easy to model. As such finding techniques that are easily ready to handle most relationship patterns discovered in these data sets has become paramount. Regression serves as a tool that could aid us in trying to better understand this evolving new data sets. The regression techniques studied in this paper serve as possible tools to address this arising issue.

### *The Level of Difficulty and Discrimination Power of the NSC Mathematics Examination Questions*

*Nombuso Zondo*
*University of KwaZulu-Natal*
*Zewotir, T, North, D (Statistics Department, University of KwaZulu-Natal)*

South Africa's National Senior Certificate examination system was introduced in 2008 as a single national system in order to facilitate fair and standardized assessment and to provide all learners with an equal chance of access to higher education. However, limited research has been done to investigate the discrimination power of the examination questions and the spread of examination question difficulty for learners from different school quintile types. The purpose of this study was to investigate differential performance of learners in the National Senior Certificate mathematics examination questions. From the analysis, results show that the discrimination power of the different examination questions is not identical for different school quintiles. Further investigation of the data reflects a considerable range of category difficulty levels, with higher (above average) ability levels being tested for learners in the quintile 1 to quintile 4 schools, whilst only learners with average abilities were being tested in the quintile 5 and independent schools.

### *The Spatio-Temporal Analysis of Under-Five Mortality in Lesotho*

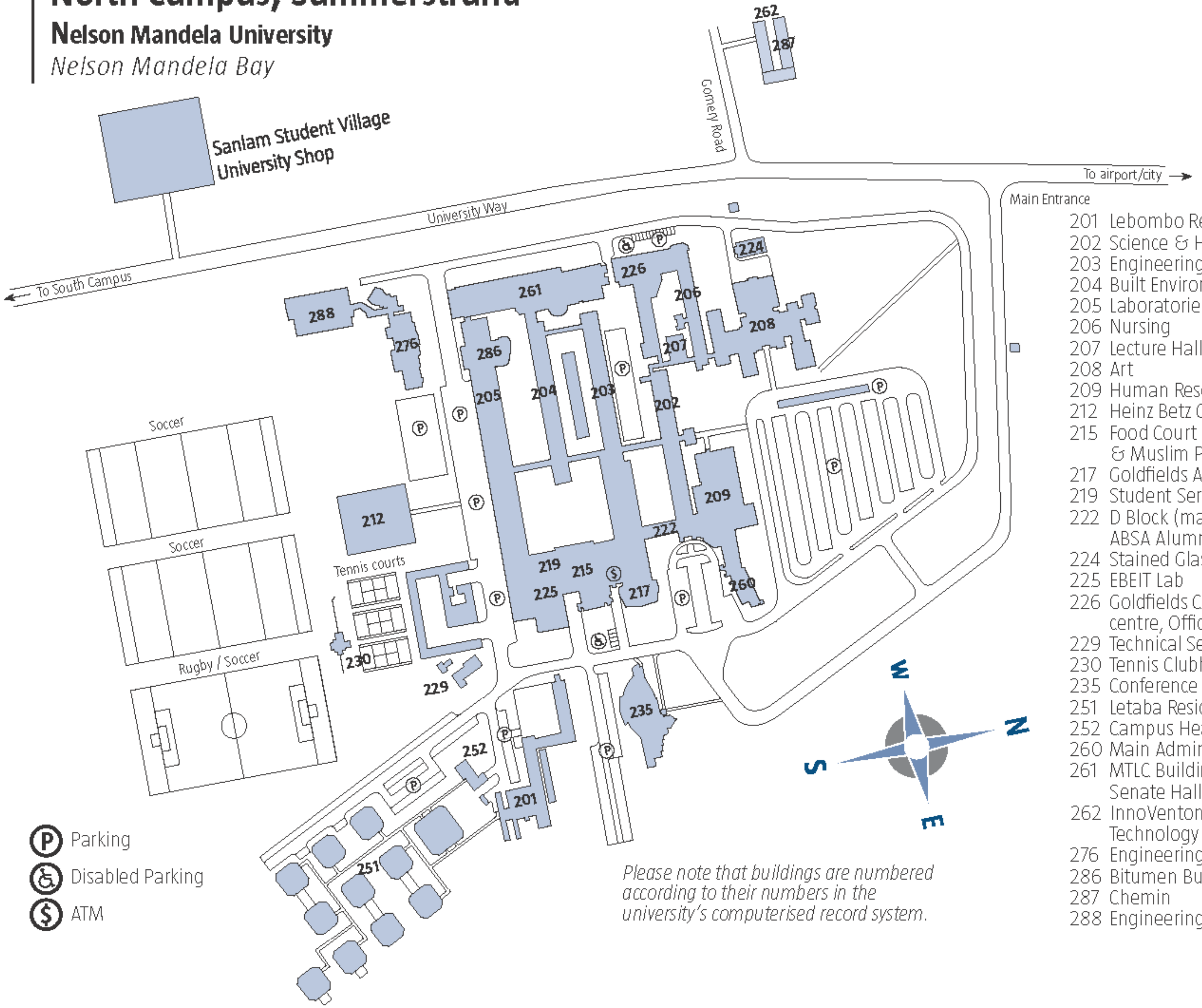
*Mthobisi Zondo*
*University of KwaZulu-Natal*

Under-five mortality (U5M) is a serious public health issue in Lesotho. Under the Sustainable Development Goal (SDGs) Lesotho aims to reduce the mortality rate to an average of 25 deaths per 1000 live births by the end of 2030. In the past, there has been very little statistical work concerning U5M in Lesotho, especially using the recently developed spatial mapping models. In this talk, datasets from the Lesotho Demographic and Health Surveys (LDHS) program conducted in 2004, 2009, and 2014 will be used. The surveys comprises 7 095, 7 624, and 6 621 women aged 15-49 for 2004, 2009, and 2014 respectively. The spatially structured and unstructured random effects models were used to construct the choropleth maps. The Spatiotemporal models were incorporated in this study to investigate how the proposed covariates vary over time. The inference was done using the full Bayesian framework through the R-INLA function. The spatial variation pattern of the U5M risk appeared to be different over time. The results further indicated that the children who are breastfed had lower odds of death overtime compared to those who are not breastfed, education also showed a significant protective effect over time on U5M, having many children aged below 5 contributes to higher odds of U5M, and age at death of under-five was shown to have a positive linear relationship with U5M.

# North Campus, Summerstrand

## Nelson Mandela University

*Nelson Mandela Bay*



Ⓟ Parking

♿ Disabled Parking

$ ATM

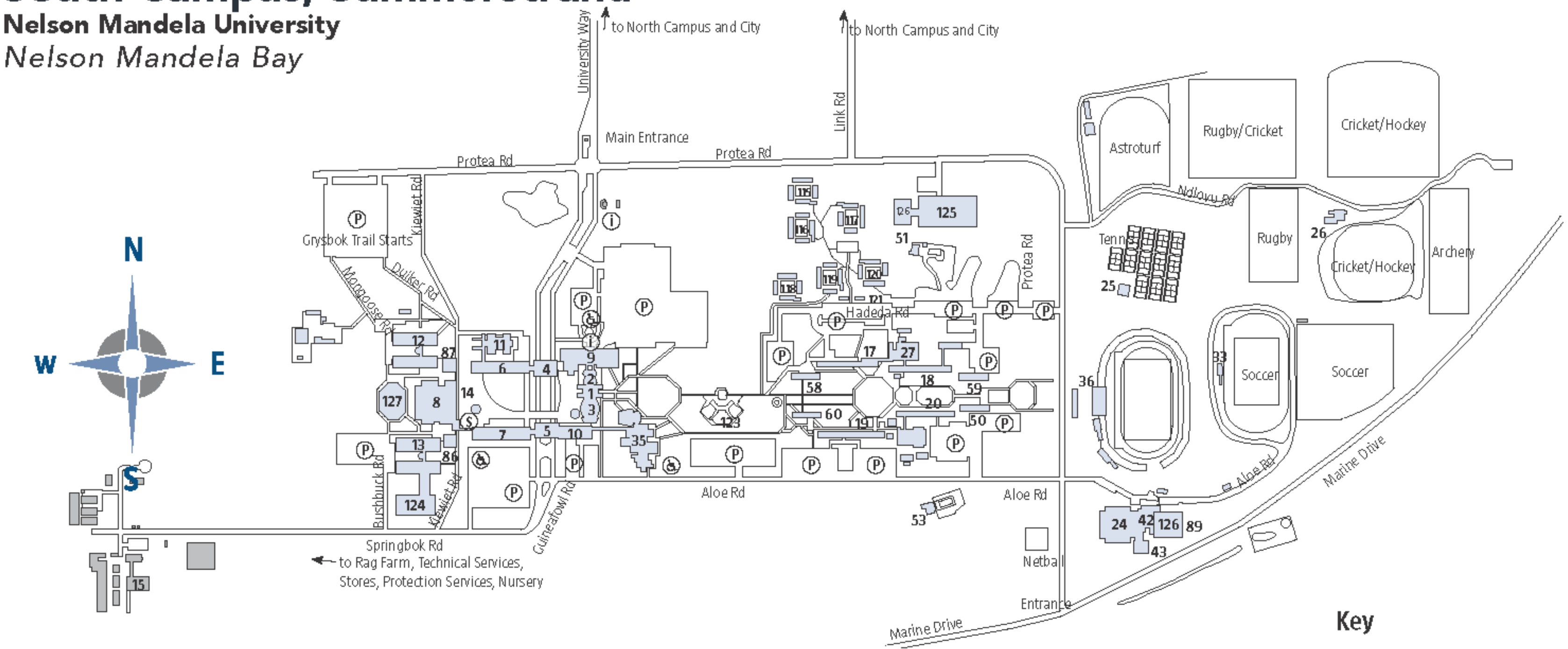
*Please note that buildings are numbered according to their numbers in the university's computerised record system.*

| | |
|---|---|
| 201 | Lebombo Residence |
| 202 | Science & Health Sciences |
| 203 | Engineering |
| 204 | Built Environment & Engineering |
| 205 | Laboratories |
| 206 | Nursing |
| 207 | Lecture Halls (N1 & N2) |
| 208 | Art |
| 209 | Human Resources & Contact Centre |
| 212 | Heinz Betz Centre |
| 215 | Food Court (cafeteria) & Muslim Prayer Room |
| 217 | Goldfields Auditorium |
| 219 | Student Services Centre |
| 222 | D Block (main entrance, cashiers, ABSA Alumni Centre, Student Counselling) |
| 224 | Stained Glass Studio |
| 225 | EBEIT Lab |
| 226 | Goldfields Computer Centre (writing centre, Offices etc.) |
| 229 | Technical Services |
| 230 | Tennis Clubhouse |
| 235 | Conference Centre |
| 251 | Letaba Residence |
| 252 | Campus Health Services |
| 260 | Main Administration Building |
| 261 | MTLC Building (eNtsa, library, CTLM, Senate Hall, lecture halls) |
| 262 | InnoVenton / Downstream Chemicals Technology Station |
| 276 | Engineering Building |
| 286 | Bitumen Building |
| 287 | Chemin |
| 288 | Engineering Building II |

# South Campus, Summerstrand

**Nelson Mandela University**

*Nelson Mandela Bay*



| | |
|---|---|
| 1 | Main Building |
| 2 | Council Chamber |
| 3 | Auditorium |
| 4 | Old Mutual Lecture Halls |
| 5 | Sanlam Lecture Halls |
| 6 | Education, Writing Centre & ABSA Computer lab |
| 7 | M & P Building |
| 8 | Library & School of Architecture |
| 9 | Embizweni |
| 10 | Music |
| 11 | Education |
| 12 | Biological Sciences |
| 13 | Physics & Chemistry |
| 14 | Food Court |
| 15 | Technical Services/Procurement |
| 17 | Unitas Main Block |
| 18 | Veritas Main Block |
| 19 | Xanadu Main Block |
| 20 | Melodi Main Block |
| 24 | Indoor Sport Centre & Sport Offices |
| 25 | Tennis Clubhouse |
| 26 | Cricket Clubhouse |
| 27 | Study Centre (Veritas) |
| 33 | Soccer Clubhouse |
| 35 | Building 35 (Universet Lecture Halls) |
| 36 | Stadium & Clubhouse |
| 50 | Melodi Annex |
| 51 | Unitas/Veritas Clubhouse & Pool |
| 53 | Xanadu/Melodi Clubhouse & Pool |
| 58 | Unitas Annex |
| 59 | Veritas Annex |
| 60 | Xanadu Annex |
| 89 | Underwater Clubhouse |
| 86 | Goldfields South |
| 87 | Goldfields North (International Office) |
| 115-120 | Renaissance Postgrad Student Village |
| 121 | Housing Administration |
| 123 | Building 123 |
| 124 | Centre for High Resolution Transmission Electron Microscopy (CHRTEM) |
| 125 | Human Movement Science |
| 126 | Dietetics |
| 127 | Life Sciences |

## Key

| | |
|---|---|
| Ⓘ | Information |
| Ⓟ | Parking |
| ♿ | Disabled Parking |
| Ⓢ | ATM |

*Please note that buildings are numbered according to their numbers in the university's computerised record system.*

Special thanks to our sponsors:

sas®

CASIO®

clindata™

NELSON MANDELA
UNIVERSITY

Research Management

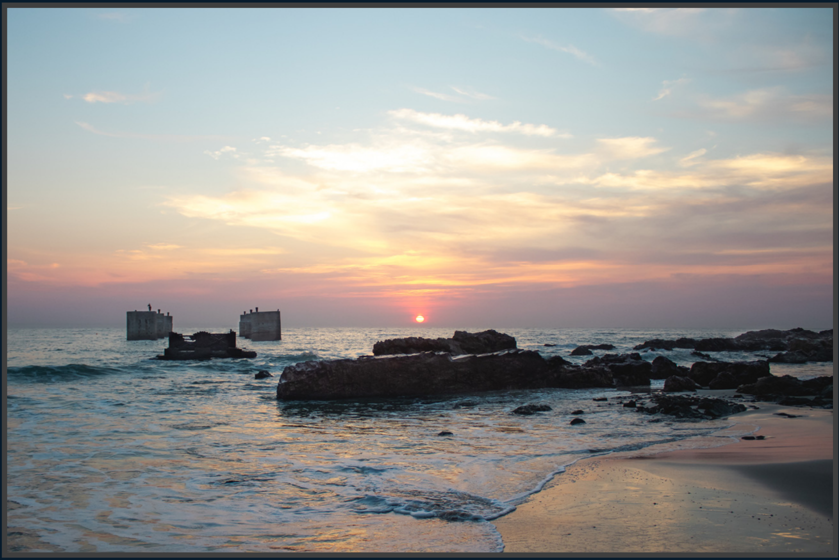
NELSON MANDELA
UNIVERSITY

Department of Statistics

NELSON MANDELA
UNIVERSITY

Faculty of Science

UNISA
university
of south africa



Seasonal adjusted value = TCI
AIN'T NUTHIN' BUT A PEANUT