



# 58<sup>TH</sup> ANNUAL CONFERENCE OF THE SOUTH AFRICAN STATISTICAL ASSOCIATION

28<sup>th</sup> November to 1<sup>st</sup> December 2016



## PROGRAMME AND ABSTRACTS





Would you  
bank on  
you? //

For more information on our graduate programmes  
visit [graduate.standardbank.com](http://graduate.standardbank.com) or email  
[Gregory.Newnham@standardbank.co.za](mailto:Gregory.Newnham@standardbank.co.za)

**[graduate.standardbank.com](http://graduate.standardbank.com)**

**Standard Bank** Moving Forward™



## CONTENTS

Welcome	Page 4
Sponsors	Page 5
General Information	Page 6
Vanue Layout	Page 10
Map	Page 11
Programme at a Glance	Page 12
Workshops	Page 14
Detailed Programme	Page 18
Poster Sessions	Page 30
Abstracts - Plenary Speakers	Page 32
Abstracts - Invited Speakers	Page 36
Abstracts - Contributed Papers	Page 39
Abstracts - Poster Sessions	Page 215
List of Delegates	Page 247



## WELCOME TO THE CONFERENCE



Dear conference delegate

We welcome you to the 58th annual conference of the South African Statistical Association. The photograph on the cover of this conference program and abstract booklet was taken on a morning during the week of the 19th September while we were all making our way by foot up the hill to campus, having parked our cars somewhere near Main Road since protesting students had blocked the entrances to campus. The dark sky and clouds are symbolic of what was to come during the next six weeks when both on our campus and many campuses around the country, students and staff were at odds with one another over fee and cultural issues that brought violence and destruction to these campuses. One consequence of this was that we had to move this conference to an off campus venue and while we are pleased that the River Club could accommodate us, we are also sad that we could not host you on our beautiful campus at the foot of Table Mountain.

However, the rainbow on this cover photograph can be seen as a sign of hope and optimism and we are pleased that we have over 350 delegates attending this conference coming from both academia and industry to share their work and ideas in the areas of Statistics and Data Science. We have an exciting group of international visitors from whom we hope to learn about the latest developments in Statistical Learning for Data Science, Correspondence Analysis, Statistical Ecology, Astronomy, Computational Developments in High Dimensional Models and Change Point Analysis. We thank our sponsors for their generous support. We hope that you will enjoy your time at the conference and your visit to Cape Town.



## SPONSORS



[www.sas.com](http://www.sas.com)

THE  
POWER  
TO KNOW.



DST-NRF CENTRE OF  
EXCELLENCE IN  
MATHEMATICAL &  
STATISTICAL SCIENCES  
M a s s

[www.wits.ac.za/coe-mass](http://www.wits.ac.za/coe-mass)



[graduate.standardbank.com](http://graduate.standardbank.com)



[www.jamesralph.com](http://www.jamesralph.com)



[www.quintiles.com](http://www.quintiles.com)



[www.cengage.co.uk](http://www.cengage.co.uk)



[www.olspes.com/predictive-analytics/](http://www.olspes.com/predictive-analytics/)



[www.wiley.com](http://www.wiley.com)



[www.quantec.co.za](http://www.quantec.co.za)



2017 International Population Conference

[ipc2017capetown.iusspp.org](http://ipc2017capetown.iusspp.org)



[www.oxford.co.za](http://www.oxford.co.za)



## GENERAL INFORMATION



### Registration

The registration desk is situated at the River Club reception, main entrance, ground floor and will be open: on Monday 28 Nov from 08h00 – 17h00 and on Tues 29 Nov – Thurs 1 Dec from 08h00 – 16h00.

### WiFi

Delegates can connect to the River Club WiFi. There is no code.

### Parking

Free secure parking is available at the River club.

### Name tags

Delegates should wear their name tags at all times to gain access to the lecture halls, lunch venue and evening functions.

### Tickets

Tickets will be issued for the welcome function, young statistician's pub quiz with drink vouchers and conference dinner. Please make sure you have these tickets on hand as they will give you access to these events.

### *Optional walk in Kirstenbosch gardens*

Delegates who would like to take part in the walk in Kirstenbosch gardens on Wednesday 30 November from 18h00 -19h00 need to sign up at the registration desk on Monday 28 and Tuesday 29 November. Delegates are to meet at the entrance to Moyo, Kirstenbosch at 18h00 in order to gain access to the Kirstenbosch gardens via the restaurant at no extra cost. If delegates wish to access the Kirstenbosch Gardens before 18h00 they will need to pay the R 60 p person fee at the Kirstenbosch entrance.

### Poster sessions

Posters have been allocated to one of three sessions (see pages 30 and 31). Posters will be displayed for the duration of the day your session is scheduled. Prestik and double-sided tape will be available from the registration desk. Posters should be up by 09:00 on the day of the poster session and be removed by 19:00 on the same day. Poster presenters are asked to be at their poster for discussion at the following times:

Poster session 1: 13:30 – 14:00 on Tuesday 29 November

Poster session 2: 12:40 – 13:40 on 30 November

Poster session 3: 12:30 – 13:30 on 1 December



## GENERAL INFORMATION

### Social events

The welcome function will take place at 17:00 on Monday 28 November in the Tugela room weather permitting, else in the Orange room on the first floor. Please check the signage on the day. Music will be provided by the Hout Bay Music Project, [www.houtbaymusic.org](http://www.houtbaymusic.org).



The young statistician's burger evening and pub quiz will take place at **18:00 in Zambezi**.

The conference dinner is preceded by a walk in the Kirstenbosch gardens at 18:00 on Wednesday 30 November. Should you wish to attend the walk, please sign up at the registration desk before Tuesday 29 November. The dinner will take place at Moyo Kirstenbosch, Kirstenbosch National Botanical Garden, Rhodes Drive, Newlands at 19:00.

### Venues

The conference takes place at The River Club, Liesbeek Parkway, Observatory. Tea, coffee and lunch will be served in the Orange room, first floor. The opening, plenary sessions and closing will take place in the Congo room, first floor. Parallel session will take place in the Congo room (first floor), Tugela, Berg, Club house, Zambezi and Kei on the ground floor. On Thursday 1 December there will also be parallel sessions in the Observatory auditorium.



## GENERAL INFORMATION

### Meetings

#### SASA Executive Committee meeting

- Monday 28 November at 16:00 in Berg

#### SASA Annual General meeting

- Tuesday 29 November at 14:40 in the Congo room.

#### Multivariate Data Analysis Group (MDAG) annual general meeting

- Wednesday 30 November at 13:40 in Kei

#### SUSAN-IBS and South Africa Region of the IBS meeting

- Wednesday 30 November at 13:40 in Zambezi

#### Industrial group meeting

- Wednesday 30 November at 13:40 in Tugela

#### Bayesian group meeting

- Wednesday 30 November at 13:40 in Berg

#### Data scientists group meeting

- Wednesday 30 November at 13:40 in Congo

### SASA 2016 organising committee

A/Prof Francesca Little (Chair)

A/Prof Sugnet Lubbe

Mr Neil Watson

Mr Stefan Britz

Mr Chun-Kai (Karl) Huang

Mr Greg Distiller

Ms Belinda Chapman (Conference organiser)

Ms Roxanne Adams (Conference organiser)

Dr Lizanne Raubenheimer, Rhodes Univ. (Finances)

### Guidelines to speakers and chairpersons

Parallel talks are 15 minutes with 5 minutes for questions and changing venues.

### Speakers

- Double check the date and time of your presentation.
- Load your presentation on the computer before the start of the session.
- Report to the chairperson of the session at least 5 min before the start of the session.
- Keep to the time allocated for your presentation.
- You are not allowed to move your presentation to any other time slot.
- Laser pointers will be available in the venue.
- *Once the chair indicates the end of your session, you must stop your presentation immediately.*



## GENERAL INFORMATION



### Chairpersons

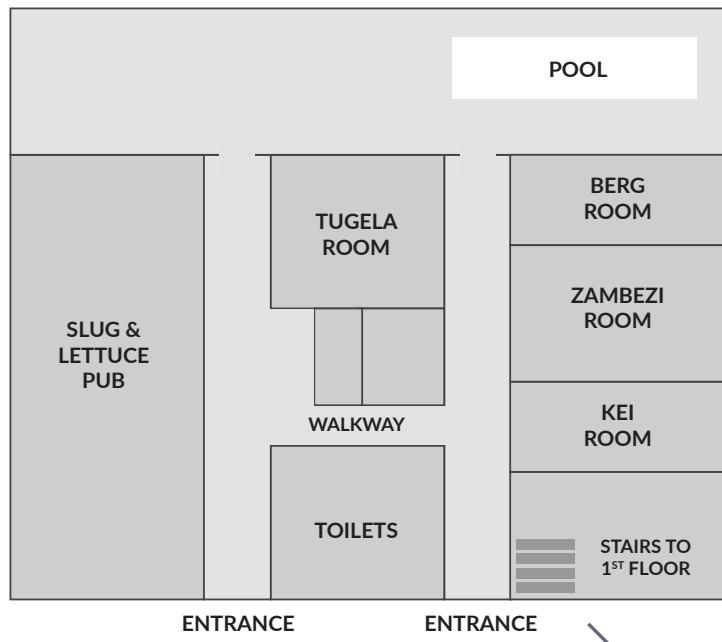
- Double check the date and time of your session.
- *Keep to the scheduled times.*
- No changes are to be made to the programme. If a speaker does not show, please wait for the next time slot, do not move other speakers earlier.
- Check the attendance of all the speakers, and ensure that all presentations have been loaded on the computer.
- Welcome delegates and speakers at the beginning of your session.
- Make the following announcements:
  - All cell phones to be switched off
  - State the programme for the session
  - Start with the first lecture
- Use the cards available in the venue to indicate to speakers the last 5 min and last 2 min of their presentation.
- Ask speakers to stop their presentation when the time is up.
- Allow questions according to time.
- Thank all speakers and delegates at the end of the session.
- Report to the registration desk if a speaker was absent.
- Report shortcomings to the session assistant.

These are only guidelines. Please use your own initiative to make your session a success.



## VENUE LAYOUT

### GROUND FLOOR



### FIRST FLOOR





MAP

## FROM AIRPORT

SAAO

# THE RIVER CLUB



## PROGRAMME AT A GLANCE

### Monday 28 November 2016

	Congo	Berg	Zambezi	Kei
8:00 - 9:00	Workshop registration			NRF/SASA mentoring session
9:00 - 10:30	Workshop 3 (Greenacre)	Workshop 5 (SAS)	Workshop 1 (Antoniadis)	Workshop 2 (Borchers)
10:30 - 11:00	Tea / Coffee			
11:00 - 12:30	Workshop 3 (Greenacre)	Workshop 5 (SAS)	Workshop 1 (Antoniadis)	Workshop 2 (Borchers)
12:30 - 13:30	Lunch			
13:30 - 16:00	Workshop 4 (Hastie)	Workshop 5 (SAS)	Workshop 1 (Antoniadis)	Workshop 2 (Borchers)
16:00 - 17:00	Registration			
17:00 - 19:00	Welcome function with music provided by Hout Bay Music Project			

### Tuesday 29 November 2016

	Congo	Tugela	Berg	Clubhouse	Zambezi
8:00 - 8:45	Registration				
8:45 - 10:00	Welcoming, Opening, SAS, Presidential address				
10:00 - 10:30	Tea / Coffee				
10:30 - 11:10	Plenary Marie Hušková				
11:20 - 13:00	Ecological statistics	Bioinformatics	Biostatistics	Theoretical statistics	
13:00 - 14:00	Lunch & Poster session 1				
14:00 - 14:30	Plenary Yigit Karabag (SAS)				
14:30 - 14:40	Awards				
14:40 - 15:40	SASA AGM				
15:40 - 16:00	Tea / Coffee				
16:00 - 17:40	Ecological statistics	Biometry and Experimental design	Biostatistics	Theoretical statistics	Time series
18:00 - 22:00	Young statistician's burger evening and Pub quiz				



## PROGRAMME AT A GLANCE

### Wednesday 30 November 2016

	Congo	Tugela	Berg	Clubhouse	Zambezi	Kei
8:20 – 10:00	Spatial statistics	Financial statistics	Ecological statistics	Theoretical statistics	Climate and extreme value theory	Time series
10:00 – 10:30	Tea / Coffee					
10:30 – 11:10	Plenary Anestis Antoniadis					
11:20 – 12:40	Business Analytics	Financial statistics	Spatial statistics	Stochastic processes	Biostatistics	Multivariate data analysis
12:40 – 13:40	Lunch & Poster session 2					
13:40 – 14:40	Society meetings					
14:40 – 16:00	Data science	Financial statistics	Bayesian statistics	General	Biostatistics	Multivariate data analysis
16:00 – 16:20	Tea / Coffee					
16:20 – 17:20	Time series	Process control	Bayesian statistics	General	Biostatistics	Multivariate data analysis
18:00 – 19:00	Walk in Kirstenbosch gardens					
19:00 – 22:00	Gala dinner at Moyo Kirstenbosch					

### Thursday 1 December 2016

	Congo	Clubhouse	Zambezi	Kei	Observatory
9:00 – 10:20	Official statistics	General	General	Clinical trials and complex surveys	Multivariate data analysis
10:20 – 10:50	Tea / Coffee				
10:50 – 12:30	Process control	Energy	Education	Biostatistics	Multivariate data analysis
12:30 – 13:30	Lunch & Poster session 3				
13:30 – 13:50	Report back on the submission to Department of Higher Education and Training				
13:50 – 14:30	Plenary Trevor Hastie				
14:30 – 15:00	Closing & prizes for competitions				
15:00 – 15:20	Tea / Coffee				
15:20 – 18:00				NRF/SASA mentoring workshop	



## WORKSHOPS

### Workshop 1: Computational Developments in High Dimensional Models

#### Workshop organiser: Anestis Antoniadis

Laboratoire Jean Kuntzmann, University Joseph Fourier, Grenoble & Department of Statistical Sciences, University of Cape Town

This workshop will focus on both theoretical and computational developments in high-dimensional statistical models. Of particular interest are models that involve high-dimensional vector or matrix parameter estimation, such as those occurring in copula models, graphical and network models, factor models, and functional data. These models usually depend on parameters with sparse patterns. High-dimensional models with low-dimensional structure are crucial for the successful implementation and theoretical analysis of such models, especially for a limited amount of data. Rich applications occur in genetics, neuroscience, economics, public health, psychology and sociology. New scientific challenges in these established areas, or in other emerging areas arise on a continual basis.

The one day workshop on High Dimensional Statistics will include lectures by internationally prominent researchers, chosen by the organising committee and will cover many aspects of this theme, motivated by many applications. The aim of the workshop is to bring together researchers and scientists interested in high dimensional statistics, its applications, and related areas. In the afternoon there will be lectures and a computer lab session on longitudinal and functional data analysis.

#### Invited speaker: Irène Gijbels

Department of Mathematics and Leuven Statistics Research Centre (LStat), KU Leuven

#### *Quantile regression in varying coefficient models*

Quantile regression is an important tool for describing the characteristics of conditional distributions. To allow for analyzing complex data situations, several flexible regression models have been introduced. Among these are the varying coefficient models, that differ from a classical linear regression model by the fact that the regression coefficients are no longer constant but functions that vary with the value taken by another variable, such as for example time. In this talk we study quantile regression in varying coefficient models for longitudinal data. The quantile function is modeled as a function of the covariates and the main task is to estimate the unknown regression coefficient functions. We approximate each coefficient function by means of P-splines. Theoretical properties of the estimators, such as rate of convergence and an asymptotic distribution are established. We further consider several other aspects of the problem studied. Firstly, we pay attention to the property that population conditional quantile functions cannot cross for different quantile orders, and hence ensure that the estimated quantile curves do not cross. Secondly, we allow for some heteroscedasticity in the error modelling, and also estimate the associated variability function. Finally, we briefly discuss other interesting problems that have been addressed, such as testing for a specific variability structure, or constrained inference for the varying coefficients. Some illustrations of the developed methodology with simulated data and real data analysis are provided.

This talk is mainly based on joint work with Yudhie Andriyana and Anneleen Verhasselt.

14



## WORKSHOPS



**Invited speaker: Jean-Michel Poggi**

University Paris Descartes and University Paris Sud, France

### *Random Forest-Based Approach for Physiological Functional Variable Selection for Driver's Stress Level Classification*

With the increasing urbanization and technological advances, urban driving is bound to be a complex task that requires higher levels of alertness. Thus, the drivers mental workload should be optimal in order to manage critical situations in such challenging driving conditions. Past studies relied on drivers performances used subjective measures. The new wearable and non-intrusive sensor technology, is not only providing real-time physiological monitoring, but also is enriching the tools for human affective and cognitive states monitoring.

This study focuses on a drivers physiological changes using portable sensors in different urban routes. Specifically, the Electrodermal Activity (EDA) measured on two different locations: hand and foot, Electromyogram (EMG), Heart Rate (HR) and Respiration (RESP) of ten driving experiments in three types of routes are considered: rest area, city, and highway driving issued from physiological database, labelled "drivedb", available online on the PHYSIONET website.

Several studies have been done on driver's stress level recognition using physiological signals. Classically, researchers extract expert-based features from physiological signals and select the most relevant features in stress level recognition. This work aims to provide a random forest-based method for the selection of physiological functional variables in order to classify the stress level during real-world driving experience. The contribution of this study is twofold: on the methodological side, it considers physiological signals as functional variables and offers a procedure of data processing and variable selection. On the applied side, the proposed method provides a "blind" procedure of driver's stress level classification that do not depend on the expert-based studies of physiological signals. This talk is mainly based on joint work with Neska El-Haouij, Raja Ghozi, Sylvie Sevestre Ghalila and Méeriem Jaidane.

**Invited speaker: Idris Eckley**

Mathematics and Statistics, Lancaster University, UK

### *Contemporary challenges in multivariate change-point analysis*

In recent years there has been a surge of interest in the development of new methods for tackling change-point problems. Much of this work has focused on the analysis of univariate time series, with contributions in this area having an impact in various disciplines, notably climate science and genomics. However with the increased use of low-cost sensors, there is a growing need to be able to efficiently and accurately identify changes in structure within highly multivariate time series. In moving to this more complex setting, the problem of detecting change-points becomes more subtle. We will outline some recent developments in this area motivated by ongoing work with industrial collaborators.



## WORKSHOPS

**Invited speaker:** Marie Hušková  
Charles University in Prague, Czech Republic

### *Change point detection in panel data*

The talk presents recent results on testing and estimating for a change in panel data. Asymptotic properties of the suggested procedures are presented. It is assumed that either the number of panels is large and the number of the observations in each panel are large or fixed (small w.r.t. the number of panels). Results of a simulation study and an application to a real data set will be presented.

This talk is based on joint work with several co-authors.

### **Workshop 2: Modelling spatial distribution from wildlife surveys**

**David Borchers**  
University of St Andrews

This one day workshop will provide participants with an overview of some recently-developed methods for modelling the spatial and spatio-temporal distribution of wildlife populations using data arising from the most widely-used wildlife survey methods. More specifically, it will cover methods of fitting Poisson point process models and log-Gaussian Cox process models to survey data from plot surveys, distance sampling surveys and capture-recapture surveys.

A feature of wildlife surveys that distinguishes them from most other kinds of surveys is that inclusion probabilities (the probabilities that individuals in the study population are included in the sample) are beyond the surveyors' control, unknown and spatially varying. This creates difficulties for spatial modelling because spatially varying inclusion probability and spatially varying density may be confounded. Wildlife survey methods have tended to focus on estimating inclusion probabilities under simplistic assumptions about spatial distribution. Spatial modelling methods, on the other hand, have focussed on realistically complex models for spatial distribution while assuming known inclusion probabilities. One would really like methods that combines the strengths of these two approaches.

This workshop gives an overview of some recently-developed methods that integrate the two approaches, allowing inferences about spatially varying inclusion probabilities to be drawn simultaneously with inferences about the spatial distribution and abundance of wildlife populations.

The workshop will include some hands-on sessions to give participants (a) a taster of R software packages implementing some of the methods and (b) a break from being lectured at. Participants should bring their own laptop computers for these sessions, with R and (preferably) RStudio installed. Details of required R libraries will be provided in advance of the workshop. Please note that there is insufficient time in the workshop to provide anything more than a taster of the packages.



## WORKSHOPS

### Workshop 3: Recent developments in Correspondence Analysis

**Michael Greenacre**

Pompeu Fabra University, Barcelona, Spain

The third edition of Michael Greenacre's book *Correspondence Analysis in Practice* will appear in 2016, published by Chapman & Hall. In this course he will present some of the new material in the book, such as the use of fuzzy coding, the treatment of missing data, contribution biplots, relationship to compositional data analysis, the analysis of matched matrices, and permutation testing. The accent is on the practical utility of these new aspects, and many illustrations will be given as well as a brief introduction to correspondence analysis. Many parts of his presentation will use dynamic graphics to illustrate the concepts and facilitate participants' understanding of the methods.

### Workshop 4: Statistical learning in data science

**Trevor Hastie**

Stanford University, USA

We give an overview of statistical models used by data scientists for prediction and inference. With the rapid developments in internet technology, genomics, financial risk modelling, and other high-tech industries, we rely increasingly more on data analysis and statistical models to exploit the vast amounts of data at our fingertips. We then focus on two important classes of tools. For wide data, we have a closer look at the lasso and its relatives, and for tall data random forests and boosting.

Most of the material can be found in "An Introduction to Statistical Learning, with Applications in R" by Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani (Springer, 2013), which is also available free in pdf format from  
[http://web.stanford.edu/~hastie/local.ftp/Springer/ISLR\\_print6.pdf](http://web.stanford.edu/~hastie/local.ftp/Springer/ISLR_print6.pdf)

### Workshop 5: Using SAS Analytics in Anger

**André Zitzke**

Senior Business Solutions Manager: Analytics, SAS

This workshop addresses various aspects of utilizing SAS Analytics in real-world scenarios. Attendees will receive insights into the modern application of Statistics in industry, including:

- Open Source integration
- Point & Click Statistics
- Data Visualization
- In-memory Statistics
- Analytics Lifecycle



## DETAILED PROGRAMME

Monday 28 November 2016			
	Congo	Berg	Zambezi
8:00 - 9:00		Workshop registration	
9:00 - 10:30	Workshop 3 (Greenacre)	Workshop 5 (SAS)	Workshop 1 (Antoniadis)
10:30 - 11:00		Tea / Coffee	Workshop 2 (Borchers)
11:00 - 12:30	Workshop 3 (Greenacre)	Workshop 5 (SAS)	Workshop 1 (Antoniadis)
12:30 - 13:30		Lunch	Workshop 2 (Borchers)
13:30 - 16:00	Workshop 4 (Hastie)	Workshop 5 (SAS)	Workshop 1 (Antoniadis)
16:00 - 17:00		Registration	Workshop 2 (Borchers)
17:00 - 19:00	Welcome function with music provided by Hout Bay Music Project		



Tuesday 29 November 2016					
8:00 - 8:45	Registration				
8:45 - 10:00	Registration, Welcoming, Opening, SAS, Presidential address; Chair Francesca Little/ Danie Uys				
10:00 - 10:30	Tea / Coffee				
10:30 - 11:10	Plenary Marie Huskova: Some recent results in change point analysis Chair James Allison				
11:20 - 13:00	Ecological statistics	Bioinformatics	Biostatistics	Theoretical statistics	
Chair	Congo Res Altwegg	Tugela Jorcache Ramjith	Berg Freedom Gumede	Berg Andriette Bekker	ClubHouse
	Clark, Altwegg, Ormerod: A Variational Bayes approach to the analysis of spatial site-occupancy models	de Vries: Multivariate Dose-Response Profiling	Kassanjee, McWalter, Welte: A general biomarker-based incidence estimator for HIV	Huang, North, Zewotir: Sampling variance of a multi-phase stratified design on partially exchangeable sequences	
	Durbach, Wolff-Piggott: Differentiating fish stocks using image classification	Lacerda, Murrell: A Population Genetics Approach to Inferring Selection from Longitudinally-Sampled HIV-1 Haplotypes	Reddy, Molenberghs, Aerts, Njagi: Prediction of the time to reach a biomarker threshold from complex longitudinal data	Minkah, de Wet, Dokument, Amponsah, Nortey: Comparison of Confidence Interval Estimators: An Index Approach	
	Haines: Maximum Likelihood: A Problem in Ecology	Sadiq, Lacerda, Murrell: Assessing the robustness of phylogenetic models to changes in selection pressures over time: A simulation study	Lesosky, Myer: Semi-parametric mixture models for skewed longitudinal HIV viral load trajectories with rebound	Steyn: Statistical Modelling with Quantile-based Methods and L-moments	



## Tuesday 29 November 2016

Borchers: Wildlife survey models: thinned spatial point processes with unknown thinning probabilities	Wolff-Piggott: Phylogeographic General Linear Models for the Evaluation of Epidemiological Predictors	Yende-Zuma, Vansteelandt, Mwambi: Adjusting the effect of integrating antiretroviral therapy and tuberculosis treatment on mortality for non-compliance: an instrumental variables analysis	Visagie: On the use of numerical optimisation procedures in parameter estimation
	Muller, Tolver: Wavelet-based Functional Mixed Models applied in Metabolomics	Glass, Myer, Hsiao, Lesosky: Latent class mixed models for classification of HIV viral load profiles in the Western Cape public sector antiretroviral treatment programme: 2009 - 2015	Buitendag: The Saddle-point Approximation and its Applications to the Hill Estimator
13:00 - 14:00			
		Lunch, Poster session 1	
		Plenary Yigit Karabag (SAS): Data Scientitit: The Alter Ego of The Statistician Chair Paul Mostert	
		14:00 - 14:30	
		14:30 - 14:40	Awards, Chair James Allison
		14:40 - 15:50	SASA AGM
		15:40 - 16:00	Tea / Coffee
16:00-17:40			
Chair: Allan Clark	Ecological statistics Congo	Biometry and Experimental design Tugela	Biostatistics Berg
	Mzabalazo Ngwenya	Maia Lesosky	Theoretical statistics Chun-Kai (Karl) Huang
17:40 - 19:00			
Tea / Coffee			
19:00 - 20:30			
Dinner			
20:30 - 22:00			
Poster session 2			
22:00 - 23:00			
Social gathering			
23:00 - 00:00			
Sleep			



## Tuesday 29 November 2016

Eri: Dynamic Factor Analysis for Ecological Time Series	Lalbahadur: Forecasting Hail Damage for Crop Insurance  Distiller, Borchers: The benefits of using a continuous-time framework for spatial capture-recapture (SCR) models	Manda, Meyer, Romeo, Masenyaete, Mwambi: Bayesian Copula-based Analysis of Multiple Longitudinal Measures and Multiple Recurrent Events  Gemechu, Debusho, Haines: Algorithmic Construction of Optimal Block Designs for Two-Colour cDNA Microarray Experiments Using the Linear Mixed Effects Model	Swanepoel, Janssen, Veraverbeke: Smooth copula-based estimation of the conditional density function and the conditional hazard rate function  Iiyambo, Schall: Fiducial inference for location-scale and log-location-scale families of distributions: Application to the Weibull distribution two-sample problem	Janse van Rensburg, Sharp: GAS models of financial time series
		Goodall, O'Connor: Use of mixture models for assessing the size structure of long-lived plants in a temporally varying environment	Turfan, Yilmaz, Aktas: Multi-criteria decision making in Taguchi orthogonal designs	Young statisticians burger evening & Pub quiz



Wednesday 30 November 2016						
8:20-10:00	Spatial statistics	Financial statistics	Ecological statistics	Theoretical statistics	Climate and extreme value theory	Time Series
	Congo	Tugela	Berg	ClubHouse	Zambezi	Kei
Chair:	Christine Kraamwinkel	Henri Moolman	Victoria Goodall	Jaco Visagie	Yehenev Kifle	Johan Ferreira
	08:00 Dziwa, Adékambi: Moment of the discounted compound renewal cash flows with dependence: The use of Farlie-Gumbel- Morgenstern Copula	Boyd, Winker, Althweg, Kerwath: Fishery population dynamics and stock assessment of geelbek (Atractoscion aequidens), a commercially important migrant fish species off the coast of South Africa	Kamwi, Koen, Kazembe: Characterization of bias in misspecified power law distribution	Musyoka, Omony, Mwalli, Achia, Gichangi, Ahmed, Joseph, Marano, Mwambi: Modeling the effect of climate on incidence of respiratory syncytial virus infections in a refugee camp in Kenya: A non-Gaussian time- series analysis	Pillay, Naidoo: Benchmarking time series using the Fernandez Optimisation Technique for South African private sector companies and municipalities	
	Fabris-Rotelli: A point process model for DPT pulses	Combrink: A robust construction of the efficiency frontier using resampling and shrinkage techniques	Geldienhuys: Species Distribution Models for spatial conservation planning of African penguins	Omachar, van Staden: The two- piece hyperbolic secant distribution	Diriba, Debusho: Random Effect Approach for Extreme Wind Speed in South Africa	Sjolander, Meades: Inference on Time Series Models for Paired Comparisons
	Maposa, Cochran, Lesaoana: Towards application of spatial statistics of extremes to the extreme flood heights in the lower Limpopo River basin of Mozambique	Huang, North, Zewotir: A conditional extreme value approach to value- at-risk estimation with exchangeable innovation	Millard, Kanfer: Mixture regression using t-distributed errors	Nemukula, Sigauke: Modelling average maximum daily temperature using r largest order statistics: An application to South African data	Kanfer, Millard: Mixture of mixed effects regression with t-distributed error and random components	Masoumi Karakani, van Niekerk, van Staden: Subjective Bayesian Analysis under the AR(1) model
	Thiart, Haines, Jacob, Prins: Determining Sample line spacing for a Marine Diamond deposit based on D-Optimality	Kruger, Rajaratnam, Huang: A Comparative Analysis of Aggregational Gaussianity for JSE-listed Shares and Indices	Hugo, Altwegg: The second Southern African Bird Atlas Project: causes and consequences of geographical sampling bias			



## Wednesday 30 November 2016

Stein: Dealing with uncertainty in spatial big data studies	Ilupéju, Chifurira, Chinhamu: Modelling South Africa's market risk using the APARCH model and heavy-tailed distributions	Photopoulou, Leos-Barajas: Using hidden Markov models for the analysis of animal acceleration data	Millard, Kanfer: Mixture regression using t-distributed errors	Sigauke: Modelling heat waves and their impact on electricity demand	MacDonald: Some models for discrete-valued time series
	Naradha, Chinhamu, Chifura, Hammujuddy: Multivariate Stable distributions and its applications for BRICS financial data	van der Merwe, Periquet: The eating habits of bat-eared foxes	Venter: Survival probability for a defended system from incomplete and Uncertain Data	Smit, Kijko, Fabris-Rotelli, van Staden: New Procedure for Probabilistic Tsunami Hazard Assessment	Viljoen: Investigating Stepwise Common Singular Spectrum Analysis and Multi-channel Singular Spectrum Analysis
10:00-10:30 Tea / Coffee					
10:30-11:10	Plenary Anestis Antoniadis; Estimation and group variable selection for additive partial linear models with wavelets and splines Chair: Linda Haines				
11:20-12:40	Business Analytics	Financial statistics	Spatial statistics	Stochastic Processes	Biostatistics
	Congo	Tugela	Berg	ClubHouse	Zambezi
Chair:	Melvin Varughese	Mardi Jankowitz	Christien Thiart	Delson Chikobvu	Greg Distiller
	Zitzke, Neuhoff: Using Text Analytics to Analyse Research Papers	Panman, Haasbroek, Pieters: Estimating correlations within operational loss data	Khuluse-Makhanya, Stein, Debra: Identifying local emission sources of particulate matter to improve geostatistical mapping of PM10	Bechelt: A Statistical First Passage Time Problem in System Replacement	Majakwara, Pal: Destructive COM-Poisson Cure Rate Model and Associated Likelihood inference under Lognormal Lifetime
	12:40-13:40 Multivariate Data Analysis Chair: Dominique Katshunga				
	Schutte, Schutte, Wissing: Applying multidimensional item response theory models in assessing dimensionality of the Mental Health Continuum Short Form (MHC-SF)				



## Wednesday 30 November 2016

Bergh: Predictive Analytics & SPSS in the business world	Verster, Venter: A predictive approach towards automatic discretization using data splitting	van Biljon, Pannan: On the quantification of model risk in statistical models of empirical data	Kraamwinkel, Fabris-Rotelli: Spatial sampling for rabies vaccination schedule	Ludick: Stochastic Processes in Sound Synthesis	Musekiwa, Manda, Mwambi: Longitudinal Meta-analysis of Multiple Effect Sizes	Crafford, Ehlers: Estimation of Multilevel Models with Interactive Generalised Least Squares
	Mavuso: Mean-Variance Hedging of Basis Risk	Kebotsmang: Pareto Optimal Space-filling Designs	Visagie: On generalised gamma convolution processes in time changed models	Chatora, Little, Barnes: Joint models for zero inflated nonlinear profiles	Hendriks, Crafford: The analysis of a grouped response variable in a multifactor design	
	Ngwenya: Building Robust Systems Adept at Handling Concept Drift	Chule: Market risk spectrum performance on the Diebold-Mariano framework using statistical arbitrage	Reynolds, Fabris-Rotelli, Loots: Design-based spatial sampling	Pienaar, Varughese: Jump-Reverting Diffusions and Parity to Continuous-Time Reversion	Ogunsakin, Lougue: Bayesian Generalized Linear Mixed Model of Breast Cancer Patients	Mazarua: A comparison of the performance of latent Dirichlet allocation and the Dirichlet multinomial mixture model on short text
	Lunch, Poster session 2					
13:40 - 14:40 Society meetings						
Data Scientists group		Industrial group	Bayesian group	SUSAN-IBS and South Africa Region of the IBS		
Chair: Congo		Tugela	Berg	ClubHouse	Zambezi	Kei
Chair: Frans Kanfer		TBA	Lizanne Raubenheimer		Iain MacDonald	Sugnet Lubbe
14:40-16:00 Data Science		Financial statistics	Bayesian statistics	General	Biostatistics	Multivariate Data Analysis
Chair: Congo		Tugela	Berg	ClubHouse	Zambezi	Kei
Chair: Miguel Lacerda		Willie Conradie	Abrie van der Merwe	Jenny Holloway	Chris Muller	Johane Nienkemper-Swanepoel



## Wednesday 30 November 2016

Varughese: Anomaly detection	Dalmeyer, Bradfield: Robust Portfolio Construction	Kirkland, de Waal, Haas: A Bayesian Approach to Green Security Games: Initial Perspectives	Brettenny, Sharp: The effect of the National Benchmarking Initiative (NBI) on the Efficiency of Water Service Provision in South Africa	Phaweni: Text Classification with Network Based Features
Diamini, Durbach: Individual identification of animals using acoustics	Lakhoo Lala: Enhanced Minimum Variance Optimisation - A pragmatic approach	Raubenheimer: Fiducial and objective Bayesian inference for the difference between two binomial proportions	Mokobane, Lesoana: Application of small area estimation methods in modelling lack of service delivery at ward level: water, sanitation and electricity in South Africa	Mugenyi, Abrams, Smith, Staecke, Hens: Estimating age-time dependent malaria force of infection accounting for unobserved heterogeneity
Bassett: Machine learning and Inference Challenges in Astronomy	Oppel, Loots, Beyers: Accelerated test with an application to too-big-too-fail portfolios	Harvey, van der Merwe: The Performance of Non-Informative Priors for the Bivariate Lognormal Distribution	Mféka, Zondo, Lougue: Grade repetition in school - The case of South Africa	Batidzirai, Manda, Mwambi: Gaussian Processes in analyzing big data: An application with multilevel models
	Raubenheimer, de Jongh, Lombard: A multiplier approach for approximating and estimating extreme quantiles of compound frequency distributions using the less extreme quantiles of the severity distribution	Maribe, Verster, Beirant: Reducing bias and MSE in estimation of heavy tails: a Bayesian approach	Obaromi, Qin, James, Adeboye, Odeyemi: Robustness and flexibility modeling of the prior for the unstructured spatial random effect in a disease mapping model	Matizirofa: Predictors of blood pressure in South Africa: Quantile regression Approach
				Lau, Fabris-Rotelli: Dependence structures in multi-dimensional arrays



## Wednesday 30 November 2016

						Tea / Coffee								
16:00-16:20		Time series and electricity		Process Control		Bayesian statistics		General		Biostatistics		Multivariate Data Analysis		
Chair:	Lienki Viljoen	Congo	Tugela	Berg	ClubHouse	Zambezi								
Pokolo, Pazvakawambwa: A Time series ARIMA model for rainfall for Katima Mulilo Town of Namibia, 2003- 2015	Mijburgh, Human, Bekker: Distributional development of ratios of gamma variables with application to a changepoint model in statistical process control	Andre Mostert	Lizanne Raubenheimer	Dorette Bekker		Henry Mwambi	Mutambayi, Ndege, Azeel, Odeyemi, Qin: Assessment of Risk Determinants in the Regularity of Malaria Using Cox Proportional Approach	Mulenga, Er, Brouze: Validity of the Psychological Capital Questionnaire in the South African context		Naseje, Justine, Mwambi, Lesosky. The effects of the splitting rule on the predictive performance of random survival forests		Ntushelo, Minnaar, Ngqumba: Multiple Factor Analysis and Principal Components Analysis with FactoMineR in exploratory data analysis of White wines		
Kapp, Litvine, Biesmans: Dating Financial Cycles - a Comparison of New and Existing Techniques	van Zyl, Lombard: Distribution-free cusums for location and scale			Burger, van der Merwe, Schall, Teise: Bayesian Statistical Inference for Ectoparasiticide Efficacy in Animal Trials	Uys: Income estimation		Azeez, Ndege, Odeyemi, Mutambayi, Obarami: A Mathematical model for transmission mechanism of TB/HIV co-infection			Adeyemi, Zewotir, Ramroop: Bayesian multinomial ordinal model to analyse the spatial patterns of childhood anaemia in Tanzania		Lubbe, Leclercq, de Vries: Multivariate analysis to find natural clusters of Tuberous Sclerosis Complex Associated Neuropsychiatric Disorders		
Gqwaka, Bretteny, Sharp: Efficiency Analysis of Electricity Distribution by South African Municipalities: A Cost Frontier Approach	Shongwe: Some important remarks regarding the zero- state and steady-state runs-type control charts			von Maltitz, Hen- Boisen, van der Merwe: Examining the impact of first-year developmental modules using the potential outcomes approach	Moolman: Variable selection advances in regression using R								Walk in Kistenbosch gardens	
													Gala dinner at Moyo's	
18:00-19:00														
19:00-22:00														



## Thursday 1 December 2016

8:40-10:20	Official statistics	General	General	Clinical trials and complex surveys	Multivariate Data Analysis
Chair:	Congo Sonali Das	ClubHouse Gina Joubert	Zambezi Ian Durbach	Kei Robert Schall	Observatory Niel le Roux
	Maremba: Methods of analysing complex sample survey data: Importance of taking sample design into account	Hlatshwayo, Chifurra, Chinhamu: Modelling Extreme Events for BRICS Exchange Rates	Muhinda, Pazvakawambwa: An evaluation of HIV Testing among women in the reproductive age-group in Namibia	9:00 start	9:00 start
	Luus, Neethling, de Wet: Simulating Complex Sampling Data for the evaluation of Survey-Weighted Linear Models	Kaombe, Manda: A review of influence measures in generalised linear mixed models	Muller: Discrepancies in the chi-square test and odds ratio confidence interval when analysing 2x2 tables	Chen: Statistical Assurance in Clinical Trial Design	Mostert, Coetzer, Lubbe: Evaluation of fault diagnosis techniques in multivariate statistical process control
	Neethling, Luus, de Wet: The Effectiveness of Weighting and Bootstrap in the Estimation of Model Parameters under Complex Sampling	Oluyede: A New Class of Distributions for Survival and Lifetime Data Analysis : Theory and Applications	Watson, Durbach, Stewart, Hendricks: Team key performance indicators in rugby union: A validation study	Lombard, Smuts, Faber: Modelling Growth in a Randomised Controlled Trial	Nienkemper-Swanepoel, le Roux, Lubbe: Multidimensional graphical comparisons of imputation methods for missing values in categorical data sets
	Oosthuizen, Neethling: Analysis of complex sample survey data with various software packages	Kleyn, Arashi, Bekker, Millard: Preliminary testing of the Cobb-Douglas production function and related inferential issues	Janse van Rensburg: Big Data opportunities and challenges in South African health sector	Nhapi Little, Kassanjee, Ward, Wessels: A GLMM analysis of data from the Sinovuyo Caring Families Program (SCFP)	Greenacre: At least 13 ways to define correspondence analysis
10:20-10:50	Tati: Addressing the deficits in urban statistics across Africa: challenges and opportunities with the 2010 and 2020 Population and housing census rounds for monitoring the SDGs urban core (Goal 11)?	Lima, Mwambi, Achia: Spatial hierarchical modeling of malaria in Angola	Stewart: Limitations to Probability Modelling in Strategic Decision Analysis and an Alternative Approach	Holloway, Duden-Throne: Statistical clustering of voting behaviour for election predictions	Tea / Coffee



## Thursday 1 December 2016

10:50-12:30	Process control	Energy	Education	Biostatistics	Multivariate Data Analysis
	Congo	ClubHouse	Zambezi	Kei	Observatory
Chair:	Albert Mijburg	Igor Litvine	Stefan Britz	Birgit Erni	Sugnet Lubbe
	Malela-Majika, Rapoo: Improved modified synthetic and runs-rules control charts combined with a Stewart Mann-Whitney chart	Bismans, Gnimassoun: Estimation of a demand electricity function for France	de Villiers: Leveraging the SAS Global Academic Program	Moiloj, Matlebjane, Maluleke, Kifile: Survival Analysis of Age Differences at Marriage and Time-to-Divorce in South Africa	le Roux, Bakk: Biplot-based visualizations of latent class models
	Malela-Majika, Chakrabarti, Graham: A comparison of the minimum and the precedence charts	Gqwaka, Brettenmy, Sharp: Efficiency Analysis of Electricity Distribution by South African Municipalities: A Cost-Frontier Approach	Diriba, Swanepoel: Improving the enrolment strategy in the Faculty of Economica and Management Sciences through an inquiry into the throughput rates of diverse enrolment and transfer streams	Karangwa, Blignaut: Using survival analysis to model the woman's waiting time to first birth after marriage in Rwanda	Papavarnavas, Fabris-Rotelli: Image denoising by principal component analysis using the Discrete Pulse Transform
	Contardo-Berning, Beukman, Ward: Statistical Process Control and the enhancement of water quality in the Western Cape	Mokilane: Density forecasting of long-term electricity demand in South Africa using Quantile Regression	Naudé: Mind the Gap: Are students prepared for higher education?	Kifile, Tsegaye: Survival Analysis of Under-Five Twins Mortality in Ethiopia using Semi-parametric Bivariate Frailty Models	Bierman, Steel: A multivariate regression approach to multi-label classification
	Muller, van den Heuvel, Stroomberg, Buydens, Jansen: Water quality of the river Rhine: Multivariate process monitoring with multi-block regression methods	Chanza, Khamfula, Mhlanga: Forecasting electricity consumption in South Africa using SARIMA models	van der Merwe, Mokhela: The Association of Statistics Anxiety, Attitude Toward Statistics and Mathematical Self-Concept with Performance in a Business Statistics Class	Lewis, Grobler, Kharsany, Cawood: Factors associated with home-based HIV counselling and testing in men in KwaZulu-Natal, South Africa	Pretorius, Bierman, Steel: A Bias-Variance Analysis of Ensemble Learning Classifiers



## Thursday 1 December 2016

Lindner, Auret: Application of Causal Analysis for Fault Diagnosis in Mineral Processes	MaPoswayo, Litvine: Subjective Measurements of Persistence of Time Series	Ehlers, Crafford, van Staden: Multilevel analysis of Grade 5 reading literacy	Singini: Diagnostics for joint models for longitudinal and survival data	Yilmaz, Aktas: Latent class models approach to the rater agreement
12:30-13:30	Loyiso Nongxa: Report back on the submission to Department of Higher Education and Training. Chair Danie Uys			
13:30-13:50		Plenary Trevor Hastie: Statistical Learning with Big Data Chair Francesca Little		
13:50-14:30		Closing & prizes for competitions: Danie Uys / Francesca Little		
14:30-15:00		Tea / Coffee		
15:00-15:20				
15:20-18:00		Mentoring workshop for NRF/SASA bursary holders		



## POSTER SESSIONS

### Poster session 1: Tuesday 29 November 2016

- Becker, Tolley, Altweeg, Measey, Slingsby: Estimating the global population size of a threatened species  
Bhgawandin, Er: Multivariate Extreme Value Theory with an Application to Climate Data  
Cervantes: A probabilistic kernel approach to analyse raptors flight trajectories observed from vantage points  
Dube: Species Distribution Modelling of Aloe Dichotoma  
Eastwood: Experimental Design for the Optimization of Microalgae Growth in Modified Cultivation Systems  
Mafuna, Makitla, Williams, North: The Effect of Cane Purity and Levels of Sucrose Extraction on Mixed Juice Purity  
Mathye: Analysing the Impact of Climate Change on Agricultural Productivity in South Africa  
Poognavanan: Uncovering the Latent Movement States for Predator State Space Model  
Stindt, Sharp, Mey: An introduction to indices used in model fit for structural equation models (SEM)  
van Niekerk, Goodall: The Influence of the Turning Angle in Hidden Markov Movement Modelling of Terrestrial Animals  
Watchurst, Goodall: Presence and Movement Modelling of Leopards

### Poster session 2: Wednesday 30 November 2016

- Akarawak, Adeleke, Olalude: Modelling health care data using Rayleigh-pareto distribution  
Ganey, Lubbe: Principal surfaces  
Lanyini, Litvine: Time Series models for Persistence in Volatility of Share Price  
Mbava, Sharp, Janse van Rensburg: Addressing the limitations of financial factor models  
Miyambu, Sseeletse: Numeric Measurement of BPO  
Mngcele: study on service industry macroeconomic variables and their effects on GDP  
Modibane, Strydom: Estimation of large dimensional covariance matrices  
Nkoana, Litvine: Causality between stock indexes and exchange rates: South Africa  
Özgül, Polat: Burnout Levels of Academic Staff: An Investigation at Hacettepe University in Turkey  
Pazi, Clohessy, Sharp: Statistical methods for the detection of non-technical losses: A case study for the Nelson Mandela Bay Municipality  
Sangqu, Litvine: Developing statistical indices for economic contribution of cultural and creative industries in South Africa



## POSTER SESSIONS



### Poster session 3: Thursday 1 December 2016

Awine: The malaria elimination agenda: Prospects for Ghana using mathematical tools

Deyzel, Clohessy, Brettenny: Non-Parametric Tolerance Intervals for the Assessment of Photovoltaic Energy Yield

Familusi: Analysis of Clustered Competing Risks with Application to a Multicentre Clinical Trial

Glass, Dalvie, Ramesar, Holtman: Genetic polymorphisms and organophosphate neurotoxicity amongst emerging farmers in the Western Cape

Joselson, Lubbe: Uses of Principal Components in Generalized Linear Modeling

Lesosky, Mukonda: Joint models for HIV viral load profiles and time to loss of viral control

Lewis, Kanfer: Exploring robust regression

Maharela, Fletcher: Survival analysis in higher education

Maposa, Blignaut: Bayesian Cox Proportional Hazards model and insights from the censored quantile regression model for paediatric and adolescent HIV/AIDS patients on antiretroviral treatment

Nwoko, Sameul: Implications of Incorrect Specifications of the Components of a Zero-inflated Poisson Model

Plumstead: Identifying Genetic Signatures using Phylogenetic Regression with a LASSO Penalty

Rodo: Characterising and Clustering TB Vaccine Candidates



## ABSTRACT - PLENARY SPEAKER



### Estimation and group variable selection for additive partial linear models with wavelets and splines

Anestis Antoiadis, University Joseph Fourier & University of Cape Town

Anestis.Antoniadis@imag.fr

Umberto Amato, CNR Napoli

Italia De Feis, CNR Napoli

Yannig Goude, University Paris Orsay & EDF France

**Abstract:** Additive partial linear models with nonparametric additive components of heterogeneous smoothness are studied. To achieve optimal rates in large sample situations we use hybrid P-splines and block wavelet penalization techniques combined with adaptive (group) LASSO-like procedures for selecting the variables in the linear part and the additive components in the nonparametric part of the models. Numerical implementations of our procedures for proximal like algorithms are discussed. Large sample properties of the estimates and of the model selection are presented and the results are illustrated with simulated examples and a real data analysis.

**Keywords:** additive partial linear models; wavelets; splines





## ABSTRACT - PLENARY SPEAKER



### Statistical Learning with Big Data

Trevor Hastie, Stanford University  
hastie@stanford.edu

**Abstract:** As our ability to gather and store data improves, we are faced with the task of analysing these ever-growing mounds of information. This has required Statisticians to gain computing and database skills, and Engineers and Computer Scientists to learn statistical modeling and data analysis. The result is a data scientist, one of the hottest job-descriptions in the tech world. In this talk I will give some examples of big data and data-science challenges, and explore some approaches in detail.

**Keywords:** big data; statistical learning



## ABSTRACT - PLENARY SPEAKER

### Some recent results in change point analysis

Marie Hušková, Charles University, Czech Republic  
marie.huskova@karlin.mff.cuni.cz

**Abstract:** The change point problem is usually treated by statistical procedures for detection of instabilities in statistical models. The problem is usually formulated in terms of hypothesis testing and estimation problem. Typically, we have observations  $X_1, \dots, X_n$  obtained at ordered time points and the basic task is to decide whether the model remains stable during the whole observational period or whether the model changes at some unknown point(s) or become generally unstable. In case of change(s) in the model being detected the further task is also to estimate the time of change and other parameters of the model in the periods where the model is stable.

Such problems are also called disorder problems or testing for presence of structural breaks (in econometrics) or testing for stability or segmented regression or switching regression in the regression setup. If all  $n$  observations are available at the beginning of the statistical analysis we speak about a retrospective setup. If observations are arriving sequentially and after each new observation we have to decide whether the observations obtained so far indicate an instability or not we have a sequential setup or monitoring. Originally, such problems were studied within statistical quality control, however nowadays there are many applications in various areas, e.g. medical research, econometrics, financial models, risk management, environmetrics, climatology. It brings a number of interesting theoretical problems.

Statistical procedures (tests and estimators) for detection of a change were developed applying various principles. As a motivation can serve the case of the known change point which leads to a variant of the two-sample problem in case of one change, while the  $k$  sample problem is related to more changes.

The talk will focus on detection of changes in regression models and time series models. Particularly, problems we present applications to meteorology (change in dependence of discharges on precipitations), medical research (two-sample change point problem, difference between sport efficiency of boys and girls), financial time series (detection of changes in capital assets prize models-CAPM). Of course, some related theoretical results will be included.

**Keywords:** change point; sequential setup



## ABSTRACT - PLENARY SPEAKER



### The Alter Ego of The Statistician

Yigit Karabag, SAS Institute

**Abstract:** In an era of fast paced change and constant disruption, there is one thing that all analysts agree on: The world needs more data scientists! But who are data scientists? What happened to statisticians, predictive modelers or analysts? How does one become a data scientist? Is it just a fancy title or does it require special skills? What does this mean for statisticians and what role will they play in the new era of advanced analytics where the technology landscape is changing rapidly? In this session, we will try to find answers to all these questions while taking a sneak peek at the future of analytics.



## ABSTRACT - INVITED SPEAKER



### Machine learning and Inference Challenges in Astronomy

Bruce Bassett, University of Cape Town  
Bruce.Bassett@uct.ac.za

**Abstract:** We review some of the recent progress in astrostatistics as well as key unsolved data science challenges in astronomy where statistical and machine learning input are required. The goal is to provide a gentle introduction to where statisticians can get involved.

**Keywords:** astrostatistics; machine learning



## ABSTRACT - INVITED SPEAKER

### Wildlife survey models: thinned spatial point processes with unknown thinning probabilities

David Borchers, University of St Andrews  
dlb@st-andrews.ac.uk

**Abstract:** Estimates of abundance and distribution of wildlife populations is crucial for conservation and management. Historically, these estimates tend to have been available only at the resolution of large spatial blocks, and conservation and management actions have been implemented at similarly coarse spatial scales. These may obscure important relationships between animal density and spatially-varying covariates that drive heterogeneous animal distribution.

Spatial point processes provide a natural statistical tool for modelling animal distribution continuously in space, and for uncovering the drivers of heterogeneity in animal distribution at fine spatial scales. However, the nature of wildlife surveys makes inference difficult. Wildlife survey data can be viewed as arising from a thinned spatial point process with unknown, spatially varying thinning probabilities. Wildlife survey methods have focused on estimating these thinning probabilities, sometimes on the basis of unrealistic assumptions about spatial distribution. Spatial modelling methods in ecology typically involve more realistic models for spatial distribution, but are not designed to deal with unknown, spatially-varying thinning probabilities. In this talk I formulate the two most widely used wildlife survey models (distance sampling and capture-recapture) as thinned spatial Poisson process or log Gaussian Cox process models, and demonstrate how the thinning probabilities can be estimated simultaneously with the spatial process parameters. I review recent developments in spatial modelling with distance sampling and capture-recapture data, illustrating with examples, and suggest directions that spatial modelling from such data might take in the near future.

**Keywords:** wildlife survey models; thinned spatial point processes



## ABSTRACT - INVITED SPEAKER



### At least 13 ways to define correspondence analysis

Michael Greenacre, Pompeu Fabra University  
michael.greenacre@upf.edu

**Abstract:** Correspondence analysis has a long and interesting 80-year old history, being discovered and re-discovered in different countries and different contexts. The version that is the most well-known is that of a technique of multivariate data visualization, a type of principal component analysis of categorical data, originating in the work of Jean-Paul Benzécri in France in the early 1960s, although the underlying theory can be traced back to the 1930s.

Each definition provides additional insight into how this fascinating method operates.

In this talk, I will take a simple cross-tabulation and demonstrate 13 different ways of defining correspondence analysis, for example as discriminant analysis, canonical correlation analysis, optimal scaling, multidimensional unfolding, and some tautologies such as correspondence analysis (CA) = co-correspondence analysis (CoCA) = canonical correspondence analysis (CCA). The more I think about this, however, the more I think that there are more than n=13 ways, so n may increase by the time this talk is given!

**Keywords:** correspondence analysis; multivariate data visualization



## ABSTRACT - CONTRIBUTED PAPER



### Bayesian multinomial ordinal model to analyse the spatial patterns of childhood anaemia in Tanzania

Rasheed Adeyemi, School of Mathematics, Statistics and Computer Science, University of KwaZulu-Natal, Pietermaritzburg campus  
adeyemira@yahoo.ca

Temesgen Zewotir, School of Mathematics, Statistics and Computer Science, University of KwaZulu-Natal, Westville, Durban campus

Shaun Ramroop, School of Mathematics, Statistics and Computer Science, University of KwaZulu-Natal, Pietermaritzburg campus

**Abstract:** This paper used self-reported data obtained from 2010 Tanzania Demographic and Health Survey. We propose a semi-parametric model that combines the fixed effects, non-linear term and spatial component in a unified framework. The fixed (linear) effects were modeled parametrically, and the non-linear effects of metrical covariates were modeled using P-splines. The spatial effect was modeled using a Markov random field prior. We explore multinomial logit models to analyse the severity of anaemia among under-five children and assess the risk factor of childhood anaemia. We run several Bayesian models via Markov Chain Monte Carlo(MCMC) simulation techniques and the models were compared using Deviance information criteria(DIC). We found the spatial residual pattern of anaemia and the risk determinants (factor). The risk factors associated with anaemia include place of residence, maternal poverty index, childhood under-nutrition, and vitamin A supplementation and infectious diseases. The findings also estimate non-linear function of continuous covariates(child's age and maternal body mass index (mbmi)) on childhood anaemia. Our method also estimate the residual spatial effects that are not captured by the underlying factors and produced probability predictive maps. Higher residual risk were identified in Northern-South of Tanzania. These spatial maps highlight high endemic regions that can assist government agency to target scarce health resource and effective policy making.

**Keywords:** Tanzania; childhood anaemia; Under-five children; Semiparametric model; Spatial effects



## ABSTRACT - CONTRIBUTED PAPER



### Quantifying the shape of social conversations

Simon Alger, TNS  
simon.alger@tnsglobal.com  
Kyle Findlay, TNS

**Abstract:** Social media has emerged as a critical platform for brand-related messaging in the digital age. Both the size and effectiveness of a brand's social presence have the potential to significantly affect a brand's power in the consumer's mind as well as brand revenues. Despite these opportunities, relatively little is understood about the actual impact of social campaigns. Many social media managers still judge success by simplistic yet opaque metrics such as number of likes, retweets and so on, which bear no clear connection to brand value or revenue. Here we present a methodology to quantify the nature of social conversations and aid in a more intuitive and useful understanding thereof.

By visualising interactions in the social space as edges in a user graph we can obtain a view of the nature of a brand's presence in a given social conversation, such as an ad campaign. While some conversations are centred on the brand's own account, other conversation networks evince organic, user-driven discussion. By introducing a means to quantify the tension between brand and user control in conversation networks, we illustrate our argument with several real brand campaigns and show how the shape of the network correlates with traditional advertising efficacy metrics. An important application of this methodology is real-time diagnostics for brands' social ad campaigns, including competitor analysis.

**Keywords:** social media; graph theory; market research



## ABSTRACT - CONTRIBUTED PAPER

### Occupancy models for species range dynamics

Res Altwegg, Statistics in Ecology, Environment and Conservation, Department of Statistical Sciences, University of Cape Town

res.altwegg@gmail.com

Guillaume Peron, Statistics in Ecology, Environment and Conservation, Department of Statistical Sciences, University of Cape Town

Kristin Broms, Colorado State University

Florent Bled, Statistics in Ecology, Environment and Conservation, Department of Statistical Sciences, University of Cape Town

James D Nichols, Patuxent Wildlife Research Centre

**Abstract:** Citizen science data allow ecologists to examine processes at large spatial scales. However, incomplete detection and variable sampling effort are two key issues for the analysis of such data. We used dynamic occupancy models adapted to bird atlas data for analysing patterns and processes in the distributions of bird species in southern Africa between 1987 and 2015. We found both clear legacies of past climatic conditions in the current species assemblages, and evidence that species ranges dynamically changed in apparent reaction to recent bush encroachment and changes in land use. Range changes led to opportunities for new ecological and evolutionary interactions between brood parasites and their host species. We also found evidence that pairs of closely related species competitively limit the range of one another. Finally, we found that protected areas are critical for the persistence of some species. Our work shows the value of citizen science projects for our understanding of large-scale biodiversity processes but also highlights the need to account for the variable observation process that affects such data sets.

**Keywords:** occupancy model; macroecology; range



## ABSTRACT - CONTRIBUTED PAPER



### Three Different Enhancements of the Generalized Estimating Equations Method in Handling Incomplete Longitudinal Binary Outcome

Omololu Aluko, University of KwaZulu-Natal

llaluko@yahoo.com

Henry Mwambi, University of KwaZulu-Natal

**Abstract:** This paper compares the performance of three techniques of analyzing incomplete longitudinal binary data when the missingness is due to dropout. It is assumed the response data are missing at random. We consider two modifications of the generalized estimation equations (GEE) based on inverse probability weighting (IPW) and multiple imputation (MI). The multiple imputation (MI) combined with GEE analysis is commonly known as MI-GEE. In the weighted (WGEE), we weight observations by the inverse of the probability of being observed. In the multiple imputation (MI), the missing observations are filled multiple times with the predicted values from the imputation model. The so-called doubly-robust (DR) technique combines the multiply imputed binary responses with IPW and then applying GEE to the completed data sets. A simulation study is first used to compare the performance of the methods followed by an application to a clinical trial data on amenorrhea.

**Keywords:** Missing at Random; Multiple Imputation; Inverse Probability Weighting; Doubly Robust



## ABSTRACT - CONTRIBUTED PAPER

### A Mathematical model for transmission mechanism of TB/HIV co-infection

Adeboye Azeez, University of Fort Hare

azizadeboye@gmail.com

James Ndege, University of Fort Hare

Akinwumi Odeyemi, University of Fort Hare

Ruffin Mutambayi, University of Fort Hare

Davies Obaromi, University of Fort Hare

**Abstract:** A novel SIRS Mathematical models that focuses on the underlying transmission mechanism of the TB and HIV co-infection is developed and analysed to help understand and predict the spread and progression of the two infectious diseases in different population. The models exhibit two equilibriums namely, the disease free equilibrium and the endemic equilibrium. It was established that if the basic reproductive number  $< 1$ , the disease free equilibrium is locally asymptotically stable using Castillo-Chavez theorem, which may not be globally asymptotically stable when  $< 1$ . If  $> 1$ , the endemic equilibrium exist which is locally asymptotically stable under certain conditions was investigated using Central Manifold Theorem. Numerical simulations suggest that the individual experiencing incident of HIV infections are at a risk of TB co-infection, compared with individuals without HIV infection. Also, the reduction of the effective contact rate of HIV can reduce the disease burden of TB co-infection. Controlling the transfer rate from HIV class to the AIDS class disease elimination is feasible. Controlling the transfer rate from the HIV exposed class to the HIV infected class disease control is also feasible.

**Keywords:** Co-infection; Equilibrium; Endemic; Jacobean theorem; Susceptible



## ABSTRACT - CONTRIBUTED PAPER

### Gaussian Processes in analyzing big data: An application with multilevel models

Jesca MerCy Batidzirai , University of KwaZulu- Natal

batidzirai@ukzn.ac.za

Samuel Manda, Medical Research Council of South Africa

Henry Mwambi, University of KwaZulu- Natal

**Abstract:** Big discrete longitudinal data that has multi- state events and competing risks are common in the health and demographic research. Such data has a hierarchical structure and observations are nested within individuals, hence multilevel models can be used to analyze such data. This study uses Gaussian Processes , in particular B- splines, to smooth out the time effect in estimating the time trend over longitudinal data. We will implement non- parametric interpolation in the discrete- time- to event analysis where there are many possible events that an individual may experience (competing risks). B-splines are going to be used where linearity is not assumed and L.O.C curves will be used for estimation of parameters.

**Keywords:** Competing Risks; B-Splines; Multilevel model; Gaussian processes



## ABSTRACT - CONTRIBUTED PAPER

### A Statistical First Passage Time Problem in System Replacement

Frank Beichelt, The University of the Witwatersrand  
Frank.Beichelt@wits.ac.za

**Abstract:** Abstract Functionals of the Brownian motion will be used to model the random cumulative repair cost  $C(t)$  caused by the maintenance of a technical system over a given time period  $[0, t]$ . Two basic maintenance policies are considered on conditions that a new system starts operating at time  $t = 0$ , the cost of a replacement is  $c$ , replacement times are negligibly small, and the maintenance-replacement process continues to infinity: Policy 1 As soon as  $C(t)$  reaches level  $x$ , the system is replaced by an equivalent new one.The efficiency of this policy is compared to the economic lifetime policy:Policy 2 The system is replaced at that time point  $t = t^*$  at which the expected total maintenance cost rate assumes its absolute minimum.Policy 1 is generalized by combining it with age-dependent replacement policies, which takes into account the advantages of both purely cost related and age-cost related maintenance policies.

**Keywords:** Brownian motion; first passage time; system replacement



## ABSTRACT - CONTRIBUTED PAPER



### Predictive Analytics & SPSS in the business world

Mike Bergh, OLRAC SPS  
mike@olsp.com

**Abstract:** The acronym SPSS started out as "Statistical Programme for the Social Sciences" more than 40 years ago, and is still regarded by many as a tool for social scientists, suitable for the desktop computer. What is perhaps less appreciated is that the well-known SPSS statistics tool is now the entry point into a battery of integrated enterprise level software products which have catapulted SPSS into the forefront of modern predictive analytics technology and the world of data science, playing a pivotal role in IBM's growth strategy. As a result SPSS is widely used for data driven decision making and predictive analytics in diverse government and business sectors including revenue services, insurance, telecommunications, banking, retail, medical aid administration and logistics. Specific applications include credit risk modelling, forecasting, market segmentation, optimisation, insurance claim segmentation, churn prediction, fraud detection, next best offer and crime and security analysis, many of which have already subsumed or threaten to subsume traditional professional activities. Therefore, universities will need to play an integral role in preparing students for this rapidly evolving new world of data science in which SPSS is a key competitor. This talk will elaborate on SPSS's position in the business environment and highlight some of the latest developments in the area of predictive analytics.

**Keywords:** predictive analytics; SPSS; business analytics



## ABSTRACT - CONTRIBUTED PAPER

### A multivariate regression approach to multi-label classification

Surette Bierman, Stellenbosch University  
surette@sun.ac.za  
Sarel J. Steel, Stellenbosch University

**Abstract:** Multi-label classification (ML classification) is an extension of binary- and multi-class classification to scenarios where several labels are associated simultaneously with each data instance. Examples of ML classification include directed marketing, functional genomics, and text/image annotation. In the literature, methods to perform ML classification are divided into three groups, viz. problem transformation approaches, algorithm adaptation approaches and ensemble methods. A comprehensive overview may be found in Madjarov et al. (2012).

In this paper we propose the use of a multivariate regression approach to ML classification. Following a regression approach offers the potential advantage of higher predictive accuracies and may facilitate investigation of other problems for ML scenarios, for example the problem of feature selection. The latter advantage stands in contrast to most ML classification approaches where interpretation of the relationship between features and labels is not addressed. We focus in our analysis on extensions of ordinary multivariate regression which account for potentially informative dependencies amongst labels. These include reduced rank regression (Izenman, 1975), filtered canonical Y-variate regression (Van der Merwe and Zidek, 1980), and curds-and-whey regression (Breiman and Friedman, 1997). ML classification by means of regression requires thresholding. We consider three approaches towards obtaining label-specific thresholds data-dependently. Regression for ML classification is empirically compared with state-of-the-art ML classification procedures, and found to perform competitively. A heatmap of the regression output highlights the advantage of a regression approach in terms of interpretation. Overall, we deem a multivariate regression approach to be a serious contender in ML scenarios.

**Keywords:** Canonical shrinkage; curds-and-whey regression; FICYREG; multi-label classification; multi-output regression; reduced rank regression.



## ABSTRACT - CONTRIBUTED PAPER



### Product of independent generalized gamma distributions

Vusi Bilankulu, University of Pretoria

vrbilankulu@gmail.com

Filipe Marques, Universidade Nova de Lisboa

Andriette Bekker, University of Pretoria

**Abstract:** The generalized gamma distribution has received much attention due to its flexibility and also for having some of well-known distributions as its special cases. This presentation entails a statistic defined as the ratio of products of independent generalized gamma random variables and shows that it can be represented as the product of independent generalized gamma random variables with some re-parametrization. By decomposing the characteristic function of the negative logarithm of the statistic and by using the distribution of the difference of two independent generalized integer gamma random variables as a basis, accurate and computationally appealing near-exact distributions are derived for the statistic. In the process, a new flexible parameter is introduced in the near-exact distributions which allows to control the degree of precision of these approximations. Furthermore, the performance of the near-exact distributions is assessed using a measure of proximity between cumulative distribution functions and, also, by comparison with the exact distribution, empirical distribution and with an approximation developed using a different method and which can only be applied in some particular cases.

**Keywords:** Generalized gamma distribution; Generalized integer gamma distribution; near-exact distributions; approximation



## ABSTRACT - CONTRIBUTED PAPER



### Estimation of a demand electricity function for France

Francis Bismans, University of Lorraine  
francis.biesmans@univ-lorraine.fr  
Blaise Gnimassoun, University of Lorraine/France

**Abstract:** This paper aims to estimate the electricity demand in the short and long term for France over the period 1990-Q1 to 2015-Q3. To this end, we first perform standard unit root tests and the Johansen multivariate co-integration tests. Then we rely on the DOLS method to robustly estimate the long-run relationship between electricity demand and its main determinants. Our results clearly validate the existence of a cointegration relationship. While the level of economic activity and the level of electricity prices are the main determinants of long-term electricity demand in France, the short-term variations of the latter depend on the variation of its price and the occurrence of economic recessions. These results are also confirmed by causality tests.

**Keywords:** Electricity demand; cointegration; causality; ECM



## ABSTRACT - CONTRIBUTED PAPER

### Fishery, population dynamics and stock assessment of geelbek (*Atractoscion aequidens*), a commercially important migrant fish species off the coast of South Africa

Danielle Winona Boyd, University of Cape Town

daniellewboyd@gmail.com

Henning Winker, South African National Biodiversity Institute

Res Altwege, University of Cape Town

Sven Kerwath, South African Department of Agriculture, Forestry and Fisheries

**Abstract:** Geelbek (*Atractoscion aequidens*) is an important species in South Africa's linefishery. Distributed from the south-west to the east coast, they are targeted throughout their range by the commercial linefishery, recreational anglers and small-scale fishers. The majority of geelbek are caught on the Agulhas Bank during austral summer. Due to current size limits of 60 cm, many are not yet mature, making the stock vulnerable to growth overfishing. Adults (>5 years, 90 cm) migrate seasonally to spawn off KwaZulu-Natal and congregate in offshore shoals. These spawning aggregations allow fishermen to catch adults extensively at night, making geelbek vulnerable to recruit overfishing. The current state of the geelbek stock is estimated to be a small fraction of their pristine state.

This study aims to improve understanding of the fishery and population dynamics of geelbek to help inform natural resource management of the geelbek linefishery. A stock assessment of South African geelbek was undertaken to fulfil this aim. Spatially explicit equilibrium and dynamic operating models were developed for geelbek, incorporating inter-annual coastal migration for all sub-populations. These models were developed using R, parameterised and calibrated with data from the National Marine Linefish System and established literature. Stock specific fisheries reference points and thresholds were obtained and used to test the probability of future stock sustainability over different harvesting regimes. Successful harvesting regimes were aligned with management scenarios to provide various operational targets to aid the South African Department of Agriculture, Forestry and Fisheries in generating a management evaluation framework for geelbek.

**Keywords:** Geelbek; *Atractoscion aequidens*; stock assessment; linefish; spatially explicit; age-structured equilibrium production model; age-structured dynamic production model



## ABSTRACT - CONTRIBUTED PAPER

# The Effect of the National Benchmarking Initiative (NBI) on the Efficiency of Water Service Provision in South Africa

Warren Brettenny, Nelson Mandela Metropolitan University

warren.brettenny@nmmu.ac.za

Gary Sharp, Nelson Mandela Metropolitan University

**Abstract:** The national benchmarking initiative (NBI) began in 2005 and ran for three years. The principle objectives of the NBI was to use benchmarking of water service authorities (WSAs) to improve performance, efficiency and sustainability of water service delivery in South Africa. The termination of the initiative in 2008 resulted in frustration amongst participating WSAs. A similar initiative, the Municipal Benchmarking Initiative replaced the NBI as of 2012. Owing to data limitations this initiative is not included in the current study. This study uses data envelopment analysis (DEA) and the Malmquist productivity index (MPI) to determine the effectiveness of the NBI in achieving its objective of improved efficiency of water service provision during the course of its implementation. The effect of the termination of the NBI is also investigated. The results of this study will inform decision makers on the benefit (or lack thereof) of such benchmarking schemes in the South African context.

**Keywords:** Efficiency; Benchmarking; Water Services; DEA; MPI



## ABSTRACT - CONTRIBUTED PAPER



# The Saddle-point Approximation and its Applications to the Hill Estimator

Sven Buitendag, Stellenbosch University  
svenbuitendag@gmail.com

**Abstract:** The saddle-point approximation is a highly accurate approximation of the distribution of a random variable. It was originally derived as an approximation in situations where a parameter takes on large values, but due to its high accuracy and good behaviour in a variety of applications not involving such a parameter, it has been generalised and applied to the distribution of any random variable with a well-behaved cumulant generating function. In this talk the theory underlying the saddle-point approximation will be discussed and illustrated with an application to approximate the distribution of the Hill estimator in extreme value theory.

**Keywords:** saddle-point approximation; saddle-point method; method of steepest descent; Laplace's method; Hill estimator; extreme value theory; sum of independent non-identical random variables



## ABSTRACT - CONTRIBUTED PAPER

### Bayesian Statistical Inference for Ectoparasiticide Efficacy in Animal Trials

Divan Aristo Burger, University of Pretoria  
divanaburger@gmail.com

Sean van der Merwe, University of the Free State

Robert Schall, University of the Free State

Chandr   Laverne Teise, University of the Free State

**Abstract:** In controlled animal trials of ectoparasiticides the efficacy of treatments is estimated based on the number of surviving parasites with which experimental animals have been infested. Guidelines for the conduct and analysis of animals trials published by international regulatory authorities require that the estimated efficacy of a test treatment (as determined by the Abbott estimator) should at least be 90%, for the treatment to be declared efficacious. This decision rule, therefore, is simply based on a point estimate of efficacy, and does not take into account the precision of the estimate, specifically, proper statistical inference on the efficacy of the test treatment in question is not required. In the proposed research project we investigate the use of appropriate statistical decision rules for the efficacy which control the Type 1 error. The statistical model for the data assumes a Beta-Binomial distribution which can accommodate the binomial over-dispersion typically associated with such data. A Bayesian approach based on integrated nested Laplace approximations will be explored for the analysis of ectoparasiticide efficacy data.

**Keywords:** Ectoparasiticide efficacy; Bayesian analysis; integrated nested Laplace approximation



## ABSTRACT - CONTRIBUTED PAPER



### Misconceptions of independent replications in agricultural experiments and the alternative application thereof

Frikkie Calitz, Agricultural Research Council - Biometry  
CalitzF@arc.agric.za

**Abstract:** In my work as a consulting Biometrist I regularly find that statistical replications are not done correctly. The purpose of this paper is to demonstrate how the assumption of independent replications are often not met when performing experiments. The experimental layout is often not statistically defendable as a result of the misunderstanding of the assumption of independent replications and the incorrect application of blocking.

One of the most common errors made in agricultural experiments is that experimental units are simply subdivided into sub-plots and then considered as experimental replications (also called random replications). Also composite treatment samples are collected from true replications and then sub-samples are drawn as replications.

Blocking in the wrong direction in field-, laboratory-, growth chamber- and green house experiments results in the within block variation being larger than the between block variation and therefore the treatment effect is not significant. Examples of the above will be demonstrated.

It is recommended that consulting statisticians must be aware of the incorrect application of replications and blocking. It is important that they familiarise themselves with the experimental layout and the data collection method before analysing the data.

**Keywords:** independent replications agricultural experiments



## ABSTRACT - CONTRIBUTED PAPER



# Forecasting Electricity Consumption in South Africa Using SARIMA Models

Martin Chanza, North-West University Mafikeng Campus

martin.chanza@nwu.ac.za

Yohane Khamfula, North-West University Mafikeng Campus

Isaah Mhlanga, Global Markets Research, Rand Merchant Bank

**Abstract:** Electricity consumption in South Africa has increased drastically and hence there is a drive by the electricity supply authority ESKOM to reduce consumption. In this study, we use SARIMA models to forecast electricity consumption in South Africa. A SARIMA(2,0,0)x(1,2,1)\_12 is fitted to electricity consumption data.

**Keywords:** ARIMA models; SARIMA models; Box Jenkins Method; Forecasting; Electricity consumption



## ABSTRACT - CONTRIBUTED PAPER



### Joint models for zero inflated nonlinear profiles

Tinashe Chatora, University of Cape Town

tdchatora@gmail.com

Francesca Little, University of Cape Town

Karen Barnes, University of Cape Town

**Abstract:** Gametocytes are the sexual form of the malaria parasite, responsible for the transmission of malaria. The gametocyte distribution is characterised as being zero inflated and right skewed, with a typically nonlinear longitudinal profile. This presentation looks at modelling the longitudinal profiles of gametocytes in the presence of censoring. A joint modelling approach, which combines longitudinal and time to event processes, will be used in this presentation. With respect to the longitudinal process, we will make use of hurdle models which combine a logit model for the presence/absence of gametocytes with a truncated normal and gamma distribution.

**Keywords:** Joint models



## ABSTRACT - CONTRIBUTED PAPER

### Statistical Assurance in Clinical Trial Design

Din Chen, Department of Statistics, University of Pretoria  
dinchen@email.unc.edu

**Abstract:** A well designed clinical trial requires an appropriate sample size with adequate statistical power to address trial objectives. The statistical power is traditionally defined as the probability of rejecting the null hypothesis with a pre-specified true clinical treatment effect. This power is a conditional probability conditioned on the true but actually unknown effect. In practice, this true effect is never fixed as a constant so a newly proposed alternative to this conventional statistical power is statistical assurance, by O'Hagan and Stevens (2001). The statistical assurance is a new paradigm in clinical trial design and is defined as the unconditional probability of rejecting the null hypothesis. It can then be obtained as an expected power where the expectation is based on the prior probability distribution of the unknown treatment effect, therefore it is a Bayesian concept. In this talk, we review the transition from conventional statistical power to assurance and discuss the computations of assurance using Monte-Carlo simulation-based approach.

**Keywords:** Statistical Power; Statistical Assurance; sample size; Clinical trials



## ABSTRACT - CONTRIBUTED PAPER



### Market risk spectrum performance on the Diebold-Mariano framework using statistical arbitrage

Siyabonga Chule, African Institute for Mathematical Sciences  
chule@aims.ac.za

**Abstract:** The statical arbitrage strategy which is the algorithmic market trading strategy of buying and selling assets to exploit the stastical mis-pricing. The arbitrage strategy is studies for the market risk performance. The associated risks viz. (i) Market spectrum risk, (ii) Timing risk,(iii) Performance accross spectrum risk. The forecasted reversion bounds of the market spectrum is analysed using the Diebold-Mariano (D-M) framework of the forecast errors. The D-M framework approach is employed in the market risk stastical arbitrage strategy is considered to evaluate the market performance accross the spectrum and to support active trading.

The strategy analysis is performed in the portfolio of asset prices in the currency market.

**Keywords:** Statical arbitrage; Mis-pricing; Diebold-Mariano framework; market spectrum performance



## ABSTRACT - CONTRIBUTED PAPER



### A Variational Bayes approach to the analysis of spatial site-occupancy models

Allan Ernest Clark, University of Cape Town, Centre for Statistics in Ecology, Environment and Conservation  
allan.clark@uct.ac.za

R. Altwegg, University of Cape Town, Centre for Statistics in Ecology, Environment and Conservation  
J.T. Ormerod, University of Sydney, ARC Centre of Excellence for Mathematical & Statistical Frontiers

**Abstract:** Detection-nondetection data are often used to investigate species range dynamics using Bayesian occupancy models which rely on the use of Markov chain Monte Carlo (MCMC) methods to sample from the posterior distribution of the parameters of the model. In this article we develop a Variational Bayes (VB) approximation to the posterior distribution of the parameters of a single-season spatial site occupancy model. This task is accomplished through the development of iterative algorithms that do not use MCMC methods. Simulations and a small practical example demonstrate the effectiveness of the proposed technique.

**Keywords:** logistic regression; species distribution model; variational Bayesian modelling





## ABSTRACT - CONTRIBUTED PAPER



### A robust construction of the efficiency frontier using resampling and shrinkage techniques

James combrink, University of Cape Town  
jamziec@gmail.com

**Abstract:** Statistical analysis within portfolio construction allows for previously inaccessible information into optimisation of performance to enhance the composition of a portfolio, increasing expected return and/or decreasing expected risk. Interrogation of the robustness within portfolio management was introduced by Michaud (1989), identifying that mean-variance optimisation methods are highly sensitive to sampling error, referring to the algorithms as 'error-maximisation'.

The inputs of these algorithms are (i) time-series expected returns (ii) cross-sectional behaviour and risk(covariance matrix) and (iii) biases of the statistician. This paper looks at identifying and removing avoidable risk though a statistical robustification of the algorithms, comparing the out-of-period results with standard industry results, which are highly sensitive and prone to sampling-error and extreme weightings.

The methodology looks at both a resampling efficiency frontier, under a covariance matrix with reduced bias from sampling error ('covariance shrinkage'). We use Monte-Carlo simulation techniques to reduce the sensitivity to the historical time-series anomalies. The work is adapted further to integrate the lack of certainty known of the true time-series after removing noise effects, in the eternal trade-off between sampling-error and specification-error.

**Keywords:** robust robustification efficiency frontier resampling shrinkage



## ABSTRACT - CONTRIBUTED PAPER

### Statistical Process Control and the enhancement of water quality in the Western Cape

Ivona Contardo-Berning, Stellenbosch University

ivona@sun.ac.za

Erika Beukman, Stellenbosch University

Christopher Ward, Stellenbosch University

**Abstract:** Traditional Statistical Process Control (SPC) techniques, such as the Shewhart control chart, Cumulative Sum (CUSUM) control charts and Exponentially Weighted Moving Average (EWMA) control charts, require that the underlying distribution of the quality characteristic of interest be known. Often, a normal distribution is required for these techniques to perform at desirable levels. In practice often no assumptions can be made regarding the underlying distribution. There are many variables with distributions that differ significantly from the normal distribution. Economic indices and other non-negative indices are typically skewed to the right, while the lifetimes of products can typically be described by Weibull distributions. Various nonparametric SPC (NSPC) control charts have been developed in order to address this problem. This study focusses on the univariate case. The goal of this study is to determine whether the parametric control charts are outperformed by their nonparametric counterparts. The monitoring of water quality can be done through the use of SPC. Water quality characteristics such as E.coli count are expected to come from a right-skewed distribution, therefore parametric SPC techniques might not perform as well as NSPC techniques.

**Keywords:** Statistical Process Control; Nonparametric Statistical Process Control; CUSUM; EWMA; Water Quality



## ABSTRACT - CONTRIBUTED PAPER



### Estimation of Multilevel Models with Iterative Generalised Least Squares

Gretel Crafford, University of Pretoria

gretel.crafford@up.ac.za

René Ehlers, University of Pretoria

**Abstract:** In this paper the focus will be to get a better understanding of multilevel analysis and the iterative generalised least squares (IGLS) procedure by making use of matrix methods. The basic philosophy of a two-level model will be demonstrated by simulating data with a specified two-level structure. This nested structure is a common phenomenon in especially the social and medical sciences where observations are grouped within certain levels. Explanatory variables will be introduced to accommodate the variation within and between levels. PROC IML in SAS will be used to simulate data and estimate the multilevel model.

**Keywords:** multilevel model iterative generalised least squares; nested structure; two-level model





## ABSTRACT - CONTRIBUTED PAPER



### Robust Portfolio Construction

Lara Dalmeyer, University of Cape Town

dlmlar001@gmail.com

David Bradfield, University of Cape Town

**Abstract:** This thesis investigates Regime Switching models and Asset Allocation in an African context. Regime switching models are used in conjunction with robust portfolio optimization methodologies to determine optimal asset allocations in identified regimes. The robust portfolio optimization methodologies include an extension of the Gols and Jones (2009) alpha weight angle in mean variance optimization, to VaR and CVaR optimization methodologies. Appropriate methods for improving the condition of asset covariance matrices are also addressed.

**Keywords:** Portfolio Optimisation



## ABSTRACT - CONTRIBUTED PAPER



### Leveraging the SAS Global Academic Program

Murray de Villiers, SAS Institute  
murray.devilliers@sas.com

**Abstract:** The SAS Global Academic Program serves more than 4000 degree granting institutions world-wide. Through this program, significant student and research success has been achieved by major local and international universities. In particular, the role of industry in academic learning is highlighted. This paper furthermore details the approach, engagement methodology and opportunities within this program. Specifically, the free SAS University Edition, SAS Academic Workshop Materials, SAS Curriculum Pathways, SAS Joint Certificate Program, SAS Data Science Graduate Program, SAS Certification as well as inter- and intra-university collaboration are addressed.

**Keywords:** SAS; analytics curriculum; industry engagement; data science; certification; university collaboration



## ABSTRACT - CONTRIBUTED PAPER



### Multivariate Dose-Response Profiling

Liesl de Vries, Hasselt University  
liesl.devries@quintiles.com

**Abstract:** This project was completed for a leading pharmaceutical company as part of the completion of my Master degree in Statistics Bioinformatics at Hasselt University, Belgium. The project is centered on the subject of analyzing dose response profiles in various settings. The company provided data from a pilot experiment, where measurements were collected for various test compounds, at various concentration levels and at various time points.

The objective was to gain an understanding of how the different compounds effect the gene expression data, in different concentrations and at various time points, i.e. to investigate firstly if the gene expression change after treatment, and secondly how the gene expression change after treatment. This is also known as measuring the pharmacokinetic/pharmacodynamics (PK/PD) biomarker potential.

Experiments can be very costly per subject in your sample size. This makes pilot studies very important as the discoveries of this study will provide insights which can direct the more extensive experiments to follow. For example: which drug compounds can be excluded, what dose levels are ineffective and also at which time point measurements should be taken.

**Keywords:** Bioinformatics; Dose-Response Profiling.



## ABSTRACT - CONTRIBUTED PAPER

### Improving the enrolment strategy in the Faculty of Economic and Management Sciences through an inquiry into the throughput rates of diverse enrolment and transfer streams

Tadele Akeba Diriba, Department of Statistics, University of Pretoria  
tadthstat@yahoo.com

Mr. Andre Swanepoel, Department of Statistics, University of Pretoria

**Abstract:** Students from diverse backgrounds increasingly enter higher education institutions. Universities also need to make sure that their programmes are responsive to these diversities. The university of Pretoria is one of the higher institutions in South Africa consisting of nine faculties. This study attempts to identify factors that could lead to improving the enrolment strategy in the Economic and Management Sciences faculty through an inquiry into the throughput rates of diverse enrolment and transfer streams. Therefore, the analysis investigates the factors leading to student success across various enrolment streams using General linear models. Among all the factors investigated, the effect of high school academic achievements appeared to be a major contributor to the academic success of students. The effect of gender, race, citizenship, admittance type and student's school type they attended, on the academic success of students was also examined in the study. The results of the analysis show that the first and second-year academic performance of female students is higher than that of male students. Similarly, academic performance of white students is higher than that of the other race groups except for Asian race group. Also, students with South African citizenship are more successful in their second year while it is the same in the first year as compared to the international students. The academic performance of students is also significantly different across varies admittance types and type of school which the students attended.

**Keywords:** General linear models; First-year GPA; second-year GPA



## ABSTRACT - CONTRIBUTED PAPER



### Random Effect Approach for Extreme Wind Speed in South Africa

Tadele Akeba Diriba, Department of Statistics, University of Pretoria

tadthstat@yahoo.com

Legesse Kassa Debusho, Department of Statistics, University of South Africa

**Abstract:** The aim of the extreme value analysis is to quantify and analyse the stochastic behaviour of extreme values. The estimation of the best appropriate distribution for extreme rainfall is done using the extreme value theory. In this paper, we have conducted random effect approach to the extreme value analysis using the GEV and GPD distribution. The annual maxima data and daily data used to the block maxima and threshold method respectively for modelling wind speed data at selected 10 weather stations in South Africa. The random effect model applied to the parameters of the GEV and GPD distributions to examine the variation that is inherent in the data. These seasonal and stations variation incorporated into the parameters of GEV and GPD has been estimated through the models allowing the sharing of information between stations and seasons. This estimation results in apparent advantages in terms of a greatly improved precision obtained in the estimation of the parameters of the models except for the scale parameter as compared to the maximum likelihood method.

**Keywords:** Random Effect; GEV; GPD; block maxima; Threshold



## ABSTRACT - CONTRIBUTED PAPER

### The benefits of using a continuous-time framework for spatial capture-recapture (SCR) models.

G Distiller, University of Cape Town

greg.distiller@uct.ac.za

D Borchers, University of St Andrews

**Abstract:** Spatial capture-recapture (SCR) models have become the standard approach used to estimate animal density from capture-recapture (CR) data. In traditional CR and SCR studies populations are sampled at discrete points in time leading to clear and well defined occasions but there are new types of detectors (such as camera traps) that sample populations continuously in time. A recently developed continuous-time SCR model treats detections as a temporal Poisson process and uses a detection hazard function to model the actual times of capture. Using a CT framework avoids having to impose an artificial construct on the data for analytical convenience, allows one to learn about animal behavior, facilitates a parsimonious and flexible way to model heterogeneity in detection, and leads to a likelihood for single-catch traps.

**Keywords:** continuous-time spatial capture-recapture; modelling animal activity



## ABSTRACT - CONTRIBUTED PAPER



### Individual identification of animals using acoustics

Gciniwe Dlamini, University of Cape Town

gci\_sims@yahoo.com

Ian Durbach, University of Cape Town

**Abstract:** The general problem being addressed in the research is the one of individual identification of animals using acoustic detection methods. The animals that are at the centre of the research are the moss frogs and crickets. Ultimately, the head count of individuals are to be used in subsequent studies to estimate the population density of animals, though this is not the focus of the research. The frog recordings are unlabelled, which dictates that an unsupervised learning approach should be taken to address the problem of individual recognition. On the other hand, the cricket recordings are labelled, which means that the problem is tackled using supervised methods. The overall system consists of several key components which include call segmentation, feature extraction and then lastly the machine learning process. Call segmentation entails isolating the parts of the recordings that contain the animals calls. From these calls, features that are commonly used in acoustic models are extracted. One of the features are the Mel-frequency cepstral coefficients (MFCC) which are widely used in human speaker and speech applications and in other animal recognition systems. The fundamental problem of recognising the individual animals is then tackled by applying machine learning techniques (either of the unsupervised or supervised kind), to the extracted call features.

**Keywords:** acoustic detection methods; individual identification; supervised and unsupervised learning



## ABSTRACT - CONTRIBUTED PAPER



### Differentiating fish stocks using image classification

Ian Durbach, University of Cape Town  
ian.durbach@uct.ac.za  
Tim Wolff-Piggott, University of Cape Town

**Abstract:** The management of a number of commercial fish species in South Africa depends on whether the species is viewed as a single homogenous stock, or as two largely independent stocks: a “Western” and a “Southern” population separated, roughly, by Cape Agulhas. We address this question by assessing whether images of otoliths - calcified structures in the inner ears of fish that exhibit stock-specific variation in shape - can be used to classify coastal origin. Otolith shape information is extracted by elliptic Fourier analysis, and the resulting coefficients used as features in machine learning classifiers. Fish length confounds the relationship between otolith shape and coastal membership. We show that existing standardization procedures retain some length information that can be exploited by machine learning approaches, and propose new resampling-based standardizations. We test these on three pelagic fish stocks found in South African waters.

**Keywords:** statistical ecology; machine learning



## ABSTRACT - CONTRIBUTED PAPER



### Multilevel analysis of Grade 5 reading literacy

Rene Ehlers, University of Pretoria

rene.ehlers@up.ac.za

Gretel Crafford, University of Pretoria

Surette van Staden, University of Pretoria

**Abstract:** Multilevel analysis is performed on the most recent data available from the Progress in International Reading Literacy Study (PIRLS) to identify and gain a better understanding of factors that are associated with reading literacy scores of Grade 5 learners in South Africa. The hierarchical structure of the data is accommodated by this type of analysis where learners are nested within schools. Two random intercept models are considered which allow for different perspectives, one where reading literacy score is entered as a continuous variable and the second where it is entered as a binary response variable. In the latter case the reading literacy score is assessed relative to the international centre point of 500 and the model is used to identify factors that are significantly related to the odds of obtaining a score greater than or equal to 500. Covariates considered are grouped into three main categories namely demographics, those used to describe Carroll's model of school learning and those for socio-economic status at both learner- and school level. Results from the two models are compared and differences discussed. The importance of considering different multilevel models to gain a better understanding of relationships is highlighted.

**Keywords:** Carroll's model of school learning; multilevel analysis; random intercept model.



## ABSTRACT - CONTRIBUTED PAPER



### Validity of the Psychological Capital Questionnaire in the South African context

Sebnem Er, University of Cape Town  
er.sebnem@gmail.com  
Chao Mulenga, University of Cape Town  
Kim Brouze

**Abstract:** In order to determine the validity of the Psychological Capital Questionnaire (PCQ) in the South African context, explanatory factor analysis (EFA) and confirmatory factor analysis (CFA) have been applied to two different datasets. Both Explanatory factor analysis (EFA) and Confirmatory factor analysis (CFA) were conducted using psych and Lavaan packages (Rosseel, 2011) for R open source package. EFA was performed using oblimin rotation for the polychoric correlation matrix and CFA was applied using weighted least squares (WLS) (asymptotically distribution free: ADF) estimation method to assess the internal consistency and factor structure stability.

**Keywords:** confirmatory factor analysis





## ABSTRACT - CONTRIBUTED PAPER



### Dynamic Factor Analysis for Ecological Time Series

Birgit Erni, University of Cape Town  
birgit.erni@uct.ac.za

**Abstract:** In dynamic factor analysis  $n$  time series are modelled in terms of  $m$ ,  $m \ll n$ , latent trends, cycles or seasonal effects. For ecological time series this is useful to identify groups of species or locations with similar trends. I will give a quick introduction to dynamic factor analysis (DFA), show how DFA is related to state-space models for time series, and illustrate the method on semi-annual waterbird count data.

**Keywords:** time series; dynamic factor analysis; counts



## ABSTRACT - CONTRIBUTED PAPER



### A point process model for DPT pulses

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

**Abstract:** The DPT is a decomposition of an image into pulses/shapes at various scales, providing a multi-scale representation. We look into fitting a non-homogeneous spatial point pattern to the pulses for pattern analysis.

**Keywords:** Image analysis; spatial statistics; pattern analysis



## ABSTRACT - CONTRIBUTED PAPER



### New bivariate gamma types with MIMO application

Johannes T Ferreira, University of Pretoria

johan.ferreira@up.ac.za

Andriette Bekker, University of Pretoria

Mohammad Arashi, Sharhood University of Technology, Iran

Jaco Visagie, University of Pretoria

**Abstract:** Bivariate gamma type distributions, including a bivariate Nakagami-m type, which originates from the diagonal elements of a complex inverse Wishart type distribution, are presented. The derived results can be applied as decision statistics for a multiple-input-multiple-output (MIMO) system with two transmit antennas. It is shown that, with the elliptical assumption, the outage probability of such wireless communication systems can be analysed in broad generality.

**Keywords:** Complex inverse Wishart; Complex t distribution; Elliptical; MIMO system; Nakagami-m; Outage probability.



## ABSTRACT - CONTRIBUTED PAPER

### Using Species Distribution Models for spatial conservation planning of African penguins

Frieda Geldenhuys, Stellenbosch University, SACEMA  
Frieda.Geldenhuys@gmail.com

**Abstract:** The African penguin is endemic to South Africa (breeding only in South Africa and Namibia). Current colonies are under threats posing a serious issue to conservation and eco-tourism. Species such as these needs to be conserved otherwise might face the threat of extinction. The penguins are also important as they play the role of an early warning system for environmental threats. By global standards, a population is considered unhealthy and in danger if it decreases to 10 percent of the former (pre-exploitation / decrease) levels. The African penguin population is currently at about 14 percent its 1950s level, when the first official census was conducted and is still on a strong downward population trajectory. Some of the drivers of change includes climate change, parasites, pollution (oiling), disease, food resources, predation risk and habitat interference. A huge component to it is the anthropogenic, human impact, especially with population expansion. A species distribution model, including a suitability map for the distribution of the penguins and modelling of the meta-community dynamics of regional colonies would be established in order to attend to these matters. It is a predictive, conceptual model of the abiotic (e.g. physical barriers, climate, lack of resources) and biotic (e.g. competition, predators, parasites) factors influencing habitat suitability controlling species distributions in space, time and scale. Other names such as habitat suitability, niche modelling, bioclimatic models, resource selection functions and spatial correlation models are used to describe the species distribution models.

**Keywords:** African penguins; conservation; species distribution model; suitability map



## ABSTRACT - CONTRIBUTED PAPER

### Algorithmic Construction of Optimal Block Designs for Two-Colour cDNA Microarray Experiments Using the Linear Mixed Effects Model

Dibaba Bayisa Gemechu, University of South Africa  
diboobayu@gmail.com

Legesse Kassa Debusho, University of South Africa  
Linda M. Haines, University of Cape Town

**Abstract:** In this study, methods for efficient construction of A-, MV-, D- and E-optimal or near-optimal block designs for two-colour cDNA microarray experiments with array as the block effect are considered. Two algorithms, namely the array exchange and treatment exchange algorithms together with the complete enumeration technique are introduced. The three methods are compared on the basis of computer time required to find the optimal design and the efficiency of the optimal designs. For large numbers of arrays or treatments or both, the complete enumeration method is highly computer intensive. The treatment exchange algorithm computes the optimal or near-optimal designs faster than the array exchange algorithm. The three methods however produce optimal or near-optimal designs with the same efficiency for the four optimality criteria considered in this study.

**Keywords:** Microarray experiment; A-, MV-, D- and E-optimal designs; complete enumeration; array exchange algorithm; treatment exchange algorithm.



## ABSTRACT - CONTRIBUTED PAPER



### Latent class mixed models for classification of HIV viral load profiles in the Western Cape public sector antiretroviral treatment programme: 2009 - 2015

Tracy Glass, University of Cape Town  
tracy.glass@uct.ac.za

Landon Myer, University of Cape Town

Nei-yuan Hsiao, National Health Laboratory Service  
Maia Lesosky, National Health Laboratory Service

**Abstract:** Viral load (VL) monitoring is an important aspect of long term care of HIV-infected individuals on antiretroviral therapy (ART). Centralised laboratory systems provide comprehensive data, allowing for analysis of viral load profiles in a population. Classification of these profiles provides important information on programme and community level success in ART provision. Multiple methods of statistical classification for longitudinal data are applied and resulting model determined clusters summarised and mean VL trajectories estimated. We reviewed the HIV viral load data from 2009-2015 tested at the two National Health Laboratory Service laboratories in the Western Cape, South Africa. These data represent a catchment area consisting of virtually all of the public sector ART clinics and hospitals in the Province. We defined virological suppression as any viral load <1000 copies/mL and analysed the VL trajectories of all adults (> 16 years of age) that had at least 3 VL tests at any time in the follow up period. We classified and described the types of viral load trajectories and fit latent class mixed models with non-standard errors to classify the longitudinal profiles. Over 120,000 individuals with more than three VL measures were identified, including more than 600,000 assays of which 84% were virologically suppressed. Five distinct VL classes were recognised: consistently elevated VL (no suppression), suppression with followup, suppression without followup, suppression followed by viral blips and suppression followed by viral rebound. Latent class mixed models were fit and contrasted.

**Keywords:** latent class; longitudinal; HIV; viral load



## ABSTRACT - CONTRIBUTED PAPER



### Use of mixture models for assessing the size structure of long-lived plants in a temporally varying environment

Victoria Goodall, Nelson Mandela Metropolitan University

victoriagoodall@gmail.com

Tim O'Connor, South African Environmental Observation Network

**Abstract:** Assessing the size structure of a plant population allows for rapid inference on possible future population performance, commonly required for management or conservation action. Current approaches rely on construction of a size class distribution by which information is lost by categorising individuals. Current analytical methods assume an ideal size structure follows a negative exponential distribution, the observed population is compared to this. A negative exponential distribution is not an expected structure for plant populations in a temporally varying environment. An alternative approach is based on independent mixture models using a log-normal distribution. A mixture of distributions are fitted to the observed plant sizes, with each distribution corresponding to an underlying latent state. Interpretation is based on the pattern and size of states and not on an individual summary statistic. The mixture model approach is illustrated using tree populations responding to rainfall variability in a semi-arid environment or to variation in fire regime in a moist environment for both relatively short- or long-lived species. Mixture models offer a statistically efficient means and a robust approach across species and environments of assessing the size structure of plant populations.

**Keywords:** independent mixture models; log-normal distribution; semi-arid savanna; woody plant



## ABSTRACT - CONTRIBUTED PAPER

### Efficiency Analysis of Electricity Distribution by South African Municipalities: A Cost Frontier Approach

Aviwe Gqwaka, Nelson Mandela Metropolitan University

s210137045@nmmu.ac.za

Warren Brettenny, Nelson Mandela Metropolitan University

Gary Sharp, Nelson Mandela Metropolitan University

**Abstract:** In South Africa, local government is mandated to provide electricity services to citizens. In recent times, the ability of the municipalities to fulfil this obligation has come under scrutiny. To remedy this, the efficiency with which municipalities distribute electricity, taking into account the costs that they incur within that process, requires evaluation. This study assesses the cost efficiency with which a cross section of 121 municipalities in 2005 and 108 municipalities in 2006 distribute electricity. This is achieved using an estimable stochastic cost function that models the frontier (theoretical best practice), to which each municipality's performance is compared. Municipalities whose performances are not optimal deviate from the frontier, where these deviations (errors) are assumed to be caused by two factors, namely stochastic random noise and cost inefficiency. This composition of the errors accounts for effects for which municipalities cannot control (stochastic random noise) and those for which it can (cost inefficiency). Distribution theory is then imposed on these factors as it is assumed that stochastic random noise is always normally distributed, whereas cost inefficiency is assumed to be one-sided and positively skewed. The study assesses cost inefficiency by using five one-sided, positively skewed distributions, namely the half normal, exponential, truncated normal, gamma and Rayleigh distributions. This is to determine whether cost efficiency estimates are robust across all distributions. Furthermore, the selection of the most suitable distribution to model cost inefficiency is considered and explored using the Vuong test.

**Keywords:** efficiency analysis; stochastic cost frontier; composed error; electricity distribution; Vuong



## ABSTRACT - CONTRIBUTED PAPER



### A comparison of the minimum and the precedence charts

Marien Graham, University of Pretoria

marien.graham@up.ac.za

Subhabrata Chakraborti, University of Alabama

Jean-Claude Malela-Majika, University of Pretoria

**Abstract:** Development of control charts that do not require a particular distributional assumption is desirable in practice. Nonparametric or distribution-free control charts can serve this purpose. Chakraborti, Van der Laan and Van de Wiel (2004) proposed a class of nonparametric Shewhart-type control charts, called the precedence charts, using some order statistic of a Phase II sample as the charting statistic and control limits constructed from a Phase I reference sample. Albers and Kallenberg (2008) proposed a similar nonparametric Shewhart-type control chart where the charting statistic is the minimum of a Phase II sample. A comparison between the minimum chart and a precedence chart is done through simulation. Following this, we enhance their performance by adding some well-known runs-rules. Theoretical properties are studied and the in- and out-of-control performances of the charts are examined. A summary and some concluding remarks are given.

**Keywords:** Nonparametric; Order Statistics; Improved Runs Rules; Improved Modified Runs Rules



## ABSTRACT - CONTRIBUTED PAPER

### Variance component estimation in linear mixed models using data augmentation

Freedom Gumedze, University of Cape Town  
freedom.gumedze@uct.ac.za

**Abstract:** The fitting of a linear mixed model requires the estimation of variance parameters, prediction of random effects, and estimation of fixed effects. This process can be very computationally demanding for very large and complex datasets with many random effects. In terms of REML estimation, these large datasets create complex models due to the highly dimensional matrices that must be inverted. The computer algorithms that are currently employed to such models are slow or are even unable to produce the reliable variance component estimates. In this paper we discuss data augmentation (imputation) as method of REML estimation of variance components in linear mixed models.

**Keywords:** Variance component estimation; REML; linear mixed model; data augmentation



## ABSTRACT - CONTRIBUTED PAPER



### Maximum Likelihood: A Problem in Ecology

Linda Haines, University of Cape Town  
linda.haines@uct.ac.za

**Abstract:** N-mixture models for estimating population size and the probability of detection of animals in a closed population from repeated counts data were introduced into the literature by Royle in 2004 and are widely used in the area of statistical ecology. The focus of this talk is on the nature of the likelihood associated with such N-mixture models. Specifically, it is shown that the infinite sum embedded in the likelihood associated with the Poisson mixing distribution can be expressed in terms of a hypergeometric function and, thence, in closed form. The resultant expression for the likelihood can be readily computed to a high degree of accuracy and is algebraically tractable. In particular, the likelihood equations can be simplified to some advantage, the concentrated likelihood in the probability of detection formulated and problematic cases identified. The results are illustrated by means of a simulation study and a real world example.

**Keywords:** repeated count data; maximum likelihood; hypergeometric functions





## ABSTRACT - CONTRIBUTED PAPER



### The Performance of Non-Informative Priors for the Bivariate Lognormal Distribution

Justin Harvey, University of Stellenbosch

jharvey@sun.ac.za

Abrie van der Merwe, University of the Free State

**Abstract:** Lognormally distributed data presents itself in a number of scientific fields. The setting for this talk is that of dependent samples from two lognormal distributions, without the possibility of zero values. The aim is to develop Bayesian procedures for computing confidence (credibility) intervals for the ratio of the means of the bivariate lognormal distribution. The same procedures can also be used to obtain credibility intervals for the ratio of variances. The choice of prior distributions is the factor of interest. Specifically the choice of different prior distributions in different parameter settings and the appropriateness of each is of primary importance. The credibility intervals will be compared with the generalized confidence intervals approach (GCI) used by Bebu and Mathew (2008) and the "method of variance estimates recovery" (MOVER) proposed by Zou et al (2009a). Results of a simulation study will be presented, together with a practical application.

**Keywords:** Bayes; Non-informative Priors; Bivariate Lognormal



## ABSTRACT - CONTRIBUTED PAPER



### The analysis of a grouped response variable in a multifactor design

Johannes Jurgens Hendriks, University of Pretoria

hendriks.jj2@gmail.com

Gretel Crafford / University of Pretoria

**Abstract:** Variables that are continuous in nature are often recorded in grouped format. This may prohibit the researcher to make use of multivariate regression techniques and valuable information in the data may be lost. In this paper the 10% sample of the South African Census 2011 will be used to explain how the grouped response variable INCOME may be modeled to explain the effect of GENDER, POPULATION GROUP, EDUCATION and AGE on INCOME. A cross tabulation of certain explanatory variables will lead to a multifactor design with T so-called cells, each containing a frequency distribution of income. Log-logistic distributions will be fitted to the T cells where the median income in the T cells will adhere to a specific model to explain the effects and trends with regard to the explanatory variables using maximum likelihood estimation under constraints.

**Keywords:** grouped response variable; maximum likelihood estimation under constraints; log-logistic distribution; multifactor design



## ABSTRACT - CONTRIBUTED PAPER

### Modelling Extreme Events for BRICS Exchange Rates

Gugu Hlatshwayo, University of KwaZulu Natal

mhaise3@yahoo.com

Retius Chifurira, University of KwaZulu Natal

Knowledge Chinhamu, University of KwaZulu Natal

**Abstract:** Extreme events also known as Black swan events, have a suicidal feature when they occur, this has increased the interest of academics and risk managers to study the extremes behavior of financial markets and discover how to manage risk during those events. This study examines the extreme events of the BRICS (Brazil, Russia, India, China, South Africa) daily exchange rate returns using the extreme value theory (EVT). In this study, we compare the relative performance of the EVT and the ARMA-GARCH-EVT models in estimating value at risk (VaR) of the BRICS exchange rate returns. The EVT framework captures the non-normality feature while the ARMA-GARCH framework captures the volatility feature of the daily BRICS exchange rate returns. The VaR estimates of the proposed models are backtested via the Kupiec likelihood ratio test to check for model adequacy and identify the robust model. The main findings indicate that the ARMA-GARCH-EVT models give better result, when compared with EVT models. The ARMA-GARCH-EVT model offers an alternative modelling tool for risk assessment and management.

**Keywords:** ARMA-GARCH; EVT; BRICS; Kupiec; Exchange Rates



## ABSTRACT - CONTRIBUTED PAPER



### Statistical clustering of voting behaviour for election predictions

Jenny Holloway, CSIR  
jhollowa@csir.co.za  
Nontembeko Duden-Tlhone, CSIR

**Abstract:** Statistical clustering of the South African voting population into segments with similar voting patterns is one of the major building blocks for election night predictions. This type of election prediction is based on the results of the first voting districts released and takes place while vote counting is still underway in the rest of the voting districts. The prediction model was developed by the CSIR and first used in the prediction of the 1999 national and provincial election results and has since been applied in the subsequent general and local government elections, including the recent 2016 local government elections. Due to the non-random nature of the voting behaviour of the electorate, which is influenced by political association, social and economic factors and previous voting history, among other reasons, the application of clustering is very relevant and the need for good (representative) clusters is critical to the accuracy of the predictions.

In this paper we present the clustering approach that we use to model election results. We also share some insights into the major considerations regarding the construction of clusters, the assumptions we make, as well as their impact on the successful prediction of the outcome of an election. We also present the results from the recent 2016 local government elections.

**Keywords:** Clustering; elections; prediction



## ABSTRACT - CONTRIBUTED PAPER



### A conditional extreme value approach to value-at-risk estimation with exchangeable innovations

Chun-Kai Huang, University of Cape Town & University of KwaZulu-Natal  
chun-kai.huang@uct.ac.za

Delia North, University of KwaZulu-Natal

Temesgen Zewotir, University of KwaZulu-Natal

**Abstract:** Extreme value theory (EVT) is commonly used for evaluating risk in financial returns. In particular, it can be amalgamated with a GARCH model, where the peaks-over-threshold (POT) method is applied to the innovations. However, this GARCH-EVT approach relies on the assumption that the innovations are independent and identically distributed. To relax this assumption, we generalise the POT method to exchangeable sequences. We apply this new approach, with the GARCH filter, to forecast one-day-ahead Value-at-Risk estimates.

**Keywords:** Extreme value; Exchangeability; Value-at-Risk; GARCH; Stationary bootstrap.



## ABSTRACT - CONTRIBUTED PAPER



### Sampling variance of a multi-phase stratified design on partially exchangeable sequences

Chun-Kai Huang, University of Cape Town & University of KwaZulu-Natal  
chun-kai.huang@uct.ac.za

Delia North, University of KwaZulu-Natal

Temesgen Zewotir, University of KwaZulu-Natal

**Abstract:** In estimating the population mean of a study variable  $y$ , we can often use a ratio-type estimator when a related auxiliary variable  $x$ , with improved accessibility, is available. In cases where  $x$  is qualitative, or may be categorised, and a double sampling plan is used, we may consider a two-phase stratified sampling design. Traditionally, it is assumed that the  $N$  variables representing the readings on  $y$  are IID within and across strata. In this paper, we relax this assumption to a judgment of exchangeable sequences within each stratum, while still maintaining the assumption of independence across strata. We propose a methodology for estimating the variance of the ratio estimator under this scenario.

**Keywords:** Multi-phase sampling; Stratification; Exchangeability; Auxiliary variable; Ratio.





## ABSTRACT - CONTRIBUTED PAPER



### A Comparative Analysis of Aggregational Gaussianity for JSE-listed Shares and Indices

Ryan Kruger, University of Cape Town

chun-sung.huang@uct.ac.za

Kanshukhan Rajaratnam, University of Cape Town

Chun-Sung Huang, University of Cape Town

**Abstract:** A topic of recent interest is whether returns on the Johannesburg Stock Exchange (JSE) exhibit aggregational gaussianity (AG). We examine index and share returns from the JSE for evidence of AG with a particular focus on whether the results differ between (a) index and share samples and (b) shares and indices of different sizes as defined by market capitalisation. Our findings suggest that TOP 40 index returns are normally distributed at monthly intervals. However, evidence of AG in the share sample is weak overall, with a greater incidence of AG in large-cap shares relative to mid- and small-cap shares.

**Keywords:** aggregational gaussianity; asset pricing; stylised fact; Johannesburg Stock Exchange; Shapiro-Wilk test.



## ABSTRACT - CONTRIBUTED PAPER

### The second Southern African Bird Atlas Project: causes and consequences of geographical sampling bias

Sanet Hugo, Centre for Statistics in Ecology, Environment and Conservation, Department of Statistical Sciences, University of Cape Town

sanethug@gmail.com

Res Altweegg, Centre for Statistics in Ecology, Environment and Conservation, Department of Statistical Sciences, University of Cape Town

**Abstract:** Since commencing in 2007, the ongoing second Southern African Bird Atlas Project (SABAP2) aims to record bird species within seven southern African countries at a five minute spatial resolution. Thus far, all of the more than 8.5 million records were contributed by volunteering citizen scientists. Sampling effort (e.g. number of surveys per grid cell) is highly dependent on when and where volunteers prefer to sample, resulting in a pronounced spatial bias in sampling effort, and spurious species distribution patterns. Species distribution and occupancy models are increasingly used to mitigate such observation bias. However, these techniques are most effective if based on repeated sampling at sites representing the full range of environmental variation (e.g. climate) in the region of interest. We examined the possible causes and consequences of spatial variation in sampling effort for SABAP2. Focusing on South Africa, we used linear regression, taking spatial autocorrelation into account, to show that spatial variation in sampling effort is best explained by distance from main city or popular birding area, distance from major road, and protected area cover. Further, we examined geographical zones with similar environments (based on climate and vegetation) by fitting species accumulation curves to the asymptotic Lomolino model. We found that environmental zones vary in terms of how well they are represented by the currently available data, and in terms of how much additional sampling is required. Building on our findings, atlas coordinators may improve SABAP2 by directing volunteers to undersampled regions, while taking into account their preferences.

**Keywords:** Citizen science; linear regression; spatial observation bias; species distribution atlas; species accumulation curves.



## ABSTRACT - CONTRIBUTED PAPER



### Fiducial inference for location-scale and log-location-scale families of distributions: Application to the Weibull distribution two-sample problem

Peter Iiyambo, University of Namibia  
ptiyambo@gmail.com  
Robert Schall, University of the Free State

**Abstract:** The Weibull distribution has many applications in engineering to model lifetime data. Inference for comparing the parameters and quantiles of location-scale and log-location-scale families of distributions is usually based on the maximum likelihood based approximate methods. However, parameter estimation using maximum likelihood can be difficult and may require extensive programming. This study compares exact, rank-based fiducial generalized confidence intervals (FGCIs) for the ratio of scale parameters, difference of location parameters and difference of quantiles of two Weibull distributions, with maximum likelihood-based FGCIs. The results show that ranked-based FGCIs are very close to the maximum likelihood-based FGCIs.

**Keywords:** Fiducial inference; Location-scale distribution; Log-location-scale distribution; Two-sample problem; Weibull distribution



## ABSTRACT - CONTRIBUTED PAPER

### Modelling South Africa's market risk using the aparch model and heavy-tailed distribution

Yetunde Ilupeju, University of KwaZulu-Natal

ilupejuyetunde@yahoo.com

Retius Chifurira, University of KwaZulu-Natal

Knowledge Chinahmu, University of KwaZulu-Natal

**Abstract:** Data analysis of stock returns, especially from emerging economies has recently attracted attention mainly because stock returns are relatively more volatile than its historical trend. In this study, we compare the relative performance of APARCH combined with Pearson Type IV distribution, Generalized Pareto Distribution (GPD), Generalized Extreme Value Distribution (GEVD), and stable distribution in estimating Value-at-Risk of South Africa JSE All share index returns. Model adequacy is checked through the backtesting procedure. The proposed models are able to capture volatility clustering, and asymmetric effect (leverage effect) through the APARCH and the heavy-tailedness through the heavy-tailed distributions. The main findings indicate that APARCH model combined with the Pearson Type IV distribution gives better results when compared in term of their predictive ability at the probability levels for long positions and high probability levels for short position. Thus, APARCH model combined with Pearson Type IV distribution model provides a good alternative for modelling stock returns.

**Keywords:** All share index (ALSI); Asymmetric power ARCH (APARCH); Generalized Extreme Value Distribution (GEVD); Kupiec test; Pearson Type IV distribution; stable distribution; Value-at-Risk (VaR).



## ABSTRACT - CONTRIBUTED PAPER

### Big Data opportunities and challenges in South African health sector

Charl Janse van Rensburg, South African Medical Research Council  
charl.jansevanrensburg@gmail.com

**Abstract:** Big Data is shaping the world we live in. In various places Big Data has been utilised in the area of medical science and public health. In 2012 the US government invested \$200 million in research of Big Data, a part of which was invested in health research specifically. The health media collaboratory in the US has practically shown the use of Big Data using one of the big drivers, social media to assess how marketing on social media influences health. The uptake of Big Data and data science in the developing world, including South Africa is yet to be taken seriously. This study presents an overview of Big Data's current uses and then identifies possible opportunities and challenges in the health sector in South Africa.

**Keywords:** Big data; health sector; South Africa



## ABSTRACT - CONTRIBUTED PAPER



### GAS models of financial time series

Stefan Janse van Rensburg, Nelson Mandela Metropolitan University  
sjansevanrensburg@gmail.com  
Gary Sharp, Nelson Mandela Metropolitan University

**Abstract:** The class of generalised autoregressive score (GAS) models offer a unified framework for developing time series models with time-varying parameters. GAS models accommodate heavier-than-normal tails and are well-suited to modelling financial time series. This paper considers several specifications of the GAS model. An application to equities listed on the Johannesburg Stock Exchange is provided.

**Keywords:** DCS; GAS; GARCH; dynamic conditional score; generalised autoregressive score; volatility



## ABSTRACT - CONTRIBUTED PAPER



### Characterization of bias in misspecified power law distribution

Silibelo Kamwi, University of Namibia

ikamwi@unam.na

Chris Koen, University of the Western Cape

Lawrence Kazembe, University of Namibia

**Abstract:** The power law distribution is often considered in many man-made and natural phenomena, where the value of observations varies inversely as a power of the measured quantity. The problems of fitting power law distributions to data assuming no measurement errors using the method of maximum likelihood are well documented. Not so well documented, however is, the level of bias that arises when there is misspecification between the homoscedastic and heteroscedastic fitted power law distributions. Although the power law distribution with homoscedastic measurement error is common, it has been posited in a number of settings that the measurement errors might depend on the value of the variable being measured and hence the need to fit the heteroscedastic power law distribution. In this article, we consider the quasi-maximum likelihood estimator (QMLE), in estimating the parameters with standard errors for the heteroscedastic power law distribution under model misspecification. Carried out as well is the asymptotic characterization of the bias arising from such model misspecification. Simulated examples are undertaken to describe and assess the finite sample performance of the estimation method.

**Keywords:** Misspecification; bias; parametric deconvolution; parameter estimation; Pareto distribution





## ABSTRACT - CONTRIBUTED PAPER



### Mixture of mixed effects regression with t distributed error and random components

Frans Kanfer, University of Pretoria

frans.kanfer@up.ac.za

Sollie Millard, University of Pretoria

**Abstract:** Dependent measurements within subjects are widely modeled using linear mixed models. Subjects are often grouped within latent structures. The talk consider finite mixture linear mixed models, identifying latent groups and simultaneously estimating a linear mixed model for each latent group. Multivariate t distributions is assumed for the error and random effects for the latent group, leading to more robust estimates as compared to the normal case. A conditional EM estimation algorithm is considered with alternative approaches for estimating the degrees of freedom. Results from simulated and real data are discussed.

**Keywords:** Mixed effects regression; multivariate t distribution; EM algorithm



## ABSTRACT - CONTRIBUTED PAPER



### A review of influence measures in generalised linear mixed models

Tsirizani Kaombe, University of Malawi - Chancellor College, Maths Dept

tsirizanikaombe@yahoo.com

Samuel Manda, South African Medical Research Council, Biostatistics Research Unit

**Abstract:** A number of diagnostic measures for model fit in standard linear models are well-known and widely implemented in statistical software. However, in the context of generalised linear mixed models these measures are not as standard. This paper discusses extensions of the various measures to assessing influential observations when fitting generalised linear mixed model.

**Keywords:** multilevel models; outliers; influential observations



## ABSTRACT - CONTRIBUTED PAPER



# A New Class of Distributions for Survival and Lifetime Data Analysis : Theory and Applications

Broderick Oluyede, Georgia Southern University

boluyede@georgiasouthern.edu

Baitshephi Mashabe, Botswana International University of Science & Technology

Adeniyi Francis Fagbamigbe, Botswana International University of Science & Technology

Shujiao Huang, Department of Mathematics, University of Houston, Houston, TX

**Abstract:** A new generalized class of distributions called the Loglogistic Extended-Weibull Logarithmic (LLoGEWL) distribution is developed and presented. Structural properties of the LLoGEWL distribution are derived. The special case of loglogistic Weibull logarithmic (LLoGWL) distribution is studied in detail. Method of maximum likelihood is used to estimate the parameters of this new distribution. A simulation study to examine the bias, mean square error of the maximum likelihood estimators for each parameter is presented. Finally, real data examples are presented to illustrate the applicability and usefulness of the proposed model.

**Keywords:** Generalized Distribution; Logarithmic Distribution; Loglogistic Distribution; Weibull Distribution; Log-logistic Weibull Distribution; L-Moments; Maximum Likelihood Estimation



## ABSTRACT - CONTRIBUTED PAPER

### Dating Financial Cycles - a Comparison of New and Existing Techniques

Konrad Kapp, Nelson Mandela Metropolitan University, Centre of expertise in forecasting (COEF), Department of statistics

s210076453@nmmu.ac.za

Igor Litvine, Nelson Mandela Metropolitan University, COEF, Department of statistics

Francis Biesmans, Nelson Mandela Metropolitan University, COEF

**Abstract:** Litvine and Biesmans (2015) proposed a new technique for dating cycles in financial time series for purposes of optimisation of buy-sell strategies. However, this proposed hierarchical method may be exceptionally impractical for large data sets as it may require unacceptably long computation time. In this paper, new procedures are introduced that are feasible for large time series data sets. A comparison is then made between these new methods and the existing technique by evaluating the performance of all procedures on different time series lengths, Hurst exponent, and number of buy-sell points. It was found that the new algorithms outperform the existing technique in terms of computation time, without a significant sacrifice in the accuracy of the solution. Time series size, Hurst exponent and number of buy-sell points are all found to have an effect on relative performance of the methods.

**Keywords:** Optimisation; Computational statistics; Computational complexity; Cycles; Time series analysis

100



## ABSTRACT - CONTRIBUTED PAPER



### Using survival analysis to model the woman's waiting time to first birth after marriage in Rwanda

Innocent Karangwa, University of the Western Cape

fokafoka15@gmail.com

Renette J. Blignaut, University of the Western Cape

**Abstract:** Waiting time to first birth is a very important social indicator which shows a woman's readiness to acknowledge the role of mother and has social and economic implications on her and her family's future. Transition to motherhood, for instance, influences women's opportunities to further their education and careers prospects and may also limit their participation in other activities such as travel and leisure. Economically, it has been shown that early first birth can lead couples to have fewer savings. Furthermore, waiting time to first birth may also give an idea about the frequency of the fecundity and preferred interval union to first birth, as well as the patterns of childbearing in a particular society. This study used survival analysis to model the time to first birth after marriage in Rwanda. The 2014 demographic health survey (DHS) containing couples' records was used for analysis. The main interest was to determine the distribution of the waiting time to the first birth after marriage and to examine the factors that influence the occurrence and timing of the first birth after marriage. The estimated Kaplan-Meier survival curves showed differences in survival functions of women' socio-demographic characteristics including the age at first marriage, type of residence, wealth index, educational level, religious affiliation and contraception use status. Furthermore, the Cox regression model indicated that contraceptive method use status, age at first marriage and educational level of women were the only factors influencing the waiting time to first birth after marriage.

**Keywords:** Waiting time to first birth; survival analysis; Kaplan-Meier; survival curve; Cox regression; Rwanda



## ABSTRACT - CONTRIBUTED PAPER

### Testing the utility of marine habitat types as surrogates for macro-infaunal biodiversity

Natasha Karenyi, Department of Statistical Sciences, University of Cape Town  
natashakarenyi@gmail.com

**Abstract:** Habitat classifications define habitat types on maps that act as surrogates for biodiversity patterns. These are developed either from data or expert opinion and may include physical variables and/or biological variables. The utility of habitat types as surrogates has recently come into focus since habitat maps are based on extrapolated data or expert opinion and require verification. Three habitat classifications, based on physical variables, biota and expert-opinion respectively, exist for the marine sediment ecosystems of the South African west coast. The aim of this research is to determine the habitat classification that best represents the marine macro-infauna biodiversity on the South African west coast continental shelf. Marine macro-infauna count data from 51 sites representing 449 species are modelled using finite mixtures to group rows and columns. This likelihood-based fuzzy clustering provides a model-based alternative to traditional multivariate clustering techniques. Habitat types based on each habitat classification are included as predictor variables in the best clustered model and information criteria are used for model comparison. We will discuss the utility of these methods in comparing surrogates for biodiversity.

**Keywords:** surrogacy; habitat types; biodiversity; finite mixture models



## ABSTRACT - CONTRIBUTED PAPER

### A general biomarker-based incidence estimator for HIV

Reshma Kassanjee, Department of Statistical Sciences, University of Cape Town  
reshma.kassanjee@uct.ac.za

Thomas A McWalter, Department of Actuarial Science, University of Cape Town

Alex Welte, The South African DST/NRF Centre of Excellence in Epidemiological Modelling and Analysis (SACEMA), Stellenbosch University

**Abstract:** South Africa carries a high burden of HIV, and the reliable estimation of incidence is essential for targeting and assessing interventions. It is one of several countries that utilises approaches for estimating HIV incidence from cross-sectional surveys, made possible by applying biomarkers that aim to distinguish 'recent' from 'non-recent' infection. Since the introduction of the surveillance approach in 1995, a number of estimators have been proposed. These aim to infer incidence from (i) the prevalence of HIV and prevalence of the biomarker-defined state of 'recent' HIV, as measured in the survey, and (ii) a few well-defined properties of the biomarkers. However, earlier estimators were derived under unrealistic assumptions about biomarker dynamics and epidemiological/demographic equilibrium, leading to confusion about the interpretation and accuracy of results. A general framework for incidence estimation has since been developed: a particular weighted average of past incidence is defined, and any bias in its estimation is explicitly quantified. The relaxing of assumptions is made possible by introducing a timescale that truncates the period of averaging. An explicit expression for the bias, arising from the inherent loss of information that occurs when inferring a historic rate from current prevalence values, allows a researcher to understand the utility of the approach in relevant contexts. Furthermore, the precision of incidence estimation provides a metric for assessing the context-dependent performance of potential tests for classifying infections. This work provides the foundations upon which several studies have since been built, to further develop methodology, characterise tests or measure incidence.

**Keywords:** Incidence; HIV; biomarkers; recent infection



## ABSTRACT - CONTRIBUTED PAPER



### A multiple functional response model with historical effects

Dominique Katshunga, University of Cape Town  
dominique.katshunga@uct.ac.za

**Abstract:** In this presentation we introduce an extended version of the function-on-function linear models with historical effects. The model accommodates more than just one covariate and describes the relationship between functional responses and covariates such that only past and current values of the covariates influence the current value of the response variable. We investigate the structure of the variance-covariance matrix for this model and discuss possible fitting methods and estimation of the bivariate coefficients.

**Keywords:** Functional principal components; Backfitting



## ABSTRACT - CONTRIBUTED PAPER



### Pareto Optimal Space-filling Designs

Kago Kebotsmang, University of Cape Town  
kbtkag002@myuct.ac.za

**Abstract:** Spatial designs for environmental problems are usually based on optimizing the covariance structure of the process under study. Unfortunately this could lead to sub-optimal designs in situations where there is little or no prior knowledge about the process. Space-filling designs which are based on geometric criteria would be ideal in such cases as they largely avoid issues with misspecified covariance structures. Efficient space-filling designs should have both good projectional properties and cover the experimental region very well. However, individual criteria such as Latin hypercube sampling do not produce desirable designs with both of these properties. Therefore multiple criteria in order to obtain better designs using either the nested or non-nested approaches are to be preferred. The nested approach restricts a candidate designs to a certain class before applying a secondary criterion to select the final design. This approach does not allow flexible trade-offs between criteria because a re-run would be required every time there is a change in criteria preference. The non-nested approach considers all the criteria simultaneously and in some way optimizes them objectively. In this paper, we adopt the non-nested approach and optimize the resulting multi-criteria function using Pareto optimality-based methods. Pareto-based methods produce a number of best candidate designs which accommodate different objectives and thus give flexible trade-offs between the criteria considered. We also propose a hybrid of simulated annealing (SA) and genetic algorithms (GA) to construct these designs. SA is good in obtaining local optimums whereas GA efficiently search large regions of the design space.

**Keywords:** Space-filling designs; Pareto Optimality; Simulated Annealing; Genetic Algorithm



## ABSTRACT - CONTRIBUTED PAPER



### Survival Analysis of Under-five Twins Mortality in Ethiopia using Semi-parametric Bivariate Frailty Models

Yehene G Kifle, University of Limpopo

yehene.kifle@ucl.ac.za

Mesfin Tsegaye, Dire Dawa University

**Abstract:** Background: Several studies confirmed that twin births contribute substantially to the rate of under-five mortality in many countries. Quite high number of infant and child mortality rates among twins, mainly in developing countries, calls for a need to identify the potential determinant factors of under-five child mortality. This study intended to identify risk factors that are significantly associated with survival of under-five twins in Ethiopia. Data and Methods: Twins under-five mortality data was extracted from the birth history of women who were included in the 2011 Ethiopia Demographic and Health Survey (EDHS). In order to account for the unobserved heterogeneity and correlation into account a semi-parametric bivariate frailty models were employed under different settings. Results: The results of the study showed that covariates such as place of residence, preceding birth interval, and birth order were significantly associated with under-five twin mortality in Ethiopia. Whereas, sex of child and age of mother at child birth were found to be statistically non-significant. The result also showed the presence of significant correlation between the lifetimes of twins in pairs.

**Keywords:** Twins; mortality; heterogeneity; frailty; semi-parametric survival models



## ABSTRACT - CONTRIBUTED PAPER



### A Bayesian Approach to Green Security Games: Initial Perspectives

Lisa Kirkland, University of Pretoria

lisakirkland25@gmail.com

Alta de Waal, University of Pretoria

Timothy Haas, University of Wisconsin-Milwaukee

**Abstract:** Rhino poaching is a persistent problem in South Africa which may potentially lead to the extinction of the white and black rhinoceros. Recent studies have shown that the Stackelberg model can be an effective tool for security games through optimal resource allocation strategies. Few of these studies focus their attention on wildlife crime (green security games) and predicting poacher behaviour within a wildlife park remains a challenge. We propose using a Bayesian network model to learn the utility functions of the poachers and possibly employ a mixture model to capture the heterogeneity of poacher types. Furthermore, we would like to investigate whether spatial analysis and stochastic animal movement models could provide insight into the location of rhino in the park. To our knowledge, incorporating rhino location into a game-theoretic algorithm has not been explored. Finally, we attempt to provide a simulation platform that emulates realistic movement of rhinos, rangers and poachers in the park, which can ultimately be used for validation of the game theory model.

**Keywords:** Bayesian networks; game theory; mixture models; security games; spatial analysis; stochastic animal movement models; wildlife protection.



## ABSTRACT - CONTRIBUTED PAPER



### Preliminary testing of the Cobb-Douglas production function and related inferential issues

Judy Kleyn, University of Pretoria

judy.kleyn@up.ac.za

Mohammad Arashi, Shahrood University of Technology, Shahrood, Iran

Andriette Bekker, University of Pretoria

Sollie Millard, University of Pretoria

**Abstract:** In this paper, we consider the multiple regression model in the presence of multicollinearity and study the performance of the preliminary test estimator (PTE) both analytically and computationally, when it is a priori suspected that some constraints may hold on the vector parameter space. The performance of the PTE is further analysed by comparing the risk of some well-known estimators of the ridge parameter through an extensive Monte Carlo simulation study under some bounded and or asymmetric loss functions. An application of the Cobb-Douglas production function is included and from these results as well as the simulation studies, it is clear that the bounded linear exponential loss function outperforms the other loss functions across all the proposed ridge parameters by comparing the risk values.

**Keywords:** BLINEX loss; Cobb-Douglas production function; LINEX loss;  
Multicollinearity; Preliminary test estimator; Ridge regression



## ABSTRACT - CONTRIBUTED PAPER



### Spatial sampling for rabies vaccination schedule

Christine Kraamwinkel, University of Pretoria  
christine.kraamwinkel@gmail.com  
Inger Fabris-Rotelli, University of Pretoria

**Abstract:** We present the background to spatial sampling via a model-based approach. The technique is applied to a rabies vaccination schedule in rural villages of Northern Tanzania.

**Keywords:** spatial statistics; spatial sampling; point patterns



## ABSTRACT - CONTRIBUTED PAPER

### A Population Genetics Approach to Inferring Selection from Longitudinally-Sampled HIV-1 Haplotypes

Miguel Lacerda, University of Cape Town

Miguel.Lacerda@uct.ac.za

Ben Murrell, University of California, San Diego

**Abstract:** With high-throughput sequencing technologies, we are now able to accurately track changes in the genetic composition of an intra-host viral population over time. Such a time series of haplotype frequencies may be used to infer the strength of natural selection acting on viral variants. To achieve this, we modelled the sampled viral haplotypes as observations from a hidden Markov model, with the hidden states representing unobserved population frequencies that are assumed to have evolved according to a multivariate Wright-Fisher process. For computational reasons, the discrete state space is usually approximated by the standard simplex and the transition density between compositional vectors is obtained in a diffusion limit that assumes negligible selection. This assumption is not appropriate for rapidly evolving pathogens such as HIV and influenza. We therefore derived alternative approximations to the transition density by matching the moments of the Dirichlet and logistic normal distributions to those obtained for the Wright-Fisher process using the delta method without making any assumptions about the strength of selection. To avoid numerical integration over the hidden state vectors, we implemented our model with constant selection in a Bayesian framework and obtain posterior distributions for the selection coefficients using MCMC. We then extended this model to infer time-varying selection by performing a multiple change-point analysis with reversible jump MCMC. We applied our models to infer the selection coefficients of HIV-1 envelope haplotypes that are associated with immune escape.

**Keywords:** Natural selection; Multivariate Wright-Fisher process; Bayesian inference; HIV



## ABSTRACT - CONTRIBUTED PAPER



### Enhanced Minimum Variance Optimisation - A pragmatic approach

Bernisha J Lakhoo Lala, University of Cape Town  
bernishal@gmail.com

**Abstract:** Since the establishment of Markowitz's theory, numerous studies have been carried out over the past six decades or so that cover the benefits, limitations, modifications and enhancements of Mean Variance (MV) optimisation. This study endeavours to extend on this, by means of adding factors to the minimum variance framework, which would increase the likelihood of outperforming both the market and the minimum variance portfolio (MVP). An analysis of the impact of these factor tilts on the MVP is carried out in the South African environment, represented by the FTSE-JSE Shareholder weighted Index as the benchmark portfolio. The main objective is to examine if the systematic and robust methods employed, which involve the incorporation of factor tilts into the multicriteria problem, together with covariance shrinkage – improve the performance of the MVP. The factor tilts examined include Active Distance, Concentration and Volume. Additionally, the constant correlation model is employed in the estimation of the shrinkage intensity, structured covariance target and shrinkage estimator. The results of this study showed that with specific levels of factor tilting, one can generally improve both absolute and risk-adjusted performance and lower concentration levels in comparison to both the MVP and benchmark. Additionally, lower turnover levels were observed across all tilted portfolios, relative to the MVP. Furthermore, covariance shrinkage enhanced all portfolio statistics examined, but significant improvement was noted on drawdown levels, capture ratios and risk. This is in contrast to the results obtained when the standard sample covariance matrix was employed.

**Keywords:** minimum variance; optimisation; covariance shrinkage; factor tilting; volume; concentration; active distance



## ABSTRACT - CONTRIBUTED PAPER



### Forecasting Hail Damage for Crop Insurance

Shravan Lalbahadur, University of Cape Town  
shravan.lalbahadur@santam.co.za

**Abstract:** Demand for crop insurance against hail damage in South Africa is correlated with the occurrence of hail for each season. This adversely affects the insurers' ability to recover losses after catastrophic seasons. Many studies have been performed to examine and forecast rainfall patterns in South Africa but none have looked at the incidence of hail. The impact of natural variability will most likely present itself as a periodic signal with noise.

Using crop insurance data from Santam Agriculture that dates back to 1982, the data were analysed for the presence hidden periodicities in order to perform short-term forecasts. This was done by using ARIMA and harmonic regression models- the predecessor to modern time series analysis. Spectral analysis of the data were also performed in order to identify any trends.

Unlike previous studies, the period parameters of the harmonic regression model were treated as random variables and were estimated simultaneously with their coefficients. Short term forecasts from the harmonic regression model appear to be superior to those from the ARMA. However, the lack of out of time data for model validation is a limitation to the study. Despite this, the harmonic model's forecasts concur with those from more sophisticated models and its operational implementation has been promising.

**Keywords:** time series; harmonic regression; arima; insurance; spectral analysis



## ABSTRACT - CONTRIBUTED PAPER

### Dependence structures in multi-dimensional arrays

Alex Lau, University of Pretoria  
alexla1712@gmail.com  
Inger Fabris-Rotelli, University of Pretoria

**Abstract:** In the process of data acquisition the information obtained are more than often contaminated by noise. To purify the data smoothers are designed to remove the noise. The LULU operators are such smoothers, more specifically, they are designed to remove impulsive noise in images. However, because their definition uses a morphological concept of a connection, the question of how complex the connectivity should be therefore arises. Using the results from correlation analysis, we explore the extent at which the pixels of an image depend on its neighbours and establish the complexity of the connectivity for LULU operators in two dimensions. In addition, as a measure of how effective the LULU smoothers remove noise, we examine the noise extractions by the operators for images.

**Keywords:** Image analysis; denoising; LULU



## ABSTRACT - CONTRIBUTED PAPER

### Biplot-based visualizations of latent class models

Niël le Roux, Stellenbosch University  
njlr@sun.ac.za  
Zsuzsa Bakk, Leiden University, The Netherlands

**Abstract:** We consider latent class (LC) analysis performed on multivariate categorical data. The most common visualization tool for visualizing LC models is the univariate profile plot. A ternary plot provides a multivariate plot of LC models containing three posterior classes. We propose using categorical analysis of distance (AoD) biplots to visualize the posterior classifications arising from LC models with no limitation on the number of LCs. Categorical AoD biplots allow the best two- or three dimensional graphical representations of any number of LCs. The profiles of multiple latent classes are thus visualized while retaining the inherent multivariate character of the data. Both the original sample points and all the categorical variables in the form of category level points are simultaneously shown in the CatAoD biplot. Specifically, both the within and between class variation are visualized together with the class sizes and the overlap or separation among the different LCs. Furthermore the visualization of the relative density of each of the data patterns associated to a class is possible. We discuss two different methods, modal allocation and proportional allocation for allocating individual samples to the respective LCs. Measures to evaluate how well the variation in the full space is approximated in the lower dimensional biplot space are briefly discussed. We illustrate this approach with real data examples of LC models with three and more classes.

**Keywords:** analysis of distance; biplot; categorical variables; latent class analysis



## ABSTRACT - CONTRIBUTED PAPER



### Semi-parametric mixture models for skewed longitudinal HIV viral load trajectories with rebound

Maia Lesosky, University of Cape Town

maia.lesosky@uct.ac.za

Landon Myer, University of Cape Town

**Abstract:** HIV viral load (VL) testing is the diagnostic of choice to measure transmission risk and treatment outcomes in HIV-infected individuals. It remains poorly understood as to when, and how often, HIV VL testing should be done in HIV-infected pregnant women to detect loss of viral control. Statistical modelling of longitudinal HIV VL measures for is incomplete, particularly in light of the high proportion of measures below the limit of detection and the strongly skewed nature of the data. Multi-compartmental semi-parametric models for class trajectories under a Bayesian framework are fit to more than 3000 VL measures in a cohort of 550 pregnant women initiating ART.

**Keywords:** longitudinal; semi-parametric; mixture models; HIV; viral load



## ABSTRACT - CONTRIBUTED PAPER



### Factors associated with home-based HIV counselling and testing in men in KwaZulu-Natal, South Africa

Lara Lewis, CAPRISA

Lara.Lewis@caprisa.org

Anneke Grobler, CAPRISA

Ayesha Kharsany, CAPRISA

Cherie Cawood, Epicentre

**Abstract:** More needs to be done to encourage regular HIV testing in individuals at risk of acquiring HIV in South Africa. It is important to explore the role that home-based HIV counselling and testing (HCT) plays in this regard, particularly as South Africa initiates the HIV Test and Treat strategy working towards the ambitious '90-90-90' target set out by UNAIDS.

The objectives of this study were to describe factors associated with home-based HCT among men aged 15-49 in the uMgungundlovu municipality of KwaZulu-Natal, South Africa. The analyses were performed using data from the second cross sectional survey of the HIV Incidence Provincial Surveillance System (HIPSS) which ran from June 2015 to May 2016. Households were randomly selected from 220 clusters defined by enumeration areas. One individual from each household was invited to complete the survey regarding demographics, health, sexual behaviour and HIV testing patterns. After completing the questionnaire, participants were offered home-based HCT. Multivariate logistic models were used to assess factors associated with home-based testing.

Of the 10236 enrolled participants, 38.9% (3985) were men, 82.2% (3275) of whom were sexually active and of these 68.0% (2708) did not know their HIV status. Of these, 1291 (47.7%) agreed to undergo HCT at home. Variables significantly associated with home HCT at  $p<0.05$  included HIV status, viral load, partner having a sexually transmitted infection, receiving or giving gifts for sex, number of sexual partners in the last 12 months, having tested in the last year and suspecting their partner of being HIV positive.

**Keywords:** HIV home-based testing survey



## ABSTRACT - CONTRIBUTED PAPER



### Spatial hierarchical modeling of malaria in Angola

Artemisa Lima, University of KwaZulu-Natal

aaslima10@gmail.com

H. Mwambi, University of KwaZulu-Natal

Thomas Achia, University of KwaZulu-Natal

**Abstract:** Nearly a half of the world's population is at risk of malaria. The most vulnerable group when infected are young children, pregnant women and non-immune travellers from malaria-free areas. The malaria prevalence among the population at risk fell by 37% globally between 2000 and 2015 due to internationally coordinated efforts to combat the disease (UNICEF (2015), WHO (2015)). Malaria is a major public health problem in Angola. It represents close to 35 percent of the demand for curative care, 20 percent of hospital admissions, 40 percent of perinatal deaths, and 25 percent of maternal mortality. We use four Bayesian hierarchical spatial models (BHMs) namely the Poisson gamma model, poisson log-normal model, the conditional autoregressive (CAR) model and the convolution model to estimate the relative risk of malaria mortality at provincial level using the 2003-2010 National Malaria Control Program (NMCP) data. The convolution model was found to be the best fitting model. Results from this model showed some considerable differences in the relative risk of malaria across the provinces. The highest relative risk were recorded in Luanda and Cuando Cubango while Bengo, Uige, Malanje,Lunda Norte and Lunda Sul recorded the lowest relative risk.

**Keywords:** Malaria Prevalence; Public health; hierachical spatial models; relative risk.



## ABSTRACT - CONTRIBUTED PAPER



### Application of Causal Analysis for Fault Diagnosis in Mineral Processes

Brian Lindner, Stellenbosch University

brianslindner@gmail.com

Lidia Auret, Stellenbosch University

**Abstract:** In large-scale chemical and mineral processes, processing units are connected to each other through material flow, energy flow, or information flow (manifested in these process by control loops). Faults (abnormal events potentially detrimental to process performance) occurring in these processes can propagate throughout the process as a result of this interconnection, causing multiple sensors to display abnormal behaviour and obscuring the root cause of the fault. Diagnosis of such faults is essential to determining the corrective action needed to return the process to its desired operating state.

Data-based causal statistics may be employed to infer the cause-effect connections between the time series of measured variables in a process, allowing identification of the propagation path of the fault. Examples of causal statistical methods include transfer entropy, an information theoretic method, and Granger causality, a regression-based method.

In the last decade research into the causal analysis approach to fault diagnosis has gained attention, and these methods have been proven effective. However, in order to provide a reliable, robust and automated methodology that will be widely accepted by the minerals processing industry, a number of shortcomings still need to be addressed. These shortcomings include: automatic selection of significance thresholds for the causal statistics, automation parameter selection (e.g. the embedding dimensions for transfer entropy), determination of the effects of different process conditions on the calculated statistics.

A fault diagnosis methodology incorporating data-based causal analysis for application in large-scale mineral processing plants is demonstrated in this work-in-progress presentation.

**Keywords:** Causal Analysis; Mineral Processing; Information Theory; Granger Causality; Transfer Entropy;



## ABSTRACT - CONTRIBUTED PAPER



### Generalised Hurst Exponent

Igor Litvine, Nelson Mandela Metropolitan University  
igor.litvine@nmmu.ac.za

**Abstract:** Hurst exponent is widely used in the time-series analysis as for various purposes, e.g. for measuring persistence.

**Keywords:** Time Series; Hurst Exponent; Persistence; Predictability

119



## ABSTRACT - CONTRIBUTED PAPER



### Modelling Growth in a Randomised Controlled Trial

Carl Lombard, South African Medical Research Council

carl.lombard@mrc.ac.za

Marius Smuts, North West University

Mieke Faber, South African Medical Research

**Abstract:** The study protocol of a randomised control trial usually indicates in some detail how the primary outcome will be analysed. This was the case in a trial evaluating the impact of complementary food supplements on the growth of young children(<1 year old) enrolled in the study. This presentation will look at the practical execution of an analysis of the growth outcome (length standardised for age) to fit the protocol specifications and the data. The final growth model was based on a mixed effects model using linear splines.

**Keywords:** trial; growth model; linear splines

120



## ABSTRACT - CONTRIBUTED PAPER

### Multivariate analysis to find natural clusters of Tuberous Sclerosis Complex Associated Neuropsychiatric Disorders

Sugnet Lubbe, Department of Statistical Sciences, University of Cape Town

Sugnet.Lubbe@uct.ac.za

Loren Leclezio, Division of Child & Adolescent Psychiatry, University of Cape Town

Petrus de Vries, Division of Child & Adolescent Psychiatry, University of Cape Town

**Abstract:** Tuberous Sclerosis Complex (TSC) is a genetic disorder associated with a wide range of TSC-Associated Neuropsychiatric Disorders (TAND). TAND affect up to 90% of those with TSC in an apparently unique, individual pattern. This uniqueness poses significant challenges for psycho-education and intervention planning. The study examines whether there are natural TAND clusters which could be used to develop a 'treatment toolbox' for each set of psychiatric disorders.

In this paper it will be shown how cluster analysis, factor analysis and classification trees are combined with suitable graphical representations. TAND items are clustered to obtain a psychiatric meaningful set of clusters. Thereafter a factor analysis model is fitted to the data. Different factor extraction methods and types of rotations are compared. The goal of the factor analysis model is to identify individuals who struggle with particular sets of psychiatric disorders. In a further step an attempt is made to find defining TAND items for each set of psychiatric disorders.

**Keywords:** Cluster analysis; Factor Analysis; Classification trees



## ABSTRACT - CONTRIBUTED PAPER



### Stochastic Processes in Sound Synthesis

Zani Ludick, University of the Free State  
ludickz@ufs.ac.za

**Abstract:** Computer-synthesized sounds and electronically emulated instruments are prevalent in contemporary music and music composition software. The program GENDYN (1992), written by the composer Iannis Xenakis (1922-2001), uses random walks to dynamically shape waveforms in the amplitude and time domains. This synthesis technique, known as dynamic stochastic synthesis, originated from Xenakis' compositional practices and theories. We consider the origin and subsequent development of dynamic stochastic synthesis as an alternative to the harmonic techniques used in Fourier synthesis.

**Keywords:** dynamic stochastic synthesis; stochastic processes; random walks; sound synthesis



## ABSTRACT - CONTRIBUTED PAPER



### Simulating Complex Sampling Data for the evaluation of Survey-Weighted Linear Models

Retha Luus, University of the Western Cape

rluus@uwc.ac.za

Ariane Neethling, Stellenbosch University

Tertius de Wet, Stellenbosch University

**Abstract:** It has been found that, in contrast to developed countries, some researchers in developing countries, such as South Africa and the rest of Africa, confuse linear regression using CS data, i.e. survey-weighted least squares (SWLS) regression, with weighted least squares (WLS) regression. The objective of this research is to illustrate how the results could differ if the researcher ignores the complex design of the data, i.e. applies ordinary least squares (OLS) regression, or naively applies WLS in comparison to the correct SWLS regression.

Since real data typically do not meet any or all of the linear model assumptions, the use of simulated data in the development and testing of new models has become invaluable. Although the simulation of i.i.d. data is well-known, the same cannot be said of the simulation of complex sampling data. In this research a multilevel model is developed and it is shown how the model can be used for the simulation of a hierarchical population from which complex samples can be selected.

Repeated sampling from the simulated population will be used to compare the OLS, WLS and SWLS results for inference concerning the linear model. Since it has been observed that benchmarking methods often result in weights having large variability which could affect the precision of any analyses where they are incorporated, a further consideration in the simulation study will be the application of different weight trimming methods to address this phenomenon.

**Keywords:** Complex sampling; Simulation; Linear models; Trimming; Benchmarking; Inference



## ABSTRACT - CONTRIBUTED PAPER



### Some models for discrete-valued time series

Iain MacDonald, University of Cape Town  
iain.macdonald@uct.ac.za

**Abstract:** Many of the time series one meets in practice are by their very nature discrete-valued. Since the 1980s there has been a significant research effort devoted to finding useful models for such series.

Many of the proposed models are ARMA analogues, inspired by the great success of Gaussian autoregressive moving-average models for continuous-valued series. But that success has (in my opinion) not been repeated in the case of discrete-valued series, where a wider range of approaches seems to be necessary. With the recent publication of a valuable Handbook of Discrete-Valued Time Series, a discussion of the topic seems timely. In this talk I shall describe some of the (many) classes of models that have been proposed, give examples of data analyses, and relate some of these models to agent-based models.

**Keywords:** discrete-valued time series; INAR process; hidden Markov model; agent-based model



## ABSTRACT - CONTRIBUTED PAPER



### Destructive COM-Poisson Cure Rate Model and Associated Likelihood Inference under Lognormal Lifetime

Jacob Majakwara, University of Witwatersrand

Jacob.Majakwara@wits.ac.za

Suvra Pal, University of Texas at Arlington, Arlington, USA

**Abstract:** In this talk, we will consider the destructive COM-Poisson cure rate model. The model assumes 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. By assuming a COM-Poisson distribution for the initial risk factors, the steps of the EM algorithm are developed in detail to calculate the MLEs of 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 both likelihood test and information based criteria. Finally a melanoma data is analyzed for illustrative purpose.

**Keywords:** COM-Poisson distribution; Competing cause scenario; Profile likelihood; EM algorithm





## ABSTRACT - CONTRIBUTED PAPER



### Bivariate beta-generated distributions with illustrations

Seite Makgai, University of Pretoria

seite.makgai@up.ac.za

Daniel de Waal, University of the Free State and University of Pretoria

Jaco Visagie, University of Pretoria

Andriette Bekker, University of Pretoria

**Abstract:** In this study, the beta-generated method is used to construct bivariate beta-generated distributions. This method was pioneered by Eugene et al. (2002). (*Comm. in Stats.-Theory and Methods*, 31(4)). We propose bivariate beta distributions through the generating function  $H(x)=F(G(x))$ , where  $x$  is a vector,  $G(\cdot)$  is a vector baseline distribution and  $F(\cdot)$  is the univariate beta distribution. We consider the bivariate exponential and normal distributions as candidates for  $G(\cdot)$ . In this speculative research, the elegance and flexibility of the generated distributions are illustrated by fitting the distributions to observed data.

**Keywords:** generator distribution; exponential distribution; normal distribution; parent distribution



## ABSTRACT - CONTRIBUTED PAPER



### Identifying local emission sources of particulate matter to improve geostatistical mapping of PM10

Sibusisiwe Khuluse-Makhanya, CSIR and University of Twente

smakhanya@csir.co.za

Alfred Stein, University of Twente

Pravesh Debba, CSIR and University of Witwatersrand

**Abstract:** An objective in mapping air quality is to determine areas that are hot spots and factors that contribute to their formation. Once identified, these factors or spatial predictors can be used to improve geostatistical mapping of PM10. Previously, housing informality at small area level was identified as a significant predictor of the annual average exceedance frequency of the South African PM10 national air quality standard (NAQS) of  $120 \mu\text{g m}^{-3}$ . In this case, fugitive dust emission sources in the neighbourhood of an air quality monitor were assessed for their strength as predictors of ambient PM10 concentrations on days characterized by strong local winds. An ensemble maximum likelihood method which involved multiple training iterations for improved accuracy of the bare soil class was developed for land cover mapping in the vicinity of an air quality station using SPOT 6 multi-spectral images. The overall accuracy of the classification routine was 78%. Using cluster analysis and a varying intercepts regression model, land cover patterns in the neighbourhood of an air quality station were found to be significant predictors of observed average PM10 concentrations on days where wind speeds were conducive for dust emissions. The variability explained was above 60 %, supporting the use of land cover data for improved prediction of PM10 at locations without air quality monitoring stations. Next, we extended on these findings by considering both land cover patterns and housing informality as predictors in a varying intercepts model.

**Keywords:** Kriging; ensemble classification; air quality; regression



## ABSTRACT - CONTRIBUTED PAPER



### Improved modified synthetic and runs-rules control charts combined with a Shewhart Mann-Whitney chart

Jean-Claude Malela-Majika, University of South Africa  
malelm@unisa.ac.za, CC: jcmaleta2005@yahoo.fr  
Eeva Rapoo, University of South Africa

**Abstract:** A control chart is one of the most important tools used in statistical process control and monitoring (SPCM) to detect changes in quality processes. This paper investigates the performance of the improved modified distribution-free synthetic and runs-rules charts combined with a Shewhart Mann-Whitney (MW) control charts, in terms of the average run length (ARL), standard deviation of the run length (SDRL) and median run length (MRL) through intensive simulation. It is observed that, the new control charts present very attractive run-length properties and outperform the competing charts in many cases. Numerical examples are given as illustration of the design and implementation of the proposed charts.

**Keywords:** distribution-free control charts; Monte Carlo simulation; synthetic MW chart; improved modified synthetic chart; improved modified runs-rules; median run-length



## ABSTRACT - CONTRIBUTED PAPER

### Bayesian Copula-based Analysis of Multiple Longitudinal Measures and Multiple Recurrent Events

Samuel Manda, South African Medical Research Council

samuel.manda@mrc.ac.za

Renate Meyer, University of Auckland

Jose Romeo, University of Santiago

Lieketseng Masengetse, University of Lesotho

Henry Mwambi, University of Kwazulu-Natal

**Abstract:** Estimating parameters of a joint model for longitudinal measurement and recurrent event data have been developed and further research is still ongoing. In some longitudinal studies we observe multiple repeated measures processes and multiple recurrent events that maybe correlated within a subject. For example, lower CD4 cell counts and higher viral loads are associated with higher risk of recurrent antiretroviral (ARV) regimen changes and opportunistic disease infections. A simple statistical procedure jointly models a single longitudinal measurement and recurrent event data, both of which are modified to include a random effect term which captures the dependence of the data within a subject. A more realistic technique needs to jointly model all the correlated processes, which may be achieved by using shared-random effects components for all the processes and process-specific random components. We show how the outcome-specific covariate effects and the correlation structure can be estimated by Bayesian estimation using copulas. The estimation method is applied to the ARV Pharmacovigilance Study data set from South Africa. Results are compared to those obtained using Gaussian quadrature based maximum-likelihood techniques.

**Keywords:** Bayesian statistics; Repeated Data; Shared Frailty Model; Survival Analysis



## ABSTRACT - CONTRIBUTED PAPER



### Towards application of spatial statistics of extremes to the extreme flood heights in the lower Limpopo River basin of Mozambique

Daniel Maposa, University of Limpopo

danmaposa@gmail.com

James J. Cochran, University of Alabama

Maseka Lesaoana, University of Limpopo

**Abstract:** We look at the possibility of application of spatial extreme value statistics to the extreme flood heights in the lower Limpopo River basin. This study is motivated by the existence of spatial variability in the lower Limpopo River basin revealed from the findings in the just completed PhD thesis, particularly with reference to the year 2000 disastrous floods.

**Keywords:** spatial extremes; statistics of extremes; extreme value theory; flood heights.



## ABSTRACT - CONTRIBUTED PAPER



### Reducing bias and MSE in estimation of heavy tails: a Bayesian approach

Gaonyalelwé Maribe, University of Free State

gao.isc@gmail.com

Andréhette Verster, University of Free State

Jan Beirlant, KU Leuven and University of the Free State

**Abstract:** Bias reduction in tail estimation has received considerable interest in extreme value analysis. Estimation methods that minimize the bias while keeping the mean squared error (MSE) under control, are especially useful when applying classical methods such as the Hill (1975) estimator. In Caeiro et al. (2005) minimum variance reduced bias estimators of the Pareto tail index were first proposed where the bias is reduced without increasing the variance with respect to the Hill estimator. This method is based on adequate external estimation of a pair of second-order parameters. Here we revisit this problem from a Bayesian point of view starting from the extended Pareto distribution (EPD) approximation to excesses over a high threshold, as developed in Beirlant et al. (2009) using maximum likelihood (ML) estimation. Using asymptotic considerations, we derive an appropriate choice of priors leading to a Bayes estimator for which the MSE curve is a weighted average of the Hill and EPD-ML MSE curves for a large range of thresholds, under the same conditions as in Beirlant et al. (2009). A similar result is obtained for tail probability estimation. Simulations show surprisingly good MSE performance with respect to the existing estimators.

**Keywords:** Extended Pareto Distribution; Peaks over threshold; extreme value index; Bayesian parameter estimation; bias reduction; posterior simulation.



## ABSTRACT - CONTRIBUTED PAPER

### Subjective Bayesian Analysis under the AR(1) model

Hossein Masoumi Karakani, University of Pretoria

hosseinstatistics@gmail.com

Janet van Niekerk, University of Pretoria

Paul J.van Staden, University of Pretoria

**Abstract:** The first-order autoregressive process, AR(1), has been widely used and implemented in time series analysis. Different estimation methods have been employed in order to estimate the autoregressive parameter. This article focuses on subjective Bayesian estimation as opposed to objective Bayesian estimation and frequentist procedures. The truncated normal distribution is considered as a prior, to impose stationarity. The posterior distribution as well as the Bayes estimator are derived. A comparative study between the newly derived estimator and other existing estimation methods (frequentist) is employed in terms of simulation and real data.

**Keywords:** Autoregressive process; Bayesian estimation; Stationarity; Subjective prior; Truncated normal distribution



## ABSTRACT - CONTRIBUTED PAPER

### Predictors of blood pressure in South Africa: Quantile regression Approach

Lyness Matizirofa, University of Johannesburg  
lmatizi@yahoo.com

**Abstract:** Background: Hypertension is known as a risk factor for cardio-cerebrovascular disease and one of the leading causes of morbidity and mortality worldwide. It is one of the top five major public health problems in South Africa. Methodology: This study aims at exploring quantile-specific associations of blood pressure with risk factors. The secondary data analysis was conducted using data from the National Income Dynamics Study (2008). A stratified two-stage cluster sampling technique was employed and yielded a sample of 9 685 adults. Data were explored with descriptive statistics .Ordinary least squares(OLS) regression analysis and quantile regression(QR) analysis, for diastolic blood pressure(DBP) and systolic blood pressure(SBP) as dependent variables respectively against age, body mass index, smoking, alcohol consumption, exercises, gender and race as independent variables was conducted using "R" version 3.12. Results: The results from the OLS regression analysis and QR analysis indicated that age, BMI and alcohol consumption statistically significantly impacts DBP whilst exercises and race did not significantly influences DBP. Gender and smoking were statistically significant and insignificant for part of the quantiles. This study also revealed that age, BMI and gender statistically significantly affected SBP whilst alcohol consumption, exercises and race did not significantly impact SBP. Smoking was statistically significant and insignificant for part of the quantiles. Conclusions: Hypertension prevalence rate was high among the adults. Quantile regression analysis provided more detailed evidence of quantile-specific associations of DBP and SBP with risk factors. Age, obesity, gender and smoking revealed statistically significant coefficients with all BP measures.

**Keywords:** diastolic blood pressure (DBP); systolic blood pressure (SBP); body mass index (BMI); ordinary least squares regression (OLS); quantile regression (QR).



## ABSTRACT - CONTRIBUTED PAPER

### Mean-Variance Hedging of Basis Risk

Melusi Mavuso, University of Cape Town  
melusi.mavuso@uct.ac.za

**Abstract:** We begin by giving an overview of some modern techniques that are used for hedging derivatives in incomplete markets, and give an outline of mean-variance hedging as a useful tool in this context. We then show how mean-variance hedging is applied to some specific examples. We first consider the standard basis risk problem where an option is written on an untradable underlying asset. We derive a mean-variance optimal dynamic hedging strategy based on a liquidly traded correlated asset, and study the strategy's performance in comparison with other trading strategies. We then generalize this example to when the underlying asset is tradable but illiquid. In both examples we derive closed form formulae for the expected squared replication error. We end by exploring how these techniques can also be applied to interest rate derivatives.

**Keywords:** Mean-Variance Hedging; Expected Squared Replication Error; Incomplete Markets; Hedging Derivatives



## ABSTRACT - CONTRIBUTED PAPER



### How hydrogen can enhance wind power competitiveness? A case study in Lorraine (France)

Cremel Maxime, Université de Lorraine  
cremel.maxime@gmail.com

**Abstract:** Hydrogen production through electrolysis represent a potential solution to store and redistribute power generated by renewable energies, especially wind power. This paper show that the implementation of electrolyzers would increase the competitiveness of wind power and change the shape of the optimal electricity capacity mix. Further, in a best-case scenario for these technologies and with an appropriate carbon pricing, we found that a carbon-free capacity-mix is possible and could thus strongly contribute in the mitigation effort for global warming.

**Keywords:** Hydrogen Renewable Energy Wind Power Electricity



## ABSTRACT - CONTRIBUTED PAPER

### A comparison of the performance of latent Dirichlet allocation and the Dirichlet multinomial mixture model on short text

Jocelyn Mazarura, University of Pretoria  
jocelyn.mazarura@yahoo.com

**Abstract:** Topic modeling is a text mining tool that is used to discover hidden topics from large collections of documents. It has proven to be very successful in summarising documents that are long, such as journal articles. However, the need for effective methods for analysing texts that are short, such as posts on Twitter and Facebook, has become of great relevance. This is due to the realisation that such stores of data could potentially hold useful information as people readily share their ideas, interests and opinions on such platforms. One of the challenges faced in applying topic models to short text, is that, unlike long text, it contains few words, which may not contain sufficiently enough meaningful words to aid the identification of the hidden topics.

Latent Dirichlet allocation (LDA) is one of the most popular probabilistic topic models due to its proven success when applied to long text. It makes the generative assumption that a document is created from different proportions of multiple topics. We posit that, the Dirichlet multinomial model (DMM), should be able to discover the hidden topics from short text more effectively than LDA. In contrast to LDA, the DMM makes the generative assumption that each document belongs to only one topic, which we believe to be a more sensible assumption about short text. In order to validate this hypothesis, the performance of both models on short text is compared based on topic stability and coherence.

**Keywords:** Dirichlet multinomial mixture model; LDA; Short text; Topic coherence; Topic modeling; Topic stability



## ABSTRACT - CONTRIBUTED PAPER

### Distributional development of ratios of gamma variables with application to a changepoint model in statistical process control

Albert Mijburgh, University of Pretoria

albert@mijburgh.co.za

Schalk Human, University of Pretoria

Andriette Bekker, University of Pretoria

**Abstract:** This research emanates from a practical problem in the statistical process control (SPC) environment, where the variance of a process is being monitored. Hawkins et al. (*Technometrics*, 47(2):164-173, 2005) proposed a changepoint model, to detect potential shifts in a process's variance, which overcomes many of the shortcomings that other traditional SPC methods poses. However one of the limitations of the changepoint model is that a closed form expression for it has never been derived. This research paves the way and derives the joint and marginal distributions that would be required to find such a closed form expression, and in doing so proposes a new distribution that can be regarded as a generalized multivariate beta distribution with unbounded domain.

The focus of this presentation is specifically on the distributional development of the bivariate density function and its relevant properties. In order to derive these properties the relationships between a compilation of well known bivariate beta density functions is investigated (including the proposed distribution). An exploratory shape analysis is performed to investigate how the parameters of this distribution affect its shape.

**Keywords:** Statistical Process Control; Changepoint model; Bivariate beta; Hypergeometric function; Gamma random variables



## ABSTRACT - CONTRIBUTED PAPER



### Mixture regression using t distributed errors

Sollie Millard, University of Pretoria  
sollie.millard@up.ac.za  
Frans Kanfer, University of Pretoria

**Abstract:** Estimation of mixture models are frequently based on the assumption of component distributions that are normally distributed, an approach usually sensitive to outliers or heavy tailed errors. Failure to accommodate outliers may jeopardise the estimation and inferential results. This talk gives an introduction to the use of a robust version of mixtures of regression models based on the t-distribution. We illustrate the use of a simple EM algorithm to perform maximum likelihood (ML) estimation of the parameters. Estimation of the degrees of freedom for the components are addressed using a profile likelihood approach. The robust mixture regression approach is illustrated on an industry dataset.

**Keywords:** Mixture regression; t distribution; outliers; EM algorithm



## ABSTRACT - CONTRIBUTED PAPER



### Comparison of Confidence Interval Estimators: An Index Approach

Richard Minkah, University of Ghana  
rminkah@ug.edu.gh  
Tertius de Wet, Stellenbosch University  
Kwabena Doku-Amponsah, University of Ghana  
Ezekiel N. N. Nortey, University of Ghana

**Abstract:** We develop two indices for comparing confidence interval estimators based on the coverage probability and confidence interval length. We show that both confidence interval indices have range of values within the neighbourhood of the range of the coverage probability, [0,1]. In addition, a good confidence interval estimator is shown to have an index value approaching 1 and a bad confidence interval has an index value approaching 0. A simulation study is conducted to assess the finite sample performance of the indices. Finally, the proposed indices are illustrated with a practical example from the literature.

**Keywords:** Confidence interval; Empirical coverage probability; Confidence Interval length; Bootstrap calibration.



## ABSTRACT - CONTRIBUTED PAPER



### Density forecasting of long-term electricity demand in South Africa using Quantile Regression

Paul Mokilane, CSIR  
PMokilane@csir.co.za

**Abstract:** Electricity demand forecasting is crucial not only in the day to day running of power systems, but also in system planning. Long-term forecasts are useful in capital planning. However, any prediction comes with uncertainties. Uncertainties in predictions imply that forecasts should ideally be probabilistic. Poor predictions could have far reaching consequences because an overestimate of long-term electricity demands could result in substantial wasted investment in the construction of excess power facilities, while underestimate of demands could result in insufficient generation and unmet future demand.

The long term hourly electricity demand was forecasted. Quantile regression (QR) was used to forecast hourly electricity demand at 0.01, 0.02, 0.03,...,0.99 quantiles of the demand distribution which effectively described the full demand distribution. QR was developed as an extension of the ordinary least-square regression for estimating rates of change in all parts of the distribution of the response variable.

In addition to point forecasts, QR gives forecasts at all percentiles of the distribution. The distributions of hourly demand and the peak daily demand seem to be shifting towards smaller demands over the years until 2023.

A portion of the data was used as out of sample forecasts to assess the model with the Mean Absolute Percentage Error and the density functions of the forecasted and actual demands. The demand forecasts at 50th percentile closely estimated the actual hourly demands. The actual demand and the forecasted demand densities seem to track each other and therefore the model seemed to be forecasting the true density demand well.

**Keywords:** Probability; Quantile Regression; Density Functions; Percentiles



## ABSTRACT - CONTRIBUTED PAPER



### Application of small area estimation methods in modelling lack of service delivery at ward level: water, sanitation and electricity in South Africa

Reshoketswe Mokobane, University of Limpopo

nkadirm@gmail.com

Maseka Lesaoana, University of Limpopo

**Abstract:** This research study presents the application of direct and indirect estimation of small area estimation (SAE) techniques. The research study is aimed at estimating the trends and the proportion of households accessing water, sanitation, and electricity at small areas of Limpopo Province over a period of seven years. The study has used modified recoded data from the General Household Survey (GHS) for the years 2009-2015 and Census 2011. The aggregated Census 2011 data in Supercross were used to validate the results. The key point is to suggest policy considerations to the South African government for future equitable provisioning of the basic services selected for this study across the country. This was triggered by the fact that Statistics South Africa does not disseminate the GHS information at ward level.

**Keywords:** basic services; small area estimation; GHS; census 2011



## ABSTRACT - CONTRIBUTED PAPER

### Survival Analysis of Age Differences at Marriage and Time-to-Divorce in South Africa

Khehla Moloi, University of Limpopo

khehla.moloi@ul.ac.za

Mokgadi Priscilla Matlebjane, University of Limpopo

Happy Maluleke, University of Limpopo

Yehene G Kifle, University of Limpopo

**Abstract:** Background: Age at marriage is one of the possible factors for the long term success of a particular marriage. This study focused on assessing the relationship between age differences of couples at marriage and duration of the marriage. Age differences at marriage was calculated by subtracting the age of the wife from that of the husband and then compare the age difference at marriage with time to divorce to check whether marital age difference has any significant on divorce rate in the context of South African data set. Hence, the overall aim of the study was to explore and then model the relationship between age differences of couples at marriage and time to divorce in South Africa. Methods: Civil marriage and divorce data were used. These data were collected during the follow up time from 2010 to 2013. 738,261 couples was used in this study for the analysis. Age, duration of marriage, survival status was recorded in the data set. The Kaplan Meier and non-parametric survival analysis techniques including log-rank test were used, also the semi parametric cox proportional hazard regression were used to explore and model age difference of the couple. Results and Conclusions: The results of log rank test, cox proportional regression showed that there is a significance association among the age difference of couples at marriage and time to divorce. The log-rank test revealed a highly significant ( $P\text{-value}=0.000$ ) difference exists between the two study categories w.r.t time to divorce.

**Keywords:** Marriage; Time-to-divorce; Age difference; Survival analysis; Cox PH; South Africa



## ABSTRACT - CONTRIBUTED PAPER



### Variable selection advances in regression using R

Henri Moolman, Walter Sisulu University  
moolman.henri@gmail.com

**Abstract:** The variable selection problem in multiple regression is well known among statisticians. Traditional variable selection methods such as stepwise, forward and backward regression, all subsets regression and selection using the Mallows selection criterion have been in use for a long time. With the development of R as a statistical software tool and programming language over the past 20+ years the number of variable selection techniques that can be applied with relative ease have increased considerably. This presentation is a discussion of some of the variable selection techniques that are available in R together with a few examples.

**Keywords:** regression; variable selection; R



## ABSTRACT - CONTRIBUTED PAPER

### Evaluation of fault diagnosis techniques in multivariate statistical process control

Andre Mostert, Group Technology, a division of Sasol South Africa (Pty) Ltd  
andre.mostert@sasol.com

Roelof Coetzer, Group Technology, a division of Sasol South Africa (Pty) Ltd  
Sugnet Lubbe, University of Cape Town

**Abstract:** The safe and optimal operation of industrial sized chemical plants is dependent on efficient and accurate fault identification and diagnosis techniques. Multivariate statistical process monitoring strategies based on historical data have become the methodology of choice due to the complexity of the processes monitored. Principal component analysis (PCA), the most popular technique employed in practice, is preferred because of its simplistic formulation. Performance statistics calculated from the PCA are used to determine when the process is operating in an out-of-control state. If abnormal process behavior is detected the fault statistics are further analyzed to identify the variables responsible for the process drift. This fault diagnosis methodology is often referred to as contribution analysis. Contribution analysis however has been shown to suffer from fault smearing i.e., the correct variables responsible for process drifts are not always identified. In this discussion the different fault identification methodologies will be explored and compared. Improvements on the current methods will then be proposed.

**Keywords:** Multivariate statistical process control; Fault diagnosis; Principal component analysis



## ABSTRACT - CONTRIBUTED PAPER



### Estimating age-time dependent malaria force of infection accounting for unobserved heterogeneity

Leviciatus Mugenyi, Center for Statistics, Hasselt University, Diepenbeek, Belgium  
l.mugenyi005@gmail.com

Steven Abrams, Center for Statistics, Hasselt University, Diepenbeek, Belgium

David Smith, Department of Zoology, University of Oxford and Sanaria Institute for Global Health and Tropical Medicine, UK & Sanaria

Sarah Staedke, London School of Hygiene & Tropical Medicine, United Kingdom, Infectious Diseases Research Collaboration, Kampala, Uganda

Niel Hens, Center for Statistics, Hasselt University, Diepenbeek, Belgium, Centre for Health Economics Research and Modelling Infectious Diseases, Vaccine and Infectious Disease Institute, University of Antwerp, Antwerp, Belgium

**Abstract:** Despite well-recognized heterogeneity in malaria transmission, key parameters such as the force of infection (FOI) are generally estimated without accounting for the intrinsic variability in individual infection risks. Given the potential impact thereof on the estimation of the malaria FOI, we estimate the malaria FOI explicitly accounting for both observed and unobserved heterogeneity. We use cohort data obtained by testing children aged 0.5-10 years for the presence of malaria parasites from August 2011 to August 2014 at three sites in Uganda, i.e., Walukuba (peri-urban, n = 300), Kihiihi (rural, n = 355), and Nagongera (rural, n = 334). Assuming a Susceptible-Infected-Susceptible model, we show how the FOI relates to the point prevalence, enabling the estimation of the site- and individual-specific FOI by modeling the prevalence using a generalized linear mixed model. The resulting FOI varies significantly with age and time and is estimated to be highest among children aged 1-2 years, after which it decreases within the first 4 to 5 years and remains constant thereafter. The infection rate is highest in Nagongera while no significant differences are found for Kihiihi and Walukuba. Heterogeneity is larger between households (variance = 0.88, P<0.001) than within households where it is non-significant (variance=0.06, P=0.223). This indicates that, aside from the individual's age, heterogeneity in malaria FOI can be attributed to household conditions. Accounting for observed and unobserved heterogeneity in malaria acquisition is important when estimating the FOI and gives important information for developing malaria spread models.

**Keywords:** Point prevalence; SIS compartmental model; Generalized linear mixed model; Random effects model



## ABSTRACT - CONTRIBUTED PAPER



### Discrepancies in the chi-square test and odds ratio confidence interval when analysing 2x2 tables

Chris Muller, Stellenbosch University  
cmuller@sun.ac.za

**Abstract:** Two-by-two contingency tables play an important role in epidemiological studies. These tables are frequently used to analyse how the presence or absence of a risk factor influences a disease being present or absent in a group of patients under study.

Two of the most widely used techniques to assess if there is an association between the risk factor and the disease is the Chi-square test (or Fisher's exact test if a non-parametric approach is required) and the odds ratio (OR) with its corresponding confidence interval. In this talk I'll be examining situations where using these two approaches, the hypothesis test approach and the confidence interval approach, may lead to different conclusions when using the same 2x2 table.

**Keywords:** Chi-square; odds ratio; contingency table





## ABSTRACT - CONTRIBUTED PAPER

### Wavelet-based Functional Mixed Models applied in Metabolomics

Martie Muller, Department of Mathematical Sciences, University of Copenhagen, Denmark  
m.muller@science.ru.nl

Anders Tolver, Department of Mathematical Sciences, University of Copenhagen, Denmark

**Abstract:** Metabolomics studies the ‘unique chemical fingerprints’ that cellular processes create. These ‘fingerprints’ can be measured in biofluids like blood and urine and can be used to study the influence of different factors on human metabolism. For each sample that is chemically analysed with nuclear magnetic resonance (NMR) we obtain a spectrum containing hundreds of peak patterns (metabolites) constructed by tens of thousands of data values. These data are typically analysed with standard chemometric methods and do not explore the rich functional nature of spectral data.

We explore functional data analysis (FDA) as a method to analyse NMR spectral data in metabolomics. In FDA, the data are curves instead of data points. We apply wavelet-based functional mixed model methodology. We use bootstrap based inference to estimate the difference in means between groups in the longitudinal functional model. This approach allows us to respect the study design, while modelling the NMR spectra as functions. We model nonparametric fixed and random effect functions that enable us to incorporate covariates and repeated measurements in one model. We investigate NMR spectral regions that are significantly different between groups.

FDA provides access to many functional equivalents of methods commonly used in multivariate statistics, with the benefits of no strong assumptions regarding neighboring observations. The use of FDA in metabolomics can make a valuable contribution to the discovery of biomarkers and identification of metabolic phenotype.

**Keywords:** Functional data analysis; Wavelets; Functional mixed models; Metabolomics



## ABSTRACT - CONTRIBUTED PAPER



### Water quality of the river Rhine: Multivariate process monitoring with multi-block regression methods

Martie Muller, Department of Analytical Chemistry and Chemometrics, Radboud University, Nijmegen, The Netherlands

m.muller@science.ru.nl

Edwin van den Heuvel, Department of Mathematics and Computer Science, Technical University Eindhoven, The Netherlands

Gerard Stroomberg, RIWA-Rhine, Association of River Water Works, Nieuwegein, The Netherlands

Lutgarde Buydens, Department of Analytical Chemistry and Chemometrics, Radboud University, Nijmegen, The Netherlands

Jeroen Jansen, Department of Analytical Chemistry and Chemometrics, Radboud University, Nijmegen, The Netherlands

**Abstract:** Environmental and industrial processes are typically monitored by measuring multiple variables over time. The aim is to reduce product / output variability, improve product / output quality, detect out of control situations and diagnose process faults. Multivariate statistical process control (mspc) based on latent variables is able to deal with the high dimensionality and high collinearity of modern process data. Nevertheless, confounding variation in the process may have a serious impact on process monitoring and may mask subtle changes in the process. In turn this may limit the ability of early, specific and robust detection of "out of control" situations in the system. To take into account confounding variation in the process, we use multi-block partial least squares (pls) regression methods. The approach will be demonstrated by modelling water quality data from the river Rhine. By sequential use of pls regression together with orthogonalisation, (so-pls) the additional or incremental contribution of a new block of data (e.g. Hydrological data, General physio-chemical data or toxic compound data) can be estimated (Næs et al., 2013).

**Keywords:** Sequential Orthogonalized Partial Least Squares. Multi-block regression. Multivariate Process Monitoring.



## ABSTRACT - CONTRIBUTED PAPER

### Longitudinal Meta-analysis of Multiple Effect Sizes

Alfred Musekiwa, University of KwaZulu-Natal School of Mathematics, Statistics and Computer Science  
Alfred.Musekiwa@gmail.com

Samuel Manda, South African Medical Research Council, Biostatistics Unit

Henry Mwambi, University of KwaZulu-Natal School of Mathematics, Statistics and Computer Science

**Abstract:** In recent past, advances in meta-analysis include multivariate meta-analysis where multiple effect sizes are jointly synthesized. Longitudinal meta-analysis is a special case of multivariate meta-analysis where each study reports an effect size at each of multiple time points. Random effects models using both study and random time effects with different covariance structures have been proposed for longitudinal meta-analysis. However, in cases where each longitudinal study reports multiple effect sizes at each of multiple time points, the joint meta-analysis of such studies have not yet been developed. In this study, modelling and estimation approaches of longitudinal meta-analysis where the effect sizes are multiply measured at each time point, are given. A practical example data set is used to show the performance of the models proposed.

**Keywords:** longitudinal meta-analysis; multivariate; multiple effect sizes



## ABSTRACT - CONTRIBUTED PAPER

### Modeling the effect of climate on incidence of respiratory syncytial virus infections in a refugee camp in Kenya: A non-Gaussian time-series analysis

Raymond Musyoka, Center for Disease Control and Prevention (CDC,Kenya)  
uys6@cdc.gov

Jimmy Omonyi, University of Groningen, Netherlands

Samuel Mwalili, Center for Disease Control and Prevention (CDC,Kenya)

Thomas Achia, Center for Disease Control and Prevention (CDC,Kenya)

Anthony Gichangi, Jhpiego, Kenya

Jamal Ahmed, Center for Disease Control and Prevention (CDC, Kenya)

Rachael Joseph, Center for Disease Control and Prevention (CDC, Kenya)

Nina Marano, Center for Disease Control and Prevention (CDC, Atlanta)

Henry Mwambi, University of KwaZulu- Natal

**Abstract:** Respiratory syncytial virus (RSV) is a major cause of acute lower respiratory tract infections in children. RSV infections often occur seasonally in temperate and tropical climates. We compared four models to identify the best one to describe the relationship between climatic factors and RSV incidence among children less than 5 years of age for the period September 2007 to August 2011 in Dadaab refugee camp, Kenya. A generalized linear model (GLM) with a sinusoidal component accounting for seasonal variation was compared to a generalized additive model (GAM) with smoothing cubic splines, both models with and without decomposition of covariates. Data on climatic factors was obtained from World Weather Watch Program and wind speed, rainfall, temperature, mean dew point, and visibility were included in the models as covariates. The explanatory, predictive power and goodness of fit of the models were compared. Models with decomposed covariates fit RSV incidence data better than those without. The Poisson GAM with decomposed covariates of climatic factors fit the data well and had higher explanatory and predictive power than GLMs. For Poisson GAM, the trend 2.80(4.023) and seasonal -11.78(4.023) effects of wind speed, seasonal effect of rainfall -3.34(2.017), trend effects of mean dew point 4.94(1.148) and visibility -6.43(1.492) all significantly explained RSV incidence. This model facilitates further assessment and validation of the use of climatic data to predict RSV incidence in the tropics. Our approach has applications for using climatic factors to predict incidence of other diseases.

**Keywords:** Respiratory syncytial virus; time series; seasonal; climate; modeling.



## ABSTRACT - CONTRIBUTED PAPER

### Assessment of Risk Determinants in the Regularity of Malaria Using Cox Proportional Approach

Ruffin Mpiana Mutambayi, University of Fort Hare

mmutambayi@ufh.ac.za

James Ndege, University of Fort Hare

Adeboye Azeez, University of Fort Hare

Akin Odeyemi, University of Fort Hare

YongSon Qin , University of Fort Hare

**Abstract:** Malaria is a life-forbidding blood sickness promoted by parasites spread to human beings through the bite of the Anopheles mosquito. The source of the mosquito responsible of malaria parasite vary from generally living in unhygienic conditions such as poor environmental conditions as frequently exercised in Lubumbashi. Cox Proportional Hazard concepts was utilized to diagnose the risk factors influencing the relapsing of malaria patients.

The model  $h(t)=h_0(t)\exp(1.91613X_{\text{dump}}-0.49633X_{\text{spr}}+0.81466X_{\text{inf1}})$  was found to fit the data better (Likelihood Ratio: 18.2264 with  $p<0.0004$ , Score: 17.6569 with  $p<0.0005$  and Wald: 19.3975 with  $p<0.0002$ ). The covariates, 'dumping' ( $p<0.0106$ , 95% HR C.I: 1.545, 29.451), 'spray' ( $p<0.0220$ , 95% C.I: 0.391, 0.915), and 'information' ( $p<0.0012$ , 95% C.I: 1.380, 3.725) are affecting the time to relapse.

Since the risk to relapse depends on the exposure time to risk factors, the improved model with covariates 'pit' ( $p\text{-value} = 0.0009$ , 95% HR C.I: 11.4133, 1.243e+04), 'tdwel\_spr' ( $p\text{-value} = 0.0257$ , 95% HR C.I: 1.0016, 1.030e+00), 'tdwel\_inf1' ( $p\text{-value} = 0.0122$ , 95% HR C.I: 0.9691, 9.964e-01), 'maint\_dump' ( $p\text{-value} = 0.0005$ , 95% HR C.I: 1.0239, 1.087e+00), 'maint\_pit' ( $p\text{-value} = 0.0007$ , 95% HR C.I: 0.9179, 9.774e-01), 'dump\_spr' ( $p\text{-value} = 0.0027$ , 95% HR C.I: 0.9308, 9.853e-01), 'dump\_inf1' ( $p\text{-value} < .0001$ , 95% HR C.I: 0.9849, 9.947e-01) and 'spr\_inf1' ( $p\text{-value} = 0.0005$ , 95% HR C.I: 1.0119, 1.043e+00) was found fitting much better the data than the standard model (AIC value improved model was 754.46 and AIC value of the standard model was 767.428).

**Keywords:** cox determinants malaria; modelling; risk



## ABSTRACT - CONTRIBUTED PAPER



### Multivariate Stable distributions and its applications to BRICS financial data

Kimera Naradh, University of KwaZulu-Natal

kimmi.naradh@gmail.com

Knowledge Chinhamu, University of KwaZulu-Natal

Retius Chifura, University of KwaZulu-Natal

Jahvaid Hammujuddy, University of KwaZulu-Natal

**Abstract:** Brazil, Russia, India, China and South Africa (BRICS) are regarded as the five major emerging economies where all BRICS members are a part of a select group of developing industrialized countries. In the financial industry various models are used for the description and analysis of financial trends, one of these models is stable distribution which allows for the skewness and heavy tails that are frequent in financial data. The main objective of this study is to investigate the fit of stable distributions for exchange rates of each of the BRICS countries against the U.S Dollar in both the univariate and multivariate cases. The data set consists of exchange rate data from the period January 2011 to January 2016. Nolans S\_0 -parameterization stable distribution was fitted using maximum likelihood method in the univariate case and in a fitted stable model where a GARCH (1, 1) filter was applied to the returns (Stable-GARCH(1,1)). The Kolmogorov-Smirnov test and the Anderson-Darling test show that stable distributions adequately fit the returns of BRICS financial data. Value-at-Risk (VaR) calculations and VaR in sample backtesting using Kupiec Likelihood Ratio test and the Christoffersens Conditional Coverage test were applied as per the International Basel Regulatory where the robustness of each model describing the financial data was evaluated. Thereafter we proceeded to fit bivariate elliptical stable models using the Rachev-Xin-Cheng method after analyzing the scatterplot matrix of BRICS countries. This study validates the usefulness of stable distributions for modelling BRICS financial data.

**Keywords:** BRICS; stable distribution; exchange rates; GARCH; VaR; backtesting; bivariate elliptical stable model.



## ABSTRACT - CONTRIBUTED PAPER

### The effects of the splitting rule on the predictive performance of random survival forests

Naseije Justine, University of KwaZulu-Natal

justinenaseije@gmail.com

Prof. Henry Mwambi, University of KwaZulu-Natal

Maia Lesosky, University of Cape Town

**Abstract:** Random survival forests are known to be robust in analyzing survival data. Several split-rules have been suggested in literature and some of these have been used in building survival trees for random survival forests. Each split-rule is known to produce a different random survival forest model. However, the effects of a splitting rule on the predictive performance of random survival forest models remain widely unknown.

In this study, we instigate the predictive performance of random survival forest models consisting of survival trees built with three split-rules, that is, the log-rank, log-rank score and pure random split-rules on survival data. The three models are compared on simulated survival data-sets with an increasing, decreasing and constant hazard. Real data sets are also used in this study.

**Keywords:** Survival analysis; split-rules; survival trees; and random survival forests  
Survival analysis; split-rules; survival trees; and random survival forests



## ABSTRACT - CONTRIBUTED PAPER



### Mind the Gap: Are students prepared for higher education?

Andries Naude, University of the Free State  
NaudeAM@ufs.ac.za

**Abstract:** This is a pilot survey of statistic students to investigate what is the influence of students study habits formed in high school that rolled over to their university study habits. The study is conducted on the first year, first semester students at the University of the Free State. The survey compares the student's prior perceptions with their experiences after the first semester. I consider the literature relating to preparedness and students retention in the semester. Although the perception that students not bridging the gap between school and university quickly and effectively, most students appear to have managed the transition into university life successfully. They did not experience the academic, personal and practical difficulties that they expected. There is, however, a considerable minority, between 20% – 30%, who consistently experience academic and personal problems. For these students coming to University has been a negative experience. These students are at risk, if not from drop-out, then from under-performance and lack of fulfilment.

**Keywords:** Mathematics; Statistics; Preparation; Perceptions; Style of Teaching



## ABSTRACT - CONTRIBUTED PAPER



### Bayesian data augmentation for modeling cancer drug intake in SA

Thamsanqa Innocent Ndlela, University of KwaZulu-Natal  
ndlelathamsanqa90@gmail.com  
Siaka Lougue, University of KwaZulu-Natal

**Abstract:** The collection of household data about cancer is a very difficult task. Most existent data on the issue are highly affected by the problem of missing data. Data augmentation techniques are rarely used compared to other statistical imputation methods. Yet they are designed to tackle such problems. This study aims to improve the quality of statistical analysis of data related to cancer drug intake through unbiased estimates and good precision of results using data augmentation methods. The Bayesian data augmentation is implemented as an alternative for missing data problem and the change made in the quality of outputs is evaluated. The General Household Survey (GHS) 2014 is the main source of data in this study. All the analyses are made using the software R and more precisely the package cat. In this study, we have find that Bayesian data augmentation can solve the problem of missing data in cancer drug intake data. We have find that the model that is already imputed missing data is performing very well compared to the one with missing data. The Bayesian data augmentation performs very well in improving modelling of cancer drug affected by missing data.

**Keywords:** data augmentation; EM algorithm; Missing data; Cancer; Bayesian approach



## ABSTRACT - CONTRIBUTED PAPER



### The Effectiveness of Weighting and Bootstrap in the Estimation of Model Parameters under Complex Sampling

Ariane Neethling, University of the Free State and Stellenbosch University  
ariane\_neethling@yahoo.com

Retha Luus, University of the Western Cape  
Tertius de Wet, Stellenbosch University

**Abstract:** Linear models have underlying assumptions that have to be met in order to assure a good model is fitted. However, complex survey data rarely meet these assumptions. As such it was decided to investigate the use of resampling methods as an alternative to the standard methods, especially for the estimation of the variances of the estimators of the model parameters and the construction of confidence intervals.

The use of different sets of weights will be considered through a comparison of the results obtained in the linear modeling of person income from various explanatory variables identified from the Income and Expenditure Survey of 2005/2006. Since it has been observed that benchmarking methods often result in weights having large variability which could affect the precision of any analyses where they are incorporated, a further consideration in the simulation study will be the application of different weight trimming methods to address this phenomenon.

**Keywords:** Complex sampling; Linear models; Weighting; Trimming; Benchmarking; Bootstrap



## ABSTRACT - CONTRIBUTED PAPER



### Modelling average maximum daily temperature using r largest order statistics: An application to South African data

Murendeni Maurel Nemukula, University of Limpopo and University of the Witwatersrand  
murendeni.nemukula@gmail.com  
Caston Sigauke, University of Venda

**Abstract:** The paper discusses an application of the block maxima approach of extreme value theory in modelling temperature data from South Africa over the years 2000 to 2010. The generalized extreme value distribution for the  $r$  largest order statistics is used to estimate extreme high temperatures which result in high demand of electricity due to use of cooling systems. The maximum likelihood method is used to estimate the parameters. The estimation of the shape parameter reveals evidence that the Weibull class of distributions is a good fit to the data. Extreme quantiles for specified return periods are then estimated. A frequency analysis of extreme temperatures is carried out and the results show that most of the extreme temperatures are experienced in January, February, November and December. This modelling approach is useful to decision makers in Eskom, South Africa's power utility company as it is during the non-winter period that they plan for maintenance of their power plants.

**Keywords:** Extreme value theory; generalized extreme value distribution; temperature; order statistics.



## ABSTRACT - CONTRIBUTED PAPER

### Building Robust Systems Adept at Handling Concept Drift

Mzabalazo Z. Ngwenya, Agricultural Research Council - Biometry  
NgwenyaM@arc.agric.za

**Abstract:** Concept drift appears in learning problems over time such as in the classification of streaming data generated by earth observation satellites. The presence of concept drift means the statistical properties, such as the distribution, of the target variable may suddenly change. This presents a unique challenge to established machine learning algorithms which usually fail and have low predictive accuracy in these instances. Specialised learners and strategies are therefore needed for the task. One approach for handling drift is to train a learner incrementally as data comes in. Here we outline the challenges and design issues that must be addressed in developing such a learner. We also present the Gaussian process classifier (GPC) and sketch a framework of using GPCs as the base learner in building concept drift adept systems that are also robust to noise. This system is tested on an artificial data stream and preliminary results are presented.

**Keywords:** Concept drift; Online learning; Gaussian Process Classifier; Self-adaptive systems



## ABSTRACT - CONTRIBUTED PAPER



### A GLMM analysis of data from the Sinovuyo Caring Families Program (SCFP)

Raymond Nhapi, University of Cape Town

Raymond.Nhapi@alumni.uct.ac.za

Francesca Little, University of Cape Town

Reshma Kassanjee, University of Cape Town

Catherine Ward, University of Cape Town

Inge Wessels, University of Cape Town

**Abstract:** We present an analysis of the data from a longitudinal randomized control trial that assesses the impact of an intervention program aimed at improving the quality of childcare within families. The SCFP was a group-based program implemented over two separate waves conducted in Khayelitsha and Nyanga. The data were collected at baseline, post-test and at one-year follow-up via questionnaires (self-assessment) and observational video coding. Multiple imputation (using chained equations) procedures were used to impute missing information. Generalized linear Mixed Effect Models (GLMMs) were used to assess the impact of the intervention program on the responses, adjusted for possible confounding variables. These summed scores were often right skewed with zero-inflation. All the effects (fixed and random) were estimated through the method of maximum likelihood. Primarily, an intention-to-treat analysis was done after which a per-protocol analysis was also implemented with participants who attended a specified number of the group sessions. All these GLMMs were implemented in the imputation framework.

**Keywords:** randomized control trial; longitudinal; group-based; multiple imputation; GLMM; zero-inflation; intention-to-treat; per-protocol



## ABSTRACT - CONTRIBUTED PAPER

### Multidimensional graphical comparisons of imputation methods for missing values in categorical data sets

Johané Nienkemper-Swanepoel, Stellenbosch University

nienkemperj@sun.ac.za

Niël le Roux, Stellenbosch University

Sugnet Lubbe, University of Cape Town

**Abstract:** Missing values in simulated categorical data sets are imputed using a variety of imputation methods. The difference between the imputation methods are evaluated using exploratory analysis. In the case of multiple imputation, multiple plausible completed data sets are available for further analysis. Several multiple correspondence analysis (MCA) biplots are constructed and optimally aligned using generalised orthogonal Procrustes analysis (GOPA). Rubin's combining rules for estimates obtained from multiple imputed data sets are used as a guideline to combine the aligned multiple MCA biplots into a single average configuration.

The natural groupings of sample points with respect to the category level points (CLPs) are compared throughout the displays in order to evaluate the similarities / dissimilarities of the different configurations obtained from the imputation methods. Furthermore, the efficiency of a final predicted data matrix is explored by using the distances between the sample points and CLPs of the combined average configuration. This will result in a final combined data set for further analysis, as opposed to multiple data sets with separate analyses.

Finally MCA biplots of the predicted data sets are compared to the MCA biplot of the original complete simulated data set using orthogonal Procrustes analysis (OPA). Different measures for the goodness of fit within the Procrustes framework will be used to validate the proposed procedures.

**Keywords:** multiple imputation; multiple correspondence analysis; Procrustes analysis; biplots



## ABSTRACT - CONTRIBUTED PAPER

### Multiple Factor Analysis and Principal Components Analysis with FactoMineR in exploratory data analysis of White wines

Nombasa Ntushelo, Agricultural Research Council

ntushelon@arc.agric.za

Phillip Minnaar, Agricultural Research Council

Zama Ngqumba, Agricultural Research Council and Cape Peninsula University of Technology

**Abstract:** Multiple Factor Analysis (MFA) is recognised as an extension of Principal Components Analysis (PCA) in continuous data and an extension of Multiple Correspondence Analysis (MCA) in Categorical data. The Multiple Factor Analysis (MFA) falls under the Advanced Methods whereas the Principal Components Analysis (PCA) and the Multiple Correspondence Analysis (MCA) fall under the Classical Methods. MFA is applied in data sets where a set of individuals or observations are described by more than one set of variables in a data set. The different sets of variables could be continuous and some could be categorical. An example of a data set where the Multiple Factor Analysis can be applied is a wine data set. Usually, in a wine data set, a set of individuals, as wines, is described by two sets / groups of variables which is chemical variables and sensory attributes. Most of the time, Principal Component Analysis is used to analyse wine data where each set of variables (chemical data/ sensory data) are analysed separately. The advantage of the Multiple Factor Analysis over the Principal Components Analysis is that the MFA produces a combined analysis of results of chemical data and sensory data leading to a combined interpretation of results of chemical data, sensory data together with wines. A comparison of Principal Components Analysis and the Multiple Factor Analysis of wine data will be presented. Both MFA and PCA were performed in R using a package called FactoMineR.

**Keywords:** Multiple Factor Analysis; Principal Components Analysis



## ABSTRACT - CONTRIBUTED PAPER

### Robustness and flexibility modeling of the prior for the unstructured spatial random effect in a disease mapping model

Abiodun Davies; Obaromi, University of Fort Hare, Eastern Cape Province  
daviesobaromi@gmail.com

Yongong Qin, University of Fort Hare, Eastern Cape Province

Ndege James, University of Fort Hare, Eastern Cape Province

Azeez Adeboye, University of Fort Hare, Eastern Cape Province

Akinwumi Odeyemi, University of Fort Hare, Eastern Cape Province

Applications of disease mapping in epidemiology and public health are very vast as a result of an increase in interest in the subject. The main and key model normally used in disease mapping is the Besag, York and Mollie (BYM) model, which combines two random effects, a spatially structured and a spatially unstructured random effect. Usually, the normality (Gaussian) assumption on the spatially unstructured random effect is commonly applied. Evidences of the effects of skewness and kurtosis in any data can influence or reduce the over restrictive assumption of normality. Therefore, in this work, we investigate a more robust and heavy-tailed distributions for the spatially unstructured random effect by considering two special cases of the Generalised Hyperbolic distribution in the disease mapping problem. The generalised hyperbolic (GH) distribution is a hopeful and optimistic distribution to use when data is either skewed on both tails or has kurtosis less than or greater than 3. A significant aspect is that GH distributions incorporate many special cases and limiting distributions. Some examples of these are the hyperbolic, the normal inverse Gaussian (NIG), the (skew) Student's t, and the normal distributions. We focus here on two of the special cases, the NIG and the skew Student's t-distribution in this paper. The inferences from these models are carried out under the Bayesian approach. In this paper we showed that the NIG and the skew Student's t-distribution can produce better results when the normality assumption is violated due to kurtosis parameter.

**Keywords:** spatial; disease mapping; generalized hyperbolic; skew student's t; normal inverse Gaussian.



## ABSTRACT - CONTRIBUTED PAPER

### Bayesian Generalized Linear Mixed Model of Breast Cancer Patients

Ropo Ebenezer Ogunsakin, School of Mathematics, Statistics and Computer Science,  
University of KwaZulu Natal

re.korede@yahoo.com

Siaka Lougue, School of Mathematics, Statistics and Computer Science, University of KwaZulu Natal

**Abstract:** Breast cancer is a specific case of a deadly disease for which prevention strategy needs urgent strengthening to reduce the growth of the sickness in the World. In Africa and particularly Nigeria, Breast cancer is challenging citizens, authorities and to those in the health sector as well as other researchers. Preventing the menace that breast cancer pose in the country is a major concern. The main constraint in the management of breast cancer patients in Nigeria is the limitation of available equipment which makes patients bear the burden of paying the unfordable amount for cancer treatment in a low resource country. Available statistics shows that most of the breast cancer is diagnosed in advanced stages in the developing nations and thus result in a poor outcome. This study aims to test whether the treatment patients received in two different hospitals depends on their socio-demographic and economic background. Analysis of this study is based on the data collected from the Federal Medical Teaching Hospital and State Teaching Hospital in South-Western Nigeria. The software R will be used for classical statistical analysis and the software WinBUGS for the Bayesian analysis. A generalized linear mixed model (GLMM) with a logit link function and binomial family and hospitals at stage two (2) level of analysis is performed for both Classical techniques as well as Bayesian technique. Several diagnosis tests will be performed to answer convergence of the MCMC. The diagnostic test will also be performed in WinBUGS and CODA/BOA packages in R.

**Keywords:** Breast Cancer; Bayesian techniques; Hospital; GLMM; WinBUGS; CODA/BOA



## ABSTRACT - CONTRIBUTED PAPER



### The two-piece hyperbolic secant distribution

Brenda Vuguza Omachar, University of Pretoria  
breshyvugzy@yahoo.com  
Paul van Staden, University of Pretoria

**Abstract:** The hyperbolic secant distribution is a symmetric distribution with a density function that has heavier tails compared to the density functions of the normal and logistic distributions. This paper uses the half hyperbolic secant distribution to generate a two-piece hyperbolic secant distribution, based on the skewing methodology introduced by Balakrishnan et al. (2015). Flexibility in the distributional shape of the two-piece hyperbolic secant distribution, specifically with respect to skewness, is attained through a single shape parameter. The distributional shape characteristics as well as the relationship between the half hyperbolic secant distribution and its two-piece counterpart are discussed.

**Keywords:** Heavy-tailed distribution; hyperbolic secant distribution; skewing mechanism



## ABSTRACT - CONTRIBUTED PAPER



### Analysis of complex sample survey data with various software packages

Wallina Oosthuizen, University of the Free State

[oosthuizenw@ufs.ac.za](mailto:oosthuizenw@ufs.ac.za)

Ariane Neethling, University of the Free State

**Abstract:** For analysing complex sample survey data, the nature of the sample design and weights need to be taken into consideration in order to estimate variances and apply statistical inference. The assumption of independent and identically distributed observations is not met in data from complex sample surveys. This has the effect that commonly used statistical analyses might result in incorrect estimates and conclusions.

The use of different software packages, including SAS, SPSS, STATA and R, for the analysis of complex survey data will be discussed in this presentation. Various methods for variance estimation exist under complex sampling. Estimates obtained from the packages will be compared based on real South African data.

The application of statistical inference under complex survey data will also be addressed and the use of different weights or no weights will be tested.

**Keywords:** Complex sampling; Variance estimation; Statistical inference; Weighting.



## ABSTRACT - CONTRIBUTED PAPER



### Accelerated test with an application to too-big-too-fail portfolios

Anel Oppel, University of Pretoria

anel oppel@gmail.com

M. T. Loots, University of Pretoria

C. Beyers, University of Pretoria

**Abstract:** The event of a default for too-big-to-fail financial portfolios has received much attention due to economic crises such as the USA housing bubble in 2006. The lack of sufficient default information complicates the protection of such portfolios. Default protections have typically, in the past, relied on extreme value theory. The focus of this research is the application of accelerated test techniques to the present problem, where smaller portfolios serve as stressed cases of these large portfolios. In doing so a flexible framework for modelling large portfolio defaults is provided.

**Keywords:** Accelerated test; Binary data; Life-stress function; Market capitalisation; Single stress factor



## ABSTRACT - CONTRIBUTED PAPER



### Estimating correlations within operational loss data

Kevin Panman, Standard Bank Group  
kevin.panman@standardbank.co.za  
Leendert Haasbroek, Standard Bank Group  
Wihan Pieters, Standard Bank Group

**Abstract:** We investigate the differences in the values of correlations based on different aggregation periods of time series loss data. The aggregation periods considered for this study were annual, quarterly and monthly, ie, the losses were binned in one year, one quarter and one month buckets, respectively. We conducted a simulation study in order to cover a wide spectrum of frequencies (sample sizes), severity distributions and dependencies between the severities, choosing the parameters of the simulation study to obtain severity and frequency distributions popular in operational risk loss modeling. Our main conclusion is that the difference in values of the correlation coefficients calculated from aggregate loss severities only becomes material when the inherent correlation in the loss-generating process exceeds approximately 0.5. From a risk management perspective, where annual aggregation is desired due to loss horizons typically being annual, this result implies that aggregation periods shorter than annual can be used, which will increase the number of observations to improve the stability of correlation estimates, and the diversification benefit due to estimating correlation values using a shorter aggregation period will not result in a material misstatement of the diversification benefit, since the differences in the values of the correlations are minimal.

**Keywords:** Correlation estimation; loss aggregation period; time series data; operational risk; capital calculation



## ABSTRACT - CONTRIBUTED PAPER

### Image denoising by principal component analysis using the Discrete Pulse Transform

Christine Papavarnavas, University of Pretoria  
chrissypapav@gmail.com  
Inger Fabris-Rotelli, University of Pretoria

**Abstract:** There are various methods to smooth an image: one method is Principal Component Analysis (PCA) which provides an effective image denoising scheme. The implementation of PCA in statistical signal processing ensures that an image's local features are effectively preserved and the noise removed. The PCA based denoising scheme is spatially adaptive since the image is transformed by skillfully computing a locally fitted basis [1]. Classical PCA reduces the dimensionality of a data set, transforming the original dataset thus preserving only the predominantly significant principle components hence removing noise and trivial information from the image. We investigate using the DPT with PCA for noise removal.

References: [1] L Zhang, W Dong, D Zhang, and G Shi. Two-state image denoising by principal component analysis with local pixel grouping. Pattern Recognition, 43(4):1531-1549, 2010."

**Keywords:** Image analysis; denoising; DPT; LULU



## ABSTRACT - CONTRIBUTED PAPER



### Text Classification with Network Based Features

Thembani Phaweni, University of Cape Town  
thembani@live.co.uk

**Abstract:** We investigate features constructed from various network representations of a collection of text documents, and assess whether using these features improves classification accuracy in document classification tasks. Each document as a network in which nodes are words and edges represent the distance between words. Summary network statistics can be obtained for each document and used, respectively, as direct features or to weight the frequencies in the traditional term-document matrix. We find that the aggregate structural properties fail to improve classification accuracy, but weighting by document-specific network statistics leads to small improvements in accuracy.

**Keywords:** Networks; Text Classification; Machine Learning; Neural Networks



## ABSTRACT - CONTRIBUTED PAPER

### Using hidden Markov models for the analysis of animal acceleration data

Theoni Photopoulou, University of Cape Town  
theoni.photopoulou@uct.ac.za  
Vianey Leos-Barajas, Iowa State University

**Abstract:** Animal-borne accelerometers provide a means of measuring activity in a meaningful and quantitative way for wildlife studies. These data allow ecologists to identify important correlates and drivers of animal activity and hence behaviour. Devices typically record acceleration at high resolution resulting in a natural dependence between observations of behaviour, although this feature has been largely ignored. Hidden Markov models (HMM) can be used to explicitly model serial dependence in acceleration data, either for state prediction or to make inferences about drivers of behaviour. If the objective is state prediction, HMMs can be used in a supervised learning framework and trained to classify unlabelled acceleration data into a finite set of pre-defined categories. To make inferences about the drivers behind observed behaviour, HMMs can be used in an unsupervised learning framework to infer new aspects of animal behaviour when biologically meaningful response variables are used. Although the unsupervised approach is better suited to animal activity measured in the field, an important caveat is that the states may not map to specific behaviours. We outline an application of unsupervised learning approach with an example based on data from a black eagle, where we used an HMM to infer the effects of atmospheric and topographic conditions on animal behaviour. The HMM framework is well suited to deal with the main features commonly observed in accelerometer data, and can easily be extended to suit a wide range of types of animal activity data.

**Keywords:** hidden Markov model; acceleration; animal movement



## ABSTRACT - CONTRIBUTED PAPER



### Jump-Reverting Diffusions and Parity to Continuous-Time Reversion

Etienne Pienaar, University of Cape Town

etiennead@gmail.com

Melvin Varughese, University of Cape Town

**Abstract:** Diffusion models are a flexible class of continuous-time stochastic processes that have found application in numerous fields of science. Using various forms of drift and diffusion functionals it is possible to formulate models with features that accurately mimic the dynamics of real-world phenomena. One such feature that has been particularly useful in fields such as biology and finance is the so-called ‘mean-reversion’ property – a mechanism by which the process tends to revert toward a long-run mean trajectory. Typically, diffusion models which exhibit this property have reversion mechanisms that operate as a continuous force on the trajectory of the process by way of its drift function. Despite the ubiquity of such models in the literature, the behaviour of the mean reversion mechanism itself is rarely discussed beyond alluding to the property. For these purposes, we investigate an alternative mean reversion mechanism for diffusion models at the hand of a jump diffusion specification. Using jump diffusion models with state-dependent jump coefficients, we show that it is possible to formulate models with reversion mechanisms that operate discretely in time and for which the reversion mechanism may have distributional properties which do not depend purely on the diffusion trajectory. Furthermore, we show that standard mean reversion mechanisms such as linear drift specifications can be recovered as a special case of such jump-reverting models, thus establishing a form of parity between the two classes.

**Keywords:** Stochastic Differential Equations; Jump Diffusion Processes; Markov Processes



## ABSTRACT - CONTRIBUTED PAPER



### Benchmarking time series using the Fernandez Optimisation technique for South African private sector companies and municipalities

Sagaren Pillay, Statistics South Africa

sagarenp@statssa.gov.za

Patrick Naidoo, Statistics South Africa

**Abstract:** Statistical data are often compiled at different frequencies. When analysing high and low frequency data on the same variable one often encounters consistency problems. In particular, the lack of consistency between quarterly and annual data makes it very difficult for time series analysis. This paper discusses the processes and challenges for the alignment of the data from the quarterly and annual financial statistics surveys for the private sector and that of municipalities. The process consists of three phases, the initial editing, to deal with large inconsistencies, a presentation of the methodology using the quarterly related series to interpolate the annual series, and an analysis of the results. In the initial editing phase the large differences are resolved by manually editing the input data and imputing for missing data. The statistical temporal disaggregation/benchmarking technique used is based on the Fernandez optimisation method of allowing random drift in the error process. The main characteristic of this method is that quarter- to- quarter movements are preserved while quarterly-annual alignment is achieved. The diagnostics performed indicate that the Fernandez random walk model method produces plausible results.

**Keywords:** disaggregation; benchmarking; Fernandez random walk model private sector; municipalities



## ABSTRACT - CONTRIBUTED PAPER



### A Time series ARIMA mode for rainfall for Katima Mulilo Town of Namibia, 2003-2015

Lydia Ismelda Peneyakulo Pokolo, University of Namibia, Department of Statistics and Population Studies  
ismeldamadix@gmail.com

Lillian Pazvakawambwa, University of Namibia, Department of Statistics and Population Studies

**Abstract:** Katima Mulilo in North East Namibia has a great potential for agricultural activities. An understanding of the rainfall patterns and trends can guide decision makers on strategies pertaining to planting, irrigation, and harvesting scheduling. This paper used monthly rainfall data from July 2003 to December 2015 from the Namibia Meteorological Services to model and forecast rainfall up to December 2020 using the Box-Jenkins ARIMA procedure. The best model was seasonal ARIMA(0,0,0)(2,0,0)[12]. Model diagnostics suggested a good fit.

**Keywords:** Katima Mulilo; rainfall; ARIMA; Namibia; forecasting



## ABSTRACT - CONTRIBUTED PAPER

### A Bias-Variance Analysis of Ensemble Learning Classifiers

Arnu Pretorius, Stellenbosch University

arnu@ml.sun.ac.za

Surette Bierman, Stellenbosch University

Sarel J. Steel, Stellenbosch University

**Abstract:** Consider a supervised learning problem where accurate prediction of a response variable is of interest. Training data consisting of measurements on the inputs and response are used to estimate a function  $f$ . Subsequently the estimated function is used to predict the response. The accuracy of the estimated  $f$  is measured in terms of a loss function. In regression problems the most commonly used loss function is squared-error loss. If an additive error model is assumed, a decomposition of the expected prediction error can be derived as the sum of irreducible error, squared bias and variance. Although the search for good predictors can be restricted to the class of unbiased procedures, it is well known that improved accuracy can often be achieved by trading off a small increase in bias for a larger decrease in variance. Examples of procedures obtained from such an approach are ridge regression and the lasso in linear regression. It is clear that a decomposition of the expected prediction error into bias and variance components is useful when investigating the accuracy of a predictor. The focus in this talk is on bias-variance decompositions of prediction error in classification problems. It will become clear that such decompositions are not nearly as straightforward as for squared-error loss regression problems. In fact, the literature contains many different definitions of bias and variance in classification problems. Some of these will be reviewed and an empirical study illustrating a bias-variance analysis for several ensemble classifiers will be presented.

**Keywords:** Bias-variance analysis; classification; ensemble learning.



## ABSTRACT - CONTRIBUTED PAPER



### Fourier-type monitoring procedures for strict stationarity

Sangyeol Lee, Seoul National University, South Korea

Charl.Pretorius@nwu.ac.za

Smios G Meintanis, National and Kapodistrian University of Athens, Greece

Charl Pretorius, North-West University

**Abstract:** We propose model-free monitoring procedures for strict stationarity of an arbitrary time series. The new criteria are formulated as L<sub>2</sub>-type statistics incorporating the empirical characteristic function. Asymptotic as well as Monte Carlo results are presented. The new methods are also employed in order to test for possible stationarity breaks in time series data from the financial sector.

**Keywords:** Strict stationarity; Change-point; Empirical characteristic function; Block bootstrap



## ABSTRACT - CONTRIBUTED PAPER

# An application and comparison of the Fine & Gray competing risks regression model and the Wei, Lin & Weissfeld marginal model to model time to first episode of childhood pneumonia, severe versus non-severe: The Drakenstein child health study

Jordache Ramjith, University of Cape Town

jo.ramjith@uct.ac.za

Landon Myer, University of Cape Town

Heather J Zar, University of Cape Town

Francesca Little, University of Cape Town

**Abstract:** Pneumonia is one of the leading causes of death in children under the age of five in developing countries. The revised WHO classification and treatment of childhood pneumonia document differentiates between mild pneumonia and severe pneumonia emphasizing that the majority of death due to pneumonia are from children with severe pneumonia. Therefore it is important to understand and differentiate risk factors that affect time to first mild and first severe pneumonia.

This is usually modelled using a competing risks framework, like the Fine and Gray (FG) model for the subdistribution of a competing risk, which investigates the cumulative incidence of mild pneumonia and severe pneumonia and the associated risk factors. This framework assumes that competing events are mutually exclusive, however, even though a child will never get mild pneumonia and severe pneumonia at the same time, the child can still get both the first event of mild pneumonia and the first event of severe pneumonia. The Wei, Lin & Weissfeld (WLW) model may be more informative as it allows each individual to experience both events and keeps children at risk for both events. We aim to apply the FG model and the WLW model on data from the Drakenstein birth cohort to investigate the effect of risk factors on time to first mild and first severe pneumonia. The results and model diagnostics will be discussed and compared to understand the impact of using the competing risks model even in the case where an individual can experience both events of interest.

**Keywords:** survival analysis; competing risks; marginal model; cox; childhood pneumonia



## ABSTRACT - CONTRIBUTED PAPER



### Fiducial and objective Bayesian inference for the difference between two binomial proportions

Lizanne Raubenheimer, Rhodes University  
L.Raubenheimer@ru.ac.za

**Abstract:** Estimating the difference between two binomial proportions will be considered. Two non-informative priors will be used, the Jeffreys prior and the probability matching prior. A probability matching prior is a prior distribution under which the posterior probabilities of certain regions coincide with their coverage probabilities. Fiducial inference can be viewed as a procedure that obtains a measure on a parameter space while assuming less than what Bayesian inference does, i.e. no prior. Fisher introduced the idea of fiducial probability and fiducial inference. In some cases the fiducial distribution is equivalent to the Jeffreys posterior. The performance of the Jeffreys prior and the probability matching prior will be compared to a fiducial method and other classical methods of constructing confidence intervals for the difference between two independent binomial parameters.

**Keywords:** Binomial proportions; Coverage; Fiducial; Jeffreys prior; Probability matching prior.





## ABSTRACT - CONTRIBUTED PAPER



### A multiplier approach for approximating and estimating extreme quantiles of compound frequency distributions using the less extreme quantiles of the severity distribution

Helgard Raubenheimer, Centre for BMI, North-West University

Helgard.Raubenheimer@nwu.ac.za

Riaan de Jongh, Centre for BMI, North-West University

Fred Lombard, Centre for BMI, North-West University

**Abstract:** A popular method in modelling the aggregate loss distribution in risk and insurance is the Loss Distribution Approach (LDA). For example many banks currently use the LDA for estimating economic and regulatory capital for operational risk under Basel's Advanced Measurement Approach. The aggregate loss distribution is a compound distribution resulting from a random sum of losses, where the losses are distributed according to some severity distribution and the number (of losses) distributed according to some frequency distribution. This paper studies the approximation of extreme quantiles of the aggregate loss distribution. A key application of this approximation is the estimation of the economic or regulatory capital in a particular operational risk category (ORC). We propose an approach to approximate the extreme quantile of the compound distribution using a combination of a multiplier and the less extreme quantile of the underlying severity distribution. The proposed approximation is assessed via a simulation study.

**Keywords:** Quantile approximation; multiplier; compound distribution



## ABSTRACT - CONTRIBUTED PAPER



### Prediction of the time to reach a biomarker threshold from complex longitudinal data

Tarylee Reddy, Biostatistics Unit, Medical Research Council, I-BioStat, Universiteit Hasselt, Belgium  
tarylee.reddy@mrc.ac.za

Geert Molenberghs, I-BioStat, Universiteit Hasselt, Belgium, I-BioStat, KU Leuven, Belgium

Marc Aerts, I-BioStat, Universiteit Hasselt, Belgium

Edmund Njeru Njagi, I-BioStat, Universiteit Hasselt, Belgium

**Abstract:** In longitudinal studies of biomarkers, an outcome of interest is the time at which a biomarker reaches a particular threshold. Due to the inherent variability of biomarkers, a single measurement below a relevant threshold should be interpreted with caution. Several studies have applied persistence criteria, designating the outcome as the time to the occurrence of two consecutive measurements less than the threshold. We propose a two stage approach to estimate the time to threshold in the presence of persistence criteria, which takes into account three sources of variability: between subject variability, stochastic variability and measurement error. In the first stage, a linear mixed model is fitted to the longitudinal measurements, resulting in patient-specific predicted values that are a function of the fixed-effects and empirical Bayes estimates. In the second stage, the probability of experiencing two consecutive measurements less than a relevant threshold at each time point is computed and substituted into the expression for the expected time to threshold. Applying Bayes theorem and assuming that the first order Markov property holds, this probability can be expressed as a function of cumulative bivariate and univariate normal probabilities. We apply the methodology to a cohort of HIV positive individuals in South Africa, where the time for a patient to reach a CD4 count threshold less than 350 cells/mm<sup>3</sup> is of interest. Further applications to the area of abnormal aneurysm screening are also presented.

**Keywords:** Threshold modeling; linear mixed model; biomarker



## ABSTRACT - CONTRIBUTED PAPER



### Design-based spatial sampling

Hayley Reynolds, University of Pretoria

u12044700@tuks.co.za

Inger Fabris-Rotelli, University of Pretoria

Theodor Loots, University of Pretoria

**Abstract:** The design-based spatial sampling approach and traditional sampling are compared using a rabies vaccination schedule data set. The advantages and disadvantages of each technique will be highlighted.

**Keywords:** spatial statistics; spatial sampling; traditional sampling

180



## ABSTRACT - CONTRIBUTED PAPER

### Anaemia in children: The contribution of malaria to the burden

Danielle Roberts, University of KwaZulu-Natal  
danjader@gmail.com

**Abstract:** Anaemia is a major public health problem faced by many developing countries, particularly in Africa. According to the World Health Organization (WHO), in 2011 the highest prevalence of anaemia was in children under the age of five years old at 42.6%. This translated to 273.2 million children suffering from anaemia globally. In Africa, the prevalence of anaemia in children was estimated at 62.3% in 2011. The adverse health consequences of childhood anaemia include altered cognitive function, impaired motor development and growth, poor school performance, poor immune function and susceptibility to infections. The cause of anaemia is multifactorial, however studies in East Africa have shown that malaria and iron deficiency account for the majority of cases in children. The objective of this study is to investigate the prevalence anaemia in children under the age of five years old in Uganda, as well as to examine the contribution of malaria to the burden of anaemia in this country. This study makes use of data collected from the nationally represented 2014 Malaria Indicator Survey conducted in Uganda, where a multistage cluster sampling design was utilized. Due to this complex survey design, a generalized linear mixed model is used to account for the hierarchical nature of the data, as well as possible correlations that may exist in the observations.

**Keywords:** Anaemia; malaria; generalized linear mixed models



## ABSTRACT - CONTRIBUTED PAPER

### Assessing the robustness of phylogenetic models to changes in selection pressures over time: A simulation study

Hassan Sadiq, University of Cape Town

sdqhas001@myuct.ac.za

Miguel Lacerda, University of Cape Town

Ben Murrell, The University of California, San Diego

**Abstract:** Simulation studies are necessary for sensitivity analyses and the assessment of phylogenetic models. Consequently, these studies represent an important component of model development. Unfortunately, they are often misused. For example, the accuracy of model estimates are commonly assessed using data generated with the models under investigation. It becomes very difficult to identify the faults of the concerned model, because such procedures tend to lead to inferences that are biased in favour of the model. One of the aims of this research is to use population genetics principles in a phylogenetic context to infer natural selection. A second aim is to prompt a renewed interest in the development of more rigorous phylogenetic models that are capable of describing the complex evolutionary patterns of the recent wave of pathogen outbreaks. The main objective is to develop a genetic sequence simulator based on a mutation-selection model framework. This simulator allows multiple parameter settings which include adaptive and time-heterogeneous models of evolution. This enables the simulation of genetic sequences that accurately mimick pathogen evolution. Such sequences were analysed using some of the conventional codon models. It was shown that these models can provide inaccurate estimates of selection pressure, particularly when selection is strong.

**Keywords:** Mutation-Selection model; Natural selection; Population genetics; Phylogenetic model; Simulation; Time-heterogeneous model



## ABSTRACT - CONTRIBUTED PAPER

### Applying multidimensional item response theory models in assessing dimensionality of the Mental Health Continuum Short Form (MHC-SF)

WD Schutte, North-West University

wd.schutte@nwu.ac.za

Lusilda Schutte, North-West University

Marie P. Wissing, North-West University

**Abstract:** In this study we assessed the psychometric properties and latent factor structure of the Mental Health Continuum Short Form (MHC-SF), which is postulated to comprise of three related sub-dimensions, using multidimensional item response theory (MIRT) and confirmatory factor analysis (CFA). In the literature it is often found that unidimensional item response theory (UIRT) models are applied to multidimensional instruments. This culminated in a further aim of this study: The assessment of the consequences of fitting UIRT models to multidimensional instruments, compared to the application of MIRT models. Based on the correlation structures obtained from MIRT models fitted to the MHC-SF, a simulation study was performed to compare item and ability parameter recovery for UIRT vs. MIRT models.

The results obtained from the CFA and MIRT analysis of the MHC-SF confirmed the intended 3-factor structure of the instrument. Response categories indicative of low levels of the latent trait (mental health) seemed to be redundant and it was suggested that less response categories may yield improved psychometric properties. Information curves provided evidence that the scale yielded the most information for low levels of the latent trait, while most respondents attained high levels of the trait. This suggests that the scale would be less sensitive to detect change at high levels of the latent trait, which has implications when this instrument is used to assess the effectiveness of interventions in the general population. The simulation study revealed that there are advantages when MIRT is applied compared to UIRT.

**Keywords:** item response theory; psychometrics; simulation study



## ABSTRACT - CONTRIBUTED PAPER



### Some important remarks regarding the zero-state and steady-state runs-type control charts

Sandile Shongwe, University of Pretoria  
sandile.shongwe@up.ac.za

**Abstract:** There are a variety of methods in the statistical process control and monitoring literature used to account for the zero-state and steady-state modes. In this presentation, we discuss these and then give the corresponding run-length properties' expressions using Markov chain approach. Next, we take a sample of results from some published articles and illustrate how the authors could have easily obtained the same run-length results using our derived expressions instead of simulation and / or other complicated methods.

**Keywords:** Zero-state; Steady-state; Run-length; Markov chain; Control chart; Runs-rules.



## ABSTRACT - CONTRIBUTED PAPER



### Modelling heat waves and their impact on electricity demand

Caston Sigauke, University of Venda  
csigauke@gmail.com

**Abstract:** Recently the sub-Saharan region has experienced extreme heat waves. This phenomenon requires use of extreme value theory distributions in predicting the frequency of occurrences of these hot spells. In this paper we explore the use of boundary corrected extremal mixture models in modeling heat waves. We show that as temperature increases and converges to its upper bound, the marginal increases in electricity demand also converge. This modeling approach helps system operators in scheduling and dispatching of electricity during the heat wave period.

**Keywords:** Extremal mixture models; electricity demand; heat waves; nonlinear detrending



## ABSTRACT - CONTRIBUTED PAPER



### Diagnostics for joint models for longitudinal and survival data

Isaac Singini, University of Cape Town  
sngisa001@myuct.ac.za

**Abstract:** Joint models for longitudinal and survival data are a class of models that jointly analyse the behaviour of repeated measurements and event times. These models have two practical applications firstly focusing on survival outcome whilst accounting for time dependent covariate measured with error and secondly focusing on the longitudinal outcome while controlling for non-random dropout. There has been considerable interest in the estimation of these joint models in the past two and a half decades. However, little attention has been given towards the development of diagnostics for these joint models. The available diagnostic tools have mainly been based on separate analysis of residuals for the longitudinal and survival sub-models which could be sub-optimal. In this study we explore extensions to existing residual based diagnostics, multiple imputation residuals methodology which is due to the fact that reference distribution of residuals when there is non-random dropout in longitudinal response is inconsistent. The idea proposed by Dimitris(2010), is inspired by augmenting the observed data with randomly imputed data points that would otherwise have been provided had the subject not dropped out hence creating a complete data from which residuals can be calculated and plots produced. We will also explore Cook's distance for the separate sub-models. This exploration is conducted using data from a multi-centre clinical trial on TB pericarditis.

**Keywords:** Joint model; diagnostics.



## ABSTRACT - CONTRIBUTED PAPER



### Inference on Time Series Models for Paired Comparisons

Morné Sjölander, University of the Free State  
sjolanderM@ufs.ac.za  
Paul Meades, University of the Free State

**Abstract:** The Method of Paired Comparisons is a technique that is used to rank objects with respect to a latent trait, by a panel of judges who record the preference on objects presented to them in pairs. The field has its roots in Psychometrics, by the author G.T. Fechner (1860) and later L.L. Thurstone (1927) gave it a more Mathematical foundation. Research on Time Series Models for Paired Comparisons was done by M.E. Glickman (1993) and by M.R. Sjölander (2012). Sjölander extended the Bradley-Terry model of R.A. Bradley and M.E. Terry (1952) to make the model time dependent in a linear, log-linear and sinusoidal fashion. In our research, we will be doing various types of inference on the models by Sjölander, as this has not been done yet. We will look specifically at various types of hypothesis test on the models, as well as the respective resulting confidence intervals that follow. We will be testing the results only theoretically at this stage.

**Keywords:** Time Series; Paired Comparisons; Bradley-Terry model; Inference



## ABSTRACT - CONTRIBUTED PAPER

### New Procedure for Probabilistic Tsunami Hazard Assessment from Incomplete and Uncertain Data

Ansie Smit, University of Pretoria Natural Hazard Centre, University of Pretoria  
ansie.smit@up.ac.za

Andrzej Kijko, University of Pretoria Natural Hazard Centre, University of Pretoria

Inger Fabris-Rotelli, Department of Statistics, University of Pretoria

Paul J. van Staden, Department of Statistics, University of Pretoria

**Abstract:** The 2004 Sumatra and 2011 Japan mega earthquakes and tsunamis focused the world's attention on their devastating impact on infrastructure and economic stability. Tsunamis are a natural phenomenon with a low probability of occurring, but with a high impact factor. Observed tsunamis do not occur often with many events known only through geological investigations and historical narratives. Many tsunami-threatened coastal areas therefore have long historic records available containing information of the largest and catastrophic tsunami occurrences. Traditional probabilistic tsunami hazard analysis (PTHA) procedures require earthquake and/or tsunami observations spanning hundreds of years, but are not always capable of accommodating poorly populated catalogs. We introduce a new, empirical technique for PTHA which permits the estimation of the key tsunami distribution parameters when the tsunami uncatalogued consists of the historic and the most recent, instrumentally recorded tsunami events. The three recurrence parameters are, according to the Soloviev frequency-intensity ( $i$ ) relation, defined as the mean tsunami activity rate  $\lambda$ , the parameter  $b$ , which describes the relation between weak and strong events, and the coastline-characteristic maximum possible tsunami intensity  $i_{\max}$ . Estimates for these parameters are calculated by taking into consideration incompleteness of the catalog, uncertainty in the tsunami intensity determination, and the uncertainty associated with the applied tsunami occurrence models. Occurrence model uncertainty is introduced by assuming that the recurrence parameters are random variables, each described by the gamma distribution. This approach results in the extension of the frequency-intensity relation and the Poisson distribution of the number of tsunamis with their compounded counterparts.

**Keywords:** incomplete and uncertain data; uncertain; probabilistic tsunami hazard assessment



## ABSTRACT - CONTRIBUTED PAPER



### Dealing with uncertainty in spatial big data studies

Alfred Stein, University of Twente  
a.stein@utwente.nl

**Abstract:** There is a current research interest on big data. This interest is particularly relevant in Earth Observation studies, where satellite images are increasingly available for free. Still, uncertainty remains and spatial statistics is exceptionally suited to handle such spatial big data. It offers opportunities to summarize the data, and express measures of variation and uncertainty. Spatial statistics depends upon the notion of spatial (and spatio-temporal) dependences, and such dependence in turn depends upon the notion of distance between points. For n observations, including their coordinates in space or space and time, evaluating distances requires inspection of  $n^2$  pairs of points, and here steps should be made to be able to do this efficiently. The current data structures as such are usually able to handle the big data as well. Specific procedures have to be developed that are able to address issues that are relatively novel (such as combining data in the space-time domain) or that have to address specific questions and problems, i.e., to select data from a big data set for a particular model application. A particular way ahead may be that classification of the data into multiple classes is done in the form of metadata. In such a way it is possible to make the big data of relevance in a wide range of practical applications. This would require an improved database structure and in particular a very much adaptive spatial statistical analysis procedure.

In this presentation an overview of roadblocks and challenges for spatial statistical methods in a big data analysis will be presented. The presentation will be illustrated with a range of examples on bushfires, slum identification and urban land use.

**Keywords:** big data; Earth Observation studies



## ABSTRACT - CONTRIBUTED PAPER

### Limitations to Probability Modelling in Strategic Decision Analysis and an Alternative Approach

Theodor Stewart, University of Cape Town  
theodor.stewart@uct.ac.za

**Abstract:** Standard decision analysis is based on the axioms of expected utility theory (EUT), also generalized to multiattribute utility theory (MAUT) for multicriteria outcomes. The validity of the conclusions arising from EUT/MAUT is dependent upon the availability of a proper probability distribution on the performance measures describing outcomes from alternative courses of action, even when the distribution is subjectively assessed. In the analysis of policy alternatives in strategic planning, however, the identification of an appropriate probability distribution is severely hampered by the existence of "deep" uncertainties, i.e. of future states for which a meaningful sample space is not even conceptually realizable.

Strategic planners often represent such deep uncertainties in terms of a small number of "scenarios". The set of scenarios does not begin to approximate a sample space. Thus attempts at defining "probabilities" on scenarios is not a valid process, and in any case pointwise local densities ("likelihoods") may be grossly inappropriate for calculation of expectations. We have proposed an alternative approach to decision analysis for deep uncertainties, based on multiattribute value theory (MAVT). Simple evaluations of interval scaled ordering of policy alternatives in terms of each combination of scenario and decision criterion are elicited. These input preference orderings can be aggregated additively using the principles of MAVT into a final preference ranking of the policies, where the required weights would be elicited by "swing-weighting" methods which are well-established in MAVT. These weights are not probabilities, but rather the importance to management of being prepared for each scenario.

**Keywords:** Decision analysis; scenario planning



## ABSTRACT - CONTRIBUTED PAPER

### Statistical Modelling with Quantile-based Methods and L-moments

ML Steyn, Stellenbosch University  
17070759@sun.ac.za

**Abstract:** This talk explores quantile-based methods used for estimation and inference on a probability distribution. The quantile function is defined and fundamental theorems of it are proven. Furthermore, the empirical quantile function is defined and it is briefly indicated how non-parametric statistics of location, scale and shape can be constructed with the aid of this function.

Analogous to product moments, Hosking (1990) defined linear moments (L-moments). Properties of L-moments make them useful for statistical modelling. These quantities are defined and an overview of the most relevant properties and characteristics of L-moments is given. It is shown that L-moments can be represented as a linear combination of probability weighted moments or as an L-functional with weight-function equal to a shifted Legendre polynomial. Additionally, an introductory overview of using L-moments to describe the location, scale and shape of a probability distribution is given.

The properties of L-moments make these measures useful for estimation and inference. It is therefore discussed how linear moments can be calculated from a sample. Three estimators of sample L-moments are given. These are the unbiased estimator of linear moments, the plotting position estimator of linear moments, both proposed by Hosking (1990), and the Wang (1996) scheme. Properties including the limiting distribution (Hosking, 1990), the exact variance structure and a distribution-free unbiased estimator of the variance of L-moments (Elamir & Seheult, 2004) are discussed. The talk is concluded with some practical applications of linear moments to motivate their wide range of uses.

**Keywords:** Quantiles; Quantile function; Order statistics; L-moments; Probability weighted moments; L-functional



## ABSTRACT - CONTRIBUTED PAPER

### Smooth copula-based estimation of the conditional density function and the conditional hazard rate function

Jan WH Swanepoel, North-West University

Jan.Swanepoel@nwu.ac.za

Paul Janssen, Hasselt University, Belgium

Noël Veraverbeke, Hasselt University, Belgium

**Abstract:** Some recent papers deal with smooth nonparametric estimators for copula functions and copula derivatives. These papers contain results on copula-based Bernstein estimators for conditional distribution functions and related functionals such as regression and quantile functions. See Janssen, Swanepoel, Veraverbeke (2012, 2014, 2016).

The focus in this talk is on new copula-based smooth Bernstein estimators for the conditional density and the related conditional hazard rate function. The classical nonparametric kernel estimator for the conditional density function goes back to Rosenblatt (1969). It is the ratio of the joint and the marginal kernel density estimator. Our approach avoids going through separate density estimation of numerator and denominator. The drawbacks of the kernel-based ratio estimator motivated Faugeras (2009) to consider a product-type estimator of a marginal density and a copula density estimator. All these estimators, however, are not smooth. Our estimator is defined as a smoother of the copula based Bernstein estimator of the conditional distribution function. In a similar spirit, our conditional hazard rate estimator is not obtained from the ratio of the conditional density and conditional survival function, but as a smoother of the conditional cumulative hazard function estimator. The latter is based on Bernstein estimation.

We establish asymptotic properties of bias and variance and discuss the asymptotic mean squared error in terms of the smoothing parameters. We also obtain the asymptotic normality of the new estimators. In a simulation study we show the good performance of the new estimator in comparison with the estimators proposed in Faugeras (2009).

**Keywords:** Asymptotic distribution; Bernstein estimation; Copula; Conditional density; Conditional hazard rate



## ABSTRACT - CONTRIBUTED PAPER

# Addressing the deficits in urban statistics across Africa: challenges and opportunities with the 2010 and 2020 Population and housing census rounds for monitoring the SDG's urban core (Goal 11)?

Gabriel Tati, Department of Statistics and Population Studies, University of the Western Cape  
gtati@uwc.ac.za

**Abstract:** Africa is regarded as the continent with the fastest growing urban population in the world. Yet most African countries do not have in place well performing statistical system to collect, compile and produce on regular basis statistics on their cities, making it difficult to compare the trends within and across countries and monitor urbanisation harmoniously. Many countries rely on estimates and projections produced elsewhere (UN and World Bank mainly) to obtain even the very rudimentary indicator of urbanisation, namely the urban population percent. The communication provides an overall assessment of the current practices in that domain and the deficits in the production of pertinent urbanisation indicators. The assessment informs that most of the deficits in urban statistics emanate from the inadequate consideration of urbanisation indicators to derive from the conventional sources of official statistics (population census, household surveys and administrative data) but also from the more digitised ones (GIS). The demands and needs for appropriate urbanisation indicators in line with the urban core (Goal 11) of the Sustainable Development Goals are huge in all the dimensions of urbanisation (demographic, social, economic and environmental). This requires data for each urban locality, treated as a unit of analysis regardless of the size, in order to capture their functions and patterns of development within the national urban system. Adding to the previous censuses and household surveys, the 2010 and 2020 population census rounds come with challenges and opportunities in meeting those demands and needs for urban statistics as they provide more refined variables.

**Keywords:** Urbanisation tempo; spatial planning; urban system; spatial data; official statistics; National Statistical System; African cities.



## ABSTRACT - CONTRIBUTED PAPER

### Determining Sample line spacing for a Marine Diamond deposit based on D-Optimality

Christien Thiart, Department of Statistical Sciences, University of Cape Town  
christien.thiart@uct.ac.za

Linda Haines, Department of Statistical Sciences, University of Cape Town

Jana Jacob, Namdeb Diamond Corporation, Oranjemund, Namibia

Chris Prins, Anglo American:MinRes, Johannesburg

**Abstract:** The Southern Namibian onshore diamondiferous linear beach environment has been delineated between the 1930s and 1960s using sampling trenches comprising 1mx5m samples. The sampling trenches are perpendicular to the coast and spaced roughly 500m apart, covering a coastal strip about 100km northwards from the Orange River mouth. The onshore diamond resource has virtually been mined out and the current exploration focus has shifted to the submerged beaches extending directly offshore from the mined-out area. An offshore strip of 100kmx4km parallel to the coast and with very limited data is the target area for future resource development. This strip will be sampled with trenches a certain distances apart. Research is currently in progress to establish amongst other objectives, an optimal sample line spacing in order to provide a well-defined variogram structure.

Prins and Jacob (2014) fitted a variogram to the existing onshore sample line data and using this data and the structure and parameters of the variogram, an optimal line spacing needs to be determined, for exploring the offshore target. The covariance structure and the variance components are thus assumed known and are used as a proxy for the offshore target area. By maximizing the determinant of the information matrix for the variance components we find a D-optimal line spacing. This work extends the work of Lyman (2007), by introducing the concept of pillars to stabilize the spacing. We also investigate the role of the range, nugget and sill in a factorial setup by means of a scaled down, fictitious example.

**Keywords:** Marine Diamond deposit; D-Optimality; variogram; Line spacing



## ABSTRACT - CONTRIBUTED PAPER

### Multi-criteria decision making in Taguchi orthogonal designs

Derya Turfan, Hacettepe University

deryaturfan@gmail.com

Ayfer Ezgi Yilmaz, Hacettepe University

Serpil Aktas, Hacettepe University

**Abstract:** Taguchi method is a robust and multi-parameter optimization statistical technique which employs less number of experiments to identify and optimize parameters to reach the optimum response. Taguchi method utilizes full fractional design called orthogonal arrays that are the minimum set of experiments which represents the various combinations of factors. While Taguchi method optimizes the single response cases, it is not a proper method for optimizing a multi-response problem. Multi-criteria decision making methods enable to select the best option among a large number of alternatives and criteria such as the factor levels in Taguchi orthogonal designs. In this paper, the optimum levels of the factors are estimated and compared using the desirable functions and multi criteria decision making methods on a real data set. The signal-to-noise ratio for each criterion is calculated as well.

**Keywords:** Taguchi orthogonal designs; multi-criteria decision making; experimental design; optimal design; robust design



## ABSTRACT - CONTRIBUTED PAPER



### Income estimation

Daniel Uys, Stellenbosch University  
DWU@sun.ac.za

**Abstract:** Income estimation is a challenging and complex problem. Through the years several econometric modelling procedures try to answer this estimation problem, some of these with little success. Due to the reluctance of the general public to specify their income as a monetary value, income is usually recorded as a categorical variable. In this talk statistical learning techniques, such as classification trees and random forests, attempt to predict an individual's income category based on several independent variables such as age, gender, employment status and sector of employment. Data from various data sources are used for this purpose.

**Keywords:** income; classification trees; random forests

196



## ABSTRACT - CONTRIBUTED PAPER



### On the quantification of model risk in statistical models of empirical data

Liesl van Biljon, Standard Bank  
liesl.vanbiljon@standardbank.co.za  
Kevin Panman, Standard Bank

**Abstract:** Real-world data is often represented by statistical models, for example representing observed credit or operational loss data using a parametric distribution. These models are by definition simplifications of complex real-world phenomena since these models are used to describe the intricate nature of an underlying process, for example the loss generating process in financial risk analyses. Therefore model risk is deliberately created through this simplification. The magnitude of model errors depends amongst other on the complexity of the underlying truth that the models attempt to mimic. In financial risk management a relatively small error in the model could lead to a significant financial loss or reputational damage. It is therefore necessary to measure the model error to appreciate the inherent model risk.

This paper focuses on loss models used in the financial industry. The major contributing factors of model risk are highlighted. Using a simulation study, the model risk due to parameter uncertainty and model specification is quantified. The results of the simulation study are presented. Lastly, future research is also proposed with the intention to improve modelling practices in the financial industry.

**Keywords:** Model risk; quantification; financial risk models



## ABSTRACT - CONTRIBUTED PAPER



### The Association of Statistics Anxiety, Attitude Toward Statistics and Mathematical Self-Concept with Performance in a Business Statistics Class

Linda van der Merwe University of the Free State  
vdmerwel@ufs.ac.za  
Moletenyane Tankiso Mokhele, University of the Free State

**Abstract:** Statistics anxiety is a pervasive problem in many fields of study. A large proportion of students identify statistics courses as the most anxiety-inducing courses in their curriculum. Statistics anxiety is defined as a specific anxiety that arise when taking a statistical course or by working with statistical analysis which includes gathering, processing and interpreting the data.

It is important to investigate students' statistics anxiety as it can negatively affect students' performance and their overall psychological and physiological condition. Furthermore, understanding about a student's level of anxiety may help teachers find ways to reduce the level of anxiety and enhance the learning experienced by students.

This study examined the relationship between statistics anxiety, attitude toward statistics, and mathematics self-concept as well as their effect on performance in an introductory business statistics course. In addition, the study aimed to determine whether statistics anxiety differs by gender and to investigate the experiences and opinions of students regarding statistics anxiety by means of interviews. Statistics anxiety and attitude toward statistics was measured using the Statistics Anxiety Rating Scale (STARS). Ten questions were added to the STARS to measure mathematics self-concept. Performance measures included two tests and final examination marks. Face-to-face, semi-structured interviews were conducted after the examination was written.

The above mentioned relationships, their effect on performance, gender differences regarding statistics anxiety and students' opinions and experiences regarding statistics anxiety will be discussed.

**Keywords:** statistics anxiety; attitude toward statistics; academic performance; mathematical self-concept; gender differences



## ABSTRACT - CONTRIBUTED PAPER

### The eating habits of bat-eared foxes

Sean van der Merwe, University of the Free State

vandermerwes@ufs.ac.za

Stéphanie Périquet, University of the Free State

**Abstract:** Foxes were observed over a period of time and their prey was carefully noted. We investigate various new multivariate regression models in order to determine the effect of the moon and the seasons on the types and sizes of prey consumed by the foxes.

The complex multivariate nature of the data eliminates the use of standard software packages. Instead we implement models using the following interface options in R: Dirichlet regression, using DirichletReg package. Negative Binomial with inflated zeros and multivariate random effects, using INLA package (Integrated nested Laplace approximation). Censored LogNormal regression using custom BUGS models, using R2OpenBUGS package.

We find that the models produce consistent results. For example, the foxes appear to eat smaller prey, especially termites, in bright moon conditions, and in Autumn.

**Keywords:** Multivariate Regression; Bayes; INLA; Dirichlet; Bat-eared foxes; BUGS



## ABSTRACT - CONTRIBUTED PAPER

### Multivariate characterization: multivariable vs multi-block methods in DSA

Marieta van der Rijst, Agricultural Research Council - Biometry  
vanderrijstm@arc.agric.za

Tormod Næs, Nofima Norway

Nina Muller, Stellenbosch University

**Abstract:** Descriptive sensory analysis (DSA) uses humans as instruments to generate quantitative data on the perceived sensory characteristics of a product to describe its profile. However, human responses are complex responses to a mixture of several visual, chemical or structural components of a product. The purpose of statistical analysis on DSA data is to reflect human perception of a product by revealing perceived dimensions latent in the product, composed of the contribution the product descriptors.

Data generated from DSA typically consists of different groups or types of descriptors, for example appearance (visual), aroma, flavour (chemical) and texture (structural) descriptors. Traditionally DSA data has been analysed using univariate Anova. However, the intricate nature of DSA data requires statistical models that can extract core information from large data matrices, to present results in a simplified, more interpretable format. The rapid development of multivariate statistical techniques in recent years has resulted in a large spectrum of methods, with very different applications, being available. In this paper methods that split data into conceptually meaningful blocks, as opposed methods that use all variables simultaneously in one large data matrix, will be discussed and compared, using DSA data as illustration.

**Keywords:** descriptive sensory analysis; product profile; multivariate statistical techniques; multi-block methods



## ABSTRACT - CONTRIBUTED PAPER

### Multivariate Bayesian computation

Janet van Niekerk, University of Pretoria

janet.vanniekerk@up.ac.za

Andriëtte Bekker, University of Pretoria

Mohammad Arashi, Shahrood University of Technology, University of Pretoria

**Abstract:** Higher dimensional data is the reality of today's technologically advanced world. The implementation of Bayesian analysis for this type of data and the associated computational aspects proves very challenging, more so with non-conjugate complicated prior structures. In this paper algorithms for computing Bayes estimators and posterior characteristics are proposed and evaluated, specifically for the scale matrix of a multivariate t model. Convergence of the proposed algorithms are investigated. There is currently no software that can accommodate multivariate Bayesian analysis with non-conjugate priors and this paper addresses exactly this issue.

**Keywords:** Bayesian inference; convergence; scale matrix; multivariate Gibbs sampling.



## ABSTRACT - CONTRIBUTED PAPER



### Distribution-free cusums for location and scale

Corli van Zyl, North-West University  
vanzylcorli@gmail.com  
Freek Lombard, North-West University

**Abstract:** We consider self-starting cusums for location and scale changes based on the sequential ranks of data that arrive singly over time. The cusums have the same form as a standard normal cusum. However, the control limit guaranteeing a nominal in-control average run length does not depend upon the type of the underlying distribution. Thus, our cusum procedures do not require a parametric specification of the underlying distribution density and are, in this sense, distribution-free. In contrast to other rank based cusum procedures, we are able to establish a link between the target out of control parameter value and an appropriate measure of dispersion, which in the present instance is the Fisher information. This enables us to make a rational choice of the cusum reference value that would be appropriate for quickest detection of a targeted out-of-control value. The measurement of the performance of our cusums is based on their out-of-control average run length properties which we gauge via theory based calculations supplemented by Monte Carlo simulation. The use of the proposed procedures is illustrated by application to some industry data. Where appropriate, comparisons are made with comparable parametric procedures.

**Keywords:** Statistical process control; Cusum procedures; distribution-free; sequential ranks; distribution-free; location change; scale change.



## ABSTRACT - CONTRIBUTED PAPER

### Anomaly detection

Melvin Varughese, IBM Research Lab, University of Cape Town  
melvin.varughese@gmail.com

**Abstract:** The timely detection of anomalies is important in many contexts. Anomaly detection techniques can be applied in a variety of areas: from fraud detection to finding early signs of mechanical failure to the discovery of as yet undiscovered classes of astronomical objects. In contrast to supervised learning techniques, an anomaly classifier must be constructed with only the normal class(es) being well-sampled and characterized by the training data. As such, anomaly detection is a form of semi-supervised learning. This talk will compare and contrast the main approaches to anomaly detection. A discussion of three pressing problems within anomaly detection will be undertaken: first, what is the optimal way to combine classifiers in order to maximize the accuracy of anomaly detection? Secondly, how can the machine learning process be made interactive? That is, how should the algorithm be constructed to selectively ask a human for feedback on data points that would be particularly informative if labelled? Finally, given the ever-growing volume of available data, how can novelty classifiers be adapted to work in an online or distributed environment?

**Keywords:** Semi-supervised learning; Machine Learning; Anomaly Detection; Novelty Detection; Outliers



## ABSTRACT - CONTRIBUTED PAPER



### Survival probability for a defended system

Jan-Paul Venter, University of the Free State  
janpaulventer@gmail.com

**Abstract:** We derive an exact survival probability, where the current literature only considers approximate solutions, for a large non-repairable “main” system that is defended by a repairable “defense” system. The system is subject to two types of attacks: those against the defense system and those against the main system. The effectiveness of the defense system is dependent on its state of repair. In order to obtain the exact survival probability we first determine conditional probabilities and expectations using unordered shocks, and then apply the law of total expectation to find the unconditional probabilities and expectations.

**Keywords:** Extreme shock model · State-dependent shock model · Probability of survival · Poisson process



## ABSTRACT - CONTRIBUTED PAPER

### A predictive approach towards automatic discretization using data splitting

Tanja Verster, North-West University

Tanja.Verster@nwu.ac.za

Hennie Venter, North-West University

**Abstract:** The concept of discretization is known by many names: discretization, binning, classing, grouping and quantization. Discretization is the transformation of continuous data into discrete bins or alternatively the mapping of categorical data from a large set to a smaller set of bins. Discretization is an important pre-processing step in most predictive models, and considered a basic data preparation step in building a credit scorecard. Credit scorecards are mathematical models which attempt to provide a quantitative estimate of the probability that a customer will display a defined behavior (e.g. default) with respect to their current credit position with a lender.

Among the practical advantages of discretization are removal of the effects of outliers and a way to handle missing values. Many discretization methods exist, for example one of the simplest method to discretize a continuous variable is to partition it into equal-width intervals.

Data splitting is the method of splitting a sample in two and then developing a hypothesis or estimation method on the basis of one part and testing it on the other part. Often discretization is a time-consuming process.

We propose a new method that automatically discretizes explanatory variables. This method first handles special cases (for example missing values) and then bins the variables using a non-parametric estimate of the log likelihood function. The method uses data splitting to obtain optimal discretization/groupings. Since our aim is prediction, we formulate the discretization problem in terms of prediction and compare the new method with existing methods.

**Keywords:** Discretization; Data Splitting; Binning; Prediction; Credit scorecards;



## ABSTRACT - CONTRIBUTED PAPER



### Investigating Stepwise Common Singular Spectrum analysis and Multi-channel Singular Spectrum Analysis

Lienki Viljoen, Stellenbosch University  
lienki@sun.ac.za

**Abstract:** Two multivariate extensions of Singular Spectrum Analysis (SSA) were compared in terms of the forecast error namely, Horizontal Multi-channel SSA (H-MSSA) and Stepwise Common SSA (Stepwise CSSA). The recurrent forecasting procedure was used. In a simulation study different signal structures, defined in terms of trend, period, amplitude and phase, were investigated. Horizontal MSSA was outperformed particularly in cases where different trends were considered.

**Keywords:** Multivariate time series; Signal estimation; Singular value decomposition



## ABSTRACT - CONTRIBUTED PAPER



### On the use of numerical optimisation procedures in parameter estimation

Jaco Visagie, University of Pretoria  
jaco.visagie@up.ac.za

**Abstract:** A multitude of parameter estimation techniques are available to choose from. These techniques typically entail the optimisation of a specified goal function. For example, maximum likelihood estimation entails maximising the likelihood. Other parameter estimates are obtained by choosing the parameters so as to minimise some distance measure between certain analytical properties of the specified distribution and the corresponding sample quantities. In all but the simplest cases it is not possible to obtain analytical expressions for the parameter estimates and numerical optimisation procedures are required in order to find optima. In this talk, I discuss various parameter estimation techniques, each based on optimising some goal function. I demonstrate various optimisation procedures using numerical examples.

**Keywords:** Parameter estimation; numerical optimisation.





## ABSTRACT - CONTRIBUTED PAPER

### On generalised gamma convolution processes in time changed models

Jaco Visagie, University of Pretoria  
jaco.visagie@up.ac.za

**Abstract:** The generalised gamma convolutions are a subclass of infinitely divisible distributions. In this talk, I discuss some properties of this class of distributions as well as those of the corresponding Lévy processes. I demonstrate the use of these processes as subordinators in time changed Brownian motion models. Special attention is paid to the Pareto and lognormal distributions as well as the corresponding Lévy processes.

**Keywords:** Generalised gamma convolutions; Lévy processes; time changes.



## ABSTRACT - CONTRIBUTED PAPER

### Examining the impact of first-year developmental modules using the potential outcomes approach

Michael Johan von Maltitz, University of the Free State

vmaltitzmj@ufs.ac.za

Lana Hen-Boisen, University of the Free State

Sean van der Merwe, University of the Free State

**Abstract:** This research discusses the causal effect of developmental modules on first-year student success at the University of the Free State. The challenges associated with the data are highlighted, and potential solutions are offered. Using sequential regression multiple imputation, we multiply impute potential outcomes for each (sampled) student under their unobserved 'treatment' grouping. We then compare students' observed outcomes with their unobserved potential outcomes to understand the effects of the developmental modules on their average grades.

**Keywords:** Causal Inference; Multiple Imputation; Bayesian Statistics; Education



## ABSTRACT - CONTRIBUTED PAPER

### Team key performance indicators in rugby union: A validation study

Neil Watson, University of Cape Town

nm.watson@UCT.ac.za

Ian Durbach, University of Cape Town

Theodor Stewart, University of Cape Town

Sharief Hendricks, University of Cape Town

**Abstract:** A validation study is conducted to determine which historically statistically significant rugby union team performance indicators continue to be valid differentiators between winning and losing teams, and predictors of team success. We first examine key performance indicator validity within and between five international competitions across 2013, 2014 and 2015. Whilst a large proportion of the historically significant key performance indicators do continue to differentiate between winning and losing teams in at least one competition, only a handful do so across all five of the competitions considered. The focus then shifts to the validity of key performance indicators as predictors of team success. Particular emphasis is placed on the challenge of developing parsimonious models that attain a high degree of predictive accuracy in the context of having a large number of candidate predictors. The relevance of a few traditional, regression-based variable selection approaches is critically evaluated and contrasted with elastic net regularization. Finally, some important methodological issues are raised and discussed in the broader context of performance analysis in rugby union.

**Keywords:** Rugby union; key performance indicator



## ABSTRACT - CONTRIBUTED PAPER

### Phylogeographic General Linear Models for the Evaluation of Epidemiological Predictors

Tim Wolff-Piggott, University of Cape Town  
twolffpiggett@gmail.com

**Abstract:** Phylogeography allows for the inference of core aspects of the spatial and temporal processes involved in the distribution and mutation of organisms, a key application of which is modelling the spread of viruses. Spatial diffusion is typically discretised by partitioning the locations of the observed data into finite set of geographical sites, between which spatial diffusion is modelled as a Continuous Time Markov Chain.

The established methodology enables the identification of predictors of the geographic spread of pathogens by modelling spatial diffusion rates as log-linear functions of epidemiological predictors- phylogeographic general linear models (GLMs). However, the GLM formulations in the literature model spatial transition rates as deterministic functions of sets of predictors [see e.g. Lemey et al. "Unifying Viral Genetics and Human Transportation Data to Predict the Global Transmission Dynamics of Human Influenza H3N2", PLOS Pathogens 10.2 (2014)]. We argue that the omission of a random component from the current phylogeographic GLM formulation leads to spurious results for epidemiological predictors. Our research augments phylogeographic GLMs with an explicit error parameterisation.

Robust simulation studies demonstrated that the current GLM formulation in the phylogeographic literature produces spurious significant results for randomly generated predictors. Additionally, the proposed stochastic GLM was demonstrated not to result in false positives. A detailed power simulation confirmed that the stochastic GLM does not lose power relative to the formulation in the literature.

**Keywords:** Bio-informatics; Phylogeography; General Linear Model; GLM; Epidemiology



## ABSTRACT - CONTRIBUTED PAPER

### Adjusting the effect of integrating antiretroviral therapy and tuberculosis treatment on mortality for non-compliance: an instrumental variables analysis

Nonhlanhla Yende-Zuma, CAPRISA

nonhlanhla.yende@caprisa.org

Stijn Vansteelandt, Ghent University

Henry Mwambi, University of KwaZulu-Natal

**Abstract:** Background: Since the magnitude of the effect of ART initiation during TB treatment on mortality is less well understood due to non-compliance, we used instrumental variables (IV) analyses. Methods: We studied 642 HIV-TB co-infected patients who were randomly assigned to start ART either early or late during TB treatment or after the TB treatment completion. We analysed data using two-stage predictor substitution (2SPS) and two-stage residuals inclusion (2SRI) methods under both additive and proportional hazards regressions. Results: The 2SPS results showed that on average seven deaths ( $\beta=-0.065$ , 95% CI: (-0.124, -0.013)) were prevented for each year of follow-up in each 100 patients with full exposure to ART during TB treatment (as opposed to 100 patients with no exposure). Corresponding 2SRI results showed slightly weaker effects ( $\beta=-0.047$ , 95% CI: (-0.114, 0.012)). ITT results showed that patients in early and late arms had a reduced hazard of -0.051 (95% CI: -0.093, -0.007) and -0.061 (95% CI: -0.105, -0.015) respectively compared to sequential arm. The 2SPS and 2SRI results from proportional hazards regression showed that full exposure to ART during TB treatment reduced the hazard of death by 64.5% (HR: 0.355, 95% CI: 0.137, 0.852) and 73.7% (HR: 0.263, 95% CI: 0.094, 0.700) respectively. ITT results showed that the hazard of death was reduced by 48.9% (hazard ratio (HR): 0.511, 95% CI: 0.285, 0.917) in early and 53.8% (HR: 0.462, 95% CI: 0.259, 0.825) in the late integrated arm compared to sequential arm.

**Keywords:** Instrumental variable; non-compliance; Cox proportional hazards; additive hazards.



## ABSTRACT - CONTRIBUTED PAPER



### Latent class models approach to the Rater agreement

Ayfer Yilmaz, Hacettepe University  
a.ezgiyilmaz@gmail.com  
Serpil Aktas, Hacettepe University

**Abstract:** Cohen's kappa coefficient measures the chance-corrected nominal scale agreement between two raters for nominal scale. The weighted kappa statistic is used as an agreement index for ordinal data. Latent class models also investigate inter-rater agreement and provide a powerful and flexible tool and allow us to analyze agreement among more than two raters simultaneously. The latent class model is a finite mixture model in which the component distributions are assumed to be multi-way cross-classification tables with all variables mutually independent. This model treats a contingency table as a mixture of unobserved qualitative latent variable. When the latent class model fits to the data, the model parameters can be used to assess the rater agreement. In this study, it is purposed to describe the latent class approach in the agreement context and, to compare this approach to the classical agreement indices on various numerical examples.

**Keywords:** contingency tables; rater agreement; latent class models



## ABSTRACT - CONTRIBUTED PAPER



### Using Text Analytics to Analyse Research Papers

Andre Zitzke, SAS  
andre.zitzke@sas.com  
Louis Neuhoff, Avon

**Abstract:** Corporate Board data from 479 research documents is analysed in terms of key topics and the categories contained in those documents. The exploratory analysis utilised the default settings of the text analysis software to gain initial insight, with expert guidance explaining word clouds and content categories. This is followed by an analysis focusing on specific entities, which yielded interpretable results. It is found that some industry sectors have a higher frequency of board-member overlap occurrence, indicating that they might be influential in interlocking. Certain references to moral issues and positions in corporates are evident in the research, indicating the wide range of topics being covered by the research. The initial research is supplemented by an additional set of targeted research documents being added to the corpus, yielding additional insights to the original results.

**Keywords:** text analytics; interlock

214



## ABSTRACT - CONTRIBUTED PAPER



### Grade Repetition in School – The case of South Africa

Nombuso Zondo, University of KwaZulu-Natal

nombusogm@gmail.com

Khayelihle Mfeka, University of KwaZulu-Natal

Siaka Lougue, University of KwaZulu-Natal

**Abstract:** Grade repetition is amongst the most serious problems in the education system. This study then makes use of the General Household Survey data conducted by Statistics South Africa to analyse the provincial differences in grade repetition over the years. We identify the factors influencing repetition and make use of the multilevel logistic approach to investigate how provincial factors affect grade repetition in school.

**Keywords:** grade repetition; multilevel models



## ABSTRACT - POSTER SESSION

### Estimating the global population size of a threatened species

Francois Becker, South African National Biodiversity Institute, Kirstenbosch Research Centre, Cape Town, Statistics in Ecology, Environment and Conservation, Department of Statistical Sciences, University of Cape Town

phrogbecker@gmail.com

Krystal Tolley, South African National Biodiversity Institute, Kirstenbosch Research Centre, Cape Town, Department of Botany and Zoology, Stellenbosch University

Res Altweegg, Statistics in Ecology, Environment and Conservation, Department of Statistical Sciences, University of Cape Town

John Measey, Centre for Invasion Biology, Department of Botany & Zoology, Stellenbosch University

Jasper Slingsby, South African Environmental Observation Network (Fynbos Node), Centre for Biodiversity Conservation, Cape Town

**Abstract:** Although population size estimates are valuable for the conservation of threatened species, absolute population size is generally difficult to quantify. However, in the case of strongly clustered species, capture-mark-recapture (CMR) methods can provide reasonable approximations of population size. *Capensibufo rosei* is small toad endemic to the Cape Peninsula of the Western Cape. Due to its limited distribution and loss of its native habitat, this species is currently under threat of extinction, occurring in only three known populations, each occupying a very small range (ca. < km 2). Estimating population size is therefore a conservation research priority for this species. Because the three known populations are very localised, it was possible to estimate absolute population sizes using closed population CMR models, and thus the global population size as a summation thereof. We hypothesized that the global population size of *C. rosei* is small. We also used long-term CMR data to estimate population growth over several years, as a function of recruitment and survival, using Pradel models. Population growth was then used to project absolute population size backwards in time. We expected population size to fluctuate considerably over time, consistent with past observations. The closed population models estimated a global population size of 3,958 (LCL = 3,604, UCL = 4,327) adult individuals for the species. Pradel models showed that population size fluctuates considerably over time. Due to the small population size and substantial population fluctuations, this species may be under considerable threat of extinction.

**Keywords:** population size; population models; threatened; amphibian



## ABSTRACT - POSTER SESSION



### Multivariate Extreme Value Theory with an Application to Climate Data

Lipika Bhagwandin, University of Cape Town

lipika.bhagwandin@gmail.com

Sebnem Er, University of Cape Town

**Abstract:** The understanding of the past and current weather conditions can aid in identifying trends and changes that have occurred in weather patterns. This is particularly important as certain weather conditions can have both a positive and negative impact on various aspects in any region. Together with an ever-changing climate it has become markedly noticeable that there is an upward trend in extreme weather conditions. The aim of the study is to use multivariate extreme value theory as a modelling tool on historic climate data in the Western Cape region of South Africa. The interest is to jointly model and analyse the extremes of weather variables. Component-wise maxima, threshold excesses and point process models are used on the weather data, specifically on rain, wind speed and temperature maxima. The performance of the models is compared to each other. It is found that the component-wise models do not perform well in capturing the dependence between these weather extremes. The threshold excesses and point process models lead to improved results for jointly capturing the relationship between the maxima of different weather variables at a single location.

**Keywords:** Multivariate extreme value theory; Generalised extreme value distribution; Weather variables; Dependence; Maxima; Fréchet margins; Component-wise; Threshold excesses; Point Process.



## ABSTRACT - POSTER SESSION



### A probabilistic kernel approach to analyse raptors flight trajectories observed from vantage points

Francisco Cervantes, University of Cape Town  
f.cervantesperalta@gmail.com

**Abstract:** Direct observation of flight trajectories is often the most accessible means for studying the behaviour of raptors and other large birds. These observations are usually collected from vantage points overlooking an area of interest. The simplest way to analyse vantage point data is to treat observations as point counts, assuming a homogeneous distribution of flights over the entire plot. In this way we can estimate abundance or use intensity for the plot. However, occasionally, it is desirable to conduct a spatial analysis at a finer scale, especially when the plot surveyed is large and the landscape heterogeneous, certain behaviours may show meaningful patterns at such scale. Ecological impact studies – evaluation of wind energy facilities, for instance – would benefit from a fine scale analysis, which could provide local information at the landscape level. In order to produce a utilization distribution for an observed plot, we used a non-parametric probabilistic approach, based on kernel density estimators. We present the adaptation of a traditional kernel model, based on point samples, to one built using continuously observed flight trajectories, which are represented as line segments. The recurrent problem of selecting an appropriate bandwidth for the kernels is also addressed.

**Keywords:** kernel; vantage point; flight trajectories; raptor; utilization distribution



## ABSTRACT - POSTER SESSION



### Species Distribution Modelling of *Aloe Dichotoma*

Qobo Dube, University of Cape Town  
qobo.dube@alumni.uct.ac.za

**Abstract:** Species occurrence patterns in the geographic space respond to environmental change by altering to follow optimal ecological areas or by remaining isolated in unchanged areas, potentially leading to extirpation or worse, global extinction. Species distribution models (SDMs) are used to predict or explain why a species occurs where it does in the geographic space based on known distributions of the environmental space. With additional life-stage information, using machine learning based SDMs this paper aims to produce distributional maps of habitat suitability indicating probabilistically where *Aloe dichotoma* (quiver trees) are most likely to occur. The quiver tree is a large, iconic tree species appearing in the arid parts of Southern Africa whose greatest threat is global climate change. Presence/absence, count and compositional models are produced.

**Keywords:** Species Distribution Modelling; *Aloe Dichotoma*; Machine Learning

219



## ABSTRACT - POSTER SESSION

### Experimental Design for the Optimization of Microalgae Growth in Modified Cultivation Systems

Kirstie Eastwood, Nelson Mandela Metropolitan University  
kirstpe@hotmail.com

**Abstract:** InnoVenton, a formally registered Research Institute at the Nelson Mandela Metropolitan University, is currently developing operational processes to increase the production of microalgae as part of their microalgae-to-energy project. The project consists of three components, namely the large-scale cultivation of microalgae, the conversion of microalgae biomass into high-value products and renewable energy from microalgae biomass. InnoVenton is investigating the performance of their modified cultivation system which combines the traditional raceway cultivation system and vertical column photo-bioreactors. Various experimental designs were explored in order to optimize the production of microalgae as well as investigate the robustness of the modified cultivation system. However, practical considerations of the system placed restrictions on multiple variables of interest. This posed certain challenges during the experimental design setup phase.

**Keywords:** InnoVenton; microalgae; raceway; photo-bioreactor; experimental design



## ABSTRACT - POSTER SESSION



### The Effect of Cane Purity and Levels of Sucrose Extraction on Mixed Juice Purity

Anathi Mafuna, CSIR/University of KwaZulu-Natal

anathimafuna@gmail.com

Ishmael Makitla, CSIR

Quentin Williams, CSIR

Delia North, University of KwaZulu-Natal

**Abstract:** This paper employs linear regression statistical method to characterise the effects of cane purity and level of sucrose extraction on the purity of mixed juice. Understanding these effects on the purity of mixed juice is important in order to understand at what level of sucrose extraction does the mixed juice purity start dropping. Using linear regression model on the data that was provided by the Sugar Milling Research Institute (SMRI), it has been shown that the more sucrose is extracted, the more the percentage of sucrose in the resulting mixed juice and which translates to high purity. When the mixed juice purity increases by one unit (i.e. 1 percent), sucrose extraction -on average- increases by 2.24 percent.

**Keywords:** Linear regression; Sugar milling; Mixed juice purity; Sucrose extraction



## ABSTRACT - POSTER SESSION

### Analysing the Impact of Climate Change on Agricultural Productivity in South Africa

Ayanda Mathye, Nelson Mandela Metropolitan University  
s211164569@nmmu.ac.za

**Abstract:** This study analyses the impact of climate change on agricultural productivity in South Africa, focusing on the provinces that are major producers of both wheat and maize. Literature indicates that the production of field crops is sensitive to marginal changes in temperature as compared to changes in precipitation. The study uses correlation calculations to investigate the relationship between climatic variables and the production of both wheat and maize. Climatic data is compiled from different weather stations across the country using Wolfram Mathematica. Reliable statistics will assist in choosing the appropriate climatic variables and analytical tools to use in the study and relation to production of field crops such as maize and wheat.

**Keywords:** climate change

222



## ABSTRACT - POSTER SESSION

### Uncovering the Latent Movement States for Predator State Space Model

Jenicca Poongavanam, Nelson Mandela Metropolitan University  
s212469258@nmmu.ac.za

**Abstract:** Quantifying kill rates remains a key limitation to addressing many predator-prey questions. To date the most common approach to identify lion kill sites is done by identifying clusters of locations obtained using Global Positioning System (GPS) collars on predators. However, if clusters were determined by different decision rules, comparison across studies were not possible. This study investigated Hidden Markov Models (HMMs) as a predictive modelling technique. The Hidden Markov Model (HMM) is a stochastic model in which the system is modelled by a Markov process with "hidden" states. The states of the Markov chain can be interpreted as providing rough classifications of the animal behavioural dynamics. HMMs have been previously used in the analysis of wolves (*Canis lupus*), where aerial surveys were conducted to locate wolf-killed ungulate carcasses. The objectives of this project was to evaluate whether HMMs can predict observer-confirmed kill sites from GPS lion relocation data and also provide additional insight into lion behaviour. GPS radio-collars have been used to obtain movement data of for six lions (*Panthera leo*) in the Kruger National Park, South Africa. HMMs are fitted for each lion and a set of interpretable states were predicted, from which the behaviours were inferred. This state sequence was used to identify potential kill sites, and the actual kill data was then used to confirm these sites. The observer-confirmed kill sites provide a unique opportunity to validate predicted kill sites from HMM modelling using GPS tracking data.

**Keywords:** Predator-prey; Hidden Markov Model; Animal movement; Animal behaviour; Lion; *Panthera Leo*; kill sites; GPS



## ABSTRACT - POSTER SESSION

### An introduction to indices used in model fit for structural equation models (SEM)

Carmen Stindt, Nelson Mandela Metropolitan University

s212256483@nmmu.ac.za

Gary D. Sharp, Nelson Mandela Metropolitan University

Michelle M. Mey, Nelson Mandela Metropolitan University

**Abstract:** An integral part of any statistical analysis is the opportunity to assess how well the data fits the selected model. In structural equation modelling (SEM) there are several indices that have been developed over the years to test model adequacy. The choice of indices used is still largely disputed by researchers. This study aims to identify a composite index for use in SEM. The researcher will also compare and report on the effects of using randomised versus ordered questions within the questionnaire.

**Keywords:** structural equation modelling; SEM; goodness-of-fit; model adequacy

224



## ABSTRACT - POSTER SESSION

### The Influence of the Turning Angle in Hidden Markov Movement Modelling of Terrestrial Animals

Bracken van Niekerk, Nelson Mandela Metropolitan University

brack.z@hotmail.com

Victoria Lucy Goodall, Nelson Mandela Metropolitan University

**Abstract:** Hidden Markov Models are widely used to analyse the patterns of animal movement data. They optimise a sequence of hidden states from a sequence of observations, taking the serial correlation between these observations into account. These latent states are inferred to be the behavioural states of the animal. Either the successive displacement between locations, the turning angle, or both are used as inputs for the models. The outputs consist of the stationary distribution parameter values inferred to be the proportion of time the animal spends in a particular state, the transitional probability matrix of probabilities of transitioning from one state to the next, and distribution parameter values. From the literature there is little mention of the reasoning behind the choice of input. We investigated the statistical and ecological significance of including the turning angle in the model input when applied to terrestrial animal tracking data. Results were obtained for tracking data observed at different time scales for different species, including carnivores and herbivores. The inclusion of the turning angle in the modelling process is thought to overcomplicate the model unnecessarily, and does not influence the model outputs. The ecological interpretations of the models are unchanged and hence the inclusion on the turning angle does not seem to be necessary or provide additional benefit in many cases.

**Keywords:** hidden Markov models; turning angle; terrestrial animals



## ABSTRACT - POSTER SESSION



### Presence and Movement Modelling of Leopards

Lee Watchurst, Nelson Mandela Metropolitan University

watchurstlee@gmail.com

Victoria Goodall, Nelson Mandela Metropolitan University

**Abstract:** Animal movement tracking and analysis has become increasingly popular as technology advances. A hidden Markov model (HMM) is a type of dependent mixture model with two components, observed data and latent states (Zucchini et al. 2009). The latent states are unobservable (interpreted in terms of the animal's behaviour) and generated by a Markov process. The observed data points are seen as the result of a stochastic process over the unobserved states (Tucker et al. 2005). This study makes use of data from 27 Leopards (*Panthera pardus*) in the Phinda and uMkhuze Game Reserves, Kwazulu-Natal. HMMs are fitted using the step length between successive locations obtained from global positioning system (GPS) collars placed on each animal. The number of latent states is determined using statistical and ecological criteria. The aim of the research was to compare the differences in behaviour exhibited by each individual animal (of the same species) in the same area. This research investigated the space use overlap and behavioural patterns of a vulnerable, solitary species.

References: Tucker, B.C. and Anand, M., 2005. On the use of stationary versus hidden Markov models to detect simple versus complex ecological dynamics. Ecological Modelling, 185(2), pp.177-193. Zucchini, W. and MacDonald, I.L., 2009. Hidden Markov models for time series: an introduction using R (Vol. 150). CRC press.

**Keywords:** hidden Markov models; terrestrial animal movement; Leopard



## ABSTRACT - POSTER SESSION



### Modelling health care data using Rayleigh-Pareto distribution

Eno Akarawak, University of Lagos, Akoka, Nigeria  
eakarawak@unilag.edu.ng

Ismaili Adeleke, University of Lagos, Akoka, Nigeria  
Gbenga Olalude, Federal Polytechnic, Ede, Nigeria

**Abstract:** This study introduces the Rayleigh-Pareto Distribution (RPD) using the Transformed-Transformer method and investigates its capability in modelling health care claims data. The health claims data used for this study is secondary data obtained on health utilization claims from 2011 to 2013 from a major HMO in Nigeria. It is found that the new distribution has potentials in modelling claims size, estimating expected claims costs as well as expected cost per enrollee. This appraisal will contribute immensely towards achieving affordable and sustainable national health care delivery system in Nigeria.

**Keywords:** Rayleigh-Pareto Distribution; Transformed-Transformer Method; Health Claims Data; Claims cost; Cost per enrollee



## ABSTRACT - POSTER SESSION

### Principal surfaces

Raeesa Ganey, University of Cape Town  
raeesa.ganey@wits.ac.za  
Sugnet Lubbe, University of Cape Town

**Abstract:** Principal surfaces are smooth two-dimensional surfaces that pass through the middle of a p-dimensional data set. They minimize the distance from the data points, and provide a non-linear summary of the data. The surfaces are non-parametric and their shape is suggested by the data. The formation of a surface is found using an iterative procedure which starts with a linear summary, typically with a principal component plane. Each successive iteration is a local average of the p-dimensional points, where an average is based on a projection of a point onto the surface of the previous iteration. Principal surfaces are a useful data visualization tool particularly in cases where it is impossible to identify certain structures in data. The application of principal surfaces could identify underlying structures in data, especially in the time of big data analysis. This poster will provide details on what a principal surface is, how it is constructed and an application of a principal surface to a real-world example.

**Keywords:** principal surfaces; principal curves; nonparametric principal component analysis

228



## ABSTRACT - POSTER SESSION



### Time Series models for Persistence in Volatility of Share Price

Anela Layini, Nelson Mandela Metropolitan University

s213447533@nmmu.ac.za

Igor Litvine, Nelson Mandela Metropolitan University

**Abstract:** Volatility in the stock return is an integral part of stock market with the alternating bull and bear phases. In the bullish market, the share prices soar high and in the bearish market share prices fall down and these ups and downs determine the return and volatility of the stock market. Volatility of returns in financial markets can be a major stumbling block for attracting investments. Due to a number of its applications in financial market, volatility has deserved plentiful studies for accurate estimation and forecast. Although there has been a huge number of studies that focused on estimating stock price volatility, the emerging capital markets has been paid relatively little attention, compared to developed capital markets. In this study we use the Generalized Autoregressive Conditional Heteroscedasticity (GARCH) model to model volatility. We also investigate whether volatility persists, that is, if the forecast of future volatility will depend upon today's information about returns. The analysis is done using closing daily share prices 2016 of banks in South Africa traded in the Johannesburg Security Exchange(JSE) for the period June 2006 to June 2016.

**Keywords:** GARCH Model;Volatility;JSE

229



## ABSTRACT - POSTER SESSION



### Addressing limitations of financial factor models

Zukisa Mbava, Nelson Mandela Metropolitan University

mbavazp@gmail.com

Gary Sharp, Nelson Mandela Metropolitan University

Stefan Janse van Rensburg, Nelson Mandela Metropolitan University

**Abstract:** This paper discusses the limitations of factor models relating size and value to returns. An alternative model appropriate to analyzing such scenarios is provided in this research.

**Keywords:** Financial Factor Models

230



## ABSTRACT - POSTER SESSION



### Numeric Measurement of BPO

Gezani Miyambu, Sefako Makgatho Health Sciences University

gezani.miyambu@smu.ac.za

Solly Seeletse, Sefako Makgatho Health Sciences University

**Abstract:** The paper describes a simple, straightforward method to measure progress of business process optimisation (BPO). The aim is to derive measures of the degree of BPO attainment in order to identify future priority focus for ensuing exercises. These measures can help to identify components of business that should be improved towards full optimisation of processes in business. In an ideal case of the business containing all the components, a large business scenario is assumed. However, flexibility is permissible when changes are experienced with either some business aspects missing, or new ones added. A measure of BPO progress was eventually developed based on these circumstances. A BPO measurement was described for presentation as a percentage or proportion.

**Keywords:** BPE; BPO; change management; measure; risk management; success factor

231



## ABSTRACT - POSTER SESSION



### A study on service industry macroeconomic variables and their effects on GDP

Lizalise Mngcele, Nelson Mandela Metropolitan University  
mngceles@yahoo.com

**Abstract:** This study explores the relationship between GDP growth and the service industry. The service sector is one of the most important sectors in any economy, especially for a developing country like South Africa. It consists of activities in wholesale and retail trade (including hotels and restaurants), transport, government, finance, education, health care, real estate, etc. It has been recently reported that the services sector contributes more than 60% towards South Africa's GDP. The study seeks to determine the effect of the service industry on GDP growth. Correlation and regression analysis techniques are used to accomplish this goal.

**Keywords:** GDP-gross domestic product

232



## ABSTRACT - POSTER SESSION



### Causality between stock indexes and exchange rates: South Africa

Makgoka Nkoana, Nelson Mandela Metropolitan University, Center of Expertise in Forecasting (CoEF)  
s211185434@nmmu.ac.za

I.N. Litvine, Nelson Mandela Metropolitan University, Center of Expertise in Forecasting (CoEF)

**Abstract:** The Goods Market Theory and the Portfolio Balance Approach, provide evidence that there are linkages between the stock market and the exchange rate. However, previously conducted research has not been conclusive. This study aims to investigate the relationship between the Johannesburg Stock Exchange(JSE) and exchange rates. The JSE is represented by Top 40 index of monthly averaged daily closing stock market indexes. The exchange rates are represented by the United States Dollar, Australian Dollar and Euro to the South African Rand (monthly averaged values). To test long run relationships between these variables the co-interrogation test is used. To determine if there is causality in any direction, the Granger causality test is carried out on the variables.

**Keywords:** Granger Causality; Long-run relationship; JSE Top 40 index; Exchange rate.

233



## ABSTRACT - POSTER SESSION



### Burnout Levels of Academic Staff: An Investigation at Hacettepe University in Turkey

Nilgün Özgül, Hacettepe University

nozgul@hacettepe.edu.tr

Esra Polat, Hacettepe University

**Abstract:** The case of an academic's being burnout is an important problem as besides him/herself, it will also affect the education and the success of the students. Because, being burnout not only causes inadequacies and illnesses in individuals, but also gives harm to the other workers and the foundation as a result of their underperformance, psychological destruction and resignation. There is little research focused on this issue in the education field, especially in Public Universities in Turkey. Therefore, 150 academicians, working in Hacettepe University in 2013-2014 educational year, are selected with "Stratified Simple Sampling" method and the personal information form and Maslach Burnout Inventory (MBI) are implemented. The aim of this study is to identify the burnout levels of academic staff and to evaluate the relationship between some variables of burnout levels. This study is important to recognize the problems of academic staff and to contribute to be productive and successful in their work.

**Keywords:** Burnout Levels; Maslach Burnout Inventory; Multivariate Analysis.



## ABSTRACT - POSTER SESSION



### Statistical methods for the detection of non-technical losses: A case study for the Nelson Mandela Bay Municipality

Sisa Pazi, Nelson Mandela Metropolitan University

s209067822@live.nmmu.ac.za

Chantelle Clohessy, Nelson Mandela Metropolitan University

Gary Sharp, Nelson Mandela Metropolitan University

**Abstract:** Electricity theft is on the rise in the world with electric utilities losing billions in revenue. Practical application of statistical methods can help in identifying and predicting electricity theft. Electricity theft is one of the components of non-technical losses encountered in electrical power systems. This study presents a proposed model for the detection of non-technical losses in Nelson Mandela Bay Municipality. The proposed model was designed using Support Vector Machines (SVM), Naive Bayes (NB) and K-Nearest Neighbour (KNN) classification algorithms. The model's ability to detect fraudulent activities in electricity consumption, which are correlated to non-technical losses, was tested and evaluated using fraud detection rate and error rate. The SVM classifier achieved the highest fraud detection rate of 82.51% and error rate of 14.35%, followed by KNN classifier with 75% and 13.69%, and the least performing classifier was the NB with fraud detection rate of 73.40% and error rate of 18.01%.

**Keywords:** Non-technical losses; Support Vector Machines; Naive Bayes; K-Nearest Neighbour



## ABSTRACT - POSTER SESSION



### Developing statistical indices for economic contribution of cultural and creative industries in South Africa

Melissa Sangqu, Nelson Mandela Metropolitan University

melissasangqu@gmail.com

Igor Litvine, Nelson Mandela Metropolitan University (COEF)

**Abstract:** Cultural and creative industries (CCI) have long been neglected in mainstream trade and industry policy in South Africa. These industries however, are becoming an increasing driving force in the international marketplace. It is important to measure their impact on the economy, but these industries lack statistical data and scientific methodology to analyse the impact. One of the ways to evaluate CCIs is to construct relevant indices. Analysis of the indices employed in literature revealed many difficulties for employing in South African context. Reliable statistics will assist in developing statistical indices for CCIs that will measure the performance of these industries based on population participation, so as to determine how the industries affect the economy. The possible analytical tools such as correlation calculations and comparison of cultural and creative industries with general economic cycles are tested in relation to GDP.

**Keywords:** Cultural and Creative Industries

236



## ABSTRACT - POSTER SESSION

### The malaria elimination agenda: Prospects for Ghana using mathematical tools

Timothy Awine, University of Cape Town  
awinetimothy@gmail.com

**Abstract:** Malaria still remains a serious health challenge in most of sub-Saharan Africa. It is estimated that, 3.2 billion people globally were at risk of malaria infection with an estimated 214 million cases while resulting deaths were 438,000 in 2015. Most of these cases and deaths occurred in sub-Saharan Africa. In recent years, there has been a renewed and concerted global effort to control the transmission of malaria drawing experiences from earlier attempts. In Ghana, the National Malaria Control Program has developed intervention packages in the past decades with set targets in an attempt to reduce the prevalence of the disease. Though some modest gains have been made over the years, there still remains a lot to be achieved. In their latest 2015 to 2020 strategy, the set target is to reduce the burden of malaria by 75% in the population. Another target is to intensify the deployment of already existing interventions across the country. However, mathematical models to support these efforts are in limited use. Data from Ghana will be used to examine the incidence of malaria across climatic seasons and different epidemiological zones. The estimates derived will be used to develop mathematical models that are able to describe the transmission dynamics of malaria in Ghana. These models will be used to investigate the impact of various interventions on malaria incidence and also investigate the prospects of elimination. User friendly interface of these models will also be developed to support planning, monitoring and evaluation activities.

**Keywords:** Malaria; models; impact; interventions



## ABSTRACT - POSTER SESSION

### Non-Parametric Tolerance Intervals for the Assessment of Photovoltaic Energy Yield

Jani Deyzel, Nelson Mandela Metropolitan University

janideyzel@gmail.com

Chantelle Clohessy, Nelson Mandela Metropolitan University

Warren Brettenny, Nelson Mandela Metropolitan University

**Abstract:** South Africa is a country with numerous renewable energy resources at its disposal. The use of solar energy (photovoltaic) systems in South Africa is growing rapidly. The Nelson Mandela Metropolitan University Centre of Energy Research has proposed the installation of a 1MW photovoltaic system on the Summerstrand South Campus in Port Elizabeth. This study assesses the potential energy output (MWh) of the proposed system. Non-parametric tolerance intervals using a variance component model are implemented to assess the expected energy yields of the photovoltaic system.

**Keywords:** Photovoltaic system; Tolerance interval; Variance component model



## ABSTRACT - POSTER SESSION

### Analysis of Clustered Competing Risks with Application to a Multicentre Clinical Trial

Mary Ajibola Familusi, University of Cape Town  
fmlmar001@myuct.ac.za

**Abstract:** In time-to-event data analysis, if subjects fail from multiple mutually-exclusive causes, data are said to have competing risks. For competing risks data, the Fine and Gray proportional hazards model for sub-distributions has gained popularity due to its convenience in directly assessing the effect of covariates on the cumulative incidence function. Furthermore, sometimes competing risks data cannot be considered as independent because of a clustered design, for instance, in registry cohorts or multi-centre clinical trials. The Fine and Gray model has been extended to the analysis of clustered time-to-event data, by including random-centre effects or frailties in the sub-distribution hazard. This study focuses on the analysis of clustered competing risks with an application to the investigation of the management of pericarditis clinical trial (IMPI) dataset. IMPI is a multicentre clinical trial carried out with the principal objective of assessing the effectiveness and safety of adjunctive prednisolone and Mycobacterium indicus pranii immunotherapy, in reducing the composite outcome of death, constriction or cardiac tamponade, requiring pericardial drainage in patients with probable or definite tuberculosis pericarditis. The objectives in this study is therefore to determine the risk factors associated with these outcomes and to examine the effect of the prednisolone and M. indicus pranii while adjusting for these risk factors and considering centres as a random effect. Using Cox proportional hazards model, it was found that age, weight, New York Heart Association (NYHA) class, hypotension, creatinine, and peripheral oedema show a statistically significant association with the composite outcome.

**Keywords:** Clustered Competing Risks



## ABSTRACT - POSTER SESSION

### Genetic polymorphisms and organophosphate neurotoxicity amongst emerging farmers in the Western Cape

Tracy Glass, University of Cape Town

tglass060@gmail.com

MA Dalvie, University of Cape Town

R Ramesar, University of Cape Town

Z Holtman, University of Cape Town

**Abstract:** Organophosphates (OPs) are the most widely used class of pesticide and a known neurotoxin. Long-term toxic effects include generalised weakness, tremors and a decrease in cognitive functions, such as impairment in memory and concentration. Evidence, largely animal studies, suggest that these neurotoxic effects of OPs may be modulated by genetic polymorphisms of xenobiotic metabolising enzymes (XMEs). A cross-sectional study of 301 emerging farmers was conducted in the Western Cape of South Africa. Neurotoxicity was assessed using tests drawn from the World Health Organisation Core Test Battery and vibration sensitivity testing. Long-term OP information was collected using a questionnaire and blood samples were analysed for genetic polymorphisms of the following XMEs, glutathione S-transferases (GST), N-acetyltransferases (NAT) and Paraoxonase (PON1). Multiple logistic regression models were fitted to assess whether these enzymes modified the relationship between OP exposure and neurotoxicity and to determine whether vulnerable genetic groups could be identified. Median age was 39 (30-48) and most had 9 years of education or less (65.5%). There was a low prevalence of the GST null genotype (GSTT-1% and GSTM-16%) and the GA and GG genotype for NAT (10%). No main effect was seen for OP exposure and the neurobehavioural outcomes. The only significant association was observed between past pesticide poisoning and impaired vibration sensitivity. Results showed modulation of OP exposure and neurotoxic outcome relationships by enzymes NAT and PON1. The strongest evidence of modification was by NAT, on the association between pesticide poisoning and impaired vibration sense ( $OR=4.7$ , 95% CI=1.28-16.94).

**Keywords:** genetic polymorphisms organophosphate



## ABSTRACT - POSTER SESSION

### Uses of Principal Components in Generalized Linear Modeling

Nathaniel Joselson, University of Cape Town

jslhat001@myuct.ac.za

Sugnet Lubbe, University of Cape Town

**Abstract:** Multicollinearity between explanatory variables in linear regression and generalized linear modeling makes extracting accurate estimates of regression coefficients very difficult. Principal component analysis is a multivariate technique to create linear combinations of the original variables which are uncorrelated. If these correlated variables are replaced by a full or partial subset of principal component scores in the models, this can solve the multicollinearity problem. If these principal components have logical interpretations then inference can be performed with the coefficients on the principal component scores. If, however, the coefficients have no meaningful interpretation, accurate estimations of coefficients for the original variables can be computed through back transformation using the singular value decomposition. This technique is often applied to linear regression models. Here it is extended to generalised linear models where the response variable is distributed according to a distribution in the exponential family.

**Keywords:** Multivariate; Principal Component Analysis; Singular Value Decomposition; Generalized Linear Models; Linear Regression; Confidence Intervals



## ABSTRACT - POSTER SESSION

### Joint models for HIV viral load profiles and time to loss of viral control

Maia Lesosky, University of Cape Town

maia.lesosky@uct.ac.za

Elton Mukonda, University of Cape Town

**Abstract:** Understanding factors associated with loss of viral suppression in individuals on anti-retroviral treatment (ART) is an important component to meeting treatment targets with respect to long term viral suppression. Joint models for time-to-event outcomes are useful to account for outcome-related, endogenous, longitudinal biomarkers. HIV viral load (VL) measured as a longitudinal biomarker presents numerous difficulties in modelling in light of the high proportion of measures below the limit of detection, the strongly skewed nature of the data and the magnitude dependent measurement error. Latent class joint models are applied to an extensive set of over 3000 longitudinal VL measures taken during pregnancy and post-partum in women initiating ART. The joint models, with latent class mixture models for longitudinal outcomes and semi-parametric models for time-to-event outcomes, estimate the time to loss of viral suppression, which occurs in nearly 25% of individuals. Distinct latent profiles of viral loads emerged from the latent class analysis, emphasising heterogeneity in viral load profile. The joint model suggests duration on ART and its interaction with pre-ART viral load are among the significant factors associated with the longitudinal biomarker whereas the time to loss of viral suppression is partially explained by the gestational age at initiation of ART and the pre-ART viral load.

**Keywords:** joint models; latent class mixture models



## ABSTRACT - POSTER SESSION



### Survival analysis in higher education

Iketle Maharela, University of Pretoria  
iketle.maharela@up.ac.za  
Lizelle Fletcher, University of Pretoria

**Abstract:** The South Africa government invests R25 billion on funding higher education, it is therefore vital for universities to model the survival time of enrolled students. Student survival time at universities is a major concern due to various reasons. Firstly, funding: the longer the duration of studies the more funding required. Secondly, university facilities for instance computer labs, lecture halls and libraries are limited to a certain number of students, meaning that students not graduating within the prescribed time will result in university facilities being spread thin amongst students. Lastly, degree completion within the prescribed time can serve as a performance indicator for universities building a credible reputation in the academic world. The above mentioned concerns are relevant to South African universities which are facing increasing number of students due to an escalation in demand for higher education. This paper reviews survival analysis methods applied to educational research, particularly with respect to the duration of studies at tertiary level. Survival analysis models the time to event data where the outcome variable is the time until occurrence of the event of interest. The findings of previous research in survival analysis in education have shown that gender and the field of study play major roles in predicting graduation time. However, there are several other factors that can be considered in explaining and predicting student graduation which are but not limited to: demographics, social class, and student's living situation.

**Keywords:** Duration of study; Higher education; Survival analysis; Survival time



## ABSTRACT - POSTER SESSION

### Bayesian Cox Proportional Hazards model and insights from the censored quantile regression model for paediatric and adolescent HIV/AIDS patients on antiretroviral treatment

Innocent Maposa, Namibia University of Science and Technology

imaposa@nust.na

Rénette J. Blignaut, University of the Western Cape

**Abstract:** Survival analysis techniques are often used in biostatistics, epidemiological and clinical research to model time until event data. The purpose of this study is to fit a Bayesian proportional hazards model and a censored quantile regression (CQR) model to paediatric data and then compare the results in relation to inferences on the effect of the different prognostic risk factors on anti-retroviral treatment patients' survival times. A retrospective cohort study design was conducted for children who initiated anti-retroviral treatment between 01 January 2006 and 31 December 2010. The sampling frame constituted 1605 paediatric patients and from these a final sample of 813 children was selected. The results from a Bayesian proportional hazards model indicate that not being an infant had a positive effect on survival time. Patients initiating treatment in clinical stage II instead of stage IV had a significant positive effect on survival time. The results from the censored quantile regression model are more revealing, highlighting that initiating in clinical stage II had a significant positive effect on survival time during the early periods of initiation compared to initiating in clinical stage IV. The effect reduces towards the 80th quantile, that is, towards the end of the follow-up period. The results also reveals that patient gender has a significant effect during the early periods of starting ART but not significant at any other point during treatment. The conclusion from this study is that more insights are obtained from using censored quantile regression models as compared to the proportional hazards models.

**Keywords:** Bayesian proportional hazards; censored quantile regression; survival data; risk factors



## ABSTRACT - POSTER SESSION

### Implications for Incorrect Specifications of the Components of a Zero-inflated Poisson Model

Onyekachi Esther Nwoko, University of Cape Town  
nwkony001@myuct.ac.za  
Iddi Samuel, University of Ghana.

**Abstract:** Count outcomes are often modeled with the Poisson regression. However, this model imposes a strict mean-variance relationship that is unappealing in many contexts. Several studies in the life sciences result in count outcomes with excessive amount of zeros that introduces extra dispersion in the data which cannot be accounted for by the traditional Poisson regression. The zero-inflated Poisson (ZIP) and zero-inflated negative binomial (ZINB) models are popular alternatives. The ZIP model is composed of two components, a logistic and Poisson component. Both components allow the inclusion of covariates. Civettini and Hines (2005) investigated misspecification effects in the zero-inflated negative binomial regression models. However, the interest of this study lies in investigating the effects of misspecification of components of the ZIP regression model in relation to bias, standard error (precision) of estimates and Mean square error (MSE) while varying the sample size. Two different incorrect specification of the components of a ZIP model were considered: the Omitted and Misspecified models. Results of a Monte Carlo simulation are reported. It was observed that omissions in both parts of the models lead to biases in the estimated parameters. The intercept parameters were the most severely affected. Furthermore, in all the types of omissions, parameters in the zero-inflated part of the models were much affected compared to the Poisson part in terms of both bias and MSE. Generally, bias, standard errors and MSE decrease as sample sizes increases for all parameters.

**Keywords:** Poisson model; Maximum likelihood estimation; Misspecification; Logistic model; Omission; Simulations; Zero-Inflation.



## POSTER SESSION



### Identifying Genetic Signatures using Phylogenetic Regression with a LASSO Penalty

Emma Plumstead, University of Cape Town  
emmaplums@gmail.com

**Abstract:** Virologists are often interested in identifying sites in genetic sequences that are associated with a particular phenotype. Identifying such genetic signatures is problematic for two reasons: (1) Genetic sequences are related through descent from a common ancestor (i.e. the observations are not statistically independent), and (2) There are often very many sites (variables) available, of which only a small subset are expected to be associated with the phenotype. Each of these problems have been tackled separately, by incorporating the phylogenetic correlation structure into a regression model, and through regularised regression methods that can perform variable selection, such as the LASSO. In this project, these two approaches will be combined into a single framework.

**Keywords:** Phylogenetic regression; Lasso



246





## LIST OF DELEGATES

### A

Mr Rasheed Adeyemi  
Dr Eno Akarawak  
Dr Serpil Aktas  
Mr Simon Alger  
Prof James Allison  
Dr Res Altwegg  
Mr Omololu Aluko  
Mr Albert Kwabena Amuzu  
Prof Anestis Antoniadis  
  
Mr Timothy Awine  
Mr Adeboye Azeez

University of KwaZulu Natal  
University of Lagos  
Hacettepe University  
TNS Global Brand Equity Centre  
North West University  
University of Cape Town  
University of KwaZulu-Natal  
Shardfo-Tech Company Limited  
University Joseph Fourier, Grenoble  
& University of Cape Town  
University of Cape Town  
University of Fort Hare

adeyemira@yahoo.ca  
eakarawak@gmail.com  
serpilaltunay@gmail.com  
simon.alger@tnsglobal.com  
James.Allison@nwu.ac.za  
res.altwegg@gmail.com  
llaluko@yahoo.com  
shardfotechcomltdgh@gmail.com

Anestis.Antoniadis@imag.fr  
awinetime@ymail.com  
azizadeboye@gmail.com

### B

Prof Bruce Bassett  
Ms Jesca Mercy Batidzirai  
Mr Francois Becker  
Prof Frank Beichelt  
Prof Andriette Bekker  
Mrs Dorette Bekker  
Dr Alphonse Bere  
Dr Tom Berning  
Ms Lipika Bhagwandin  
Dr Surette Bierman  
Prof Francis Biesmans  
Mr Vusi Bilankulu  
Prof Renette Blignaut  
Mr Alexander Boateng  
Prof David Borchers  
Mr Henri Boshoff  
Mrs Sonja Bosman  
Mr Gordon Botha  
Ms Danielle Winona Boyd  
Mr Warren Brettenny  
Mr Stefan Britz  
Mrs Erin Bromley-Gans  
Mr Humphrey Brydon  
Mr Sven Buitendag  
Dr Divan Aristo Burger

University of Cape Town  
University of KwaZulu Natal  
University of Fort Hare  
University of the Witwatersrand  
University of Pretoria  
  
University of Venda  
Stellenbosch University  
University of Cape Town  
Stellenbosch University  
University of Lorraine  
University of Pretoria  
University of the Western Cape  
University of Limpopo  
South African Statistical Association  
ABSA  
Seed Analytics  
University of Cape Town  
University of Cape Town  
Nelson Mandela Metropolitan University  
University of Cape Town  
DSV South Africa  
University of the Western Cape  
Stellenbosch University  
University of Pretoria

Bruce.Bassett@uct.ac.za  
batidzirai@ukzn.ac.za  
phrogbecker@gmail.com  
frank.beichelt@wits.ac.za  
andriette.bekker@up.ac.za  
dbekker@mhg.co.za  
alphonce.bere@univen.ac.za  
tberning@sun.ac.za  
lipika.bhagwandin@gmail.com  
surette@sun.ac.za  
francis.biesmans@univ-lorraine.fr  
vrbilankulu@gmail.com  
rblignaut@uwc.ac.za  
alexander.boateng@ul.ac.za  
dlb@st-andrews.ac.uk  
henri.boshoff@absa.co.za  
sonja@seedanalytics.info  
gbot300@gmail.com  
daniellewboyd@gmail.com  
warren.brettenny@nmmu.ac.za  
steviebritz@gmail.com  
EGans@go2uti.com  
hbrydon@uwc.ac.za  
svenbuitendag@gmail.com  
divanaburger@gmail.com

### C

Mr Jon Calder  
Mr Frikkie Calitz  
Mr Fillimon Cebekulu Quintiles  
Mr Francisco Cervantes  
Mr Martin Chanza  
Mr Tinashe Chatora  
Mr Retius Chifurira  
Dr Delson Chikobvu  
Mr Knowledge Chinhamu

Agricultural Research Council - Biometry  
  
University of Cape Town  
North-West University  
University of Cape Town  
University of KwaZulu Natal  
University of the Free State  
University of KwaZulu Natal

jon.calder@derivco.co.za  
CalitzF@arc.agric.za  
fillimon.cebekulu@quintiles.com  
f.cervantesperalta@gmail.com  
martin.chanza@nwu.ac.za  
tdchatora@gmail.com  
chifurira@ukzn.ac.za  
chikobvu@ufs.ac.za  
Chinhamu@ukzn.ac.za



## LIST OF DELEGATES

### C

Ms Mutsa Chinyamakobvu	Rhodes University	m.chinyamakobvu@ru.ac.za
Mr Raymond Chiruka	University of Fort Hare	rchiruka@ufh.ac.za
Dr Siyabonga Chule	South African Institute for Mathematical Sciences	chule@aims.ac.za
Prof Hülya Çingi	Hacettepe University	hcingi@hacettepe.edu.tr
Mr Paul Claassen		paulw.claassen@gmail.com
Mr Allan Clark	University of Cape Town	allan.clark@uct.ac.za
Mr Graeme Clements	DataProphet	graeme@dataprophet.com
Mr AJ Cloete	DataProphet	andre@dataprophet.com
Ms Chantelle Clohessy	Nelson Mandela Metropolitan University	chantelle.clohessy2@mamu.ac.za
Ms Frances Coetzer	Stellenbosch University	coetzerfrances@gmail.com
Mr James Combrink	University of Cape Town	jamziec@gmail.com
Prof Willie Conradie	Stellenbosch University	wjc@sun.ac.za
Mrs Ivona Elizabeth		
Contardo-Berning	Stellenbosch University	ivona@sun.ac.za
Mr Alex Conway	University of Cape Town	alexander.m.conway@gmail.com
Dr Gretel Crafford	University of Pretoria	gretel.crafford@up.ac.za
Mr Frans Cronje		fransmcronje@gmail.com
Mr Jacq Crous	University of Pretoria	jacq.crous@gmail.com

### D

Ms Lara Dalmeyer	University of Cape Town	dlmlar001@gmail.com
Prof Sonali Das	CSIR	sdas@csir.co.za
Mrs Louise De Koker	Stellenbosch University	louisedekoker@woolworths.co.za
Dr Margaret de Villiers	Stellenbosch University	margaret.devilliers@sas.com
Mr Murray de Villiers	SAS Institute	murray.devilliers@sas.com
Ms Liesl De Vries	Quintiles	liesl.devries@quintiles.com
Prof Tertius de Wet	Stellenbosch University	tdewet@sun.ac.za
Prof Pravesh Debba	CSIR	pdebba@csir.co.za
Prof Legesse Kassa Debusho	Department of Statistics	debuslk@unisa.ac.za
Mrs Marion Delport	University of Pretoria	marion@bfap.co.za
Ms Jani Deyzel	Nelson Mandela Metropolitan University	janidehyzel@gmail.com
Mr Tadele Akeba Diriba	University of Pretoria	tadthstat@yahoo.com
Mr Greg Distiller	University of Cape Town	Greg.Distiller@uct.ac.za
Mr Ross Dix-Peek	Nelson Mandela Metropolitan University	s212286552@mamu.ac.za
Ms Gciniwe Dlamini	University of Cape Town	gci_sims@yahoo.com
Ms Nothile Dlamini	Gauteng Provincial Legislature	ndlalmini@gpl.gov.za
Mrs Susan Dowse	DigiOutsource	susan.dowse@digioutsource.com
Mr Qobo Dube	University of Cape Town	qobo.dube@alumni.uct.ac.za
Mrs Nontembeko Dudenii-Tlhone	CSIR	nddueni@gmail.com
Dr Ian Durbach	University of Cape Town	ian.durbach@uct.ac.za
Mr Simba Dziwa	University of Johannesburg	Simba@dziwainc.com

### E

Mr Werner Earle	Quintiles	werner.earle@quintiles.com
Ms Kirstie Eastwood	Nelson Mandela Metropolitan University	kirstpe@hotmail.com
Prof Idris Eckley	South African Statistical Association	i.eckley@lancaster.ac.uk
Dr Rene Ehlers	University of Pretoria	rene.ehlers@up.ac.za
Dr Sebnem Er	University of Cape Town	er.sebnem@gmail.com
Dr Birgit Erni	University of Cape Town	birgit.erni@uct.ac.za



## LIST OF DELEGATES

### F

Dr Inger Fabris-Rotelli  
Dr Adeniyi Francis Fagbamigbe  
Ms Mary Ajibola Familusi  
Prof Paul Fatti  
Mr Johan Ferreira  
Mr Kyle Findlay  
Ms Lauren Frederick

University of Pretoria  
University of Ibadan  
University of Cape Town  
University of the Witwatersrand  
University of Pretoria  
TNS Global Brand Equity Centre  
DSV Africa

inger.fabris-rotelli@up.ac.za  
franstel74@yahoo.com  
fmlmar001@myuct.ac.za  
paulfatti@gmail.com  
johan.ferreira@up.ac.za  
kyle.findlay@tnsglobal.com  
lfrederick@go2uti.com

### G

Ms Frieda Geldenhuys  
Dr Dibaba Gemechu  
Prof Irene Gijbels  
Ms Tracy Glass  
Dr Victoria Goodall  
Mr Aviwe Gqwaka  
Dr Marien Graham  
Prof Michael Greenacre  
Dr Teresa Gridley  
Dr Freedom Gumede  
Ms Nomonde Gwebushe

Stellenbosch University  
UNISA  
South African Statistical Association  
University of Cape Town  
Nelson Mandela Metropolitan University  
Nelson Mandela Metropolitan University  
University of Pretoria  
South African Statistical Association  
University of Cape Town  
University of Cape Town  
Medical Research Council

Frieda.Geldenhuys@gmail.com  
dibooobayu@gmail.com  
irene.gijbels@wis.kuleuven.be  
tracy.glass@uct.ac.za  
victoriagoodall@gmail.com  
aviwe.gqwaka@nmmu.ac.za  
marien.graham@up.ac.za  
michael.greenacre@gmail.com  
nam.dolphin.project@gmail.com  
freedom.gumede@uct.ac.za  
Nomonde.Gwebushe@mrc.ac.za

### H

Prof Linda Haines  
Ms Anneli Hardy  
Dr Justin Harvey  
Prof Trevor Hastie  
Mr Jurgens Hendriks  
Ms Gugu Hlatshwayo  
Mr Michel Hoareau  
Mrs Jenny Holloway  
Mr Euan Hope  
Mr Chun-Kai (Karl) Huang  
Mr Chun-Sung Huang  
Dr Sanet Hugo  
Prof Marie Huskova

University of Cape Town  
University of Cape Town  
Stellenbosch University  
South African Statistical Association  
University of Pretoria  
University of KwaZulu Natal  
  
CSIR  
Barclays Africa  
University of Cape Town  
University of Cape Town  
University of Fort Hare  
Charels University, Prague

linda.haines@uct.ac.za  
anneli.hardy@uct.ac.za  
jharvey@sun.ac.za  
hastie@stanford.edu  
hendriks.jj2@gmail.com  
mhaise3@yahoo.com  
mchl842@gmail.com  
jhollowa@csir.co.za  
Euan.hope@absa.co.za  
chun-kai.huang@uct.ac.za  
chun-sung.huang@uct.ac.za  
sanethug@gmail.com  
huskova@karlin.mff.cuni.cz

### I

Mr Peter liyambo  
Mrs Yetunde Ilupeju

University of Namibia  
University of KwaZulu-Natal

ptiyambo@gmail.com  
ilupejuyetunde@yahoo.com

### J

Mrs Rechelle Jacobs  
Prof Mardi Jankowitz  
Mr Charl Janse van Rensburg  
Mr StefanJanse van Rensburg  
Mr Nathaniel Jolelson  
Prof Gina Joubert  
Mr Ido Jubane  
Prof June Juritz

University of Western Cape  
UNISA  
South African Medical Research Council  
Nelson Mandela Metropolitan University  
University of Cape Town  
University of the Free State  
SASA  
University of Cape Town

rejacobs@uwc.ac.za  
jankomd@unisa.ac.za  
charl.jansevanrensburg@mrc.ac.za  
sjansevanrensburg@gmail.com  
jslnat001@myuct.ac.za  
gnbsgi@ufs.ac.za  
ijubane@ufh.ac.za  
christien.thiart@gmail.com



## LIST OF DELEGATES

### K

Dr Frans Kanfer  
Mr Tsirizani Kaombe  
Prof Konrad Kapp  
Mr Yigit Karabag  
Dr I Karangwa  
Dr Natasha Karenyi  
Dr Reshma Kassanjee  
Mr Dominique Katshunga  
Mr Kago Kebotsamang  
Prof Yehene Getachew Kifle  
Ms Lisa Kirkland  
Dr Judy Kleyn  
Mr Gerhard Koekemoer  
Dr Etienne Koen  
Ms Christine Kraamwinkel  
Mr Prince Asamoah Kusi

University of Pretoria  
University of Malawi; Chancellor College  
Nelson Mandela Metropolitan University  
SAS Institute  
University of the Western Cape  
University of Cape Town  
University of Cape Town  
University of Cape Town  
University of Cape Town  
University of Limpopo  
University of Pretoria  
University of Pretoria  
SASOL  
Barclays Africa  
University of Pretoria  
Shardfo-Tech Company Limited

frans.kanfer@up.ac.za  
tsirizaniakoembe@yahoo.com  
s210076453@nmmu.ac.za  
Yigit.Karabag@sas.com  
fokafoka15@gmail.com  
natashakarenyi@gmail.com  
reshma.kassanjee@uct.ac.za  
dominique.katshunga@uct.ac.za  
kago.kebotsamang@gmail.com  
yehene.kifle@ul.ac.za  
lisakirkland25@gmail.com  
judy.kleyn@up.ac.za  
Gerhard.Koekemoer3@sasol.com  
etienne.koen@absa.co.za  
christine.kraamwinkel@gmail.com  
shardfotechcomltd@gmail.com

### L

Dr Miguel Lacerda  
Ms Bernisha Lakhoo Lala  
Mr Shravan Lalbahadur  
Prof Sophie Lamber-Lacroix  
Dr Morne Lamont  
Mr Brendon Lapham  
Mr Abduraghiem Latief  
Mr Alex Lau  
Mr Anela Layini  
Prof Niel Le Roux  
Ms Maria Mokgadi Lekganyane  
Mr Barend Leonard  
Dr Maia Lesosky  
Ms Ilke Lewis  
Mrs Lara Lewis  
Mr Shawn Liebenberg  
Ms Artemisa Lima  
Prof Brian Lindner  
Prof Francesca Little  
Prof Igor Litvine  
Mr Leon Lombard  
Dr Carl Lombard  
Prof Sugnet Lubbe  
Ms Zani Ludick  
Mrs Retha Luus

University of Cape Town  
University of Cape Town  
University of Cape Town  
South African Statistical Association  
Stellenbosch University  
RGA  
University of the Western Cape  
University of Pretoria  
Nelson Mandela Metropolitan University  
Stellenbosch University  
University of the Witwatersrand  
DataProphet  
University of Cape Town  
University of Pretoria  
CAPRISA  
Northwest University  
University of KwaZulu Natal  
Stellenbosch University  
University of Cape Town  
Nelson Mandela Metropolitan University  
Quintiles  
South African Medical Research Council  
University of Cape Town  
University of the Free State  
University of the Western Cape

Miguel.Lacerda@uct.ac.za  
bernishal@gmail.com  
shravan.lalbahadur@santam.co.za  
Lambert@imag.fr  
mmcl@sun.ac.za  
blapham@rgare.com  
alatief@uwc.ac.za  
alexlau1712@gmail.com  
anelalayini076@gmail.com  
njlr@sun.ac.za  
mmlekanyane@gmail.com  
bennie@dataprophet.com  
lesosky@gmail.com  
ilkelewis1@gmail.com  
lara.lewis@caprisa.org  
shawn.liebenberg@nwu.ac.za  
aaslima10@gmail.com  
blindner@sun.ac.za  
francesca.little@uct.ac.za  
igor.litvine@nmmu.ac.za  
Leon.Lombard@quintiles.com  
carl.lombard@mrc.ac.za  
sugnet.lubbe@uct.ac.za  
ludickz@ufs.ac.za  
rluus@uwc.ac.za

### M

Prof Iain MacDonald  
Mr Masakheke Mafu  
Ms Anathi Mafuna  
Ms Iketle Maharela

University of Cape Town  
Rhodes University  
CSIR  
University of Pretoria

iain.macdonald@uct.ac.za  
g11m7251@campus.ru.ac.za  
anathimafuna@gmail.com  
iketle.maharela@up.ac.za



## LIST OF DELEGATES

### M

Mr Jacob Majakwara	University of the Witwatersrand	Jacob.Majakwara@wits.ac.za
Mr Event Makamo		sanelemakamo26@gmail.com
Ms Seite Makgai		seite.makgai@up.ac.za
Mrs Sibusewise Makhanya		smakhanya@csir.co.za
Mr Paul Malan		paul.j.malan@quintiles.com
Mr Ben Magabashelle Malope		magabashellemalope@gmail.com
Mr Kgwidzi Michael Mampana		mampam1@unisa.ac.za
Prof Samuel Manda		samuel.manda@mrc.ac.za
Ms Raeesa Manjoo		manjoo.raeesa@gmail.com
Mr Zuzumuzi Maponya		muzimaponya@woolworths.co.za
Mr Innocent Maposa	Namibia University of Science and Technology	imaposa@nust.na
Dr Daniel Maposa	University of Limpopo	danmaposa@gmail.com
Mr Jan Marais		maraisjandre9@gmail.com
Mr Chioneso Marange	University of Fort Hare	cmarange@ufh.ac.za
Mr Thanyani Maremba	Statistics South Africa	thanyanimar@statssa.gov.za
Mr Gaonyalelwé Maribe	University of Free State	gao.isc@gmail.com
Mr Hossein Masoumi Karakani	University of Pretoria	hosseinstatistics@gmail.com
Mr Thami Eric Mathebula	SASA	mathebulae@arc.agric.za
Ms Ayanda Mathye	Nelson Mandela Metropolitan University	s211164569@mnu.ac.za
Mrs Lyness Matzirofa	University of Johannesburg	lmatizi@yahoo.com
Ms Tshepo Matsose	Sefako Makgatho Health Sciences University	tshepo.matsose@smu.ac.za
Mr Melusi Mavuso	University of Cape Town	melusi.mavuso@uct.ac.za
Mr Cremel Maxime	Université de Lorraine	maxime.cremel@univ-lorraine.fr
Ms Jocelyn Mazarura	University of Pretoria	jocelyn.mazarura@up.ac.za
Mr Zukisa Mbava	Nelson Mandela Metropolitan University	mbavazp@gmail.com
Ms Aude Ines Mbonda Tiekwé	Stellenbosch University	ines@aims.edu.gh
Mr Albertus Meiring		albertus.meiring@gmail.com
Dr Vincent Micali	Stats4Buz (Pty) Ltd	vince.micali@stats4buz.co.za
Mr Albert Mijburgh	University of Pretoria	albert@mijburgh.co.za
Mr Sollie Millard	University of Pretoria	sollie.millard@up.ac.za
Mr Richard Minkah	University of Ghana	rminkah@ug.edu.gh
Mr Gezani Richman Miyambu	Sefako Makgatho Health Sciences University	gezani.miyambu@smu.ac.za
Mr Lizalise Mngcele	Nelson Mandela Metropolitan University	mngceles@yahoo.com
Ms Masego Modibane	University of Pretoria	masegomodibane@gmail.com
Mr Moletenyane Mokhele	University of the Free State	moletenyane@gmail.com
Mr Paul Mokilane	CSIR	PMokilane@csir.co.za
Prof Reshoketswe Mokobane	University of Limpopo	nkadirm@gmail.com
Mr Khehla Daniel Moloi	SASA	khehla.moloi@ul.ac.za
Ms Kirshnee Moodley	Barclays Africa	kirshnee.moodley@absa.co.za
Dr Henri Moolman	Walter Sisulu University	moolman.henri@gmail.com
Mr Andre Mostert	SASOL	andre.mostert@sasol.com
Prof Paul Mostert	Stellenbosch University	pjmos@sun.ac.za
Mr Levicatus Mugenyi	Hasselt University	levicatus.mugenyi@uhasselt.be
Mr Elton Mukonda	University of Cape Town	elton.mukonda@uct.ac.za
Dr Chris Muller	Stellenbosch University	cmuller@sun.ac.za
Dr Martie Muller	Radboud University	m.muller@science.ru.nl
Mr Raymond Musyoka	Centre for Disease Control and Prevention	uys6@cdc.gov
Mr Ruffin Mpiana Mutambayi	University of Fort Hare	mmutambayi@ufh.ac.za
Prof Henry Mwambi	University of KwaZulu Natal	mwambih@ukzn.ac.za
Mr Tsepang Mzamane		tsepangmz@yahoo.com
Ms Anathi Mafuna		anathimafuna@gmail.com
Ms Iketle Maharela	CSIR	ikettle.maharela@up.ac.za
	University of Pretoria	



## LIST OF DELEGATES

### N

Ms Priyanka Nagar	University of Pretoria	priyanka.nagar194@gmail.com
Dr Patrick Naidoo	Statistics South Africa	patrickn@statssa.gov.za
Ms Kimera Naradhb	University of KwaZulu Natal	kimmi.naradh@gmail.com
Mrs Justine Nasejje	University of KwaZulu Natal	justinenasejje@gmail.com
Mr Dries Naude	University of the Free State	naudeam@ufs.ac.za
Dr James Ndege	University of Fort Hare	jndege@ufh.ac.za
Mr Thamsanqa Innocent Ndlela	University of KwaZulu Natal	ndlelathamsanqa90@gmail.com
Mr Lethani Mbongeni Ndwandwe	University of KwaZulu Natal	returnndwandwe@gmail.com
Ms Ané Neethling	University of Pretoria	annaneut@gmail.com
Dr Ariane Neethling	University of the Free State	ariane_neethling@yahoo.com
Ms Riette Nel	University of the Free State	gnbsmn@ufs.ac.za
Mrs Elouise Nell	DigiOutsource	Elouisenell@live.de
Mr Trevor Nell	DigiOutsource	trevorn@forwardslash.com
Mr Murendeni Maurel Nemukula	University of Limpopo	murendeni.nemukula@gmail.com
Mr Mmbengeni Victor		
Netshidzivhani	University of Limpopo	mmbengeni.netshidzivhani@ul.ac.za
Mr Nyiko Herman Ngobeni	UNISA	ngobeni@unisa.ac.za
Ms Cynthia Ngwane	Agricultural Research Council - Biometry	NgwaneC@arc.agric.za
Mr Mzabalazo Ngwenya	Agricultural Research Council - Biometry	NgwenyaM@arc.agric.za
Mr Raymond Nhapi	University of Cape Town	Raymond.Nhapi@alumni.uct.ac.za
Mrs Johané		
Nienkemper-Swanepoel	Stellenbosch University	nienkemperj@sun.ac.za
Mrs Efa Nkgabutle		efankg@gmail.com
Mr Makgoka Nkoana	Nelson Mandela Metropolitan University	s211185434@nmmu.ac.za
Ms Thobeka Nombebe	SASA	23800364@nwu.ac.za
Ms Nombasa Ntushelo	SASA	ntushelon@arc.agric.za
Ms Onyekachi Esther Nwoko	University of Cape Town	nwkony001@myuct.ac.za

### O

Mr Davies Obaromi	University of Fort Hare	daviesobaromi@gmail.com
Mr Akinwumi Sunday Odeyemi	University of Fort Hare	aodeyemi@ufh.ac.za
Dr Kanayo Ogujiuba	University of the Western Cape	kannyog@gmail.com
Mr Ropo Ebenezer Ogunsakin	University of KwaZulu-Natal	re.korede@yahoo.com
Prof Broderick Oluyede	Georgia Southern University	boluyede@georgiasouthern.edu
Ms Brenda Omachar	University of Pretoria	breshyvugzy@yahoo.com
Ms Wallina Oosthuizen	University of the Free State	oosthuizenw@ufs.ac.za
Mrs Anel Oppel	University of Pretoria	anel oppel@gmail.com
Ms Chane Orsmond	Telkom	clorsmond@telkom.co.za
Dr Nilgün Özgül	Hacettepe University Department of Statistics	nozgul@hacettepe.edu.tr

### P

Dr Kevin Panman	Standard Bank Group	kevin.panman@standardbank.co.za
Ms Christine Papavarnavas	University of Pretoria	chrissypapav@gmail.com
Mr Sisa Pazi	Nelson Mandela Metropolitan University	s209067822@live.nmmu.ac.za
Dr Lillian Pazvakawambwa	University of Namibia	lpazvakawambwa@unam.na
Mr Thembani Phaweni	University of Cape Town	Thembani@live.co.uk
Dr Theoni Photopoulou	University of Cape Town	theoni.photopoulou@uct.ac.za
Mr Etienne Pienaar	University of Cape Town	etiennepi@gmail.com
Dr Sagaren Pillay	Statistics South Africa	sagarenp@statssa.gov.za
Mr Shawn Pilusa	SAS	shawn.pilusa@sas.com
Ms Emma Plumstead	University of Cape Town	emmaplums@gmail.com



## LIST OF DELEGATES

### P

Prof Jean-Michel Poggi  
Prof Daniel Polakow  
Dr Esra Polat  
Ms Jenicca Poongavanam  
Mr Sihle Poswayo  
Mr Charl Pretorius  
Mrs Marelize Pretorius  
Mr Arnu Pretorius

South African Statistical Association  
University of Cape Town  
Hacettepe University  
Nelson Mandela Metropolitan University  
Nelson Mandela Metropolitan University  
North-West University  
North-West University  
Stellenbosch University

Jean-Michel.Poggi@math.u-psud.fr  
daniel.polakow@uct.ac.za  
espolat@hacettepe.edu.tr  
s212469258@nmmu.ac.za  
s212222910@nmmu.ac.za  
Charl.Pretorius@nwu.ac.za  
21632464@nwu.ac.za  
arnu@ml.sun.ac.za

### R

Mr Estian Rademeyer  
Prof Sarah Radloff  
Mr Mabu Isaac Ramathoka  
Mr Jordache Ramjith  
Dr Lizanne Raubenheimer  
Prof Helgard Raubenheimer  
Ms Tarylee Reddy  
Ms Hayley Reynolds  
Ms Danielle Roberts  
Mr Miguel Rodo  
Ms Tessa Louise Rodseth  
Dr David Rose  
Mr Louis Rossouw  
Mr Adriaan Rowan  
Mr Brett Rowland

University of Pretoria  
Rhodes University  
Limpopo Department of Agriculture  
University of Cape Town  
Rhodes University  
North-West University  
South African Medical Research Council  
University of Pretoria  
University of KwaZulu Natal  
University of Cape Town  
University of the Witwatersrand  
University of Cape Town  
University of Pretoria

estian@mweb.co.za  
S.Radloff@ru.ac.za  
Ramathoka.M@agric.limpopo.gov.za  
jo.ramjith@uct.ac.za  
L.Raubenheimer@ru.ac.za  
Helgard.Raubenheimer@nwu.ac.za  
tarylee.reddy@mrc.ac.za  
u12044700@tuks.co.za  
danjader@gmail.com  
miguelrodo2106@gmail.com  
tess.rodseth@gmail.com  
david.rose@wits.ac.za  
lrossouw@genre.com  
adriaan.rowan@genre.com  
u12032906@tuks.co.za

### S

Mr Hassan Sadiq  
Ms Melissa Sangqu  
Dr Leonard Santana  
Prof Robert Schall  
Dr WD (Willem) Schutte  
Prof Solly Matshonisa Seeletse  
Ms Eunice Lebogang Sesale  
Mr Prenil Sewmohan  
Mr Sandile Shongwe  
Dr Caston Sigauke  
Mr Odwa Sihlobo  
Mr Isaac Singini  
Dr Morné Sjölander  
Ms Ansie Smit  
Mr Neill Smit  
Mr Melnick Smith  
Mr Marius Smuts  
Ms Jacqueline Sommerville  
Mr Richard Southey  
Ms Robyn Steenkamp  
Mrs Lizelle Steenkamp  
Mr Paul Swanepoel

University of Cape Town  
Nelson Mandela Metropolitan University  
North West University  
University of the Free State  
North-West University  
Sefako Makgatho Health Sciences University  
Sefako Makgatho Health Sciences University  
University of Pretoria  
University of Pretoria  
University of Venda  
Prescient  
University of Cape Town  
University of the Free State  
University of Pretoria Natural Hazard Centre  
North West University  
Quintiles  
North West University  
University of Pretoria  
Rhodes University  
University of Cape Town  
Discovery  
Barclays Africa

sdqhas001@myuct.ac.za  
melissasangqu@gmail.com  
leonard.santana@nwu.ac.za  
schall@ufs.ac.za  
wd.schutte@nwu.ac.za  
solly.seeletse@smu.ac.za  
lebogang.sesale@smu.ac.za  
prenil.s@gmail.com  
sandile.shongwe@up.ac.za  
csigauke@gmail.com  
odwa.sihlobo@prescient.co.za  
singisa001@myuct.ac.za  
sjolanderM@ufs.ac.za  
ansie.smit@up.ac.za  
neillsmit1@gmail.com  
Melnick.Smith@quintiles.com  
22996168@nwu.ac.za  
jaqui.sommerville@up.ac.za  
g12s1490@campus.ru.ac.za  
robs.steenekamp@gmail.com  
lizelles2@discovery.co.za  
paul.swanepoel@absa.co.za



## LIST OF DELEGATES

### T

Prof Gabriel Tati  
Ms Chandré Teise  
Prof Christien Thiart  
Ms Tlaleng Tshabalala  
Mrs Derya Turfan

University of Western Cape  
University of the Free State  
University of Cape Town  
University of the Western Cape  
Hacettepe University

gtati@uwc.ac.za  
ctk681@gmail.com  
christien.thiart@uct.ac.za  
3307536@myuwc.ac.za  
deryaturfan@hacettepe.edu.tr

### U

Prof Daniel Uys

Stellenbosch University

dwu@sun.ac.za

### V

Ms Liesl van Biljon  
Mr Calven van der Byl  
Dr Linda van der Merwe  
Mr Sean van der Merwe  
Prof Abrie van der Merwe  
Mrs Marieta van der Rijst  
Dr Frederik van der Walt  
Mr Stephan Van Der Westhuizen  
Prof Francois van Graan  
Mr Carl van Heerden  
Ms Bracken van Niekerk  
Mrs Janet Van Niekerk  
Ms Corli van Zyl  
Dr Melvin Varughese  
Mr Jan-Paul Venter  
Mrs Elmarie Venter  
Dr Els Vermeulen  
Dr Andrehette Verster  
Prof Tanja Verster  
Dr Lienki Viljoen  
Dr Jaco Visagie  
Dr Michael Johan von Maltitz

Standard Bank Group  
DSV South Africa  
University of the Free State  
University of the Free State  
University of the Free State  
Agricultural Research Council - Biometry  
Absa  
Stellenbosch University  
Department of Statistics  
  
Nelson Mandela Metropolitan University  
University of Pretoria  
North-West University  
IBM Research Lab  
University of the Free State  
  
University of Pretoria  
University of the Free State  
North-West University  
Stellenbosch University  
University of Pretoria  
University of the Free State

liesl.vanbiljon@standardbank.co.za  
cvanderbyl@googlemail.com  
vdmerwel@ufs.ac.za  
vandermerwes@ufs.ac.za  
matheeme@ufs.ac.za  
VanDerRijstM@arc.agric.za  
frederikvdw@gmail.com  
stephanvdw89@gmail.com  
Francois.VanGraan@nwu.ac.za  
cjvanheerden@gmail.com  
brack.z@hotmail.com  
Janet.vanniekerk@up.ac.za  
vanzylcorli@gmail.com  
melvinv@za.ibm.com  
janpaulventer@gmail.com  
elmarie@seedanalytics.info  
elsvermeulen5@gmail.com  
verster@ufs.ac.za  
Tanja.Verster@nwu.ac.za  
lienki@sun.ac.za  
jaco.visagie@up.ac.za  
vmaltitzmj@ufs.ac.za

### W

Ms Li Wang  
Ms Lee Watchurst  
Mr Neil Watson  
Mr Tim Wolff-Piggott

University of Stellenbosch  
Nelson Mandela Metropolitan University  
University of Cape Town  
University of Cape Town

18461271@sun.ac.za  
watchurstlee@gmail.com  
nm.watson@UCT.ac.za  
twolffpigott@gmail.com

### Y

Ms Nonhlanhla Yende-Zuma  
Ms Ayfer Ezgi Yilmaz

CAPRISA  
Hacettepe University

nonhlanhla.yende@caprisa.org  
ezgiyilmaz@hacettepe.edu.tr

### Z

Mr Andre Zitzke  
Ms Nombuso Zondo

SAS  
University of KwaZulu-Natal

andre.zitzke@sas.com  
nombusogm@gmail.com



# CASIO.

## Solve complicated financial calculations with ease!



### Brand new calculators for the professionals!

#### Exclusive modes make calculations easy

##### Exclusive mode keys:

The direct mode keys allow quick access to each financial calculation display.



##### New operation method:

A full-dot, 4-line display allows up & down scrolling between parameters, making input, confirmation, change or correction so simple.



Easy to input numerals and quick confirmation with the scroll key.

#### Main functions

##### FC-200V & FC-100V

Simple interest calculation / Compound interest calculation / Investment appraisal (Cash Flow) / Amortization / Conversion between nominal interest rate and effective interest rate / Day calculation / Cost/Sell/Margin / Statistics

##### FC-200V ONLY

###### Depreciation

The amount of depreciation can be calculated by the straight-line method, fixed percentage method, sum-of-the-year's digits method or declining balance method.

###### Bond calculation

Bond purchase price and annual rate of yield can be calculated by Date input/Period input.

###### Break-even point calculation

Sales and leverage within the break-even point can be calculated.

###### Shortcut key

Parameter values or settings once used in calculation can be memorized and the stored data can be recalled for use in similar calculations. A very efficient function for repeated calculations.



JAMES RALPH (PTY) LTD

#### Exclusive distributors for Casio in Southern Africa

See our website for more details: [www.jamesralph.com](http://www.jamesralph.com)  
or phone us on 011 314 8888, 0861 CASIO1 or  
0861 318 888 for your nearest dealer

Reliable & Durable



## The customer journey just took a turn for the better.

Take it. Leave it. Keep looking. Understanding a customer's journey can feel elusive. You track behavior, deliver offers - but it's not enough. Customers want to be appreciated and wowed. If your marketing approach has customers thinking "same as it ever was," it's time for a refresh.

With SAS® Customer Intelligence, see each customer in a new light. Create a 360-degree customer view with unified digital data and traditional data sources. Gain a deeper understanding with imbedded predictive analytics. Connect with customers across all channels. And create compelling experiences tailored to each journey.

Give your customers more than they expect. Before they know they want it.

SAS® Customer Intelligence  
[sas.com/ci](http://sas.com/ci)

  
THE POWER TO KNOW®

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. \* indicates USA registration. Other brand and product names are trademarks of their respective companies. © 2016 SAS Institute Inc. All rights reserved. G28381US.0516